

MTA doktori értekezés

# **A rendezetlen fehérjék bioinformatikai vizsgálatai**

**Dosztányi Zsuzsanna**



ELTE TTK Biológiai Intézet Biokémia Tanszék

Budapest, 2023

# TARTALOMJEGYZÉK

<b>1 BEVEZETÉS</b>	<b>5</b>
<b>2 IRODALMI ÁTTEKINTÉS</b>	<b>6</b>
2.1 A fehérje rendezetlenség	6
2.2 Rendezetlen fehérjék funkcionális jellemzői	7
2.3 Szerkezet predikciós eljárások	9
2.4 Rendezetlen fehérjék adatbázisai	10
2.5 Rendezetlen fehérjék predikciója	12
2.6 Rendezetlen kötőhelyek predikciója	14
2.7 A rendezetlen fehérjék általános jellemzése	15
<b>3 CÉLKITŰZÉSEK</b>	<b>18</b>
<b>4 MÓDSZEREK</b>	<b>19</b>
<b>5 EREDMÉNYEK</b>	<b>20</b>
5.1 Rendezetlenség predikció	20
5.1.1 Az energia becslő eljárás	20
5.1.2 Az IUPred módszer	22
5.1.3 Az IUPred webszerver	23
5.2 A rendezetlen fehérjék kölcsönhatásainak jellemzése	23
5.2.1 A rendezetlenség szerepe a fehérje kölcsönhatási hálózatok központi fehérjéiben	23
5.2.2 A rendezetlen fehérje kölcsönhatások molekuláris alapelvei	24
5.3 A rendezetlen kötőhelyek predikciója	25
5.3.1 Az ANCHOR módszer	25
5.3.2 Az energiabecsülő eljáráson alapuló módszerek webszervere	27
5.4 A rendezetlenség és a rák kapcsolata	29
5.4.1 A rendezetlenség biológiai kockázata	29
5.4.2 Rákban gyakran mutálódó rendezetlen szegmensek	30
5.4.3 A rákban mutálódó rendezetlen régiók evolúciós eredete	31
<b>6 KONKLÚZIÓ</b>	<b>33</b>
<b>7 TÉZISEK</b>	<b>34</b>
<b>8 HIVATKOZÁSOK</b>	<b>35</b>
<b>9 AZ ÉRTEKEZÉS ALAPJÁUL SZOLGÁLÓ KÖZLEMÉNYEK</b>	<b>42</b>

## KÖSZÖNETNYILVÁNÍTÁS

Mindenekelőtt köszönettel tartozom Simon István akadémikusnak, korábbi csoportvezetőmnek, aki már egyetemi éveim, majd PhD munkám során is témavezetőm volt az Enzimológiai Intézetben és akinek szakmai és emberi támogatására mindig számíthattam.

Szeretném megköszönni a Simon csoport tagjainak, elsősorban Fiser Andrásnak, Tusnády Gábornak, Fuxreiter Mónikának, Magyar Csabának a sok segítséget és az inspiráló légkört. Az Enzimológiai Intézet többi kutatójáról sem szeretnék megfeledkezni. Külön szeretném kiemelni Tompa Pétert, akinek a rendezetlen fehérje kutatás is, és személy szerint én is nagyon sokat köszönhetek.

Hálás vagyok Nyitray Lászlónak, aki támogatta a Lendület pályázatomat, amellyel elindíthattam önálló kutatócsoportomat az ELTE Biokémiai Tanszékén. Köszönet illeti Kovács Mihályt és a tanszék valamennyi munkatársát, hogy befogadtak és mindenben támogattak.

Köszönettel tartozom korábbi és jelenlegi diákjaimnak, elsősorban Mészáros Bálintnak, Pajkos Mátyásnak és Erdős Gábornak, a közös munkálkodásért és együtt gondolkodásért.

Szeretnék köszönetet mondani minden szerzőtársamnak itthon és külföldön, akik iránymutatásukkal és támogatásukkal hozzájárultak kutatóvá válásomhoz, valamint részt vettek a dolgozatban is bemutatott eredmények elérésében.

Végül, de nem utolsósorban megkülönböztetett hála illeti Szüleimet, akik mindenben támogattak, és páromat, Markot, akinek bátorítása és türelme nélkül nem jutottam volna el idáig.

# 1 BEVEZETÉS

A fehérjék mind szerkezeti mind funkcionális tulajdonságait tekintve hihetetlenül változatosak. Ugyanakkor ennek az elképeztő változatosságnak az alapját egy viszonylag egyszerű kémiai felépítés adja: a fehérjék 20 különböző aminosav egymáshoz kapcsolódásával kialakuló láncmolekulák, melyek egyediségét az aminosavak sorrendje, az aminosav szekvencia adja. A fehérjekutatás egyik legalapvetőbb kérdése, hogy hogyan határozza meg az aminosav sorrend a fehérjék térszerkezetét és funkcióját. Ennek a kérdéskörnek az alapjait több mint ötven éve fektették le, többek között Christian Anfinsen, a ribonukleáz refolding vizsgálatával, Cyrus Levinthal, a fehérje feltekeredés, a folding, alapvető paradoxonának megfogalmazásával, illetve John Kendrew az első fehérje, a mioglobín szerkezetének meghatározásával [1]. Bár az azóta eltelt időszakban jelentős előrelépések történtek a szekvencia - szerkezet - és funkció összefüggésének megértésében, mind a mai napig újabb és újabb dolgokat tanulunk a fehérjék sokszínűségéről.

Az egyik alapvetően új irány a rendezetlen fehérjékhez kapcsolódik. A sokáig általánosan elfogadott nézet az volt, hogy a fehérjék megfelelő működéséhez elengedhetetlen, hogy egy jól definiált szerkezettel rendelkezzenek. Ezt a paradigmát írta át a rendezetlen fehérjék funkcionális fontosságának felismerése. Bár ez a jelenség a kezdetekben jelentős vitákat váltott ki [2], a 2000-es évek kezdetétől megindult a rendezetlen fehérjék szisztematikus vizsgálata. Ezekben - a rendezetlen fehérjék kísérletes vizsgálatának nehézségei miatt - döntő szerep jutott a bioinformatikai megközelítéseknek.

Doktori értekezésemben a rendezetlen fehérjék bioinformatikai vizsgálata során elért eredményeimet foglaltam össze. Közel két évtizedes tevékenységem során többek között új bioinformatikai eszközöket fejlesztettem ki, melyekkel felismerhetők a rendezetlen régiók illetve azok kötőhelyei az aminosav szekvenciából. A saját és mások által kifejlesztett módszerek alkalmazásával elemeztem a rendezetlen fehérjék kölcsönhatási tulajdonságait és betegségben betöltött szerepüket. Összességében, vizsgálataim alapvetően járultak hozzá ezen új fehérje osztály jobb megismeréséhez.

## 2 IRODALMI ÁTTEKINTÉS

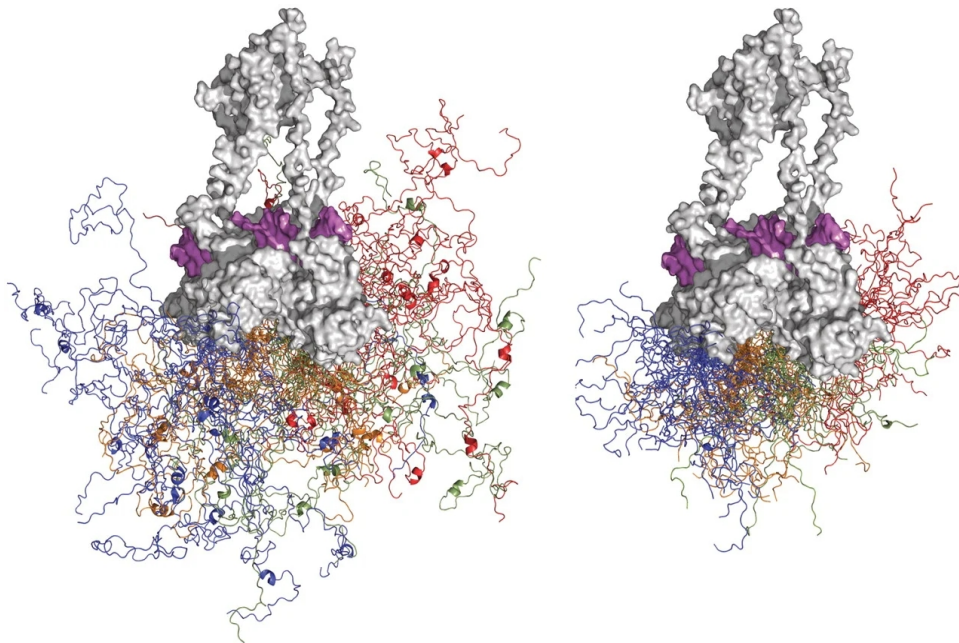
### 2.1 A fehérje rendezetlenség

Bár a klasszikus szerkezet-funkció paradigma szerint a fehérjék funkciójához elengedhetetlen a jól definiált térszerkezet megléte, a múlt század végére egyre nyilvánvalóbbá vált, hogy ez a kép nem teljes: nem minden biológiailag funkcionális fehérje vesz fel spontán rendezett szerkezetet, ugyanakkor ezek a fehérjék illetve fehérje szegmensek számos biológiai folyamatban kulcsfontosságúak. Ezt felismerve vetette fel 1999-ben Jane Dyson és Peter Wright, hogy a szerkezet-funkció paradigmát felül kell vizsgálni, és vezette be az eredendően szerkezet nélküli fehérjék (intrinsically unstructured proteins) fogalmát [3]. Ezzel párhuzamosan Keith Dunker és kollégái felismerték, hogy a röntgenszerkezet hiányzó szegmensei gyakran tartalmaznak olyan funkcionális elemeket, amelyek rendezetlenek [4,5]. Ők végezték el az első bioinformatikai elemzéseket is ezen a területen. Magyarországon Tompa Péter indította el a rendezetlen fehérjék vizsgálatát [6]. Sokáig vita volt arról, hogy hogyan nevezzék el ezt az újfajta fehérje osztályt, végül az eredendően rendezetlen fehérje (intrinsically disordered protein/IDP) elnevezés vált általánosabban elfogadottá [7].

Bár az ismert fehérje térszerkezetek egy statikus képet sugallnak, a rendezett fehérjék is dinamikusak és ez elengedhetetlen a funkciójukhoz [8]. A globuláris szerkezet tartalmazhat rövidebb-hosszabb flexibilis, illetve rendezetlen szegmenseket is, melyek jellemzően a hurok régiókban illetve a szerkezet terminális részein található. Azonban a globuláris fehérjékre alapvetően jellemző egy egyensúlyi szerkezet, amely körül fluktuálnak [7]. Ezzel szemben a rendezetlen fehérjéket csak egy, nagymértékben különböző konformációkat tartalmazó szerkezeti sokasággal lehet jellemezni [7,9]. Ennek a sokaságnak a részletes tulajdonságai nagyon heterogének lehetnek [10]. Vannak olyan rendezetlen fehérjék, melyek teljesen véletlenszerűen, random coil-ként viselkednek. Számos esetben megfigyelhetőek azonban lokális szerkezeti preferenciák vagy hosszútávú tranziens kölcsönhatások. A rendezetlen fehérjék egy része a globuláris fehérjék kitekeredése során megfigyelt, úgynevezett olvadt gombóc (molten globule) állapothoz hasonlítható, melyet nagyobb mennyiségű lokális szerkezeti elem és kompaktság jellemez a random coil állapothoz képest. Bár vannak teljesen rendezett vagy teljesen rendezetlen fehérjék is, a legtöbb fehérje mindkét féle régiót tartalmaz. Ezek szerkezeti jellemzőit befolyásolhatják egymással való kölcsönhatásaik, a sejten belüli környezeti tényezők (pl. pH, redox potenciál, hőmérséklet), poszttranszlációs módosítások, illetve más partnerrel való interakciók is [11].

A rendezetlen fehérjék jellegzetes molekuláris tulajdonságai, mint például a megnövekedett molekuláris méret, a denaturált állapotba való átmenet hiánya, a periódikus másodlagos szerkezetek hiánya vagy a konformációs heterogenitás számos kísérleti módszerrel megragadható [12]. Indirekt módon, a fehérje térszerkezeteket összegyűjtő Protein Data Bank (PDB) adatbázis is szolgáltat bizonyítékot a

rendezetlenség jelenlétéről. Ez röntgenszerkezetek esetében a hiányzó elektronsűrűség, NMR-szerkezetek esetében pedig a nagy szerkezeti variabilitás formájában jelentkezik. A rendezetlen fehérjék konformációs preferenciáiról a legrészletesebb információ NMR segítségével nyerhető [13–15]. A konformációs sokaság teljeskörű jellemzéséhez azonban többféle technikára, például különböző típusú NMR-mérések és kisszögű röntgen szórás kombinációjára van szükség, melyeket számítógépes számításokkal kell kiegészíteni a modellalkotáshoz [16]. Az egyik legjobban ismert és jellemzett rendezetlen fehérje a p53 N-terminális rendezetlen részének modellje látható az 1. ábrán. A kapott modell jól mutatja az N-terminális részen a lehetséges konformációk változatosságát és rámutat egy tranziens  $\alpha$ -hélix jelenlétére is [17].



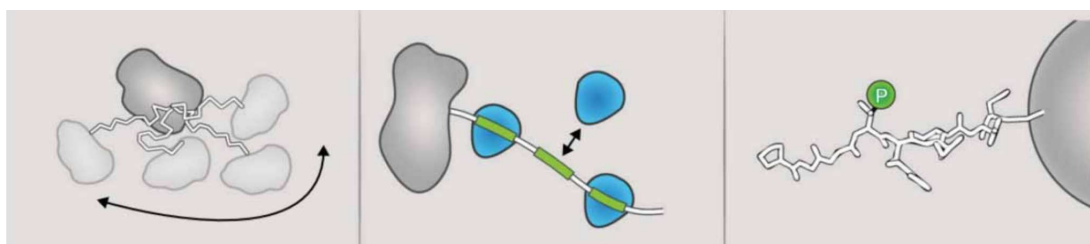
1. ábra. Az N-terminális rész modellje a p53 tetramer és DNS (lila színnel jelölve) komplexében, amit NMR spektroszkópia, kisszögű röntgenszórás, molekuladinamikai szimulációk kombinációjával határoztak meg [17].

## 2.2 Rendezetlen fehérjék funkcionális jellemzői

Az eredendően rendezetlen fehérjék alapvető fontosságának felismerése drasztikusan alakította át a fehérjék szerkezete és működése közötti összefüggésről alkotott elképzelésünket [12]. Általánosságban elmondható, hogy a globuláris fehérjékre olyan feladatokban van szükség, amelyek az aminosavak pontos orientációját igénylik, mint például az enzimek katalitikus aktivitása [18]. Ezzel szemben a rendezetlen régiók a bennük rejlő dinamikus tulajdonságukat és a plaszticitásukat használják fel a működésük során [12,19].

Az ismert példák alapján a rendezetlen fehérjék funkcionális szempontból több kategóriába sorolhatók [19,20] (2. ábra). Legjellegzetesebb működési módjuk a közvetlenül a rendezetlen állapotról eredő entrópikus lánc funkció. Ebbe a kategóriába

tartoznak például a linkerek, amelyek lehetővé teszik a domének egymáshoz viszonyított mozgását; vagy a távtartók, amelyek a funkcionális modulok közötti távolság szabályozásáért felelősek. A rendezetlen fehérjék másik tipikus funkciója molekuláris felismerésből ered, melynek során specifikus kötőpartnerekhez, például egy másik fehérjéhez, RNS-hez vagy DNS-molekulához kötődnek. A kölcsönhatás eredményeképpen befolyásolhatják a partner molekulák működését vagy nagyobb komplexek összeszerelődését. A rendezetlen régiók gyakran célpontjai különböző poszttranszlációs módosításoknak (PTM). Nyújtott és hozzáférhető szerkezetük miatt különösen alkalmasak kis méretű ligandumok, például ionok és szerves vegyületek tárolására vagy neutralizációjára. A rendezetlen szegmensek fontos szerepet játszanak fehérje és RNS chaperonok működésében is. A funkcionális kategóriák köre a közelmúltban új taggal bővült: ez egy rendkívül izgalmas jelenség, a fázisszeparáció révén kialakuló biológiai kondenzátumokhoz kapcsolódik és az ebben részt vevő rendezetlen fehérjéket gyűjti össze [21].



2. ábra. A rendezetlen régiók összeköthetnek doméneket, ahol rugalmasságuk lehetővé teszi, hogy a fehérje többféle konformációt vegyen fel; a bennük lévő lineáris motívumok fehérje kölcsönhatásokat közvetíthetnek; az aminosavaik poszttranszlációs módosítása lehetővé teszi az információ kódolását és dekódolását [22].

A rendezetlen fehérjékre jellemző konformációs szabadság nagyfokú képlékenységgel jár együtt, ami lehetővé teszi a különböző partnerek felszínéhez való igazodást [23]. A rendezetlen fehérjék gyakran tartalmaznak kompakt, néhány aminosavból álló rövid lineáris motívumokat, melyek specifikus doménnel való kölcsönhatást közvetítenek [24]. Azonban számos rendezetlen fehérje hosszabb szegmensen keresztül kötődik [10]. A partner molekulával való kölcsönhatás indukálhatja egy jól-definiált szerkezet kialakítását, az úgynevezett csatolt feltekeredés és kötődés (coupled folding and binding) során [25]. Felvetették, hogy az ilyen típusú kölcsönhatások általában gyenge, de specifikus interakciót tesznek lehetővé [3,26]. Ennek oka az, hogy a kölcsönhatás révén kialakuló entalpia nyereséggel összemérhető a rendeződéssel járó entrópia veszteség. Azonban a kölcsönhatás kinetikai és termodinamikai paramétereit befolyásolják a nem-kötött állapotban jelen levő szerkezeti preferenciák és a kötött állapotban is megmaradó szerkezeti heterogenitás, az úgynevezett bolyhosság/fuzziness [27]. Összességében, a rendezetlen fehérje szegmensek által kialakított kölcsönhatások kinetikai és termodinamikai tulajdonságai széles skálán mozognak [28,29]. A rendezetlen fehérjék nagy sűrűségben tartalmazhatnak különböző funkcionális modulokat és poszttranszlációs helyeket,

melyek kombinációja révén molekuláris kapcsolók összetett hálózata alakulhat ki [30]. Ezek a tulajdonságok különösen előnyösek különböző szabályozó és jelátviteli funkciókban, melyek a sejt- és környezeti jelzésekhez való gyors alkalmazkodást igényelnek [31].

Funkcionális jelentőségük felismerésével, a rendezetlen fehérjék vizsgálata részévé vált a szerkezeti bioinformatikai kutatásoknak is.

## 2.3 Szerkezet predikciós eljárások

A szerkezeti bioinformatikai kutatások fő fókusza sokáig a globuláris fehérjék térszerkezetének predikciója volt. Ennek kiindulópontja az ismert térszerkezetek, amiket a PDB adatbázis gyűjt össze [32]. A kísérletileg meghatározott, nagy felbontású szerkezetek száma az adatbázis 1976-os elindítása óta jelentős mértékben emelkedett, mára megközelíti a 200000-et. Azonban még ezek száma is elenyésző az ismert fehérje szekvenciák robbanásszerű növekedéséhez képest. Ez még inkább felerősítette azt az igényt, hogy képesek legyünk megjósolni, hogy milyen szerkezet tartozik egy adott fehérje szekvenciához.

Az Anfinsen által megfogalmazott termodinamikai hipotézis szerint a fehérjék háromdimenziós szerkezetét a Gibbs-féle szabadenergia határozza meg, és a normál fiziológiás körülmények között a natív szerkezet a legalacsonyabb, vagy az egyik legalacsonyabb energiájú állapotnak felel meg [33,34]. A szerkezet kialakításában számos tényező vesz részt, az aminosavak között nagyszámú van der Waals, elektrosztatikus kölcsönhatás, hidrogénhid alakulhat ki. A globuláris szerkezet kialakulásának fontos tényezője a hidrofób effektus, ami a vizes közegben az apoláros aminosavak eltemettségét eredményezi. A kölcsönhatások eredményeképpen egy kompakt szerkezet alakul ki, ami ellen dolgozik a konformációs entrópia, ami a feltekeredés következtében jelentősen lecsökken. A globuláris fehérjék szerkezetének stabilitásában résztvevő tényezők modellezhetőek fizikai erőterekkel, amelyekkel feltérképezhető a fehérjék energiafelszíne számítógépes szimulációkon keresztül. Azonban ezen megközelítés alapján továbbra is csak néhány kisebb fehérjék szerkezetét sikerült modellezni [1]. A legtöbb esetben pontosabban és gyorsabban lehet szerkezeti modellt generálni olyan bioinformatikai módszerekkel, amelyek csak részben vagy egyáltalán nem használnak fizikai megközelítéseket.

A legegyszerűbb predikciós módszerek nem magát a szerkezetet, hanem csak valamilyen szerkezeti tulajdonságot próbáltak meg jósolni közvetlenül az aminosav szekvenciából, mint például másodlagos szerkezeti elemekben való előfordulás, torziós szögek, az egyes aminosavak eltemettsége/hozzáférhetősége [35]. A kezdeti módszerek egyszerű aminosav tulajdonságokon alapultak, amiket egyre komplexebb statisztikai modellek, illetve különböző gépi tanulási eljárások (mesterséges neuronhálózatok, support vector machine-ok, rejtett Markov modellek) követtek. Jelentős előrelépést értek el ezek a módszerek azáltal, hogy nem egyetlen szekvenciát használtak bemenetként, hanem többszörös szekvencia illesztést [36]. A módszerek fejlődéséhez nagyban hozzájárult az is, hogy egyre több adat áll rendelkezésünkre a



módszerek betanításához a szekvencia és térszerkezeti adatok növekedésével [37].

A fehérjeszerkezetek számának növekedésével olyan általános szabályszerűségek is feltárásra kerültek melyek szintén kiaknázhatóak a szerkezetbecslésben. Az egyik általános megfigyelés szerint bizonyos típusú térbeli kölcsönhatások gyakrabban fordulnak, mint ahogyan azt véletlenszerű eloszlás alapján várni lehet. A kölcsönhatások megfigyelt gyakorisága energiaszerű mennyiségekké alakítható a Boltzmann-eloszlás alapján [38]. A legegyszerűbb formában ezek a statisztikai potenciálok egy 20 x 20-as mátrixban fejezhetők ki, melynek értékei az egyes aminosavpárok közötti kölcsönhatást jellemzik. A kapott energia jellegű mennyiségek nem feleltethetők meg direkt fizikai tényezőknek és nem értelmezhetők közvetlenül szabadenergiaként. Azonban általában felülmúlják a fizika alapú energiafüggvényeket a threading (felfűzés) eljárásban, amelynek célja az adott szekvenciához tartozó szerkezet megtalálása az ismert szerkezetekre felfűzve kapott alternatív konformációk között, annak kedvező energiája alapján [39]. A statisztikus potenciálok legkritikusabb eleme a normalizáláshoz használt referenciaállapot [40]. Ennek problémájára javasolt egy elegáns megoldást Ken Dill and Paul Thomas, ami egy iteratív algoritmus alapján megkerüli a referencia állapotra vonatkozó közvetlen feltételezéseket [41]. Összességében, a statisztikus potenciálok különböző formái rendkívül hasznosnak bizonyultak a szerkezetbecslés területein [42,43].

Az évek során számos módszert fejlesztettek ki a fehérje szerkezetek modellezésére [44]. Ennek motorjává vált a kétévente megrendezésre kerülő CASP, ami megalapozta a módszerek független értékelését és lehetőséget teremtett a terület fejlődésének nyomon követésére [45,46]. A probléma nehézsége szerint a szerkezet predikciókat két csoportra lehet osztani attól függően, hogy létezik-e a célfehérje szerkezetéhez hasonló ismert, úgynevezett templát szerkezet. Az első kategóriában már viszonylag korán megbízható minőségű modelleket lehetett generálni, homológia modellezés vagy - távolabbi vagy indirekt kapcsolat esetén - a threading eljárás alapján. Ezzel szemben a templát nélküli célpontok modellezésében sokáig csak szerény javulást sikerült elérni, még az olyan új módszerek, mint a fragmens könyvtáron alapuló szerkezet összeszerelés vagy a kontaktus predikciók alapján is [1,47]. Az elmúlt években azonban gyökeresen átalakult a helyzet, elsősorban a számos területen áttörő eredményeket elérő mélytanulós eljárásoknak köszönhetően [48]. A legkiemelkedőbb eredményt a Google DeepMind cég által bevezetett AlphaFold2 módszere érte el a CASP14 során, elsőként produkálva a kísérletes módszerekkel összevethető minőségű modelleket a térszerkezet predikciók teljes skáláján [49].

## 2.4 Rendezetlen fehérjék adatbázisai

A globuláris fehérjékre kidolgozott predikciós módszerek fejlődésével párhuzamosan indult el a rendezetlen fehérjék bioinformatikai vizsgálata. A korábbi megközelítések számos módon hatottak a rendezetlenség predikciókra, de az új problémakör lehetőséget teremtett eredeti megközelítések megalkotására is. Ezekhez azonban szükség volt speciális adatszettekre is.

A rendezetlenségre vonatkozó kísérletes adatok alapvetően két forrásból származnak. Az egyik halmazt a PDB szerkezetekben azok a régiók jelentik, melyekhez nem rendelhető elektronsűrűség, ezért hiányoznak a megoldott szerkezetből. Ehhez nagyon hasonlóan viselkednek az NMR szerkezetekben a nagyobb szerkezeti variabilitás alapján kigyűjtött szegmensek [50]. A PDB-ből kigyűjtött rendezetlen szegmensek általában a fragmens terminális részén találhatóak, és viszonylag rövidebb szakaszok (short disorder). A másik csoportot az egyéb kísérletes módszerrel igazolt rendezetlen szegmensek jelentik. Ezek általában hosszabb szakaszok, melyek fontos biológiai funkcióval rendelkeznek (long disorder). Ugyanakkor ez a halmaz sokáig jóval kevesebb adatot tartalmazott, és zajosabb volt, mivel több kísérletes módszer sem rendelkezik aminosav szintű felbontással. A kísérletesen igazolt rendezetlen fehérjék összegyűjtésére hozta létre Keith Dunker a DisProt adatbázist [51], melynek fenntartását 2017-ben Silvio Tosatto vette át [21]. Az adatbázis folyamatosan bővül, jelenleg (2022 végén) 2469 fehérje 5568 régiójára tartalmaz információt [52]. Ezek annotációjában több kutatócsoport, többek között a mi csoportunk is részt vett.

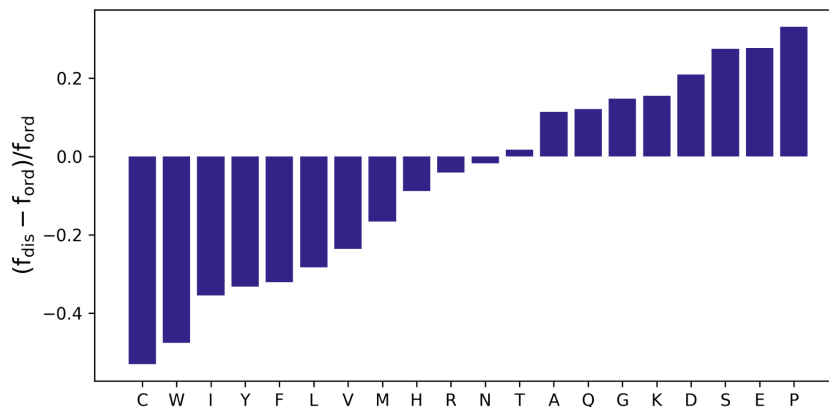
A rendezetlen fehérjék vizsgálatának másik fő fókusza kölcsönhatásainak, és ezen belül is az általuk kialakított komplexek szerkezetének vizsgálata. Azokat a rendezetlen fehérje szegmenseket melyek a kötődés határára rendezett szerkezetet vesznek fel szokás MoRF-nak (molecular recognition feature) is hívni [53]. Mohan és kollégái ezeket a szerkezeteket különböző csoportba osztotta a kialakított másodlagos szerkezetek típusa szerint, melyek között az  $\alpha$ -hélixnek megfelelő konformációt felvevő  $\alpha$ -MoRF-ok voltak a leggyakoribbak [54]. A DisProt adatbázis révén lehetővé vált azoknak a szerkezeteknek a kigyűjtése, amelyek esetében a rendezetlen státuszt kísérleti úton igazolták és a rendezetlen-rendezett átmenetet a komplex ismert szerkezete támasztja alá [55]. Ezek száma folyamatosan bővült, az általunk 2018-ban létrehozott DIBS adatbázis már 773 ilyen példát tartalmazott [28]. Egy hasonló adatbázis, az MFIB adatbázis középpontjában a kizárólag két vagy több rendezetlen fehérje között kialakult komplexek állnak [56]. A kötött állapotukban is jelentős konformációs heterogenitást tartalmazó bolyhos komplexekre a FuzDB adatbázisban gyűjtöttek példákat [57]. A kísérletesen igazolt rövid lineáris motívumok központi adatbázisa az ELM [58]. A különböző adatbázisok megléte már önmagában mutatja a rendezetlen fehérjék kölcsönhatásának a sokszínűségét.

A különböző klasszifikációs eljárások kifejlesztéséhez nagyon fontosak a negatív példák is. A rendezetlenség predikciók esetén komplementer halmaznak a PDB alapján az ismert szerkezettel rendelkező régiókat szokás tekinteni. Azonban ezek között - komplexek formájában - rendezetlen fehérjék is előfordulnak. Mi ezért

speciálisan monomer szerkezetekből álló adatszetteket használtunk [59,60]. A kötőrégiók predikciója esetén még nagyobb kihívást jelent a negatív halmaz definiálása [61] és ez különösen igaz a rendezetlen kötőhelyek predikciójára. Ebben az esetben negatív példáknak tekinthetjük a globuláris fehérjéket általában, illetve a rendezetlen fehérjék azon régióit, melyek nem vesznek részt közvetlenül kölcsönhatásban. Azonban a rendezetlen fehérjék esetén, a kölcsönhatások jelentős része még nem ismert [62]. Az ismert rendezetlen kötőrégiók viszonylag kis száma, és a megbízható negatív adathalmaz hiánya jelentős limitáló tényező, ami különösen a kezdetekben visszavetette a specifikus módszerek kifejlesztését.

## 2.5 Rendezetlen fehérjék predikciója

A rendezetlen fehérjék kutatásának egyik alapkérdése, hogy mi a jelenség biofizikai alapja és ez hogyan van kódolva a szekvenciában. Ennek vizsgálatához kezdetben, a több száz globuláris fehérjéhez képest, csak néhány tucat rendezetlen régió állt rendelkezésre. Azonban már ezek alapján is feltűnő volt a két csoport aminosav összetétele közötti alapvető különbség. Aszerint, hogy egyes aminosavak inkább a rendezett vagy rendezetlen szegmensekben gyakoribbak, megkülönböztethetünk rendezettséget, illetve rendezetlenséget elősegítő aminosavakat [63]. A jelenlegi DisProt adatbázis elemzése szerint a prolin és a glutaminsav mutatja a legerősebb rendezetlenséget elősegítő tendenciát, melyet a szerin és az aszparaginsav követ a sorban (3. ábra) [64]. Az aminosav összetételre vonatkozó alapvető megfigyelések alapján javasolta Vladimir Uversky, a rendezetlen fehérjékre az alacsony hidrofóbítás és a magas nettó töltés kombinációja jellemző [65]. Ezen fehérjeosztály másik jellemzője, hogy gyakran fednek át alacsony komplexitású szekvenciákkal [66]. Ezen megfigyelés teljesen új értelmet nyert a biológiai kondenzátumok létrejöttében szerepet játszó fázisszeparáció jelenségének tükrében. Ez a folyamat ugyanis gyakran a rendezetlen fehérjéken belüli alacsony komplexitású régiók által létrehozott multivalens kölcsönhatások révén alakul ki [67]. Az általános tendenciák mellett azonban megfigyeltek kisebb eltéréseket is a szekvenciális jellemzőkben, például a rendezetlenség meghatározásához használt kísérleti körülményektől, vagy a szegmensek hosszától függően [68]. Ennek alapján vetették fel, hogy a rendezetlenség - heterogén jellegének megfelelően - különböző „ízekkel” rendelkezik [69]. Ennek bizonyítékeként, összefüggést találtak a rendezetlen fehérjék szekvenciális és konformációs tulajdonságai között [70].



3. ábra. A rendezetlen régiók aminosav gyakoriságának ( $f_{dis}$ ) relatív eltérése a rendezett régiók aminosav gyakoriságától ( $f_{ord}$ ) a DisProt adatbázis fehérjéin ([52]) számolva [64].

A rendezett és rendezetlen fehérjék aminosav összetételében megfigyelt alapvető eltérések arra utalnak, hogy nemcsak a szerkezet, hanem a rendezetlenség is kódolva van a szekvenciában, és ez alapján predikciós módszerek készíthetők. Az elmúlt körülbelül 20 év alatt több mint 100 rendezetlenség becslő módszert fejlesztettek ki, melyek a különböző megközelítések széles skáláját fedik le mögöttes elv, komplexitás, szükséges futási idő és pontosság tekintetében is [71–73]. A legegyszerűbb módszerek egyszerű aminosav tulajdonsági skálákon alapulnak (pl. GlobPlot [74], FoldUnfold [75], TOP-IDP [63]). Ezen módszerek fő előnye könnyű értelmezhetőségük és egyszerű, gyors használhatóságuk.

A rendezetlenség predikciós módszerek másik jelentős csoportja gépi tanuláson alapul. Ezek a módszerek nagyban építettek a szerkezeti bioinformatikában korábban bevezetett megközelítésekre. Többféle algoritmust is adaptáltak a rendezetlenség predikcióra, többek között egyrétegű mesterséges neuron hálózatokat, support vector machine-okat, random forest klasszifikációt. A gépi tanulási módszerek korábbi generációjában szokásos módon, bemenetként kézzel összeválogatott tulajdonságok (pl. aminosav-összetétel, hidropátia, töltés és flexibilitás) szolgáltak egy adott szekvencia ablakon belül [76]. A másodlagos szerkezet predikciós módszerek nyomdokain haladva, PSI-BLAST-ból származó evolúciós profilt is figyelembe vettek a predikció során [18]. Bár az evolúciós információ némi előnyt jelent a szekvencia-alapú módszerekkel szemben, ennek ára a megnövekedett számítási erőforrás és idő. A gépi tanuláson alapuló módszerek azonban általában kevésbé átlátható módon, alapvetően fekete dobozként működnek, és érzékenyebbek a tanító halmazban esetlegesen előforduló hibákra. A legújabb predikciós eljárások egyre inkább támaszkodnak fejlett mélytanulási technikákra [77]. A legújabb eredmények szerint, a térszerkezet predikciókban áttörést elérő AlphaFold2 módszer a rendezetlenséget is kiválóan képes előre jelezni a modellek megbízhatóságára kifejlesztett pLDDT érték alapján [78].

A rendezetlenség predikciós módszerek hatékonyságának összehasonlító értékelésére és a terület fejlődésének nyomon követésére több nagyszabású vizsgálatot is végeztek [71–73]. A módszerek teljesítménye függ a kiértékeléshez használt adatbázistól és a használt metrikától is. A leggyakrabban használt mérőszámok az átlagos pontosság, és a ROC (Receiver Operating Characteristic/vevő működési jellemző) görbe, illetve a görbe alatti terület (AUC). Átmenetileg a rendezetlenség predikció része volt a térszerkezet predikciós módszerek értékelésére létrehozott CASP kezdeményezésnek, ami a térszerkezet predikciók értékelésénél használt szerkezetek hiányzó részein alapultak [79,80]. Néhány év után megszűnt ez a kategória, a rendezetlenségre vonatkozó kellő mennyiségű adat hiányában. A 2021-ben közzétett Critical Assessment of Intrinsic Protein Disorder (CAID) kísérlet volt az eddigi legnagyobb kezdeményezés a rendezetlenség predikció értékelésére [81]. Ebben 32 módszer előrejelzési pontosságát és futási idejét értékelték, amihez a DisProt adatbázisban újonnan annotált fehérjéit használták.

## 2.6 Rendezetlen kötőhelyek predikciója

A DisProt adatbázisban rendelkezésre álló adatok azt mutatják, hogy a funkcionálisan annotált rendezetlen régiók leggyakoribb funkciója más makromolekulákhoz való kötődés, ezen belül is elsősorban fehérje-fehérje kölcsönhatás kialakítása. Azonban az ismert rendezetlen kötőrégiók száma csak lassan növekszik. Ezért nagy szükség van olyan módszerekre, melyek a szekvenciából képesek felismerni a rendezetlen kötőhelyeket. Korábban felvetették, hogy bizonyos rendezetlen fehérje predikciós módszerek esetén a kimeneti profil jelezheti a kötőrégiók meglétét [54]. Azonban ez a megközelítés csak korlátozottan használható. A jelenleg elérhető rendezetlen kötőhely predikciós módszerek kiindulópontjai a PDB adatbázisból kigyűjtött, kísérletesen is igazolt rendezetlen fehérje szegmensek, amelyek kötődés során rendeződnek. A komplex kialakításában részt vevő rendezetlen szegmensek szerkezeti és szekvenciális tulajdonságainak vizsgálatai alapján ezek a régiók számos specifikus tulajdonsággal rendelkeznek [82], amelyek kiaknázhatóak a predikciójuk során.

Az egyik legelső próbálkozásként, a komplexben  $\alpha$ -hélix konformációt kialakító régiók felismerésére hoztak létre egy speciális módszert, az  $\alpha$ -MoRf prediktort [83]. Azonban az első általános módszer rendezetlen kötőhelyek felismerésére az általunk kifejlesztett ANCHOR módszer volt [55]. Különböző gépi tanulós módszereket is használtak a kötődés során rendezetlen-rendezett átmenetet mutató régiók azonosítására. Ezek közül a MoRFCHiBi módszer a support vector machine technikát használta, ami jól alkalmazható, ha a tanítás során csak viszonylag kis adathalmaz áll rendelkezésre. A módszer kombinálta a rendezetlenség predikciót a helyi szekvenciajellelmzőkkel, nagy hangsúlyt fektetve arra, hogy a kötődő szegmenseket megkülönböztesse szekvenciális környezetüktől [84]. Ennek társaként létrehoztak egy olyan módszert is, ami az evolúciós információk felhasználásával javítja tovább a

predikciókat [85]. A CAID kiértékelte a rendezetlen kötőrégiók jóslására szolgáló módszerek teljesítményét is, a DisProt kötőrégiókra vonatkozó annotációit használva tesztalmozként [81]. Ez az adathalmaz kevesebb mint 250 fehérje régiót tartalmazott, azonban még ezek az annotációk is gyakran hiányosak illetve pontatlanok voltak. A kategóriában mindössze tíz módszer szerepelt, és a probléma nehézsége a módszerek viszonylag szerényebb teljesítményében is tükröződött.

Az utóbbi időben több módszer is bővítette a rendezetlen fehérjék funkcionális jellemzésére használható módszerek körét. Egy új módszer a bolyhos (fuzzy) fehérjék kölcsönhatási régiót ismeri fel a szekvenciából [86]. A DisoRDPbind és az újabb DeepDISOBind többszintű gépi tanulós eljárás alapján jósol nemcsak fehérjekötő-, hanem DNS és RNS kötő rendezetlen szegmenseket is [87]. A DisoLipPred módszer célja a lipidet kötő részek azonosítása a rendezetlen fehérjékben belül [88]. A linker régiók egy másik alapvető kategóriát jelentenek a rendezetlen fehérjék funkcióin belül. Ezek felismerésére specializálódott a DFLpred és az APOD módszerek [89,90]. Összességében, ez a terület kevésbé kiforrott a rendezetlenség predikciókhoz képest is.

## 2.7 A rendezetlen fehérjék általános jellemzése

Az elmúlt évek erőfeszítései ellenére a kísérletesen jellemzett rendezetlen fehérjék száma még mindig erősen limitált, mivel speciális tulajdonságaik nagyban megnehezítik részletes vizsgálatukat. Ugyan történtek erőfeszítések a rendezetlen fehérjék nagyskálás feltérképezésére [91,92], ezek egyelőre nem váltották be a hozzájuk fűzött reményeket. A különböző bioinformatikai módszerek, köztük az IUPred és az ANCHOR, azonban lehetőséget teremtenek a rendezetlenség nagyobb léptékű vizsgálatára. A rendezetlenség predikciós módszerek a szerkezet meghatározási folyamat szerves részévé váltak, lehetővé téve a fehérje konstrukciók optimalizálását [93]. A rendezetlenségre vonatkozó információ fontos kiindulópontja a sötét proteomhoz tartozó fehérjék (olyan fehérjék, melyek nem modellezhetők a jelenleg ismert szerkezetek alapján) jellemzésének is [94]. Ezt felismerve, a fehérjék központi adatbázisa, a Uniprot is közlést tesz a rendezetlenségre vonatkozó információt, az IUPred-et is magában foglaló MobiDB-lite konszenzus módszer alapján [95]. Ezek mellett az alapvető fontosságú alkalmazások mellett, a predikciós módszerek azt is lehetővé tették, hogy betekintést nyerjünk a rendezetlen fehérjék általános tulajdonságaiba is.

A rendezetlenség általánosan elfogadottá válásához nagyban hozzájárult, hogy a különböző genom szekvenciák által kódolt fehérjék vizsgálata alapján széles körben elterjedt jelenségről van szó. A legelső elemzés 34 genomot vizsgált a PONDR család módszereivel, meghatározva a teljesen rendezetlen, illetve a legalább 30 (illetve 40 és 50) hosszú rendezetlen régiót tartalmazó fehérjék arányát [96]. A későbbiek során már jóval több genomot vizsgáltak, más eszközökkel is, azonban az általános tendenciák nem változtak. Eszerint a rendezetlen fehérjék az életfa minden ágán jelen vannak és előfordulásuk erősen korrelál az organizmusok komplexitásával [97,98]. A DISOPRED2 módszer alapján a hosszú (>30 aminosav hosszú) rendezetlen

szegmensek az archea fehérjék 2,0%-ában, az eubakteriális fehérjék 4,2%-ában és az eukarióta fehérjék 33,0%-ában fordul elő [18]. Az *E. coli* és az *S. cerevisiae* teljes proteomjának elemzése az IUPred programmal hasonló eredményekre vezetett. Eszerint, az élesztő-proteom lényegesen több rendezetlen szegmenst tartalmaz, mint az *E. coli*, a fehérjék ~50%-a, illetve 15%-a tartalmazott rendezetlen szakaszokat [99]. Érdekes módon a vírusfehérjék, különösen az RNS-vírusok fehérjei között is gyakoriak a rendezetlen fehérjék [100]. Például a SARS-CoV-2 vírusban a rendezetlen régiók fontos szerepet játszanak a vírusgenom csomagolásában [101]. Összességében, a fehérje rendezetlenség az evolúció fontos találmánya [102], amely bizonyos típusú funkciókban jelentős előnyöket biztosít a globuláris doménekkal szemben.

A bioinformatikai elemzésekkel lehetett jellemezni a rendezetlenséghez kapcsolódó specifikus funkciókat is. A génontológiai annotációk és a SwissProt adatbázis kulcsszavainak elemzése hasonló eredményekre vezetett, és megerősítette a rendezetlenség fontos szerepét jelátviteli és szabályozó folyamatokban [18,103]. A rendezetlenség hiányával leginkább az enzimmatalízishez kapcsolódó funkcionális kulcsszavak korreláltak. Az újabb vizsgálatok szerint a rendezetlenséghez kapcsolódó fő funkcionális kategóriák a differenciáció, transzkripció és szabályozása, spermatogenezis, DNS kondenzáció, sejt ciklus, mRNS processzálas és splicing, mitózis és apoptózis [104]. A rendezetlenség gyakori a sejtmag fehérjei között, és a membrán-nélküli organellumokban is. A rendezetlen fehérjék megnövekedett kölcsönhatási potenciálja megmutatkozott fehérje interakciós hálózatok vizsgálatán keresztül is [105,106]. Az alternatív splicing révén létrejövő izoformák rendezetlen részei hozzájárulnak a kölcsönhatási hálózatok szövetspecifikus újrarahuzalozásához is [107]. Egy fontos eredmény volt annak megmutatása, hogy a rendezetlen fehérjék szintje szorosan szabályozott, transzkripciós, RNS- és fehérje szinten is [108].

A rendezetlen fehérjék speciális funkcionális és szerkezeti tulajdonságai evolúciós tulajdonságaikban is tükröződnek. Általánosságban, a rendezetlen fehérjék esetén a szerkezeti megkötések hiánya nagyobb evolúciós változékonyságot tesz lehetővé, ami megmutatkozik magasabb evolúciós rátájukban is [108,109]. A részletesebb vizsgálatok azonban ennél árnyaltabb képet mutattak, három lehetséges scenáriót felvázolva [110]. Az első esetben sem a rendezetlenség, sem a szekvencia nem konzervált, ami a faj illetve kládspecifikus funkcionális modulok megjelenéséhez társul. A második esetben a szekvencia nem konzervált, de a rendezetlenség megmarad. Ugyanakkor a rendezetlen részekben a szekvencia illesztések miatti nehézségek miatt gyakran rejtve marad, hogy ezeken belül is megjelenhetnek egyéb specifikus molekuláris jellemzők amelyek megőrződnek az evolúció során, például lineáris motívumok kulcspozíciói, vagy poszttranszlációs módosítások helyei [111]. A harmadik scenárió az, amikor a rendezetlen régiókra jól detektálható, szekvenciálisan erős konzerváltság jellemző. Ennek példaként több olyan lineáris motívum ismert, melyek akár az összes eukarióta fajon átívelő konzerváltságot mutat [112], de találhatunk konzervált rendezetlen szegmenseket a PFAM szekvenciacsaldók között is [113]. Ezeknek a konzervált rendezetlen szakaszoknak az egyik lehetséges funkciója a DNS- és RNS-kötés, de több szomszédos és/vagy egymásba ágyazott lineáris motívum is alkothat domén-méretű evolúciósan konzervált funkcionális modult [110,114].

Össességében, a rendezetlen szegmensek aminosavainak esetében az evolúciós kényszerek ugyanolyan erősek lehetnek, mint a globuláris domének funkcionális pozíciói esetében, ami alátámasztja alapvető biológiai jelentőségüket.

A rendezetlen fehérjék funkcionális jelentősége alapján feltételezhető, hogy hibás működésük komoly biológiai következményekkel jár [115]. Ezt a kapcsolatot erősítik egyedi példák is, például az  $\alpha$ -szinuklein neurodegeneratív betegségekben, a p53 a ráktípusok széles skáláján, vagy a CFTR fehérje a cisztikus fibrózis betegségben betöltött szerepe [116]. Általánosabb szinten is, a rendezetlen fehérjék nagyobb arányát figyelték meg a rákhoz kapcsolódó fehérjék között. A SwissProt adatbázis adatai alapján a rákhoz társítható humán fehérjék 79%-a volt rendezetlen, szemben az összes eukarióta fehérje 47%-ával [117]. Hasonló eredményre vezetett egyéb betegségek vizsgálata is [118]. A rák és a rendezetlenség közötti kapcsolatot erősítette rák különböző formáiban gyakori kromoszóma átrendeződések vizsgálata, melyek gyakran érintenek rendezetlen régiókat [119]. A rendezetlen fehérjék nagyobb arányát figyelték meg a dózisérzékeny fehérjék között [120], összhangban azzal, hogy a rendezetlen fehérjék, kölcsönhatási promiskuitásuk és fázisreparációban játszott szerepük miatt, sokkal érzékenyebbek lehetnek a fehérjeszint változására [120,121]. Megmutatták, hogy a rákos mutációk előfordulhatnak rövid lineáris motívumokon belül [122], és leírtak egy konkrét esetet, ahol a mutáció révén kialakult új kölcsönhatási motívum vezetett tumorigenezishez [123]. Mindezen vizsgálatok ellenére, néhány kivételtől eltekintve, még most sem ismerjük pontosan a rendezetlen fehérjék szerepét a rák kialakulásában. Ennek a kapcsolatnak a feltérképezését új szintre emelheti a rák genom projektek révén egyre nagyobb mértékben rendelkezésre álló mutációs adatok vizsgálata [124].



### 3 CÉLKITŰZÉSEK

Kutatásom során alapvető célkitűzésem a rendezetlenség vizsgálatával kapcsolatosan új bioinformatikai eszközök kifejlesztése volt, illetve ezek alkalmazása a rendezetlen fehérjék funkcionális és betegséggel kapcsolatos tulajdonságainak jobb megértése érdekében. Ezeket, a közlemények időrendi megjelenését nagyjából követve, 4 főbb célkitűzésben foglaltam össze:

1. Rendezetlenség predikciós eljárás kifejlesztése egy újfajta energia becselő eljárás alapján
  - a. Az energiabecslő eljárás kifejlesztése és alkalmazása rendezett és rendezetlen fehérje szekvenciák megkülönböztetésére
  - b. Az energiabecslés alapján a rendezetlenség predikcióra szolgáló IUPred módszer kifejlesztése
  - c. Az IUPred módszer elérhetővé tétele webszerveren keresztül
2. A rendezetlen fehérjék kölcsönhatásainak jellemzése az interaktóm és a komplexek szintjén
  - a. A rendezetlenség szerepének vizsgálata a fehérje kölcsönhatási hálózatok központi fehérjeiben
  - b. A rendezetlen fehérje kölcsönhatások molekuláris alapelveinek feltárása a rendezetlen és rendezett fehérjék közötti komplexek szerkezetének elemzése révén
3. Predikciós módszer kifejlesztése a rendezetlen kötőhelyek felismerésére
  - a. A rendezetlen kötőhelyek predikciójára szolgáló ANCHOR módszer kifejlesztése
  - b. Az energiabecslésen alapuló IUPred és ANCHOR módszerek elérésének biztosítása modern webszerver és programcsomag formájában
4. A rák és a rendezetlenség kapcsolatának vizsgálata mutációs adatok alapján
  - a. Különböző mutációk eloszlásának vizsgálata rendezett és rendezetlen fehérje szegmensek között
  - b. A rákban gyakran mutálódó rendezetlen fehérje szegmensek azonosítása és ezek funkcionális és rendszer-szintű tulajdonságainak elemzése
  - c. A rákban mutálódó rendezetlen fehérje szegmensek evolúció eredetének vizsgálata

## 4 MÓDSZEREK

Kutatásaimhoz a szerkezeti bioinformatika széles eszköztárát használtam, többek között szekvencia illesztést és keresést, illetve különböző szekvencia alapú predikciós módszereket. Nagyban támaszkodtam a fehérjék központi adatbázisára, a UniProt-ra [125], a fehérje térszerkezeti adatbázisra, a PDB-re [126], illetve különböző rendezetlenséghez kapcsolódó adatbázisokra, köztük a DisProt adatbázisra [52]. A rendszerszintű vizsgálatokhoz elsősorban GO annotációkat [127] és fehérje-fehérje kölcsönhatási adatbázisokat, például az IntAct [128] adatait használtam fel. A mutációs adatokhoz a COSMIC adatbázist [129] és a UCSC Genome Browsert használtam [130]. Elemzéseinkhez alapvetően saját készítésű programokat használtunk, melyeket C, Perl és PYTHON nyelven írtam illetve írtunk. Eredményeim jelentős részét képezi predikciós módszerek kifejlesztése, aminek során nagy gondot fordítottam a megfelelő, független adatokon történő tesztelésre. A predikciós módszerek alkalmazásával kapott eredmények megerősítése érdekében gyakran alkalmaztam több különböző programot, vagy ezek konszenzusát. A bioinformatikai elemzések során kapott általános megfigyeléseket igyekeztem egyedi példákon és irodalmi adatokon keresztül is alátámasztani. Fontos küldetésünknek tekintjük, hogy az általunk készített módszereket és adatszeteket hozzáférhetővé tegyük a tudományos közösség számára is, ezért több webszerver és adatbázis kifejlesztésében is részt vettünk, az itt közölt eredményeken túlmenően is [28,52,131–133]. Az alkalmazott módszerek és eszközök részletes leírása a csatolt közleményekben található.

## 5 EREDMÉNYEK

Az itt bemutatott eredmények majdnem húsz év munkássága során elért legfontosabb eredményeit foglalják össze a rendezetlen fehérjék bioinformatikai vizsgálatának területén. Ezek alapjául 11 eredeti közlemény szolgált, melyek mindegyikében meghatározó (első vagy utolsó) szerző voltam. Az egyes alfejezetek mögötti számok jelölik a saját közlemények megfelelő hivatkozását.

### 5.1 Rendezetlenség predikció

#### 5.1.1 Az energia becslő eljárás (1)

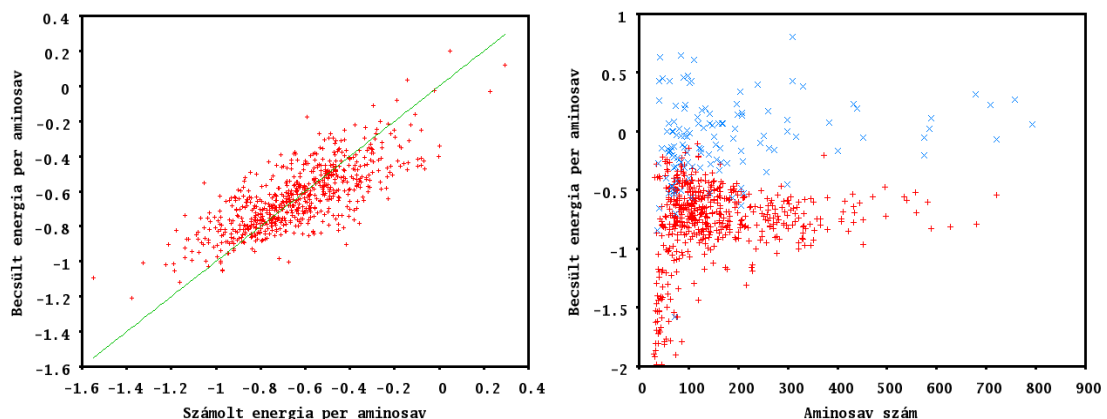
A rendezetlenség predikciók kezdeti időszakában még sok bizonytalanság övezte az újonnan felfedezett jelenséget, és viszonylag kisszámú, zajos halmaz állt csak rendelkezésre vizsgálatukhoz, ami nagyban megnehezítette megfelelő predikciós módszerek készítését. Éppen ezért megközelítésünk egyik nagy előnye volt, hogy mi a globuláris fehérjék alapvető tulajdonságaiból indultunk ki. Ezeknek a szerkezeteknek a létrejötte feltételezi a láncon belüli energetikailag kedvező kölcsönhatások kialakítását, amelyek összességében képesek kompenzálni a fehérje feltekeredéssel együtt járó entrópia veszteséget. Azonban nem minden aminosav szekvencia képes egy ilyen jól definiált szerkezet kialakítására. A mi alapvető feltételezésünk az volt, hogy azok a fehérjék, vagy fehérje szegmensek, amelyek olyan aminosavból állnak, amelyek nem képesek elegendő energetikailag kedvező kölcsönhatás kialakítására, azok rendezetlenek lesznek.

Az energia értékeket ismert fehérje térszerkezetekből számoltuk durvaszemcsés/ alacsony felbontású megközelítést használva, ahol az egyes aminosav párokhoz rendelünk kölcsönhatási értékeket. A párkölcsönhatási energia számoláshoz Thomas és Dill algoritmust használtuk [41]. A kapott 20x20-as energiamátrix értékei jellemzik, hogy mennyire szeret két aminosav kontaktust kialakítani az ismert térszerkezetekben. Ezek az energia-jellegű mennyiségek általában jól leírják az olyan alapvető összefüggéseket, mint például hogy a hidrofób aminosavak gyakrabban alkotnak kölcsönhatásokat a szerkezeten belül, vagy hogy az azonosan töltött aminosavak kontaktusa energetikailag kedvezőtlen. Ezekről a tényezőkről ismert volt, hogy fontosak a rendezett és rendezetlen fehérjék megkülönböztetésében [65]. Az eredeti algoritmus alapján az energia értékeket újraszámoltuk egy, az eredetinel jóval nagyobb adatbázison. Ez alapján a szerkezet minden pozíciójához rendelhetünk egy párkölcsönhatási energiát az aminosav szekvencia és a hozzá tartozó konformáció függvényében, összegezve a statisztikus potenciál mátrix megfelelő elemeit a kontaktusban lévő aminosavakra.

A statisztikai potenciálok alkalmazása során a számítások mindig egy meghatározott konformációra vonatkoztak. Azért, hogy túl tudjunk lépni ezen a megkötésen, kidolgoztunk egy energiabecslő eljárást, amely közvetlenül a

szekvenciából képes meghatározni egy adott szekvenciához tartozó párkölcsönhatási energiát. Az energiabecslő eljárás során azzal a durva feltételezéssel élünk, hogy egy adott pozíció energiáját alapvetően az adott aminosav és a szekvenciális környezetében lévő aminosavak típusa határozza meg. A feltételezésünk szerint, hogy ha egy adott fehérje szekvenciája olyan aminosavakból áll, melyek kedvezőbb kölcsönhatásokat tudnak kialakítani a szekvenciális környezetükben lévő más aminosavval, akkor annak várhatóan kedvezőbb lesz az energiája. Matematikailag a legegyszerűbb formula, ami ezt le tudja írni az egy kvadrátikus kifejezés az aminosav összetételre nézve, aminek a kulcsa egy 20x20-as energia prediktor mátrix. Ennek elemei kapcsolatot teremtenek az aminosav összetétel vektor tagjai és a várható energia között az aminosav típustól függően. Az energia prediktor mátrix elemeit a legkisebb négyzetek módszerével határoztuk meg, minimalizálva a szerkezetből számolt energiák és a szekvenciából becsült energiák eltérését az adatbázisban lévő fehérjék összes pozícióját figyelembe véve.

A szekvencia és az energia prediktor mátrix alapján tetszőleges fehérje energiáját meg tudjuk becsülni. Megmutattuk, hogy az ismert szerkezettel rendelkező fehérjék esetén a számolt és a becsült energiák jól korreláltak egymással ( $r^2$  érték 0.58, korreláció 0.76) (4. ábra). Továbbá azt is igazoltuk, hogy fehérje szinten a rendezett fehérjék becsült energiája általában kedvező (negatív) volt, ehhez képest a rendezetlen fehérjékre az esetek döntő többségében magasabb energia volt jellemző (4. ábra). Ezek az eredmények alátámasztották a feltételezést, hogy a rendezett és rendezetlen fehérjék megkülönböztethetők a becsült párkölcsönhatási energiájuk alapján, és rámutatott a fehérje rendezetlenség fizika alapjára.



4. ábra. A globuláris fehérjék szerkezete alapján számolt és a szekvenciájuk alapján becsült energiák korrelációja (bal). A becsült energia a globuláris (piros) és rendezetlen (kék) fehérjékre szekvenciahosszuk függvényében. Minden pont egy fehérjét jelöl.

### 5.1.2 Az IUPred módszer (1)

Az energiabecslő eljárás gyakorlati használhatóságához fontos volt, hogy predikció pozíció specifikus legyen. Ennek érdekében kismértékben módosítottuk az eljárást. Mivel sok fehérje nem teljesen rendezett vagy rendezetlen, minden pozíciót külön tekintettünk és az aminosav összetételt is külön számoltuk minden pozícióra, a 2-100 szekvenciális távolságban lévő aminosavat figyelembe véve, ami nagyjából megfelel egy átlagos domén méretnek. Ezek alapján újraszámoltuk az energiabecslő mátrixot is. Az így kapott energia értékeket simítottuk egy 21-es ablakkal. A globuláris fehérjék eloszlásából meghatároztuk az 5%-os fals pozitív predikciós értéket, vagyis azt az értéket, ahol a rendezett halmazból a pozíciók 5%-a rendelkezett ennél magasabb becült energia értékkel. Az ennél magasabb becült energiával rendelkező pozíciókat tekintettük rendezetlennek. A két eloszlás alapján a becült energia értékeket 0 és 1 közötti értékévé konvertáltuk úgy, hogy a 0,5 feleljen meg a 5%-os fals pozitív értéknek. Az itt leírt módszert IUPrednek neveztük el.

Az IUPred módszer a rendezetlen fehérje pozíciók 76%-át jósolta helyesen rendezetlennek. Az általunk végzett tesztelés megerősítést nyert számos független értékelés során is [72,134]. A módszer összességében mindössze 212 paraméterre épül (a szimmetrikus energia prediktor mátrix 210 tagja, plusz a szekvenciális összetétel és a simítás ablakmérete). A módszer egyik legfőbb erőssége, hogy a paraméterek pusztán globuláris fehérje szerkezetekből lettek származtatva, rendezetlen fehérjékre vonatkozó információk felhasználása nélkül.

Az IUPred módszer kifejlesztésénél a fő cél a hosszabb, nagy valószínűséggel biológiai funkcióval rendelkező rendezetlen szegmensek azonosítása volt. Azonban a módszer egyéb specifikus alkalmazásokat is lehetővé tesz, a paraméterek kismértékű módosításával. Az egyik ilyen változat az úgynevezett rövid rendezetlen régiók azonosítására szolgál, mint amilyenek például a röntgen szerkezetekből hiányzó régiók. Az ilyen szegmensek azonosítására csak a legfeljebb 25 aminosavra lévő aminosavakból álló szekvenciális környezetet vettük figyelembe, és ennek megfelelően módosítottuk az energia predikciós mátrixot. Az általános megfigyelések szerint a rövid rendezetlen régiók nagyobb valószínűséggel fordulnak elő a terminális régiókban, ezt egy új taggal vettük figyelembe, ami megnöveli a rendezetlenségi tendenciát a szekvencia végeken. Egy másik lehetséges alkalmazási terület a szerkezet meghatározásnál a lehetséges targetek kiválasztása. Ennek során a nagyobb szerkezeti egységek, elsősorban globuláris domének azonosítása a cél a predikciós profilból. Ehhez először azonosítjuk a tisztán rendezetlen és rendezett szegmenseket, majd a szomszédos régiókat összevonjuk, illetve a 30 aminosavnál rövidebb globuláris részeket figyelmen kívül hagyjuk.

Az IUPred módszer mind a mai napig az egyik legnépszerűbb rendezetlenség predikciós módszer. Ennek oka abban kereshető, hogy gyorsan, viszonylag megbízható predikciókat képes generálni. A módszer beépült számos egyéb web-szerverbe, például PFAM, ELM és PDB adatbázisokba és részét képezi több konszenzus rendezetlenség jósoló módszernek is [135]. Az elmúlt években számos új módszert fejlesztettek ki a

rendezetlenség jóslására, ezek azonban általában több nagyságrenddel lassabbak, és ezáltal is csak kismértékű javulást lehet elérni [81,136].

### 5.1.3 Az IUPred webszerver (2)

Az IUPred módszer sikeréhez az is nagymértékben hozzájárult, hogy könnyen elérhetővé tettük webes felületeken és programcsomag formájában is. Az IUPred webszerver eredeti verziója a <http://iupred.enzim.hu> oldalon volt elérhető. Magát a programot C programnyelven írtam, a webszervert pedig PHP nyelven készítettem el. A webszerver bemenete egy fehérje szekvencia (FASTA formátumban, vagy csak a sima szekvencia). A kimenet minden egyes pozícióban lévő aminosavhoz rendel egy értéket ami jellemzi annak rendezetlenségre való hajlamát. Ez az érték 0 (teljesen rendezett) és 1 (teljesen rendezetlen) között lehet, általánosságban 0,5-ös érték felett tekintünk egy aminosavat rendezetlennek. Három, kis mértékben eltérő paramétereket használó opció közül választhat a felhasználó (hosszú, rövid és rendezett domén). Az alapbeállítás egyszerű szöveges (text) kimenet, de a grafikus megjelenítés is választható. Ezt a beadott szekvenciára a szerver azonnal legenerálja a JpGraph software (JpGraph, 2005) program segítségével. Elkészítettem a programcsomag letölthető verzióját is.

## 5.2 A rendezetlen fehérjék kölcsönhatásainak jellemzése

### 5.2.1 A rendezetlenség szerepe a fehérje kölcsönhatási hálózatok központi fehérjeiben (3)

Korábbi megfigyelések arra utaltak, hogy a rendezetlen fehérjék egyik fő funkciója fehérje-fehérje kölcsönhatások kialakítása. Számos olyan példa ismert, ahol a rendezetlen fehérje szegmens felelős a kölcsönhatás kialakításáért, akár több különböző partnerrel is. Felmerült a kérdés, hogy van-e összefüggés egy fehérje kölcsönhatásainak száma és szekvenciális tulajdonságai, mint például fehérje rendezetlenség, vagy alacsony komplexitású régiók megléte között. Munkánkban négy faj, a *Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Caenorhabditis elegans* és *Homo sapiens* publikusan elérhető fehérje interakciós hálózatára vonatkozó adatok alapján vizsgáltuk szisztematikusan ezt a kérdést. A fehérje rendezetlenség jellemzésére az IUPred szoftvert használtuk. Emellett vizsgáltuk az alacsony komplexitású szegmensek (SEG módszerrel [137]), illetve az ismétlődő szekvencia elemek meglétét [138].

A fehérje kölcsönhatási hálózatok általában skálafüggetlen viselkedést mutatnak, ahol a partnerek száma, az úgynevezett fokszám eloszlás, hatványfüggvényt követ [139]. Az ilyen típusú hálózatokban a legtöbb csomópont néhány magas fokszámú csomóponthoz (hubokhoz) kapcsolódik, amelyek központi szereppel bírnak a hálózat szerveződésében. A mi eredményeink is igazolták, hogy a négy faj esetén a

kölcsönhatási hálózat jól közelíti a skálafüggetlen a viselkedést, és hogy viszonylag kis számú fehérje a többihez képest jóval több partnerrel rendelkezik. Legfontosabb eredményünk az volt, hogy megmutattuk, hogy ezen központi fehérjék esetében magasabb volt a rendezetlen aminosavak száma és aránya is. Valamivel kisebb mértékben, de hasonló tendencia volt megfigyelhető az alacsony komplexitású, illetve ismétlődő szekvencia elemek esetében is, illetve a hub fehérjék általában hosszabbak is voltak. Mivel a központi fehérjék azonosítása nem egyértelmű, kétfajta definíciót is használtunk a hub fehérjék azonosítására, és megmutattuk, hogy ez az eredményeket nem befolyásolja. Hasonló eredményekre jutott Haynes és szerzőtársai a PONDR VL-XT rendezetlenség jósló módszerrel [106].

Eredményeink arra mutattak, hogy a kölcsönhatási hálózatokban lévő központi fehérjék egyik meghatározó jellemzője a fehérje rendezetlenség. A hálózat biológiában általánosan elfogadott nézet szerint a fehérje kölcsönhatási hálózatok skálafüggetlen jellege génduplikáció és a preferenciális kapcsolódás eredménye [140,141]. Az általunk javasolt alternatív magyarázat a biológiai hálózatok egy kifinomultabb evolúciós modelljére épül. Eszerint a kölcsönhatási hálózatok evolúciója során előnyt jelentett az, hogy egyes fehérjék növelni tudják képességüket más fehérjékkel való kölcsönhatásra és ezáltal a hálózat szerveződéséhez kapcsolódó funkciókra specializálódnak. Ennek egyik mechanizmusa a fehérje rendezetlenségre épül. A rendezetlen fehérjék tulajdonságai, például nyújtott konformációja, ami könnyen hozzáférhető a partnerek számára, illetve flexibilitásuk és plaszticitásuk, ami lehetővé teszi, hogy alkalmazkodjanak akár különböző partner molekula kötőfelszínéhez, hozzásegíti ezeket a fehérjéket a nagyobb számú kölcsönhatás kialakításához [121].

### 5.2.2 A rendezetlen fehérje kölcsönhatások molekuláris alapelvei (4)

A fehérje kölcsönhatások molekuláris alapelveinek feltárása alapvető fontosságú a fehérjék biológiai funkciójának megértéséhez, melybe betekintést kaphatunk kialakult komplexek térszerkezetének vizsgálatával. Ennek megfelelően, több közleményben is elemezték a globuláris fehérjék kölcsönhatási felszínének tulajdonságait [142–144]. Ezekben leírták, hogy általánosságban az interfész nagysága  $1000\text{\AA}^2$ , a felszín többi részéhez képest magasabb hidrofóbicitással rendelkezik és több esetben megfigyelték evolúciósan konzervált és a kölcsönhatási energiája szempontjából kulcsfontosságú pozíciók meglétét. Geometriai szempontból eltéréseket találtak a különböző típusú komplexek között, mint például hetero- és homodimerek, vagy enzim-inhibitor komplexek, amit összefüggésbe lehetett hozni eltérő funkcionális szerepekkel. Nussinov és mtsai több különböző fehérje osztályt vizsgált, melyek között voltak már rendezetlen fehérjék is, bár rendkívül kis számban [145]. Megállapították, hogy az egy aminosavra jutó kölcsönhatási felszín nagyobb a rendezetlen fehérjék komplexeinél és hogy a hidrofób aminosavak nagyobb aránya kerül eltemetésre a kölcsönhatás során. Hasonló eredményre jutott Mohan és kollégái [53], ők azonban olyan fehérje párok szerkezetét vizsgálták ahol az egyik lánc rövid volt (kevesebb mint 70 aminosav), ami nem jelentett garanciát a rendezetlenségre.

Munkánk volt az első, ami direkt módon kísérletesen igazolt rendezetlen

fehérjéken alapult. A DisProt adatbázis [146] alapján 39 olyan kísérletesen igazolt rendezetlen fehérjét gyűjtöttünk össze, aminek ismert volt a szerkezete valamilyen másik fehérjéhez kötődve. Ezek interfészét hasonlítottuk össze 72 globuláris fehérje komplex kölcsönhatási felszínével. Elemzésünk során azt találtuk, hogy bár a két fajta komplex esetén a kölcsönhatási felszín mérete nem tért el jelentősen, rendezetlen fehérjék esetén ez jóval nagyobb részét képezte a felszínnek. A globuláris fehérjék komplexeihez képest, a rendezetlen fehérjék interfészére magasabb hidrofóbicitás volt jellemző, relatív és abszolút értelemben is. A kölcsönhatások tekintetében is a hidrofób-hidrofób kölcsönhatások domináltak a poláros aminosavak kölcsönhatásaihoz viszonyítva. A rendezetlen fehérje komplexeiben a két lánc között nagyobb számban alakultak ki atomi kontaktusok, ami arra utal, a rendezetlen fehérje jobban tud idomulni a partner fehérjéhez. A kölcsönhatás két módja abban is különbözött, hogy a rendezetlen fehérjék esetén jellemzően egyetlen folytonos szegmens alakította ki a kölcsönhatást. Ezzel szemben a globuláris fehérjék esetén több, a folding során egymáshoz közel kerülő, de a szekvencia különböző részein lévő szegmens alakítja ki a kölcsönható felszínét. A molekuláris eltérések megmutatkoztak a kölcsönhatási energiákban is. Az IUPred módszernél is használt statisztikus potenciállal számolva, a rendezetlen fehérjék sokkal több stabilizáló energiát nyertek a intermolekuláris kontaktusok révén mint a folding során. Összességében, a komplexek két osztálya közötti eltérések arra utalnak, hogy molekuláris felismerés alapelvei eltérőek a rendezetlen fehérjék által kialakított kölcsönhatások esetében.

## 5.3 A rendezetlen kötőhelyek predikciója

### 5.3.1 Az ANCHOR módszer (5, 6)

A rendezetlen fehérjék komplexeinek speciális fiziko-kémiai és szegmentációs tulajdonságai előrevetítették, hogy a rendezetlen fehérjék esetén lehetséges a kölcsönhatásban résztvevő aminosavak predikciója a szekvenciából. Megközelítésünk ezeket a speciális biofizikai tulajdonságokat próbálja megfogni a IUPred módszerben is használt energiabecslő eljárás alapján. A rendezetlen kötőrégiók egyik fő jellemzője, hogy azok egy alapvetően rendezetlen részen belül találhatóak. Ez a tulajdonság közvetlenül jellemezhető az IUPred módszer alapján. A másik fő jellemzőjük, hogy bár saját szekvenciális környezetükkel nem, de speciális globuláris fehérjéhez kötődve kedvező kölcsönhatást tudnak kialakítani. Ennek modellezéséhez meghatározhatjuk azt a becsült energiát, ami az adott aminosav a közvetlen szekvenciális környezetével tudna létrehozni, illetve azt, ami egy átlagosan globuláris fehérjéhez kötődve tudna elérni. Az általunk kifejlesztett módszer, amit ANCHOR-nak nevezünk el, ezen tényezők kombinációja alapján jósol rendezetlen kötőhelyeket a szekvenciából. A módszer továbbfejlesztett változata, az ANCHOR2 kissé eltérő architektúrára épült, amiben figyelembe vettünk egy minimális energia-nyereséget és átlagos rendezetlenségi tendenciát, amellyel egy aminosavnak rendelkeznie kell ahhoz, hogy rendezetlen



kötőhely legyen. Ezek révén jobban ki tudtuk szűrni az ANCHOR predikciónál előforduló eseteket, ahol akkor is jóslott rendezetlen kötőhelyeket, amikor a rendezetlenség kritériuma nem teljesült, de az energia nyereség nagy volt. A paraméterek egy részének optimalizálásához már támaszkodhattunk a DIBS adatbázisra is, mely nagyobb számban tartalmazott rendezetlen kötőhelyeket [28]. Ugyanakkor a paraméterek döntő többsége továbbra is a korábban kifejlesztett energiabecslő eljárás alapján alapult.

Az ANCHOR módszer a globuláris fehérjéken mért 5% fals predikciós ráta mellett, 67%-os hatékonysággal tudta jóslni a rendezetlen kötőhelyek aminosavait, de szegmens szinten a rendezetlen kötőhelyek 70%-át helyesen felismerte. Több más, gépi tanuláson alapuló módszerrel szemben, mi nem tételeztük fel, hogy a rendezetlen fehérjék nem annotált része nem tartalmaz rendezetlen kötőhelyet. Ugyanakkor módszerünk a DisProt adatbázisban annotált rendezetlen fehérje szegmensek kevesebb, mint 50%-át jósolta kötőhelynek, tehát az általános rendezetlen részekről is meg tudta különböztetni a kötőhelyeket. Megmutattuk, hogy a módszer hatékonyságát nem befolyásolja nagymértékben a kölcsönhatásban résztvevő aminosavak típusa, illetve a szerkezetben felvett másodlagos szerkezet típusa sem. Hosszabb rendezetlen részeknél általában nem a teljes szegmenst, csak azoknak egyes részeit ismeri fel, melyek általában erősebb kölcsönhatást alakítanak ki. Az ANCHOR2 módszer jóval nagyobb adatszetten hasonló eredményt ért el: 64%-os hatékonysággal tudta jóslni a rendezetlen kötőhelyek aminosavait, de szegmens szinten a rendezetlen kötőhelyek 72%-át ismerte fel helyesen.

A rendezetlenség predikciós módszerekhez képest szerényebb eredmények ellenére, az ANCHOR és ANCHOR2 módszerek a rendezetlen fehérjék funkcionális helyeinek jellemzésének alapvető eszközeivé váltak [107,115]. Ezt igazolta a CAID értékelés eredménye is, ahol a független adatokon az ANCHOR módszer bizonyult a legjobbnak a rendezetlen kötőrégiók azonosításában [81] (5. ábra).

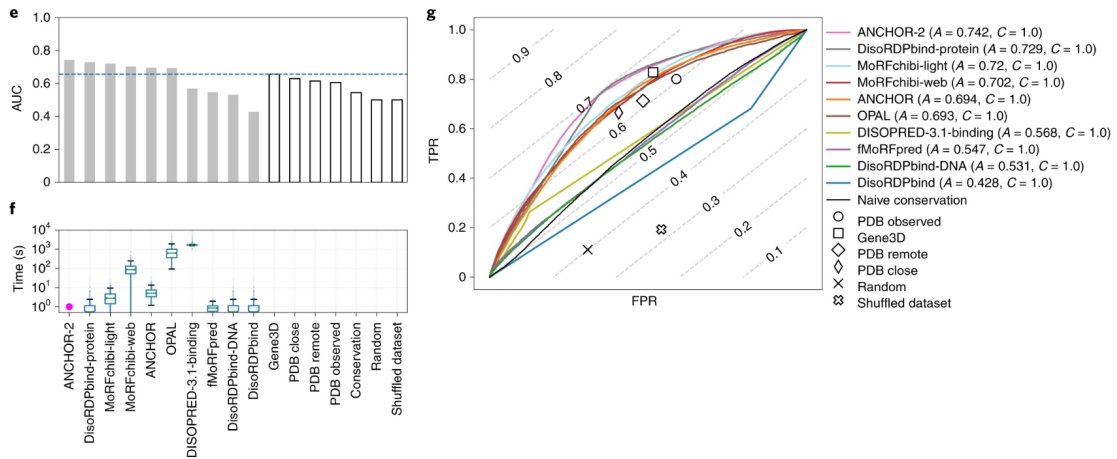


Figure 5. CAID eredménye a rendezetlen kötőhelyek predikciójára. A rendezetlen kötőrégiók predikciójának kategóriájában az ANCHOR2 és ANCHOR módszerek bizonyultak a legjobbnak [81].

### 5.3.2 Az energiabecslő eljáráson alapuló módszerek webszervere (6, 7, 8)

2017-ben a szervert átköltöztettük az ELTÉ-re, és elkészítettük a webszerverek új verzióját, az IUPred2A-t (6). Ebben egyesítettük az energiabecslésen alapuló módszereink - az IUPred és az ANCHOR - elérését. Az ANCHOR szerver eredeti verziójához, ami az <http://anchor.enzim.hu> oldalon volt elérhető, alapvetően az IUPred szerver esetében használt programokat szabtuk át, mind a bemenet, mind a kimenet kezelését tekintve (7). Az új verzióban az egyik fő változás, hogy a korábbi C, illetve Perl programnyelvek helyett áttértünk a PYTHON program nyelvre. Míg az IUPred programban csak egy kisebb hibajavítást tettünk, az ANCHOR programot teljesen átdolgoztuk, és az új ANCHOR2 verziót tettük elérhetővé. A web szervert egy további opcióval bővítettük, ami lehetővé tette redox-érzékeny rendezetlen régiók szekvencia alapú predikcióját is (6). A web szervert a Django keretrendszerben fejlesztettük, a vizualizációhoz a Bokeh PYTHON könyvtárat használtuk. A fejlesztések révén az IUPred2A egy modern megjelenésű, dinamikus web szerver lett, ami támogatja az összes jelenleg használatos web-böngészőt.

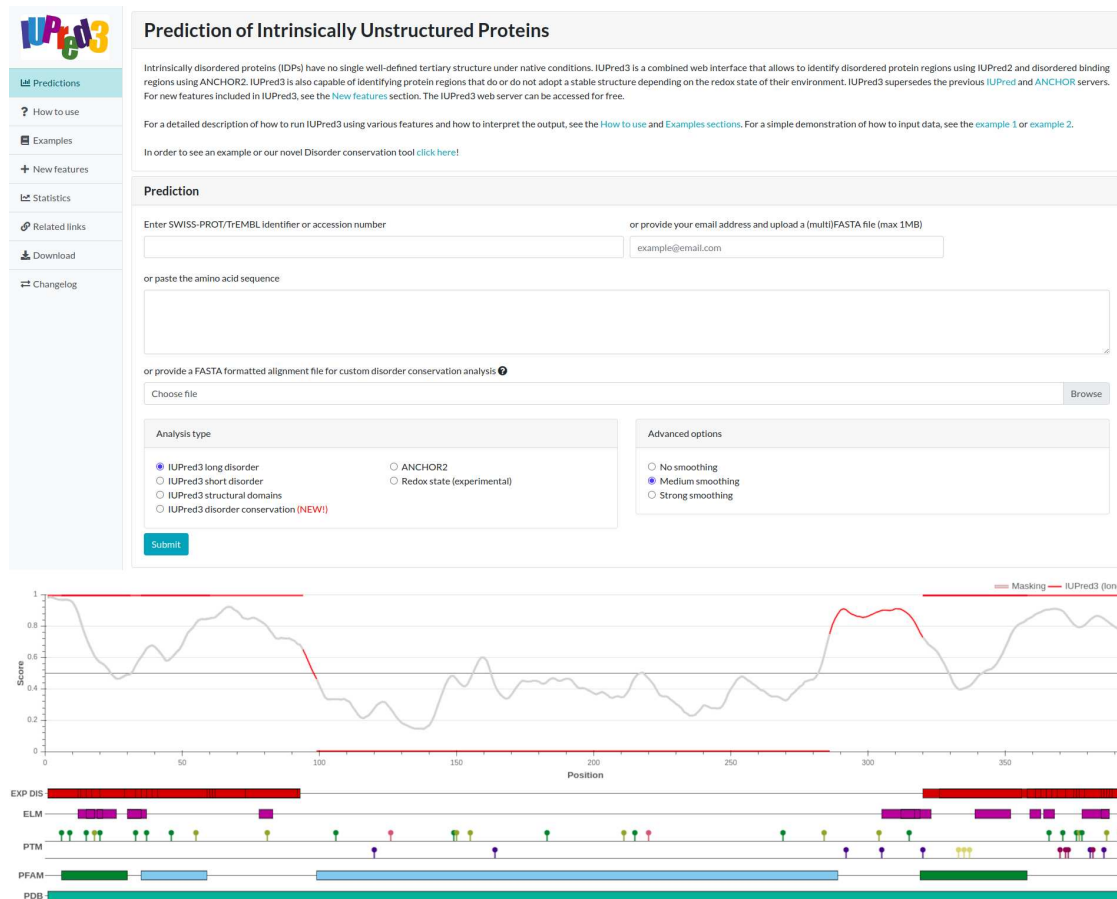


Figure 6. Az IUPred3 webszer bemeneti oldala (felül) és a predikciós kimenet a humán p53 fehérjére (UniProt azonosító: P04637) különböző annotációk ábrázolásával együtt (alul).

Új funkció volt, hogy a bemenetet nemcsak szekvencia alapján, hanem közvetlenül a Uniprot azonosító alapján is meg lehet adni. A bemeneti oldalon a felhasználó választhat a különböző predikciós opciók között, beleértve az IUPred különböző változatait és az ANCHOR módszert. Mindkét esetben a profil minden egyes pozícióhoz hozzárendel egy 0 és 1 közötti értéket, ami jellemzi annak a valószínűségét, hogy az adott aminosav rendezetlen illetve rendezetlen kötőhely része-e. A javasolt küszöbérték 0,5, e feletti pozíciókat tekintjük rendezetlennek, illetve rendezetlen kötőhelynek. Az ANCHOR esetében a kimenetnél az IUPred profilt is feltüntették. A predikciókat grafikus formában jeleníti meg a webszerver, de a vizuális megjelenítés mellett a predikciók letölthetők text és json formátumban is. A Uniprot adatbázissal való integráció révén további információk is megjelenítésre kerülnek, többek között PFAM annotációk, posztranszlációs módosítások, illetve kísérletesen igazolt rendezetlen régiók és kötőhelyek. Ezek hasznosak lehetnek a predikciós eredmények értelmezésében (6. ábra). A IUPred szerver legújabb verziójában bevezettünk simítási opciókat, illetve implementáltunk egy megjelenítőt, amelyben az adott szekvenciához tartozó szekvencia illesztés és a predikciós profil összekapcsoltan tanulmányozható eukarióta modell organizmusokban (8). A szerver legújabb verziója <https://iupred.elte.hu> illetve <https://iupred3.elte.hu> oldalakon érhető el.

## 5.4 A rendezetlenség és a rák kapcsolata

### 5.4.1 A rendezetlenség biológiai kockázata (9)

Az egyedi példák és általános elemzések megalapozták a rák és a rendezetlenség közötti kapcsolatot. Ezek alapján felmerült, hogy a rendezetlenség egyfajta biológiai kockázatot jelent [12,118]. A rendezetlenség jósló módszerek megjelenése és a különböző szekvencia variációs adatok robbanásszerű növekedése lehetővé tette ennek a kérdésnek a részletesebb vizsgálatát. Munkánkban elsők között elemeztük, hogy hogyan befolyásolják a fehérjék szerkezeti tulajdonságai a mutációk eloszlását.

Első lépésként különböző adatbázisokból gyűjtöttünk össze gyakori polimorfizmusokat, betegséggel kapcsolatos mutációkat illetve rákos mutációkat. Az összegyűjtött pontmutációkat a fehérje szekvenciákra vetítve vizsgáltuk ezek eloszlását az IUPred és egyéb rendezetlenség predikációs módszerek felhasználásával. Ennek alapján kiszámoltuk a rendezett és rendezetlen részekre eső mutációk számát, és ezeket összevetettük a várható mutációk számával, azt feltételezve, hogy azok egyenletesen oszlanak el a fehérje mentén. Eredményeink azt mutatták, hogy a neutrális polimorfizmusok gyakoribbak voltak a rendezetlen részeken, ezzel szemben a rákos mutációk sokkal inkább a rendezett részekre estek. Ez a tendencia sokkal erősebbnek mutatkozott azoknál az adatbázisoknál, melyekben a tradicionális biokémiai módszerekkel azonosított rákos mutációk domináltak (SwissProt cancer, COSMIC census adatbázisok). A rák genom projektekből származó adatok esetén ettől eltérő eredményt kaptunk, itt a mutációk kismértékben a rendezetlen részeket preferálták. Feltételezésünk szerint ez a véletlenszerűen megjelenő, úgynevezett potyautas (passenger) mutációk előfordulására vezethető vissza. Ezt igazolta, hogy ha a véletlenszerű mutációk megjelenésénél figyelembe vettük a polimorfizmusok nem egyenletes eloszlását, már minden esetben egyhangúan és statisztikailag szignifikánsan a rendezett részeken voltak a rákos mutációk túlsúlyban. A mutációk eloszlását megvizsgáltuk a rendezetlen kötőrégiókban is az ANCHOR módszerünk felhasználásával. A kötőrégiók alapvetően a rendezetlen régiókhoz hasonló tendenciát mutattak, a neutrális polimorfizmusok túl-, a rákos mutációk alul voltak reprezentálva bennük. Azonban esetükben az eltérés a globuláris részeketől kisebb volt a kötőrégióknak nem jóslt rendezetlen szegmensekhez képest, ami összhangban áll nagyobb funkcionális szerepükkel.

Elemzésünk arra is rávilágított, hogy a rákkal kapcsolatba hozott fehérjék átlagosan hosszabbak, több bennük a rendezetlenség, több kölcsönhatásban vesznek részt és speciális funkciók kapcsolódnak hozzájuk. Azonban ezek a tulajdonságok egymással is korrelálnak. A különböző tényezők között kölcsönös információt számoltunk, és megmutattuk, hogy a rendezetlenség és a rákos mutációk közötti kapcsolat indirekt, a funkción keresztül jön létre. A leggyakrabban mutálódó példák vizsgálata alapján, a p53 vagy a PTEN viselkedése összhangban van az általános trenddel. Bár a rendezetlen szegmensek alapvető funkcionális szerepet játszanak ezen

fehérjék működésében, a rákos mutációk jellemzően a rendezett részre esnek. Ugyanakkor találhatóak olyan példák is, ahol a mutációk döntő része rendezetlen részre esik. Ilyen például a  $\beta$ -catenin, vagy az APC fehérje. Hasonló példák azonosítása révén további betekintést nyerhetünk a rendezetlenség és a rák kapcsolatáról.

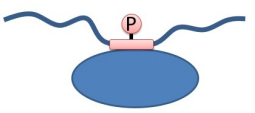
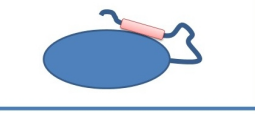
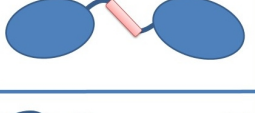


#### 5.4.2 Rákban gyakran mutálódó rendezetlen szegmensek (10)

A rák genom projektet adatainak az értelmezését nagyban megnehezíti, hogy a katalogizált mutációk döntő többsége csak véletlenszerűen előforduló potyautas mutáció, melyek gyakoribbak a rendezetlen részeken [147]. A rák kialakulásában irányító szereppel bíró, ún. irányító/driver mutációk egyik legegységelműbb jele, hogy több minta összesítése esetén a mutációk feldúsulnak a genom meghatározott szakaszán. Ez alapján több módszert is kidolgoztak rák gének azonosítására [148]. Az általunk fejlesztett iSiMPre módszer a meglévő eljárásokhoz hasonló eredményt ér el a driver gének megtalálásában, de lehetővé teszi a szignifikánsan mutálódó régiók azonosítását is, függetlenül azok méretétől vagy típusától [149].

Munkánk során a COSMIC adatbázist használtuk kiindulásként [129]. Az ebből összegyűjtött mutációkat rávetítettük a fehérje szekvenciákra, és azonosítottuk a szignifikánsan mutálódott régiókat az iSiMPre módszerrel. A kapott régiókat klasszifikáltuk rendezetlenség szerint, ehhez figyelembe vettünk kísérletes eredményeket, a domén annotációkat és a rendezetlenség profilt. Összesen 145 fehérjében találtunk 225 szignifikánsan mutálódott régiót magas szignifikancia szinttel ( $10e-6$ ). Ezek közül 47 esett rendezetlen régióra, ami kevesebb, mint amit a rendezetlenség előfordulása alapján vártunk (kb. 30%). Ez összhangban van korábbi eredményünkkel, mely szerint a rákos mutációk a rendezett részeken gyakoribbak. A legtöbb esetben a szignifikánsan mutálódó régió tartalmazta az adott fehérjére eső mutációk nagy részét, ami vagy tisztán rendezett vagy rendezetlen résszel fedett át. A találataink rákban betöltött fontos szerepét irodalmi adatok is igazolták. A rendezetlen rákos irányító gének további tulajdonságait is elemeztük molekuláris és rendszer szinten.

A rendezetlen irányító fehérjék legnagyobb csoportja rövid lineáris motívumok általi kölcsönhatások kialakításában vesz részt, és a mutációk olyan pozíciókat érintenek, amelyek kulcsfontosságúak a molekuláris felismerésben vagy annak poszttranszlációs módosítás általi szabályozásában. Azonban a rákos mutációs feldúsulása alapján egyéb típusú funkcionális modulok is fontosak lehetnek a rák szempontjából, például autoregulációs helyek, linker régiók vagy DNS és RNS kötő részek (7. ábra). Bár az azonosított példák sokfélék, közös bennünk, hogy funkciójukhoz a rendezetlenség által biztosított dinamikus szerkezeti tulajdonság alapvetően fontos. A rendezett rákos irányító fehérjékhez képest nemcsak molekuláris mechanizmus tekintetében van eltérés, hanem biológiai funkció vonatkozásában is. Ebből a szempontból kiemelkedik a fehérje degradációs rendszer, amely jellemzően a rendezetlen fehérje részeken keresztül mutálódik, például degron motívumokat érintve

[150]. A rendezetlen rákos irányító fehérjék központi szereppel bírnak a kölcsönhatási és jelátviteli hálózatokban, és perturbációjuk érintheti az ismert rákos ismertetőjegyek (cancer hallmark) [151] mindegyikét. A legtöbb betegminta tartalmazott mind rendezett mind rendezetlen részre eső mutációkat. Azonban azon minták esetében, ahol rendezetlen részre eső mutációk dominálnak, jóval kisebb az esélye hogy létezik a kezelésre jelenleg elérhető gyógyszermolekula. Ez jól mutatja, hogy nagy igény lenne olyan új megközelítésre, melyek révén a rendezetlen fehérjék is targetálhatóak lennének.

	Disordered functional unit	Tumor suppressors	Context dependent genes	Oncogenes
Linear motif / PTM		p14 <sup>ARF</sup> RSP15	EPAS1   NRF2   ESR1 FOXO1   FOXL2	CTNNB1   CCND3 MYC   MYCN   SETBP1 CD79B   MET   USP8 CSF1R   histone H3s
Auto-regulatory			EZH2	
Flexible linker			CBL	KIT   FLT3   PDGFRA
DNA/RNA binding		EIF1AX CEBPA	PAX5	FOXA1   SRSF2
Disordered domain		APC   ID3 VHL   TP53 SMARCB1	MED12	MYOD1   CARD11 CALR
Unknown	?	ASXL1   MLH1 EP300		BCL2

7. ábra. A rákban gyakran mutálódott rendezetlen fehérjék kategorizálása a funkcionális modul (lineáris motívum/PTM, autoregulációs hely, linker, DNS-RNS kötés, rendezetlen domén) típusa és rákban betöltött szerepük (tumorsuppresszor, onkogén, illetve kontextus-függő módon mindkét viselkedést mutató) alapján.

### 5.4.3 A rákban mutálódó rendezetlen régiók evolúciós eredete (11)

A rendezetlen fehérjék sajátos szerkezeti és funkcionális tulajdonságokkal rendelkeznek, és ez evolúciós tulajdonságaikban is tükröződik. A több szerkezeti megkötöttséggel rendelkező rendezett fehérjékhez képest általánosságban a rendezetlen fehérjék viszonylag friss evolúciós találmányok és szekvenciájuk nagyobb variabilitást mutat [152]. Azonban a bennük található funkcionálisan fontos régiók erősen

konzerváltak is lehetnek. Ezért felmerült a kérdés, hogy evolúciós eredetüket tekintve, mi a jellemző a rákban szignifikánsan mutálódott rendezetlen fehérje szegmensekre. Vizsgálatunkhoz a filozofiatigráfias eljárást használtuk, amely visszavezeti az egyes gének eredetét makorevolúciós átmenetekhez [153]. Korábbi vizsgálatok alapján a legtöbb rák gén megjelenése a többsejtű állatvilág megjelenéséhez kapcsolódik, de vannak még ősbibb, az egysejtű élőlények szintjére visszavezethető gének is [153]. Ezek az ősbibb esetek általában gondoskodó (caretaker) funkcióval bírnak, jellemzően a genom stabilitás biztosításában vesznek részt. A többsejtűség megjelenéséhez köthető példákra általában kapuőrző (gatekeeper) funkció köthető, melyek a sejt differenciáció, növekedés és sejthalál megfelelő működését biztosítja [154].

A korábbi vizsgálatok a teljes szekvencia szintjén történtek, amiben általában a fehérjék legkonzerváltabb régiói, a domének domináltak. Vizsgáltunk egyik fő újdonsága az volt, hogy mi nem a teljes fehérjére, hanem annak a mutációk által kitüntetett régiójára koncentráltunk. Meglepő módon, a rákban mutálódott rendezetlen fehérje régiók is ősi eredetűek voltak, a régiók többsége a legkorábbi többsejtű állatok szintjére volt visszavezethető. A legfiatalabb példa a CD79B fehérje volt, ami az immunrendszer egyéb fehérjeihez hasonlóan, a gerincesek szintjéhez volt köthető. Néhány esetben jóval ősbibb evolúciós eredetet találtunk. Ennek egyik példája az MLH1 fehérje, ami az össze nem illő bázispárok javításban játszik kulcsszerepet. A fehérje közepén található linker régiója jól ismert, azonban mi ezen belül azonosítottunk egy nagy konzerváltságot mutató funkcionális helyet, ami a rák szempontjából is fontos, azonban funkciója egészen a közelmúltig nem volt ismert [155]. Több olyan példával is találkoztunk, ahol a mutálódott régió a géncsalád eredetéhez képest később jelent meg. Létrejöttükben azonban nem gén-duplikációt követő neofunkcionalizáció volt a döntő mechanizmus, hanem nagy valószínűséggel ezek a rendezetlen részek de-novo jöttek létre. A rendezetlen régiók funkcionalitásának fontosságát támasztja alá, hogy megjelenésüket követően gyorsan rögzülnek, és a fehérje család tagjainak további duplikációja során megőrződnek. Fehérje szinten azonban, az evolúció tovább folytatódhat, megjelenhetnek újabb funkcionális modulok. Erre a példa a VHL fehérje, ami egy új N-terminális régióval bővült, vagy a humán specifikus szekvencia változások az ESR1 fehérje szabályozó funkciót betöltő régiójában. Elemzésünk világosabb képet adott a fehérjék kulcsfontosságú szabályozó elemeinek kialakulásáról, és felhívja a figyelmet a moduláris szerveződésének figyelembevételének fontosságára a fehérjék evolúciós eredetének vizsgálata során. Emellett újabb bizonyítékot adott arra vonatkozólag, hogy a rendezetlen fehérjék fontos evolúciós találmányok.

## 6 KONKLÚZIÓ

A rendezetlen fehérjék kutatása hatalmas fejlődésen ment keresztül az elmúlt két évtizedben, aminek eredményeképpen a rendezetlenség-funkció paradigma ma már szerves részévé vált a molekuláris biológiai kutatásoknak. Az egyedi fehérjék vizsgálata mellett ebben a bioinformatikai vizsgálatok is fontos szerepet játszottak. Ehhez alapvetően járultak hozzá az általam kifejlesztett predikciós módszerek és elemzések. A további kutatások egyik fő fókuszja rendezetlen fehérjék konformációs sokaságának részletes jellemzése. Szintén fontos feladat kölcsönhatásaik feltérképezése, ezen belül is új, lineáris motívumok által közvetített kölcsönhatások azonosítása, ami jelenlegi kutatásainknak is az egyik fő fókuszja. Izgalmas új terület a rendezetlen fehérjék és a fázisszeparáció kapcsolatának vizsgálata. A betegségben betöltött szerepük felismerése nyomán felmerült az igény olyan újfajta gyógyszermolekulák kifejlesztésére, melyek specifikusan rendezetlen fehérjéket céloznak meg. Ezen kérdések vizsgálatához új bioinformatikai megközelítésekre van szükség. Ezekben már várhatóan a fehérje térszerkezet predikciókban hatalmas áttörést hozó mélytanulós módszerek is döntő szerepet fognak játszani.



## 7 TÉZISEK

1. Kifejlesztettem egy újszerű energiabecslő eljárást, ami képes megbecsülni egy adott fehérjeszekvencia kompakt állapotához tartozó párkölcsönhatási energiát. Az energiabecslés alapján megmutattam, hogy a rendezetlen fehérjék becsült energiája kedvezőtlenebb, mint a rendezett fehérjéké, ami rámutatott a rendezetlenség fizikai alapjára is.
2. Az energiabecslő eljárás felhasználásával kidolgoztam az IUPred módszert, amely pozíció specifikusan jósol rendezetlenséget az aminosav szekvenciából.
3. Az IUPred módszert elérhetővé tettem webszerveren keresztül.
4. Rámutattam, hogy a fehérje kölcsönhatási hálózatok központi csomópontjaiban nagyobb arányban fordulnak elő rendezetlen fehérjék.
5. Összeállítottam egy, a rendezetlen fehérjék komplexeinek szerkezetét tartalmazó adatszettet és összevettem ezen komplexek tulajdonságait globuláris fehérjék komplexeinek jellemzőivel. Ez alapján feltártam a rendezetlen fehérje komplexek alapvető szerkezeti sajátosságait.
6. Az energiabecslő eljárás alapján kifejlesztettem egy új módszert, amellyel az aminosav szekvenciából jósolhatóak azok a rendezetlen szegmensek, amelyek más fehérjékhez kötődnek és eközben rendezett szerkezetet alakítanak ki.
7. Az energiabecslésen alapuló szekvencia predikciós módszereimet elérhetővé tettem modernizált formában webszerverként és letölthető programcsomagként is.
8. Mutációk vizsgálata alapján megmutattam, hogy a rendezetlen fehérje szegmenseken gyakoribbak a neutrális polimorfizmusok, de ritkábbak bennük a rákos mutációk a globuláris részekhez képest.
9. Azonosítottam olyan rendezetlen fehérje szegmenseket, amelyek nagyszámú rákos mutációt tartalmaztak, és ez alapján várhatóan aktív szerepet játszanak a rák kialakulásában. Elemeztem ezek jellemzőit funkcionális és rendszer szinten.
10. Evolúciós vizsgálatok alapján rámutattam arra, hogy bár a rendezetlen fehérjék általában kevésbé konzerváltak, a rákban mutálódó példák nagymértékű evolúciós konzerváltságot mutattak.

## 8 HIVATKOZÁSOK

1. Dill KA, MacCallum JL. The protein-folding problem, 50 years on. *Science*. 2012;338: 1042–1046.
2. Janin J, Sternberg MJE. Protein flexibility, not disorder, is intrinsic to molecular recognition. *F1000 Biol Rep*. 2013;5: 2.
3. Wright PE, Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol*. 1999;293: 321–331.
4. Romero P, Obradovic Z, Kissinger CR, Villafranca JE, Garner E, Guillot S, et al. Thousands of proteins likely to have long disordered regions. *Pac Symp Biocomput*. 1998; 437–448.
5. Xie Q, Arnold GE, Romero P, Obradovic Z, Garner E, Dunker AK. The Sequence Attribute Method for Determining Relationships Between Sequence and Protein Disorder. *Genome Inform Ser Workshop Genome Inform*. 1998;9: 193–200.
6. Tompa P. Intrinsically unstructured proteins. *Trends Biochem Sci*. 2002;27: 527–533.
7. Dunker AK, Babu MM, Barbar E, Blackledge M, Bondos SE, Dosztányi Z, et al. What's in a name? Why these proteins are intrinsically disordered: Why these proteins are intrinsically disordered. *Intrinsically Disord Proteins*. 2013;1: e24157.
8. Frauenfelder H, Chen G, Berendzen J, Fenimore PW, Jansson H, McMahon BH, et al. A unified model of protein dynamics. *Proc Natl Acad Sci U S A*. 2009;106: 5129–5134.
9. Marsh JA, Forman-Kay JD. Ensemble modeling of protein disordered states: experimental restraint contributions and validation. *Proteins*. 2012;80: 556–572.
10. van der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, et al. Classification of intrinsically disordered regions and proteins. *Chem Rev*. 2014;114: 6589–6631.
11. Jakob U, Kriwacki R, Uversky VN. Conditionally and transiently disordered proteins: awakening cryptic disorder to regulate protein function. *Chem Rev*. 2014;114: 6779–6805.
12. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol*. 2005;6: 197–208.
13. Dyson HJ, Wright PE. Nuclear magnetic resonance methods for elucidation of structure and dynamics in disordered states. *Methods Enzymol*. 2001;339: 258–270.
14. Sormanni P, Piovesan D, Heller GT, Bonomi M, Kukic P, Camilloni C, et al. Simultaneous quantification of protein order and disorder. *Nat Chem Biol*. 2017;13: 339–342.
15. Dyson HJ, Wright PE. Perspective: the essential role of NMR in the discovery and characterization of intrinsically disordered proteins. *J Biomol NMR*. 2019;73: 651–659.
16. Lazar T, Martínez-Pérez E, Quaglia F, Hatos A, Chemes LB, Iserle JA, et al. PED in 2021: a major update of the protein ensemble database for intrinsically disordered proteins. *Nucleic Acids Res*. 2021;49: D404–D411.
17. Wells M, Tidow H, Rutherford TJ, Markwick P, Jensen MR, Mylonas E, et al. Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain. *Proc Natl Acad Sci U S A*. 2008;105: 5762–5767.
18. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol*. 2004;337: 635–645.
19. Tompa P. The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett*. 2005;579: 3346–3354.
20. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradović Z. Intrinsic disorder and protein function. *Biochemistry*. 2002;41: 6573–6582.

21. Piovesan D, Tabaro F, Mičetić I, Necci M, Quaglia F, Oldfield CJ, et al. DisProt 7.0: a major update of the database of disordered proteins. *Nucleic Acids Res.* 2017;45: D219–D227.
22. Buljan M, Chalancon G, Dunker AK, Bateman A, Balaji S, Fuxreiter M, et al. Alternative splicing of intrinsically disordered regions and rewiring of protein interactions. *Curr Opin Struct Biol.* 2013;23: 443–450.
23. Oldfield CJ, Meng J, Yang JY, Yang MQ, Uversky VN, Dunker AK. Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. *BMC Genomics.* 2008;9 Suppl 1: S1.
24. Van Roey K, Uyar B, Weatheritt RJ, Dinkel H, Seiler M, Budd A, et al. Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation. *Chem Rev.* 2014;114: 6733–6778.
25. Dyson HJ, Wright PE. Coupling of folding and binding for unstructured proteins. *Curr Opin Struct Biol.* 2002;12: 54–60.
26. Dunker AK, Garner E, Guilliot S, Romero P, Albrecht K, Hart J, et al. Protein disorder and the evolution of molecular recognition: theory, predictions and observations. *Pac Symp Biocomput.* 1998; 473–484.
27. Tompa P, Fuxreiter M. Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem Sci.* 2008;33: 2–8.
28. Schad E, Fichó E, Pancsa R, Simon I, Dosztányi Z, Mészáros B. DIBS: a repository of disordered binding sites mediating interactions with ordered proteins. *Bioinformatics.* 2018;34: 535–537.
29. Lazar T, Tantos A, Tompa P, Schad E. Intrinsic protein disorder uncouples affinity from binding specificity. *Protein Sci.* 2022;31: e4455.
30. Van Roey K, Gibson TJ, Davey NE. Motif switches: decision-making in cell regulation. *Curr Opin Struct Biol.* 2012;22: 378–385.
31. Gibson TJ. Cell regulation: determined to signal discrete cooperation. *Trends Biochem Sci.* 2009;34: 471–482.
32. Burley SK, Bhikadiya C, Bi C, Bittrich S, Chen L, Crichlow GV, et al. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.* 2021;49: D437–D451.
33. Anfinsen CB. Principles that govern the folding of protein chains. *Science.* 1973;181: 223–230.
34. Dill KA, Ozkan SB, Shell MS, Weikl TR. The protein folding problem. *Annu Rev Biophys.* 2008;37: 289–316.
35. Bernhofer M, Dallago C, Karl T, Satagopam V, Heinzinger M, Littmann M, et al. PredictProtein - Predicting Protein Structure and Function for 29 Years. *Nucleic Acids Res.* 2021;49: W535–W540.
36. Rost B. PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol.* 1996;266: 525–539.
37. Przybylski D, Rost B. Alignments grow, secondary structure prediction improves. *Proteins.* 2002;46: 197–205.
38. Sippl MJ. Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J Comput Aided Mol Des.* 1993;7: 473–501.
39. Novotný J, Brucoleri R, Karplus M. An analysis of incorrectly folded protein models. Implications for structure predictions. *J Mol Biol.* 1984;177: 787–818.
40. Thomas PD, Dill KA. Statistical potentials extracted from protein structures: how accurate are they? *J Mol Biol.* 1996;257: 457–469.
41. Thomas PD, Dill KA. An iterative method for extracting energy-like quantities from protein structures. *Proc Natl Acad Sci U S A.* 1996;93: 11628–11633.
42. Torda AE. Perspectives in protein-fold recognition. *Curr Opin Struct Biol.* 1997;7: 200–205.

43. Shen M-Y, Sali A. Statistical potential for assessment and prediction of protein structures. *Protein Sci.* 2006;15: 2507–2524.
44. Kuhlman B, Bradley P. Advances in protein structure prediction and design. *Nat Rev Mol Cell Biol.* 2019;20: 681–697.
45. Moulton J, Pedersen JT, Judson R, Fidelis K. A large-scale experiment to assess protein structure prediction methods. *Proteins.* 1995;23: ii–v.
46. Moulton J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol.* 2005;15: 285–289.
47. Kryshchuk A, Venclovas C, Fidelis K, Moulton J. Progress over the first decade of CASP experiments. *Proteins.* 2005;61 Suppl 7: 225–236.
48. Kryshchuk A, Schwede T, Topf M, Fidelis K, Moulton J. Critical assessment of methods of protein structure prediction (CASP)-Round XIV. *Proteins.* 2021;89: 1607–1617.
49. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature.* 2021;596: 583–589.
50. Ota M, Koike R, Amemiya T, Tenno T, Romero PR, Hiroaki H, et al. An assignment of intrinsically disordered regions of proteins based on NMR structures. *J Struct Biol.* 2013;181: 29–36.
51. Vucetic S, Obradovic Z, Vacic V, Radivojac P, Peng K, Iakoucheva LM, et al. DisProt: a database of protein disorder. *Bioinformatics.* 2005;21: 137–140.
52. Quaglia F, Mészáros B, Salladini E, Hatos A, Pancsa R, Chemes LB, et al. DisProt in 2022: improved quality and accessibility of protein intrinsic disorder annotation. *Nucleic Acids Res.* 2022;50: D480–D487.
53. Mohan A, Oldfield CJ, Radivojac P, Vacic V, Cortese MS, Dunker AK, et al. Analysis of molecular recognition features (MoRFs). *J Mol Biol.* 2006;362: 1043–1059.
54. Oldfield CJ, Cheng Y, Cortese MS, Romero P, Uversky VN, Dunker AK. Coupled folding and binding with alpha-helix-forming molecular recognition elements. *Biochemistry.* 2005;44: 12454–12470.
55. Mészáros B, Simon I, Dosztányi Z. Prediction of protein binding regions in disordered proteins. *PLoS Comput Biol.* 2009;5: e1000376.
56. Fichó E, Reményi I, Simon I, Mészáros B. MFIB: a repository of protein complexes with mutual folding induced by binding. *Bioinformatics.* 2017;33: 3682–3684.
57. Hatos A, Monzon AM, Tosatto SCE, Piovesan D, Fuxreiter M. FuzDB: a new phase in understanding fuzzy interactions. *Nucleic Acids Res.* 2022;50: D509–D517.
58. Kumar M, Michael S, Alvarado-Valverde J, Mészáros B, Sámano-Sánchez H, Zeke A, et al. The Eukaryotic Linear Motif resource: 2022 release. *Nucleic Acids Res.* 2022;50: D497–D508.
59. Dosztányi Z, Csizmók V, Tompa P, Simon I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol.* 2005;347: 827–839.
60. Mészáros B, Erdos G, Dosztányi Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* 2018;46: W329–W337.
61. Trabuco LG, Betts MJ, Russell RB. Negative protein-protein interaction datasets derived from large-scale two-hybrid experiments. *Methods.* 2012;58: 343–348.
62. Tompa P, Davey NE, Gibson TJ, Babu MM. A million peptide motifs for the molecular biologist. *Mol Cell.* 2014;55: 161–169.
63. Campen A, Williams RM, Brown CJ, Meng J, Uversky VN, Dunker AK. TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. *Protein Pept Lett.* 2008;15: 956–963.
64. Erdős G, Dosztányi Z. Prediction of protein structure and intrinsic disorder in the era of deep learning. *Structure and Intrinsic Disorder in Enzymology.* Academic Press; 2023. pp. 199–224.
65. Uversky VN. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.* 2002;11: 739–756.

66. Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK. Sequence complexity of disordered protein. *Proteins*. 2001;42: 38–48.
67. Lee J, Cho H, Kwon I. Phase separation of low-complexity domains in cellular function and disease. *Exp Mol Med*. 2022;54: 1412–1422.
68. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, et al. Intrinsically disordered protein. *J Mol Graph Model*. 2001;19: 26–59.
69. Vucetic S, Brown CJ, Dunker AK, Obradovic Z. Flavors of protein disorder. *Proteins*. 2003;52: 573–584.
70. Das RK, Pappu RV. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc Natl Acad Sci USA*. 2013;110: 13392–13397.
71. He B, Wang K, Liu Y, Xue B, Uversky VN, Dunker AK. Predicting intrinsic disorder in proteins: an overview. *Cell Res*. 2009;19: 929–949.
72. Necci M, Piovesan D, Dosztányi Z, Tompa P, Tosatto SCE. A comprehensive assessment of long intrinsic protein disorder from the DisProt database. *Bioinformatics*. 2017;34: 445–452.
73. Zhao B, Kurgan L. Surveying over 100 predictors of intrinsic disorder in proteins. *Expert Rev Proteomics*. 2021;18: 1019–1029.
74. Linding R, Russell RB, Neduva V, Gibson TJ. GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res*. 2003;31: 3701–3708.
75. Galzitskaya OV, Garbuzynskiy SO, Lobanov MY. FoldUnfold: web server for the prediction of disordered regions in protein chain. *Bioinformatics*. 2006;22: 2948–2949.
76. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics*. 2006;7: 208.
77. Zhao B, Kurgan L. Deep learning in prediction of intrinsic disorder in proteins. *Comput Struct Biotechnol J*. 2022;20: 1286–1294.
78. Piovesan D, Monzon AM, Tosatto SCE. Intrinsic protein disorder and conditional folding in AlphaFoldDB. *Protein Sci*. 2022;31: e4466.
79. Monastyrskyy B, Kryshchak A, Moul J, Tramontano A, Fidelis K. Assessment of protein disorder region predictions in CASP10. *Proteins*. 2014;82 Suppl 2: 127–137.
80. Jin Y, Dunbrack RL Jr. Assessment of disorder predictions in CASP6. *Proteins*. 2005;61 Suppl 7: 167–175.
81. Necci M, Piovesan D, CAID Predictors, DisProt Curators, Tosatto SCE. Critical assessment of protein intrinsic disorder prediction. *Nat Methods*. 2021;18: 472–481.
82. Mészáros B, Dobson L, Fichó E, Tusnády GE, Dosztányi Z, Simon I. Sequential, Structural and Functional Properties of Protein Complexes Are Defined by How Folding and Binding Intertwine. *J Mol Biol*. 2019;431: 4408–4428.
83. Cheng Y, Oldfield CJ, Meng J, Romero P, Uversky VN, Dunker AK. Mining alpha-helix-forming molecular recognition features with cross species sequence alignments. *Biochemistry*. 2007;46: 13468–13477.
84. Malhis N, Gsponer J. Computational identification of MoRFs in protein sequences. *Bioinformatics*. 2015;31: 1738–1744.
85. Malhis N, Jacobson M, Gsponer J. MoRFchibi SYSTEM: software tools for the identification of MoRFs in protein sequences. *Nucleic Acids Res*. 2016;44: W488–93.
86. Miskei M, Horvath A, Vendruscolo M, Fuxreiter M. Sequence-Based Prediction of Fuzzy Protein Interactions. *J Mol Biol*. 2020;432: 2289–2303.
87. Peng Z, Wang C, Uversky VN, Kurgan L. Prediction of Disordered RNA, DNA, and Protein Binding Regions Using DisoRDPbind. *Methods Mol Biol*. 2017;1484: 187–203.
88. Katuwawala A, Zhao B, Kurgan L. DisoLipPred: Accurate prediction of disordered lipid binding residues in protein sequences with deep recurrent networks and transfer learning. *Bioinformatics*. 2021. doi:10.1093/bioinformatics/btab640
89. Meng F, Kurgan L. DFLpred: High-throughput prediction of disordered flexible linker regions in protein sequences. *Bioinformatics*. 2016;32: i341–i350.

90. Peng Z, Xing Q, Kurgan L. APOD: accurate sequence-based predictor of disordered flexible linkers. *Bioinformatics*. 2020;36: i754–i761.
91. Csizmók V, Dosztányi Z, Simon I, Tompa P. Towards proteomic approaches for the identification of structural disorder. *Curr Protein Pept Sci*. 2007;8: 173–179.
92. Galea CA, Pagala VR, Obenauer JC, Park C-G, Slaughter CA, Kriwacki RW. Proteomic studies of the intrinsically unstructured mammalian proteome. *J Proteome Res*. 2006;5: 2839–2848.
93. Huang YJ, Acton TB, Montelione GT. DisMeta: a meta server for construct design and optimization. *Methods Mol Biol*. 2014;1091: 3–16.
94. Perdigão N, Heinrich J, Stolte C, Sabir KS, Buckley MJ, Tabor B, et al. Unexpected features of the dark proteome. *Proc Natl Acad Sci U S A*. 2015;112: 15898–15903.
95. Necci M, Piovesan D, Clementel D, Dosztányi Z, Tosatto SCE. MobiDB-lite 3.0: fast consensus annotation of intrinsic disorder flavours in proteins. *Bioinformatics*. 2020;36: 5533–5534.
96. Dunker AK, Obradovic Z, Romero P, Garner EC, Brown CJ. Intrinsic protein disorder in complete genomes. *Genome Inform Ser Workshop Genome Inform*. 2000;11: 161–171.
97. Peng Z, Yan J, Fan X, Mizianty MJ, Xue B, Wang K, et al. Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. *Cell Mol Life Sci*. 2015;72: 137–151.
98. Schad E, Tompa P, Hegyi H. The relationship between proteome size, structural disorder and organism complexity. *Genome Biol*. 2011;12: R120.
99. Tompa P, Dosztanyi Z, Simon I. Prevalent structural disorder in *E. coli* and *S. cerevisiae* proteomes. *J Proteome Res*. 2006;5: 1996–2000.
100. Mishra PM, Verma NC, Rao C, Uversky VN, Nandi CK. Intrinsically disordered proteins of viruses: Involvement in the mechanism of cell regulation and pathogenesis. *Prog Mol Biol Transl Sci*. 2020;174: 1–78.
101. Cubuk J, Alston JJ, Incicco JJ, Singh S, Stuchell-Brereton MD, Ward MD, et al. The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA. *Nat Commun*. 2021;12: 1936.
102. Schlessinger A, Schaefer C, Vicedo E, Schmidberger M, Punta M, Rost B. Protein disorder--a breakthrough invention of evolution? *Curr Opin Struct Biol*. 2011;21: 412–418.
103. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Uversky VN, et al. Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. *J Proteome Res*. 2007;6: 1882–1898.
104. Bondos SE, Dunker AK, Uversky VN. On the roles of intrinsically disordered proteins and regions in cell communication and signaling. *Cell Commun Signal*. 2021;19: 88.
105. Dosztányi Z, Chen J, Dunker AK, Simon I, Tompa P. Disorder and sequence repeats in hub proteins and their implications for network evolution. *J Proteome Res*. 2006;5: 2985–2995.
106. Haynes C, Oldfield CJ, Ji F, Klitgord N, Cusick ME, Radivojac P, et al. Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS Comput Biol*. 2006;2: e100.
107. Buljan M, Chalancon G, Eustermann S, Wagner GP, Fuxreiter M, Bateman A, et al. Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. *Mol Cell*. 2012;46: 871–883.
108. Gsponer J, Futschik ME, Teichmann SA, Babu MM. Tight regulation of unstructured proteins: from transcript synthesis to protein degradation. *Science*. 2008;322: 1365–1368.
109. Brown CJ, Takayama S, Campen AM, Vise P, Marshall TW, Oldfield CJ, et al. Evolutionary rate heterogeneity in proteins with long disordered regions. *J Mol Evol*. 2002;55: 104–110.

110. Bellay J, Han S, Michaut M, Kim T, Costanzo M, Andrews BJ, et al. Bringing order to protein disorder through comparative genomics and genetic interactions. *Genome Biol.* 2011;12: R14.
111. Zarin T, Strome B, Peng G, Pritišanac I, Forman-Kay JD, Moses AM. Identifying molecular features that are associated with biological function of intrinsically disordered protein regions. *Elife.* 2021;10: e60220.
112. Davey NE, Cyert MS, Moses AM. Short linear motifs - ex nihilo evolution of protein regulation. *Cell Commun Signal.* 2015;13: 43.
113. Tompa P, Fuxreiter M, Oldfield CJ, Simon I, Dunker AK, Uversky VN. Close encounters of the third kind: disordered domains and the interactions of proteins. *Bioessays.* 2009;31: 328–335.
114. Pajkos M, Dosztányi Z. Functions of intrinsically disordered proteins through evolutionary lenses. *Prog Mol Biol Transl Sci.* 2021;183: 45–74.
115. Wright PE, Dyson HJ. Intrinsically disordered proteins in cellular signalling and regulation. *Nat Rev Mol Cell Biol.* 2015;16: 18–29.
116. Tompa P. Intrinsically disordered proteins: a 10-year recap. *Trends Biochem Sci.* 2012;37: 509–516.
117. Iakoucheva LM, Brown CJ, Lawson JD, Obradović Z, Dunker AK. Intrinsic disorder in cell-signaling and cancer-associated proteins. *J Mol Biol.* 2002;323: 573–584.
118. Uversky VN, Oldfield CJ, Dunker AK. Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu Rev Biophys.* 2008;37: 215–246.
119. Hegyi H, Buday L, Tompa P. Intrinsic structural disorder confers cellular viability on oncogenic fusion proteins. *PLoS Comput Biol.* 2009;5: e1000552.
120. Vavouri T, Semple JI, Garcia-Verdugo R, Lehner B. Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity. *Cell.* 2009;138: 198–208.
121. Babu MM. The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. *Biochem Soc Trans.* 2016;44: 1185–1200.
122. Uyar B, Weatheritt RJ, Dinkel H, Davey NE, Gibson TJ. Proteome-wide analysis of human disease mutations in short linear motifs: neglected players in cancer? *Mol Biosyst.* 2014;10: 2626–2642.
123. Meyer K, Kirchner M, Uyar B, Cheng J-Y, Russo G, Hernandez-Miranda LR, et al. Mutations in Disordered Regions Can Cause Disease by Creating Dileucine Motifs. *Cell.* 2018;175: 239–253.e17.
124. Zou H, Pan T, Gao Y, Chen R, Li S, Guo J, et al. Pan-cancer assessment of mutational landscape in intrinsically disordered hotspots reveals potential driver genes. *Nucleic Acids Res.* 2022;50: e49.
125. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* 2021;49: D480–D489.
126. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. *Nucleic Acids Res.* 2000;28: 235–242.
127. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 2000;25: 25–29.
128. Del Toro N, Shrivastava A, Ragueneau E, Meldal B, Combe C, Barrera E, et al. The IntAct database: efficient access to fine-grained molecular interaction data. *Nucleic Acids Res.* 2022;50: D648–D653.
129. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* 2019;47: D941–D947.
130. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, et al. The UCSC Genome Browser Database. *Nucleic Acids Res.* 2003;31: 51–54.
131. Mészáros B, Erdős G, Szabó B, Schád É, Tantos Á, Abukhairan R, et al. PhaSePro: the database of proteins driving liquid-liquid phase separation. *Nucleic Acids Res.* 2020;48: D360–D367.

132. Csizmadia G, Erdős G, Tordai H, Padányi R, Tosatto S, Dosztányi Z, et al. The MemMoRF database for recognizing disordered protein regions interacting with cellular membranes. *Nucleic Acids Res.* 2021;49: D355–D360.
133. Piovesan D, Necci M, Escobedo N, Monzon AM, Hatos A, Mičetić I, et al. MobiDB: intrinsically disordered proteins in 2021. *Nucleic Acids Res.* 2021;49: D361–D367.
134. Peng Z-L, Kurgan L. Comprehensive comparative assessment of in-silico predictors of disordered regions. *Curr Protein Pept Sci.* 2012;13: 6–18.
135. Dosztányi Z. Prediction of protein disorder based on IUPred. *Protein Sci.* 2018;27: 331–340.
136. Lang B, Babu MM. A community effort to bring structure to disorder. *Nature methods.* 2021. pp. 454–455.
137. Wootton JC. Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput Chem.* 1994;18: 269–285.
138. Pellegrini M, Marcotte EM, Yeates TO. A fast algorithm for genome-wide analysis of proteins with repeated sequences. *Proteins.* 1999;35: 440–446.
139. Barabási A-L, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet.* 2004;5: 101–113.
140. Eisenberg E, Levanon EY. Preferential attachment in the protein network evolution. *Phys Rev Lett.* 2003;91: 138701.
141. Pastor-Satorras R, Smith E, Solé RV. Evolving protein interaction networks through gene duplication. *J Theor Biol.* 2003;222: 199–210.
142. Jones S, Thornton JM. Principles of protein-protein interactions. *Proc Natl Acad Sci U S A.* 1996;93: 13–20.
143. Lo Conte L, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. *J Mol Biol.* 1999;285: 2177–2198.
144. Nooren IMA, Thornton JM. Diversity of protein-protein interactions. *EMBO J.* 2003;22: 3486–3492.
145. Gunasekaran K, Tsai C-J, Nussinov R. Analysis of ordered and disordered protein complexes reveals structural features discriminating between stable and unstable monomers. *J Mol Biol.* 2004;341: 1327–1341.
146. Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A, et al. DisProt: the Database of Disordered Proteins. *Nucleic Acids Res.* 2007;35: D786–93.
147. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes. *Science.* 2013;339: 1546–1558.
148. Tokheim CJ, Papadopoulos N, Kinzler KW, Vogelstein B, Karchin R. Evaluating the evaluation of cancer driver genes. *Proc Natl Acad Sci U S A.* 2016;113: 14330–14335.
149. Mészáros B, Zeke A, Reményi A, Simon I, Dosztányi Z. Systematic analysis of somatic mutations driving cancer: uncovering functional protein regions in disease development. *Biol Direct.* 2016;11: 23.
150. Mészáros B, Kumar M, Gibson TJ, Uyar B, Dosztányi Z. Degrons in cancer. *Sci Signal.* 2017;10. doi:10.1126/scisignal.aak9982
151. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell.* 2011;144: 646–674.
152. Brown CJ, Johnson AK, Dunker AK, Daughdrill GW. Evolution and disorder. *Curr Opin Struct Biol.* 2011;21: 441–446.
153. Domazet-Lošo T, Tautz D. Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. *BMC Biol.* 2010;8: 66.
154. Kinzler KW, Vogelstein B. Cancer-susceptibility genes. Gatekeepers and caretakers. *Nature.* 1997. pp. 761, 763.
155. Torres KA, Calil FA, Zhou AL, DuPrie ML, Putnam CD, Kolodner RD. The unstructured linker of Mlh1 contains a motif required for endonuclease function which is mutated in cancers. *Proc Natl Acad Sci U S A.* 2022;119: e2212870119.



## 9 AZ ÉRTEKEZÉS ALAPJÁUL SZOLGÁLÓ KÖZLEMÉNYEK

1: **Dosztányi Z**, Csizmók V, Tompa P, Simon I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol.* 2005;347:827-839.

2: **Dosztányi Z**, Csizmok V, Tompa P, Simon I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics.* 2005;21:3433-3434.

3: **Dosztányi Z**, Chen J, Dunker AK, Simon I, Tompa P. Disorder and sequence repeats in hub proteins and their implications for network evolution. *J Proteome Res.* 2006;5:2985-2995.

4: Mészáros B, Tompa P, Simon I, **Dosztányi Z**. Molecular principles of the interactions of disordered proteins. *J Mol Biol.* 2007;372:549-561.

5: Mészáros B, Simon I, **Dosztányi Z**. Prediction of protein binding regions in disordered proteins. *PLoS Comput Biol.* 2009;5:e1000376.

6: Mészáros B, Erdős G, **Dosztányi Z**. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* 2018;46:W329-W337.

7: **Dosztányi Z**, Mészáros B, Simon I. ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics.* 2009;25:2745-2746.

8: Erdős G, Pajkos M, **Dosztányi Z**. IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. *Nucleic Acids Res.* 2021;49:W297-W303.

9: Pajkos M, Mészáros B, Simon I, **Dosztányi Z**. Is there a biological cost of protein disorder? Analysis of cancer-associated mutations. *Mol Biosyst.* 2012;8:296-307.

10: Mészáros B, Hajdu-Soltész B, Zeke A, **Dosztányi Z**. Mutations of Intrinsically Disordered Protein Regions Can Drive Cancer but Lack Therapeutic Strategies. *Biomolecules.* 2021;11:381.

11: Pajkos M, Zeke A, **Dosztányi Z**. Ancient Evolutionary Origin of Intrinsically Disordered Cancer Risk Regions. *Biomolecules.* 2020;10:1115.

**JMB**Available online at [www.sciencedirect.com](http://www.sciencedirect.com)

SCIENCE @ DIRECT®



# The Pairwise Energy Content Estimated from Amino Acid Composition Discriminates between Folded and Intrinsically Unstructured Proteins

Zsuzsanna Dosztányi, Veronika Csizmók, Péter Tompa and István Simon\*

*Institute of Enzymology  
Biological Research Center  
Hungarian Academy of Sciences  
1518 Budapest, PO Box 7  
Hungary*

The structural stability of a protein requires a large number of interresidue interactions. The energetic contribution of these can be approximated by low-resolution force fields extracted from known structures, based on observed amino acid pairing frequencies. The summation of such energies, however, cannot be carried out for proteins whose structure is not known or for intrinsically unstructured proteins. To overcome these limitations, we present a novel method for estimating the total pairwise interaction energy, based on a quadratic form in the amino acid composition of the protein. This approach is validated by the good correlation of the estimated and actual energies of proteins of known structure and by a clear separation of folded and disordered proteins in the energy space it defines. As the novel algorithm has not been trained on unstructured proteins, it substantiates the concept of protein disorder, i.e. that the inability to form a well-defined 3D structure is an intrinsic property of many proteins and protein domains. This property is encoded in their sequence, because their biased amino acid composition does not allow sufficient stabilizing interactions to form. By limiting the calculation to a predefined sequential neighborhood, the algorithm was turned into a position-specific scoring scheme that characterizes the tendency of a given amino acid to fall into an ordered or disordered region. This application we term IUPred and compare its performance with three generally accepted predictors, PONDR VL3H, DISOPRED2 and GlobPlot on a database of disordered proteins.

© 2005 Elsevier Ltd. All rights reserved.

*Keywords:* intrinsically unstructured proteins; prediction of disordered proteins; low-resolution force fields; interresidue interactions; foldability

\*Corresponding author

## Introduction

Intrinsically unstructured/disordered proteins/domains (IUPs), such as p21,<sup>1</sup> the N-terminal domain of p53<sup>2</sup> or the transactivator domain of CREB,<sup>3</sup> exist in a largely disordered structural state, yet they carry out basic cellular functions.<sup>4–7</sup> Their existence defies the classical structure–function paradigm, founded on the tenet that a well-defined 3D structure is the prerequisite of protein function. The importance of protein disorder, nevertheless, is underlined by its prevalence in various proteomes<sup>8,9</sup> and by its correlation with basic

functional modes, such as signal transduction and transcriptional regulation.<sup>9,10</sup>

The identification of IUPs thus far proceeded by collecting scattered data obtained with a range of experimental techniques. As a result, available datasets are rather limited in size and are heterogeneous in terms of experimental conditions, techniques and interpretation of data. They also lack consistency, due to the absence of clear conceptual and operational definition(s) of structural disorder. All these result in false positive and false negative classifications, i.e. the inclusion of ordered segments in disorder databases and the exclusion (and inclusion in ordered reference databases) of disordered proteins/segments. Furthermore, the databases are also biased due to the overrepresentation of a few experimental techniques, such as X-ray crystallography, NMR

Abbreviation used: IUP, unstructured/disordered proteins/domains.

E-mail address of the corresponding author: [simon@enzim.hu](mailto:simon@enzim.hu)

and CD. As each technique probes different aspects of protein structure, they do not necessarily correctly identify disorder. For example, loopy proteins, which have no repetitive secondary structure,<sup>11</sup> would appear disordered by CD but ordered by the other techniques. With NMR, disorder often is concluded from poor signal dispersion, which does not distinguish between random coils and molten globules of high potential to fold in the presence of a partner. In X-ray crystallography, crystal packing may enforce certain disordered regions to become ordered, and disordered binding segments are often crystallized in complex with their partner and are classified ordered despite their lack of structure in isolation. In addition, wobbly domains would appear disordered, despite their intrinsic structural order. In consequence, predictors trained on these datasets for assessing disorder<sup>5,9</sup> reflect these uncertainties.

The basis of predicting protein disorder is the difference in sequence characteristics between folded and disordered proteins. Typically, IUPs exhibit a strong bias in their amino acid composition and even a reduced alphabet is able to recognize them at the level of complete sequences.<sup>12</sup> Other results indicate, however, that there are differences in sequence properties among different types of disordered proteins.<sup>13</sup> Various factors have been suggested to be important in terms of protein disorder, including flexibility, aromatic content,<sup>14</sup> secondary structure preferences<sup>15</sup> and various scales associated with hydrophobicity.<sup>14,16</sup> Beside low mean hydrophobicity, high net charge was also suggested to contribute to disorder.<sup>17</sup> All these different analyses, though, hint that the amino acid composition of IUPs results in their inability to fold due to the depletion of typically buried amino acid residues and enrichment of typically exposed amino acid residues,<sup>5</sup> which implies that globular proteins have specific sequences with the potential to form a sufficiently large number of favorable interactions, whereas IUPs do not. Here, we attempt to put this inference on a quantitative footing by taking an energetics point of view. On this ground, the sequences encoding for globular proteins and IUPs can be distinguished.

For globular proteins, the contribution of inter-residue interactions to total energy is often approximated by low-resolution force fields, or statistical potentials, energy-like quantities derived from globular proteins based on the observed amino acid pairing frequencies.<sup>18,19</sup> In deriving the actual potentials, different principles have been applied.<sup>18,20–23</sup> The resulting empirical energy functions are well suited to assess the quality of structural models<sup>24</sup> and have been used for fold recognition or threading,<sup>25,26</sup> but also in docking,<sup>27</sup> *ab initio* folding,<sup>28</sup> or predicting protein stability.<sup>29</sup> Their success in a wide range of applications suggests the existence of a common set of interactions, simultaneously favored in all native, as opposed to alternate, structures.

Our current formulation derives from the general

view that the primary structure of a globular protein determines its native conformation, and therefore its energy, which corresponds to the global minimum in conformational space. This energy represents the lowest level attainable by the sequence at the optimum of interresidue interactions. In this work, we introduce a novel approach to predict this optimum energy independently of a presumed structure. By applying this principle to a predefined sequential neighborhood of a particular amino acid in a sequence, this approach can be turned into a position-specific scoring scheme for disorder, termed IUPred. As IUPred has not been trained on potentially erroneous data, its unbiased assessment of the structural status of an unknown sequence/segment is of confirmatory value.

## Theory

### Estimation of the pairwise energy from amino acid composition

The pairwise energy of a protein in its native state is the function of its conformation as well as its amino acid sequence. The total energy can be calculated by taking all contacts in the protein, and weighting them by the corresponding interaction energy. In our model, the energy depends only on amino acid types, as specified by a 20 by 20 interaction matrix,  $\mathbf{M}$  (see Table 1). The pairwise energy content can be written as:

$$E = \sum_{ij=1}^{20} M_{ij} C_{ij}$$

where  $M_{ij}$  is the interaction energy between amino acid types  $i$  and  $j$ , and  $C_{ij}$  is the number of interactions between residues of types  $i$  and  $j$  in the given conformation.

We approximate  $E/L$ , the total energy per amino acid, by means of the protein's amino acid composition. Without considering the actual conformation, we rely on statistics collected from a database of globular proteins. The rationale behind this approach is that the energy contribution of a residue depends not only on its amino acid type, but also on its potential partners in the sequence. We assume that if the sequence contains more amino acid residues that can form favorable contacts with the given residue, its expected energy contribution is more favorable. The simplest formula which describes this relationship is a quadratic expression in the amino acid composition.

Let  $N_i$  denote the number of amino acid residues of type  $i$  in the sequence and  $n_i = N_i/L$  its frequency. The energy per amino acid is approximated by:

$$\frac{E_{\text{estimated}}}{L} = \sum_{ij}^{20} n_i P_{ij} n_j$$

where  $\mathbf{P}$  is the energy predictor matrix, which tells

**Table 1. M matrix**

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	-0.20	-0.44	0.16	0.26	-0.46	-0.26	0.50	-0.57	0.10	-0.36	-0.22	0.07	0.14	0.01	0.20	-0.09	-0.05	-0.42	0.05	-0.50
C	-0.44	-2.99	0.21	0.19	-0.88	-0.34	-1.11	-0.36	-0.09	-0.53	-0.43	-0.52	-0.14	-0.43	-0.24	0.13	-0.22	-0.62	0.24	-0.79
D	0.16	0.21	0.17	0.55	0.38	0.35	-0.23	0.44	-0.39	0.28	0.35	-0.02	1.03	0.49	-0.37	0.19	-0.12	0.69	0.04	0.43
E	0.26	0.19	0.55	0.60	0.55	0.65	0.18	0.37	-0.47	0.33	0.29	0.01	0.69	0.04	-0.52	0.18	0.37	0.39	0.03	0.17
F	-0.46	-0.88	0.38	0.55	-0.94	0.17	-0.40	-0.88	0.01	-1.08	-0.78	0.22	0.20	0.26	-0.19	-0.22	0.02	-1.15	-0.60	-0.88
G	-0.26	-0.34	0.35	0.65	0.17	-0.12	0.18	0.24	0.19	0.24	0.02	-0.04	0.60	0.46	0.50	0.28	0.28	0.27	0.51	-0.35
H	0.50	-1.11	-0.23	0.18	-0.40	0.18	0.42	-0.00	0.79	-0.24	-0.07	0.20	0.25	0.69	0.24	0.21	0.11	0.16	-0.85	-0.26
I	-0.57	-0.36	0.44	0.37	-0.88	0.24	-0.00	-1.16	0.15	-1.25	-0.58	-0.09	0.36	-0.08	0.14	0.32	-0.27	-1.06	-0.68	-0.85
K	0.10	-0.09	-0.39	-0.47	0.01	0.19	0.79	0.15	0.42	0.13	0.48	0.26	0.50	0.15	0.53	0.10	-0.19	0.10	0.10	0.04
L	-0.36	-0.53	0.28	0.33	-1.08	0.24	-0.24	-1.25	0.13	-1.10	-0.50	0.21	0.42	-0.01	-0.07	0.17	0.07	-0.97	-0.95	-0.63
M	-0.22	-0.43	0.35	0.29	-0.78	0.02	-0.07	-0.58	0.48	-0.50	-0.74	0.32	0.01	0.26	0.15	0.48	0.16	-0.73	-0.56	-1.02
N	0.07	-0.52	-0.02	0.01	0.22	-0.04	0.20	-0.09	0.26	0.21	0.32	0.14	0.27	0.37	0.13	0.15	0.10	0.40	-0.12	0.32
P	0.14	-0.14	1.03	0.69	0.20	0.60	0.25	0.36	0.50	0.42	0.01	0.27	0.27	1.02	0.47	0.54	0.88	-0.02	-0.37	-0.12
Q	0.01	-0.43	0.49	0.04	0.26	0.46	0.69	-0.08	0.15	-0.01	0.26	0.37	1.02	-0.12	0.24	0.29	0.04	-0.11	0.18	0.11
R	0.20	-0.24	-0.37	-0.52	-0.19	0.50	0.24	0.14	0.53	-0.07	0.15	0.13	0.47	0.24	0.17	0.27	0.45	0.01	-0.73	0.01
S	-0.09	0.13	0.19	0.18	-0.22	0.28	0.21	0.32	0.10	0.17	0.48	0.15	0.54	0.29	0.27	-0.06	0.08	0.12	-0.22	-0.14
T	-0.05	-0.22	-0.12	0.37	0.02	0.28	0.11	-0.27	-0.19	0.07	0.16	0.10	0.88	0.04	0.45	0.08	-0.03	-0.01	0.11	-0.32
V	-0.42	-0.62	0.69	0.39	-1.15	0.27	0.16	-1.06	0.10	-0.97	-0.73	0.40	-0.02	-0.11	0.01	0.12	-0.01	-0.89	-0.56	-0.71
W	0.05	0.24	0.04	0.03	-0.60	0.51	-0.85	-0.68	0.10	-0.95	-0.56	-0.12	-0.37	0.18	-0.73	-0.22	0.11	-0.56	-0.05	-1.41
Y	-0.50	-0.79	0.43	0.17	-0.88	-0.35	-0.26	-0.85	0.04	-0.63	-1.02	0.32	-0.12	0.11	0.01	-0.14	-0.32	-0.71	-1.41	-0.76

Contact potential derived from 785 proteins using the approach of Thomas & Dill.<sup>20</sup>

how the energy of amino acid  $i$  depends on the  $j$ th element of the amino acid composition vector. The parameters  $P_{ij}$ , applicable for all proteins, are determined by least-squares fitting using globular proteins. The fitting was carried out by treating each amino acid type and the corresponding row in matrix  $\mathbf{P}$  separately, to ensure that the energetic contribution is well approximated for all amino acid types. Using the additivity of the energy of pairwise interactions, we dissect the total energy of the  $k$ th protein into amino acid specific contributions  $E^k = \sum e_i^k$ , where  $e_i^k$  is the energy of all amino acid residues type  $i$  interacting with all other amino acid residues in the sequence. The  $e_i^k$  depends on the number of contacts this residue makes with other amino acid residues of type  $j$  in the sequence.

$$e_i^k(\text{calculated}) = \sum_{j=1}^{20} M_{ij} C_{ij}^k$$

This quantity is approximated by the expression:

$$e_i^k(\text{estimated}) = N_i^k \sum_{j=1}^{20} P_{ij} n_j^k$$

The parameters of the corresponding row of matrix  $\mathbf{P}$  are obtained by minimizing the function

$$Z_i = \sum_k (e_i^k - N_i^k \sum_{j=1}^{20} P_{ij} n_j^k)^2$$

Letting  $\partial Z_i / \partial P_{ij} = 0$  for all  $P_{ij}$  leads to a set of linear equations which are solved for each amino acid type by using the GSL scientific library. Only the symmetrical part of the matrix is considered, as the anti-symmetrical part is cancelled out in quadratic forms. The resulting  $\mathbf{P}$  is given in Table 2.

## Results

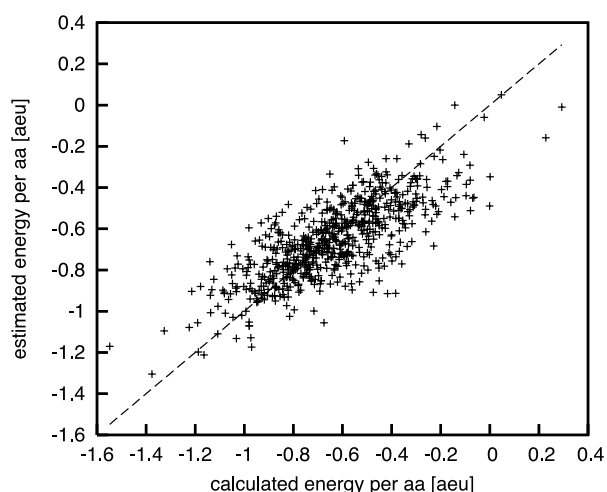
### Comparison of estimated and calculated energies for globular proteins

The validity of the energy predictor matrix was checked by comparing the energies calculated from amino acid interactions of proteins with a known structure to the energies estimated from their amino acid compositions. The fitting was carried out using 674 proteins from the Glob\_list (for the definition of this and other databases, see Materials and Methods), omitting those with high cysteine content (above 9%) as they had unusually favorable energy because of cystine pairs. The calculated energy is given in an arbitrary energy unit [aeu], with more negative values indicating more favorable interactions. Figure 1 shows that there is a clear linear relationship between calculated and estimated energies. The goodness of fit can be characterized by a correlation coefficient and the  $r^2$  value:  $r^2 = 1 - SS_{\text{reg}} / SS_{\text{tot}}$ , where  $SS_{\text{tot}}$  and  $SS_{\text{reg}}$  are the sums of the squares of distances from the mean of the calculated energies, and of estimated and

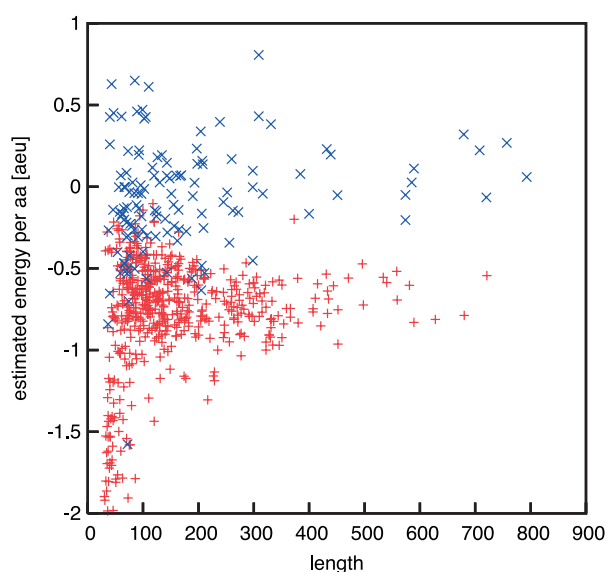
Table 2.  $\mathbf{P}$  energy predictor matrix

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	-1.65	-2.83	1.16	1.80	-3.73	-0.41	1.90	-3.69	0.49	-3.01	-2.08	0.66	1.54	1.20	0.98	-0.08	0.46	-2.31	0.32	-4.62
C	-2.83	-39.58	-0.82	-0.53	-3.07	-2.96	-4.98	0.34	-1.38	-2.15	1.43	-4.18	-2.13	-2.91	-0.41	-2.33	-1.84	-0.16	4.26	-4.46
D	1.16	-0.82	0.84	1.97	-0.92	0.88	-1.07	0.68	-1.93	0.23	0.61	0.32	3.31	2.67	-2.02	0.91	-0.65	0.94	-0.71	0.90
E	1.80	-0.53	1.97	1.45	0.94	1.31	0.61	1.30	-2.51	1.14	2.53	0.20	1.44	0.10	-3.13	0.81	1.54	0.12	-1.07	1.29
F	-3.73	-3.07	-0.92	0.94	-11.25	0.35	-3.57	-5.88	-0.82	-8.59	-5.34	0.73	0.32	0.77	-0.40	-2.22	0.11	-7.05	-7.09	-8.80
G	-0.41	-2.96	0.88	1.31	0.35	-0.20	1.09	-0.65	-0.16	-0.55	-0.52	-0.32	2.25	1.11	0.84	0.71	0.59	-0.38	1.69	-1.90
H	1.90	-4.98	-1.07	0.61	-3.57	1.09	1.97	-0.71	2.89	-0.86	-0.75	1.84	0.35	2.64	2.05	0.82	-0.01	0.27	-7.58	-3.20
I	-3.69	0.34	0.68	1.30	-5.88	-0.65	-0.71	-6.74	-0.01	-9.01	-3.62	-0.07	0.12	-0.18	0.19	-0.15	0.63	-6.54	-3.78	-5.26
K	0.49	-1.38	-1.93	-2.51	-0.82	-0.16	2.89	-0.01	1.24	0.49	1.61	1.12	0.51	0.43	2.34	0.19	-1.11	0.19	0.02	-1.19
L	-3.01	-2.15	1.14	1.30	-2.51	-0.16	-6.37	-0.01	0.49	-6.37	-2.88	0.97	1.81	-0.58	-0.60	-0.41	0.72	-5.43	8.31	-4.90
M	-2.08	1.43	2.53	1.44	-5.88	-0.75	-6.49	-0.01	0.49	-2.88	-6.49	0.21	0.75	1.90	2.09	1.39	0.63	-2.59	-6.88	-9.73
N	0.66	-4.18	0.32	0.20	0.73	0.32	-8.59	-0.82	1.61	-0.86	-0.75	1.84	0.35	2.64	2.05	0.82	-0.01	0.27	-7.58	-3.20
P	1.54	-2.13	3.31	2.67	-0.40	0.77	-0.42	2.97	1.81	0.51	0.12	-0.07	0.12	-0.18	0.19	-0.15	0.63	-6.54	-3.78	-5.26
Q	1.20	-2.91	2.67	2.67	0.77	0.77	-0.42	2.97	1.81	0.51	0.12	-0.07	0.12	-0.18	0.19	-0.15	0.63	-6.54	-3.78	-5.26
R	0.98	-0.41	-2.02	3.31	-3.13	0.81	1.54	0.12	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72
S	-0.08	-2.33	0.91	-0.65	0.94	-0.71	0.59	-0.38	1.69	-1.90	-8.80	-3.20	-4.66	0.90	1.29	-8.80	-3.20	-4.66	0.90	1.29
T	0.46	-1.84	-0.65	0.94	-0.71	0.59	-0.38	1.69	-1.90	-8.80	-3.20	-4.66	0.90	1.29	-8.80	-3.20	-4.66	0.90	1.29	-8.80
V	-2.31	0.32	-4.62	-4.46	-8.80	-3.20	-4.66	0.90	1.29	-8.80	-3.20	-4.66	0.90	1.29	-8.80	-3.20	-4.66	0.90	1.29	-8.80
W	0.32	4.26	-0.71	0.90	-1.07	0.94	-1.07	0.94	-1.07	0.94	-1.07	0.94	-1.07	0.94	-1.07	0.94	-1.07	0.94	-1.07	0.94
Y	-4.62	-4.46	0.90	1.29	-8.80	-3.20	-4.66	0.90	1.29	-8.80	-3.20	-4.66	0.90	1.29	-8.80	-3.20	-4.66	0.90	1.29	-8.80

The pairwise energy per amino acid is estimated as a quadratic form in the amino acid composition vector using the elements of this matrix.



**Figure 1.** Correlation of estimated and calculated total interaction energies of globular proteins. The total pairwise interaction energy of 674 globular proteins from Glob\_list (omitting proteins with high cysteine content), was estimated from their amino acid compositions by a method based on a quadratic formula in the amino acid composition and are shown as a function of the actual energies calculated from their known 3D structures. The energies are in arbitrary energy units, as defined in the text. The broken line represents perfect agreement between the estimated and calculated energy values.



**Figure 2.** Estimated pairwise interaction energies of globular proteins and IUPs. The total pairwise interaction energy of 559 globular proteins in Filt\_Glob\_list (red +) and 129 disordered proteins in IUP\_list (blue x) was estimated from their amino acid composition and plotted as a function of their length. Values more negative represent more stabilization due to pairwise amino acid interactions. The average pairwise interaction energy of globular proteins and IUPs are  $-0.81$  and  $-0.07$  [aeu], respectively.

calculated energies, respectively. The value of  $r^2$  can be between 0, where the average is used as an estimate, and 1, which is the ideal case. It describes how well the variance in the original data is explained by the fitted model using the least-squares approximation. In our case, the  $r^2$  value was 0.58, and the correlation coefficient was 0.76. Both values indicate a reasonable level of agreement between the estimated and calculated energies.

#### Pairwise energy content for globular proteins and IUPs

Figure 2 shows the estimated energies for globular proteins and IUPs as a function of their length. For the globular proteins of Filt\_Glob\_list the average energy is  $-0.81$  [aeu]. The estimated energies of IUPs in IUP\_list are less favorable, with an average of  $-0.07$  [aeu]. The separation between the two sets becomes more pronounced for longer sequences, while there is some overlap for shorter sequences. Based on the  $P$ -value of  $2.2 \times 10^{-16}$  obtained using the Wilcoxon rank sum test, we can reject the hypothesis that the two sets of energies are from the same distributions. Overall, the difference substantiates our assumption that the pairwise energy content is less favorable for IUPs than for globular proteins.

A corollary to this separation is that the estimated energy content may also distinguish partially

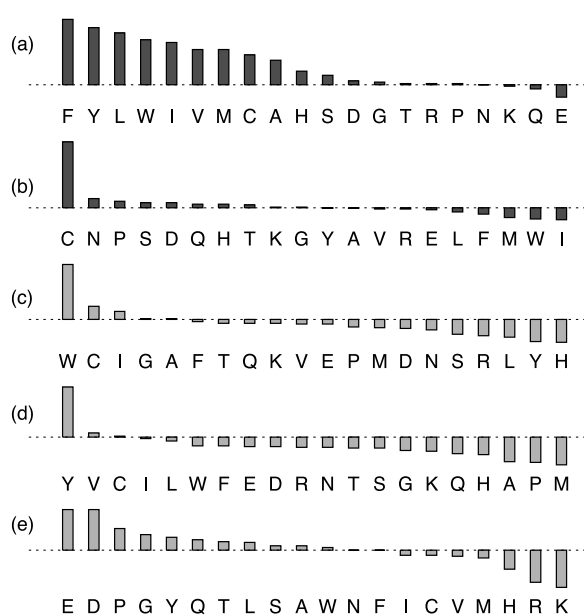
ordered IUPs, i.e. molten globules and pre-molten globules, from fully disordered (coil-like) proteins, because the former are expected to have more energetically favorable interactions. To test this assumption, 55 coil-like and 52 pre-molten globule-like proteins have been taken from Table 1 in the work done by Uversky,<sup>30</sup> and their total energy content has been estimated. These datasets, which partially overlap with IUP\_list, show a 0.3 [aeu] separation in the average energy content (data not shown). Thus, our approach is able to assess the energetic consequence of the subtle structural differences between fully and partially disordered proteins.

#### Decomposition

So far, the quadratic form for the estimated energy has been given in the natural coordinate system, each axis corresponding to one amino acid. Now we rotate the coordinate system into the one defined by the eigenvectors of the  $\mathbf{P}$  matrix, in which the expression for the estimated energy is reduced to the diagonal form:

$$E(\text{estimated}) = \lambda_1 p_1^2 + \lambda_2 p_2^2 \cdots + \lambda_{20} p_{20}^2$$

Here  $\lambda_i$  is the  $i$ th eigenvalue corresponding to the  $v^i$  eigenvector, and  $p_i$  is the corresponding coordinate of the amino acid composition vector ( $\underline{n}$ ) in the new coordinate system, calculated as a scalar



**Figure 3.** Decomposition of the energy predictor matrix to eigenvalues representing stabilizing and destabilizing interactions. The energy predictor matrix  $\mathbf{P}$ , was decomposed into (a) and (b) negative and (c)–(e) positive eigenvectors. Their corresponding eigenvalues specifying the weights in the energy function are: (a)  $-52$ , (b)  $-40$ , (c)  $24$ , (d)  $13$  and (e)  $10$  (cf. Decomposition). These vectors represent stabilizing and destabilizing contributions to the total pairwise energy content, and can be rationalized in terms of simple physical principles, such as (a) hydrophobicity, (b) cysteine abundance, (d) structure-breaking amino acid residues and (e) net charge of the protein.

product,  $p_i = n v^i$ . Since  $p_i^2$  is non-negative, terms corresponding to positive/negative eigenvalues give a positive/negative contribution to the estimated energy. Some of the individual eigenvectors can be directly interpreted in terms of physico-chemical factors and linked to stabilization or destabilization depending on its sign. Figure 3 shows the two largest negative (stabilizing) and three largest positive (destabilizing) eigenvectors. We find hydrophobicity (Figure 3(a)) and cysteine content (Figure 3(b)) as dominant factors in stabilization. The vector with the highest eigenvalue is closest to the Sweet–Eisenberg empirical hydrophobicity scale (correlation coefficient: 0.94) among more than 400 different amino acid propensities collected in the AAIndex database.<sup>31</sup> Interestingly, the same scale among hydrophobicities was found to be the best for discriminating structured proteins from IUPs in a systematic search among 265 amino acid properties.<sup>16</sup> Of particular relevance to our assessment of the determinants of protein disorder, the Sweet–Eisenberg scale is based on amino acid replaceability, which correlates with the tendency of side-chains to be buried or exposed in protein crystal structures.<sup>32</sup> As for the destabilizing factors, there is no obvious interpretation of the factor with the

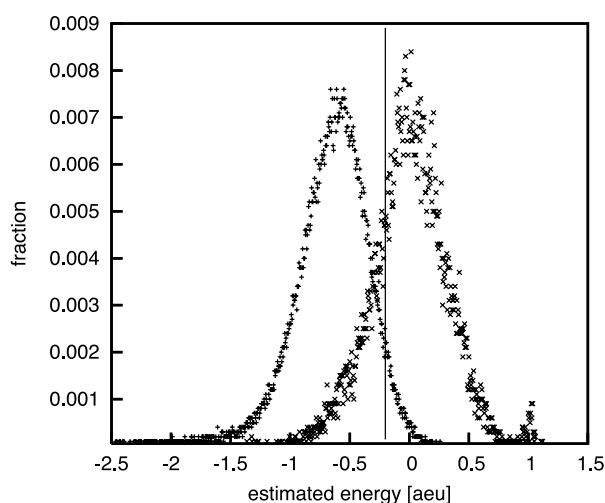
largest positive eigenvalue (Figure 3(c)), whereas the next two suggest that the abundance of structure-breaking amino acid residues like Pro, Asn, and Gly (Figure 3(d)) or high net charge (Figure 3(e)) leads to destabilization. It should be made clear, though, that these features are not incorporated in the predictor, but they can be extracted due to their importance in supporting energetically favorable/unfavorable structural states. The intriguing point is that they show good correlation with features used in previous approaches, which are knowledge-based in terms of protein disorder.<sup>14,16,17</sup>

Besides the angles between the vectors, all  $p_i$  values depend also on the norm of the amino acid composition vector ( $\text{NORM} = \sum_{i=1}^{20} n_i^2 = \sum_{i=1}^{20} p_i^2$ ). This norm takes its minimum value (0.223) for sequences with equal amino acid frequencies, and largest (1.0) when the sequence is composed of a single amino acid. For globular proteins, it varies between 0.23 and 0.39. There are 24 sequences with their norm above 0.35 in the IUP sets, including 16 out of 17 sequences that have at least 50% of residues predicted to have low complexity by the SEG program.<sup>33</sup> Thus, the norm is also a measure of the complexity of the sequence, incorporated into the estimated energy as a scaling factor. According to our model, a low complexity sequence is not necessarily disordered, if its amino acid composition is dominated by stabilizing factors, allowing the formation of favorable contacts. On the other hand, sequences with the least favorable energy are of low complexity as well, as a result of the dominance of one or a few amino acid types that have unfavorable interaction energies for each relation.

#### The estimated pairwise energy predicts protein disorder

Based on the significant separation between the estimated pairwise energies of globular and experimentally verified intrinsically unstructured proteins, this approach can be turned into a method to predict protein disorder. For this purpose it is more appropriate to consider the local sequential neighborhood only, since many proteins are not fully ordered or disordered. Thus, the original matrix  $\mathbf{P}$ , derived at the level of global sequences, was recalculated by treating each position separately, and taking into account only its predefined neighborhood in sequence. The energy and amino acid composition for each position was calculated only by considering interaction partners 2–100 residues apart. The choice of this range represents a trade-off between the intention of covering most structured domains, but separating distinct domains in multi-domain proteins. This procedure yields an estimated energy at position  $p$  of type  $i$ :

$$e_i^p = \sum_{j=1}^{20} P_{ij}^p n_j^p$$



**Figure 4.** Estimated position-specific pairwise energies of globular proteins and IUPs. The distribution of estimated position-specific pairwise energy scores is shown, calculated by considering the amino acid composition limited to a sequential neighborhood of  $\pm 100$  residues and smoothed over 21 residues. The application, termed IUPred, was applied to the Filt\_Glob\_list (+) and the IUP\_list ( $\times$ ), and the frequency of residues was plotted against their local energy content. A threshold of  $-0.2$  [aeu] provides the best separation of individual positions between these two structural classes.

where  $P^p$  is the position-specific energy predictor matrix. The position-specific estimations of energies were averaged over a window of 21 residues. This method for the prediction of protein disorder is termed IUPred.

By using IUPred, the distribution of scores for globular proteins and IUPs is as shown in Figure 4. The clear separation between the two sets is also apparent at the level of individual positions. From the distribution of globular proteins we determined a threshold where 5% of their positions were predicted as being disordered, similar to the prediction made by Ward *et al.*<sup>9</sup> This value was  $-0.2$  [aeu]: positions with energy content above this cutoff value were predicted to be disordered, whereas positions below were considered as being ordered. Using this limit, 76% of positions of IUPs were predicted to be disordered. These deviations

from complete order or disorder are fully acceptable, due to potentially disordered regions in globular proteins in solution observed to be ordered in the solid state, and the existence of significant residual structure in many IUPs.<sup>6,7,34,35</sup>

To see how the choice of the empirical force field affects the predictive power of IUPred, various 20 by 20 M scoring matrices were tested. For each M matrix the corresponding P matrix was derived as described in Theory and used for the prediction. We set the threshold to give 5% false positive predictions on the Filt\_Glob\_list, and calculated the sensitivity of the method as the percentage of predicted disorder on the IUP\_list for each interaction matrix (Table 3). The approach of Thomas & Dill<sup>20</sup> yields matrices superior to others, with the much larger dataset bringing about an improvement of almost 3%. The matrix used by Tobi *et al.*<sup>21</sup> performed comparably to the original one of Thomas & Dill in predicting disorder, but the other two showed much less ability to discriminate order from disorder.

#### Cross-validation of the method

In order to test the ability of our method to generalize on previously unseen data, we carried out a tenfold cross-validation. Glob\_list was divided into ten random subsets. One was put aside, and proteins from the remaining nine were used to calculate matrices M and P, and the cutoff value. This procedure was repeated ten times, and the goodness of fit and the amount of disorder were predicted for the proteins not used in training. It is worth noting that no cross-validation is required for IUPs, as these proteins were not included in any way in the training process.

Over the ten sets, the average of the correlation coefficient was  $0.783 \pm 0.006$  and the  $r^2$  value was  $0.600 \pm 0.070$ , compared with the values obtained for the full set, 0.786 and 0.604, respectively. Both values indicate a similar goodness of fit for globular proteins, independently of whether they were included in the training set. The amount of predicted disorder varied between 3.4% and 6.9%, with the average of  $4.96(\pm 0.97)\%$  for the training sets, compared to 5.0% for the full set.

**Table 3.** Comparison of different scoring matrices

Interaction matrix	Number of training proteins	Predicted disorder on IUP set (%) (true positives)
Thomas–Dill extended training set	785	75.95
Thomas–Dill <sup>20</sup>	37	73.25
Tobi <i>et al.</i> <sup>21</sup>	572	73.09
Mirny–Shakhnovich <sup>22</sup>	104	64.63
Miyazawa–Jernigan <sup>23</sup>	251/1661	63.64

The performance of different interaction matrices in predicting disorder and the number of proteins used to derive them. The Miyazawa–Jernigan matrix was trained on 1661 proteins including homologs, effectively representing 251 families.



**Table 4.** Performance of disorder prediction methods

Method	True positive rate		False positive rate	
	All positions (%)	Normalized positions (%)	All positions (%)	Normalized positions (%)
IUPred	76.33	67.91	5.33	5.54
PONDR VL3H	66.29	60.74	5.02	7.84
DISOPRED2	63.39	49.08	5.02	6.87
GlobPlot	32.97	30.42	18.07	19.72

Comparison of IUPred, PONDR VL3H, DISOPRED2 and GlobPlot on IUP\_list and Filt\_Glob\_list. The true positive rate was calculated as the percentage of residues predicted as disordered on the IUP\_list (sensitivity), while setting the false positive rate (percentage of predicted disordered residues on the Filt\_Glob\_set), also called specificity, to 5%, or the closest possible value (in the case of GlobPlot). These values are given averaged over all positions, and normalized by the length of the protein. This normalization weights each fragment/protein equally, independently of its length. Predictions by PONDR VL3H were collected from the server at <http://www.ist.temple.edu/disprot/predictor.php> using the default parameter (window size=1), while DISOPRED2 was downloaded from <http://bioinf.cs.ucl.ac.uk/disopred/> and run locally. GlobPlot was also run locally, but with the web server's parameters and taking the CASP-like output (<http://GlobPlot.embl.de/>).

### Comparison of different methods of disorder prediction

#### Database of disordered proteins

We compared IUPred to three widely used methods for predicting disorder, which differ not just methodologically but also conceptually due to different definitions of disorder. GlobPlot is a simple propensity-based approach evaluating the tendency of residues to be in a regular secondary structure. PONDR VL3H<sup>36</sup> was trained to distinguish experimentally verified disordered proteins from globular proteins by various machine learning approaches. In developing DISOPRED<sup>9</sup>, the definition of disorder was restrained to regions missing from X-ray structures and a support vector machine was trained to specifically recognize these. In contrast, IUPred assigns order/disorder status to residues on the basis of their ability to form favorable pairwise contacts.

To make a realistic comparison of these methods (Table 4), their cutoff values were set so that they yielded the same percentage (5%) of false positive predictions (predicted disordered when in fact ordered) on Filt\_Glob\_list. The agreement between pairs of predictions were also calculated: two predictions were said to agree when both predicted order or disorder for a given position, and the numbers of agreements were normalized by the total number of positions (Table 5). As GlobPlot was not intended as a per position prediction method, it

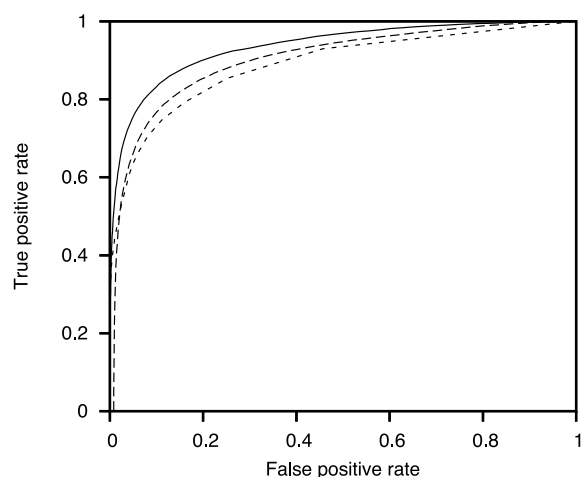
was included as a simple control to evaluate the performance of a propensity-based approach. Although the performance of the other methods, IUPred, PONDR VL3H and DISOPRED2, is comparable, there are some clear differences among them. IUPred predicted the largest amount of disorder, followed by PONDR VL3H; DISOPRED2 tended to predict the most order at the same level of false prediction rate. These differences are also apparent in the ROC curve, giving the false positive rate against true positive rate for these three methods (Figure 5). Except for very low level of false prediction rate, IUPred achieves the highest true positive rate. Intriguingly, in pairwise comparisons the three methods are very similar, each of them agreeing with the other two on about three-quarters of positions (Table 5).

The goal of this comparison was to assess the performance of IUPred in terms of predicting long disordered regions; however, it cannot be regarded as a complete benchmarking. The test set for disorder was rather small (only 129 proteins), and there could be a significant amount of local order included in this set that the various methods would treat differently. DISOPRED2 was specifically designed to predict short disordered regions in the context of globally ordered proteins, and its performance is expected to be higher on these datasets. Furthermore, some parameters (e.g. window size) could also influence the performance of the methods (VL3H, GlobPlot). Despite these limitations, the results clearly show that IUPred

**Table 5.** Similarity between disorder prediction methods

Method	Agreement (%)			
	IUPred	PONDR VL3H	DISOPRED2	GlobPlot
IUPred	100	91.61	91.97	80.76
PONDR VL3H	76.60	100	92.26	79.24
DISOPRED2	77.19	77.05	100	79.91
GlobPlot	48.07	47.84	51.31	100

The similarity between pairs of methods was calculated as the number of agreements over all positions in IUP\_list (lower triangle) and in Filt\_Glob\_list (upper triangle). Predictions were collected as given in the legend to Table 4.



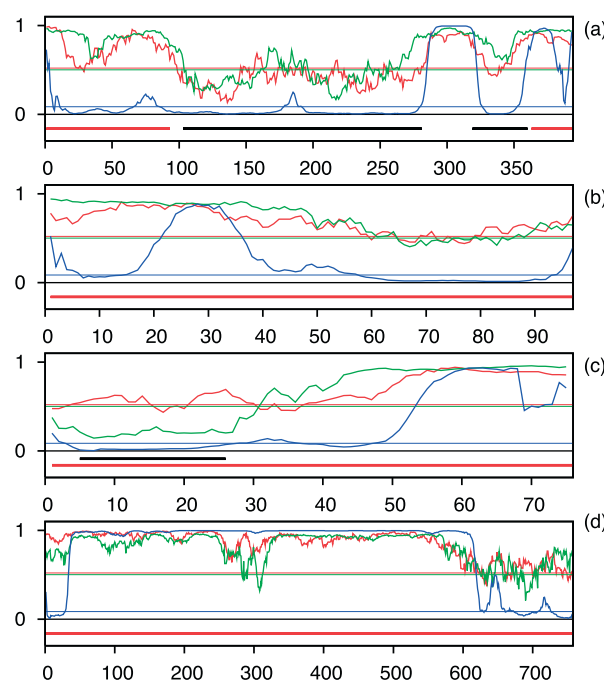
**Figure 5.** ROC curve for IUPred, PONDR VL3H, and DISOPRED2. Receiver operator characteristic (ROC) curve for IUPred (continuous), PONDR VL3H (broken) and DISOPRED2 (short broken). The true positive rate was calculated as the percentage of residues predicted as disordered on the IUP\_list (sensitivity), the false positive rate is the percentage of predicted disordered residues on the Filt\_Glob\_set, also called specificity.

is a competent predictor of protein disorder. This is achieved by considering only globular proteins during the training, without using any information on intrinsically unstructured proteins.

#### Examples of individual proteins

As a further means of comparison we present the analysis of four representative proteins, with the prediction output of the position-specific predictors IUPred, PONDR VL3H and DISOPRED2 (Figure 6). The proteins were selected because a good deal of structural information is available on the extent and mode of their disorder. p53 (Figure 6(a)) is a tumor suppressor transcription factor, the structural disorder of which has been convincingly demonstrated for the N-terminal (1–93) and C-terminal (363–393) domains.<sup>2,37</sup> FlgM, or anti-sigma-28 factor (Figure 6(b)), is one of the first proteins to be identified as intrinsically unstructured along its entire length.<sup>38</sup> Its C-terminal half, with residues 60–73 and 83–90 in particular, show some  $\alpha$ -helical preference in solution,<sup>39</sup> possibly relevant to the physiological function of this protein.<sup>40</sup> PKI-alpha (Figure 6(c)), a heat-resistant inhibitor of cAMP-dependent protein kinase, is also disordered along its entire length<sup>41</sup> with its inhibitory segment (1–13) and nuclear export signal (35–47) tending to adopt an  $\alpha$ -helical structure in the unbound state.<sup>42</sup> Microtubule-associated protein tau (Figure 6(d)) belongs to a family of heat-stable MAPs (also including MAP2 and MAP4), which are disordered along their entire length and bind microtubules *via* a C-terminal microtubule-binding domain.<sup>43</sup>

The plots by the three methods agree reasonably



**Figure 6.** Comparison of IUPred with two other predictors of disorder. IUPred scoring (red) is compared with PONDR VL3H (green) and DISOPRED2 (blue) for (a) p53, (b) FlgM, (c) PKI-alpha and (d) MAP tau. The energy values of IUPred were normalized to fall between [0,1]. Thin horizontal lines of the appropriate color represent threshold values, above which the score is characteristic of disorder (0.5 for IUPred and PONDR and 0.086 for DISOPRED2, the default values in the latter two cases). Below the scores, the region experimentally shown to be disordered (thick red line) or structured in itself or in the presence of a binding partner (thick black line) is indicated.

well, with some differences. DISOPRED2 tends to predict more order than either IUPred or PONDR VL3H, even at places where experimental evidence is for a largely disordered state, such as the N-terminal domain of p53 and PKI-alpha or the C-terminal region of FlgM. Interestingly, these regions show some tendency to be transiently ordered, as stated above. This local preference for order is probably captured by IUPred and PONDR, as witnessed by the value of their disorder score approaching or even crossing the threshold. The noted tendency of the C-terminal region of tau to be ordered is also worthy of note in light of its interaction with microtubules; this might occur *via* preformed structural elements, as demonstrated to be a general feature of IUPs.<sup>35</sup> These examples further illustrate the similarities and differences among the three prediction methods, with IUPred predicting the most disorder for fully or largely disordered proteins, and DISOPRED2 predicting the least. In addition, by looking at these examples, it is advisable to treat regions of disagreement among the predictors with caution and consider them as potential recognition sites.

## Discussion

The growing number of examples of IUPs has encouraged us to revisit the issue of the foldability of polypeptide chains. In order to understand the differences between IUPs and folded proteins better, we estimated the pairwise energy content of proteins in their native structural state by a quadratic form involving the amino acid composition vector and the energy predictor matrix. The parameters of the matrix were derived by least-squares fitting using globular proteins of known structure, which also allowed the goodness of the estimation to be tested. The robustness of this approach is quite surprising, considering that every protein structure is an intricate architecture of a multitude of interresidue contacts. We did not attempt to predict the exact pattern of these interactions, i.e. the structure, in detail, rather the compatibility of a given polypeptide with the formation of sufficient favorable interactions, as observed in globular proteins.

The success of our approach underlines some common, fundamental properties of sequences with stable folded structures. The native structure of folded proteins corresponds to a pronounced energy minimum, with no other conformations having comparable energy.<sup>44,45</sup> Ensuring this energetic separation demands the native structure to efficiently use the interactions compatible with the given sequence. As the maximum capacity for each amino acid to participate in these interactions is limited by its chemical nature, the amino acid composition can be related to the total interaction energy pertaining to the most favorable interaction pattern among all residues present in a protein. The effectiveness of our model in estimating the pairwise energy content can be attributed to folded structures being close to the optimal energy level allowed by the amino acid residues in the sequence.

The energy per residue of stably folded proteins falls into a quite narrow range, dominated by favorable interactions; the total pairwise energy estimated by our approach is consistent with this energy range. In contrast, the predicted energy of IUPs is higher. An important conceptual point is that a polypeptide with an amino acid composition compatible with a folded structure does not necessarily have a unique structure. This can be easily demonstrated by considering the random permutation of sequences of folded proteins. Although we predict the same energy for the myriad of sequences compatible with a particular amino acid composition, for most of them we expect no corresponding unique structure. Similarly, we predict globular-like energy for truncated domains or proteins, although these sequences are not likely to fold on their own. Nonetheless, these polypeptides are not IUPs either, since they exhibit some tendency to form contacts. The way around this dilemma in the present approach is that it predicts the optimum of energy, which is not generally achievable by a random sequence. Folded

structures, however, are realized by highly evolved sequences, compatible with these energies. IUP sequences are also special, selected by evolution to avoid the formation of favorable contacts in any conformation. The finding that the estimated total interaction energy reproduces the basic difference between structured and disordered proteins basically underlines the concept of protein disorder, i.e. that the lack of a well-defined 3D structure is an intrinsic property of certain evolved proteins.

Our ability to reproduce these special features depends on using the right potentials for approaching the actual interaction energies. The goodness of such extracted potentials is usually tested by their ability to identify the native structure as the lowest energy state among all the proteins in a dataset. The particular approach, proposed by Thomas & Dill,<sup>20</sup> relies on the Boltzmann relation to extract energy-like quantities from amino acid pairing frequencies, but relative to a reference state obtained through an iterative protocol to reflect the predicted ensemble of interactions. This approach aims not only at discriminating the native structure from decoys but also at giving the ratios of the interaction energies correctly. Thus, these potentials are the closest to reproducing the true energies that drive amino acid residues to form, or avoid, contacts. This could explain why the Thomas–Dill matrix outperformed other matrices in estimating the pairwise energy content of disordered proteins.

In the light of this special property of the underlying interaction matrix, we can also interpret the unexpected finding that the average energy level of IUPs is very close to zero, i.e. stabilizing and destabilizing interactions cancel. Although the absolute energy values were arbitrary, this finding is invariant to scaling, thus this energetic neutrality is a genuine property of disordered proteins. At the level of individual proteins, this neutrality may result from an overall lack of long-range interactions, but also from the balance of local organization and long-distance repulsion. For some IUPs, however, the balance appears to be set off towards net stabilization. Indeed, the predicted pairwise energy content of proteins with a molten-globule type of disorder<sup>30</sup> on average is more favorable compared to coil-like disordered proteins. It will be interesting to see how such individual structural features correlate with function.

As seen, our approach provides a realistic approximation of structural interaction energy of proteins, enabling the prediction of intrinsic structural disorder. This idea of the importance of interaction capacity has also been raised recently by work in which the average number of contacts per residue was used as an indicator of disorder.<sup>46</sup> Our quadratic formula combined with the energy predictor matrix captures the energetic aspect of this observation. By limiting the calculation to a predefined sequential neighborhood, it yields a position-specific score characteristic of the tendency of a given amino acid to fall into a structurally ordered or disordered region. This application we

term IUPred and intend to make it publicly available *via* the Internet. The logic of IUPred differs from previous prediction algorithms, which were trained on disordered proteins/segments. As already alluded to, these approaches mostly suffer from inconsistencies in the underlying databases, i.e. the inclusion of sequences of intrinsic order classified as disordered and sequences of intrinsic disorder classified as ordered. As IUPred was not trained on such data, its unbiased assessment of the structural status of an unknown sequence/segment is of confirmatory value. We have tested this conclusion by comparing IUPred with generally accepted predictors, PONDR VL3H, DISOPRED2 and GlobPlot, on disordered databases and by examining predictors on individual proteins. Although predictions were similar with the IUP dataset, IUPred predicted the most disorder and agreed best with experimental data (Table 4).

Decomposition of the energy matrix connects our model to previous attempts to predict IUPs by using simple physico-chemical properties of proteins. Some of the eigenvectors with the largest eigenvalues showed strong correlations with physico-chemical parameters, such as hydrophobicity, cysteine content, structure-breaking properties and net charge (Figure 3). Two of these, hydrophobicity and net charge have been used in the Uversky plot to separate globular proteins and IUPs,<sup>17</sup> and the importance of structure-breaking amino acid residues has also been noted.<sup>5,6</sup> We found that the eigenvector with the highest eigenvalue matches the Sweet–Eisenberg hydrophobicity scale<sup>32</sup> the best, in accordance with a previous analysis of amino acid factors discriminating structured proteins and IUPs.<sup>16</sup> This concurrence vindicates our approach, as it does not rely on prior experimental data on IUPs, still it automatically finds and combines the properties that are important for this task. Our predictor combines these factors in a quadratic function, which distinguishes it from previous, propensity-based linear predictors.<sup>13–15,46</sup> As a result of the higher-order statistics, the contribution of a given amino acid to disorder/order discrimination is context-dependent, i.e. it depends on the amino acid type as well on the amino acid composition of the sequential neighborhood of the given residue. For example, the contribution of Lys would be different if it is surrounded by other positive charges, implying an increase in the probability of unfavourable interactions, than if it is surrounded by negatively charged residues. This is in accordance with high net charge, and not simply the total charge, being one of the key determinants of disorder.<sup>17</sup> This interdependence of residues is manifest in the appearance of flavors of disorder.<sup>13</sup>

In summary, our model estimates the pairwise energy of proteins from their amino acid compositions. This allows us to test sequences for foldability, even in the absence of a structural model. By sequentially limiting the calculation, it serves as a predictor of protein disorder. By

applying this scheme for IUPs, we showed that these proteins have a special amino acid composition, which, independently of the actual sequence, does not allow the formation of sufficient favorable contacts expected for folded proteins. Given the heterogeneity and ambiguity of experimental techniques used to demonstrate the lack of structure so far, a key inference from our studies is that IUPs share a common property that distinguishes them from the class of folded proteins.

## Materials and Methods

### Databases

For the purpose of parameter fitting, the September 2001 release of the PDB-select database<sup>47</sup> with <25% sequence identity cutoff was used. Entries with resolution worse than 2.5 Å, with chain breaks or with C $\alpha$  atoms only, were omitted; the resulting dataset contained 953 protein chains. During the force field optimization, we considered the native structure for non-transmembrane sequences with length between 40 and 350, reducing the number of proteins to 785 (Glob\_list), but all structures were used as a skeleton to generate decoys.

In principle, this list could also contain IUPs, e.g. as part of multichain complexes. For the purpose of testing we created a filtered list of globular proteins with the aim to eliminate the potentially dubious cases. A newer release of PDB-select (April 2002) was used, and all entries involving multiple chains, transmembrane segments, or the binding of nucleic acid residues, heme, or metal ions were omitted, resulting in 559 proteins (Filt\_Glob\_list). The two lists (Glob\_list and Filt\_Glob\_list) are given in the Supplementary Data (Tables S1 and S2).

The IUP dataset (IUP\_list) contained 129 proteins and protein segments with experimentally verified disordered status. The complete list is given in the Supplementary Data (Table S3). The total number of residues in this set is 26,794.

### Force field optimization

A coarse-grained approach was used to describe the interactions between residues. Amino acid residues were treated as single interaction centers located at their C $\beta$  atom (virtual C $\beta$  in the case of Gly). The low-resolution energy of contacts between different amino acid residues, expressed in the form of a 20 by 20 matrix, was calculated from the observed frequencies of amino acid pairs. The interaction matrix was calculated by the iterative algorithm proposed by Thomas & Dill,<sup>20</sup> but on 785 proteins (Glob\_list) instead of the original 37. The resulting matrix **M** is given in Table 1.

---

---

## Acknowledgements

This work was supported by grants T34131 and F043609 from OTKA, Bolyai János fellowships for Zs.D. and P.T., and the International Senior Research Fellowship GR067595 from the Wellcome Trust for P.T. The fruitful discussions with Nicholas

E. Dixon and Tamas Hauer are gratefully acknowledged.

## Supplementary Data

Supplementary data associated with this article can be found, in the online version, at [doi:10.1016/j.jmb.2005.01.071](https://doi.org/10.1016/j.jmb.2005.01.071)

## References

- Kriwacki, R. W., Hengst, L., Tennant, L., Reed, S. I. & Wright, P. E. (1996). Structural studies of p21Waf1/Cip1/Sdi1 in the free and Cdk2-bound state: conformational disorder mediates binding diversity. *Proc. Natl Acad. Sci. USA*, **93**, 11504–11509.
- Dawson, R., Muller, L., Dehner, A., Klein, C., Kessler, H. & Buchner, J. (2003). The N-terminal domain of p53 is natively unfolded. *J. Mol. Biol.* **332**, 1131–1141.
- Radhakrishnan, I., Perez-Alvarado, G. C., Dyson, H. J. & Wright, P. E. (1998). Conformational preferences in the Ser133-phosphorylated and non-phosphorylated forms of the kinase inducible transactivation domain of CREB. *FEBS Letters*, **430**, 317–322.
- Wright, P. E. & Dyson, H. J. (1999). Intrinsically unstructured proteins: re-assessing the protein structure–function paradigm. *J. Mol. Biol.* **293**, 321–331.
- Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M. & Obradovic, Z. (2002). Intrinsic disorder and protein function. *Biochemistry*, **41**, 6573–6582.
- Tompa, P. (2002). Intrinsically unstructured proteins. *Trends Biochem. Sci.* **27**, 527–533.
- Tompa, P. (2003). The functional benefits of protein disorder. *J. Mol. Struct. (Theochem)*, **666–667**, 361–371.
- Dunker, A. K., Obradovic, Z., Romero, P., Garner, E. C. & Brown, C. J. (2000). Intrinsic protein disorder in complete genomes. *Genome Inform. Ser. Workshop Genome Inform.* **11**, 161–171.
- Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F. & Jones, D. T. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* **337**, 635–645.
- Iakoucheva, L., Brown, C., Lawson, J., Obradovic, Z. & Dunker, A. (2002). Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.* **323**, 573–584.
- Liu, J., Tan, H. & Rost, B. (2002). Loopy proteins appear conserved in evolution. *J. Mol. Biol.* **322**, 53–64.
- Weathers, E. A., Paulaitis, M. E., Woolf, T. B. & Hoh, J. H. (2004). Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein. *FEBS Letters*, **576**, 348–352.
- Vucetic, S., Brown, C. J., Dunker, A. K. & Obradovic, Z. (2003). Flavors of protein disorder. *Proteins: Struct. Funct. Genet.* **52**, 573–584.
- Xie, Q., Arnold, G. E., Romero, P., Obradovic, Z., Garner, E. & Dunker, A. K. (1998). The sequence attribute method for determining relationships between sequence and protein disorder. *Genome Inform. Ser. Workshop Genome Inform.* **9**, 193–200.
- Linding, R., Russell, R. B., Neduva, V. & Gibson, T. J. (2003). GlobPlot: exploring protein sequences for globularity and disorder. *Nucl. Acids Res.* **31**, 3701–3708.
- Williams, R. M., Obradovic, Z., Mathura, V., Braun, W., Garner, E. C., Young, J. et al. (2001). The protein non-folding problem: amino acid determinants of intrinsic order and disorder. *Pac. Symp. Biocomput.* **6**, 89–100.
- Uversky, V. N., Gillespie, J. R. & Fink, A. L. (2000). Why are “natively unfolded” proteins unstructured under physiologic conditions? *Proteins*, **41**, 415–427.
- Miyazawa, S. & Jernigan, R. L. (1985). Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, **18**, 534–552.
- Sippl, M. J. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* **213**, 859–883.
- Thomas, P. D. & Dill, K. A. (1996). An iterative method for extracting energy-like quantities from protein structures. *Proc. Natl Acad. Sci. USA*, **93**, 11628–11633.
- Tobi, D., Shafran, G., Linial, N. & Elber, R. (2000). On the design and analysis of protein folding potentials. *Proteins: Struct. Funct. Genet.* **40**, 71–85.
- Mirny, L. A. & Shakhnovich, E. I. (1996). How to derive a protein folding potential? A new approach to an old problem. *J. Mol. Biol.* **264**, 1164–1179.
- Miyazawa, S. & Jernigan, R. L. (1996). Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* **256**, 623–644.
- Melo, F., Sanchez, R. & Sali, A. (2002). Statistical potentials for fold assessment. *Protein Sci.* **11**, 430–448.
- Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature*, **358**, 86–89.
- Torda, A. E. (1997). Perspectives in protein-fold recognition. *Curr. Opin. Struct. Biol.* **7**, 200–205.
- Gohlke, H., Hendlich, M. & Klebe, G. (2000). Knowledge-based scoring function to predict protein–ligand interactions. *J. Mol. Biol.* **295**, 337–356.
- Kolinski, A. & Skolnick, J. (1994). Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. *Proteins: Struct. Funct. Genet.* **18**, 338–352.
- Khatun, J., Khare, S. D. & Dokholyan, N. V. (2004). Can contact potentials reliably predict stability of proteins? *J. Mol. Biol.* **336**, 1223–1238.
- Uversky, V. N. (2002). Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.* **11**, 739–756.
- Kawashima, S., Ogata, H. & Kanehisa, M. (1999). AAindex: amino acid index database. *Nucl. Acids Res.* **27**, 368–369.
- Sweet, R. M. & Eisenberg, D. (1983). Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. *J. Mol. Biol.* **171**, 479–488.
- Wootton, J. C. & Federhen, S. (1996). Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266**, 554–571.
- Uversky, V. N. (2002). What does it mean to be natively unfolded? *Eur. J. Biochem.* **269**, 2–12.
- Fuxreiter, M., Simon, I., Friedrich, P. & Tompa, P. (2004). Prefolded structural elements feature in partner recognition by intrinsically unstructured proteins. *J. Mol. Biol.* **338**, 1015–1026.
- Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., Brown, C. J. & Dunker, A. K. (2003). Predicting intrinsic disorder from amino acid sequence. *Proteins*, **53**, 566–572.

37. Bell, S., Klein, C., Muller, L., Hansen, S. & Buchner, J. (2002). p53 contains large unstructured regions in its native state. *J. Mol. Biol.* **322**, 917.
38. Daughdrill, G. W., Chadsey, M. S., Karlinsey, J. E., Hughes, K. T. & Dahlquist, F. W. (1997). The C-terminal half of the anti-sigma factor. FlgM, becomes structured when bound to its target, sigma 28. *Nature Struct. Biol.* **4**, 285–291.
39. Daughdrill, G. W., Hanely, L. J. & Dahlquist, F. W. (1998). The C-terminal half of the anti-sigma factor FlgM contains a dynamic equilibrium solution structure favoring helical conformations. *Biochemistry*, **37**, 1076–1082.
40. Dedmon, M. M., Patel, C. N., Young, G. B. & Pielak, G. J. (2002). FlgM gains structure in living cells. *Proc. Natl Acad. Sci. USA*, **99**, 12681–12684.
41. Hauer, J. A., Taylor, S. S. & Johnson, D. A. (1999). Binding-dependent disorder–order transition in PKI alpha: a fluorescence anisotropy study. *Biochemistry*, **38**, 6774–6780.
42. Hauer, J. A., Barthe, P., Taylor, S. S., Parello, J. & Padilla, A. (1999). Two well-defined motifs in the cAMP-dependent protein kinase inhibitor (PKIalpha) correlate with inhibitory and nuclear export function. *Protein Sci.* **8**, 545–553.
43. Schweers, O., Schonbrunn-Hanebeck, E., Marx, A. & Mandelkow, E. (1994). Structural studies of tau protein and Alzheimer paired helical filaments show no evidence for beta-structure. *J. Biol. Chem.* **269**, 24290–24297.
44. Shakhnovich, E. I. & Gutin, A. M. (1993). Engineering of stable and fast-folding sequences of model proteins. *Proc. Natl Acad. Sci. USA*, **90**, 7195–7199.
45. Onuchic, J. N. & Wolynes, P. G. (2004). Theory of protein folding. *Curr. Opin. Struct. Biol.* **14**, 70–75.
46. Garbuzynskiy, S. O., Lobanov, M. Y. & Galzitskaya, O. V. (2004). To be folded or to be unfolded? *Protein Sci.* **13**, 2871–2877.
47. Hobohm, U. & Sander, C. (1994). Enlarged representative set of protein structures. *Protein Sci.* **3**, 522–524.

*Edited by J. Thornton*

*(Received 27 October 2004; received in revised form 26 January 2005; accepted 28 January 2005)*

## Structural bioinformatics

**IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content**

Zsuzsanna Dosztányi\*, Veronika Csizmok, Peter Tompa and István Simon

Institute of Enzymology, BRC, Hungarian Academy of Sciences, PO Box 7, H-1518 Budapest, Hungary

Received on March 24, 2005; revised on May 27, 2005; accepted on June 13, 2005

Advance Access publication June 14, 2005

**ABSTRACT**

**Summary:** Intrinsically unstructured/disordered proteins and domains (IUPs) lack a well-defined three-dimensional structure under native conditions. The IUPred server presents a novel algorithm for predicting such regions from amino acid sequences by estimating their total pairwise interresidue interaction energy, based on the assumption that IUP sequences do not fold due to their inability to form sufficient stabilizing interresidue interactions. Optional to the prediction are built-in parameter sets optimized for predicting short or long disordered regions and structured domains.

**Availability:** The IUPred server is available for academic users at <http://iupred.enzim.hu>

**Contact:** zsuzsa@enzim.hu

**INTRODUCTION**

Intrinsically unstructured proteins exist as an ensemble of alternative conformations, in contrast to folded, globular proteins that have unique native structure. Significant fraction of known genomes encode for proteins with regions of disordered structure. In some eukaryotic genomes >20% of the coded residues are predicted as disordered (Dunker *et al.*, 2000; Ward *et al.*, 2004a). In many cases a protein is fully disordered, while in many other cases there are long disordered segments in otherwise ordered, folded proteins (Tompa, 2002; Dyson and Wright, 2005). Despite their lack of a well-defined globular structure, these proteins carry out basic functions (Iakoucheva *et al.*, 2002; Ward *et al.*, 2004a), mostly associated with signal transduction, cell-cycle regulation and transcription. Several methods have been developed to predict the disordered character from amino acid sequences. Some are based on the special amino acid composition of fully disordered proteins, i.e. the abundance of hydrophilic residues and a high net charge (Uversky *et al.*, 2000; Vucetic *et al.*, 2003), whereas others use various machine learning approaches trained on specific datasets (Obradovic *et al.*, 2003; Ward *et al.*, 2004a; Linding *et al.*, 2003b). Recently, it was suggested that these sequences do not have the capacity to properly wrap backbone hydrogen bonds (Fernandez and Berry, 2004), which has also been shown to be important for protein stability.

**BACKGROUND**

Our method is footed on the physical explanation of the ordered/disordered nature of proteins. Globular proteins make a large

number of interresidue interactions, providing the stabilizing energy to overcome the entropy loss during folding (Garbuzynskiy *et al.*, 2004). In contrast, intrinsically unstructured/disordered proteins and domains (IUPs) have special sequences that do not have the capacity to form sufficient interresidue interactions. To discriminate between ordered and disordered regions in proteins, we have developed a new approach that estimates the potential of polypeptides to form such stabilizing contacts by using a statistical interaction potential (Thomas and Dill, 1996; Dosztányi *et al.*, 2005). It was shown that the sum of interaction energies can be estimated by a quadratic expression in the amino acid composition, which takes into account that the contribution of an amino acid to order/disorder depends not only on its own chemical type, but also on its potential interaction partners (Dosztányi *et al.*, 2005).

The calculation involves a  $20 \times 20$  energy predictor matrix, parameterized by a statistical method to approach the expected pairwise energy of globular proteins of known structure. Comparing globular proteins and disordered ones, a clear separation of their energy content is found (Dosztányi *et al.*, 2005). As no training on disordered proteins is involved, this distinction underlines that the lack of a well-defined three-dimensional structure is an intrinsic property of certain evolved proteins. This approach was turned into a position-specific method to predict protein disorder by considering only the local sequential environment of residues within 2–100 residues in either direction. The score is then smoothed over a window-size of 21. This prediction method (IUPred), when tested on datasets of globular proteins and long disordered protein segments, showed improved performance over some other widely used methods, such as DISOPRED2 (Ward *et al.*, 2004a,b) and PONDR VL3H (Obradovic *et al.*, 2003).

**THE IUPred SERVER**

The web server takes a single amino acid sequence as an input and calculates the pairwise energy profile along the sequence. The energy values are then transformed into a probabilistic score ranging from 0 (complete order) to 1 (complete disorder). Residues with a score above 0.5 can be regarded as disordered. Optional is the prediction of long disorder, short disorder, and structured domains, each using slightly different parameters. The main profile of our server is to predict context-independent global disorder that encompasses at least 30 consecutive residues of predicted disorder. A different set of parameters is suited for predicting short, probably context-dependent, disordered regions such as missing residues in the X-ray

\*To whom correspondence should be addressed.

structure of an otherwise globular protein. For this application the sequential neighborhood of only 25 residues is considered. As chain termini of globular proteins are often disordered in X-ray structures, this is taken into account by an end-adjustment parameter that favors disorder prediction at the ends.

The dependable identification of ordered regions is a crucial step in target selection for structural studies and structural genomics projects (Linding *et al.*, 2003a). Finding putative structured domains suitable for structure determination is another potential application of this server. In this case the algorithm takes the energy profile and finds continuous regions confidently predicted ordered. Neighboring regions close to each other are merged, while regions shorter than the minimal domain size of at least 30 residues are ignored. When this prediction type is selected, the region(s) predicted to correspond to structured/globular domains are returned.

The core program to calculate the pairwise energy profile and disorder probability is written in C, the web server is written in PHP. The calculation of the energy profile is based on single sequence, without time-consuming alignment calculations. To further facilitate the easy accessibility for scripting, a simple text output is generated on default. However, the user can also request a graphical output. The plot shows the disorder tendency of each residue along the sequence. The plot is generated by the JpGraph software (JpGraph, 2005, <http://www.aditus.hu/jpgraph/>) on the fly, without storing the graphical images on the local machine. When the prediction type of structured domains is selected, these are highlighted on the plot by thick lines. For long sequences, the graph is shown for fragments of user-defined fixed length, 500 on default.

## ACKNOWLEDGEMENTS

This work has been sponsored by grants GVOP-3.1.1.-2004-05-0143/3.0, OTKA F043609, T049073, and NKFP MediChem2 1/A/005/2004. Z.D. and P.T. were supported by the Bolyai János

Scholarship. P.T. would like to acknowledge the support of the International Senior Research Fellowship GR067595 from the Wellcome Trust.

*Conflict of Interest:* none declared.

## REFERENCES

- Dosztányi,Z. *et al.* (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.*, **347**, 827–839.
- Dunker,A.K. *et al.* (2000) Intrinsic protein disorder in complete genomes. *Genome Inform. Ser. Workshop Genome Inform.*, **11**, 161–171.
- Dyson,H.J. and Wright,P.E. (2005) Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.*, **6**, 197–208.
- Fernandez,A. and Berry,R.S. (2004) Molecular dimension explored in evolution to promote proteomic complexity. *Proc. Natl Acad. Sci. USA*, **101**, 13460–13465.
- Garbuzynskiy,S.O. *et al.* (2004) To be folded or to be unfolded? *Protein Sci.*, **13**, 2871–2877.
- Iakoucheva,L.M. *et al.* (2002) Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.*, **323**, 573–584.
- JpGraph (2005) *JpGraph*. Aditus Consulting.
- Linding,R. *et al.* (2003a) GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res.*, **31**, 3701–3708.
- Linding,R. *et al.* (2003b) Protein disorder prediction: implications for structural proteomics. *Structure (Camb)*, **11**, 1453–1459.
- Obradovic,Z. *et al.* (2003) Predicting intrinsic disorder from amino acid sequence. *Proteins*, **53** (Suppl. 6), 566–572.
- Thomas,P.D. and Dill,K.A. (1996) An iterative method for extracting energy-like quantities from protein structures. *Proc. Natl Acad. Sci. USA*, **93**, 11628–11633.
- Tompa,P. (2002) Intrinsically unstructured proteins. *Trends Biochem. Sci.*, **27**, 527–533.
- Uversky,V.N. *et al.* (2000) Why are ‘natively unfolded’ proteins unstructured under physiologic conditions? *Proteins*, **41**, 415–427.
- Vucetic,S. *et al.* (2003) Flavors of protein disorder. *Proteins*, **52**, 573–584.
- Ward,J.J. *et al.* (2004a) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.*, **337**, 635–645.
- Ward,J.J. *et al.* (2004b) The DISOPRED server for the prediction of protein disorder. *Bioinformatics*, **20**, 2138–2139.



## Disorder and Sequence Repeats in Hub Proteins and Their Implications for Network Evolution

Zsuzsanna Dosztányi,<sup>†</sup> Jake Chen,<sup>‡,§</sup> A. Keith Dunker,<sup>||</sup> István Simon,<sup>†</sup> and Peter Tompa<sup>\*,†</sup>

*Institute of Enzymology, Biological Research Center, Hungarian Academy of Sciences, 1518 Budapest, Hungary, School of Informatics, Indiana University, Indianapolis, Indiana 46202, Department of Computer and Information Science, Purdue School of Science, Indianapolis, Indiana 46202, and Center for Computational Biology and Bioinformatics, School of Medicine, Indiana University, Indianapolis, Indiana 46202*

Received April 14, 2006

Protein interaction networks display approximate scale-free topology, in which hub proteins that interact with a large number of other proteins determine the overall organization of the network. In this study, we aim to determine whether hubs are distinguishable from other networked proteins by specific sequence features. Proteins of different connectednesses were compared in the interaction networks of *Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Homo sapiens* with respect to the distribution of predicted structural disorder, sequence repeats, low complexity regions, and chain length. Highly connected proteins (“hub proteins”) contained significantly more of, and greater proportion of, these sequence features and tended to be longer overall as compared to less connected proteins. These sequence features provide two different functional means for realizing multiple interactions: (1) extended interaction surface and (2) flexibility and adaptability, providing a mechanism for the same region to bind distinct partners. Our view contradicts the prevailing view that scaling in protein interactomes arose from gene duplication and preferential attachment of equivalent proteins. We propose an alternative evolutionary network specialization process, in which certain components of the protein interactome improved their fitness for binding by becoming longer or accruing regions of disorder and/or internal repeats and have therefore become specialized in network organization.

**Keywords:** disordered protein • unstructured protein • protein–protein interaction • interaction network • hub protein

### Introduction

Protein function at the cellular level has a significant contextual component determined by the multitude of interactions among proteins in the living cell. Therefore, a primary focus of post-genomic molecular biology has been to catalog and interpret all the interactions of the proteome. As a result of large-scale proteomic efforts, significant progress has been made. Interaction network data have been generated for several organisms, including *Saccharomyces cerevisiae* (*S. cerevisiae*),<sup>1</sup> *Drosophila melanogaster* (*D. melanogaster*),<sup>2</sup> *Caenorhabditis elegans* (*C. elegans*),<sup>3</sup> and *Homo sapiens* (*H. sapiens*).<sup>4</sup> Although these distinct interactomes differ in many details, all seem to display a scale-free topology, or at least an approximation thereof,<sup>5</sup> characterized by a power-law distribution of the degree of connectivity.<sup>6–8</sup> In such networks, most proteins (the network nodes) are connected to relatively fewer, highly connected proteins (the hubs). These hub proteins play es-

sential roles in organizing the network. The presence of hubs explains the salient features of biological networks such as robustness, because random removal of nodes is much better tolerated in a scale-free network than in a random network. This robustness resulting from the scale-free topology is of prime importance in cell survival.<sup>8</sup>

Because protein interactomes share the scale-free topology with many other networks in nature, it has been suggested that their emergence has been governed by the same underlying principles, i.e., steady and random growth and preferential attachment to already highly connected nodes.<sup>6</sup> While this random-growth (gene-duplication) model agrees with a variety of considerations and observations,<sup>9,10</sup> it oversimplifies the biology of protein interactions. In particular, this model does not consider differences among proteins nor the potential of proteins to adapt their interaction capacity to their specialized functions through molecular evolution. In fact, recently it has been formally demonstrated that scale-free topology in protein interactomes can arise from varying fitness values of nodes<sup>11</sup> and can be explained by simple genetic events, without assuming a selective pressure on network topology itself.<sup>12</sup> Furthermore, it has been suggested that the large interaction capacity of hubs might be directly manifested in discernible

\* Corresponding author. Telephone: (361) 279-3143. Fax: (361) 466-5465. E-mail: tompa@enzim.hu.

<sup>†</sup> Hungarian Academy of Sciences.

<sup>‡</sup> School of Informatics, Indiana University.

<sup>§</sup> Purdue University.

<sup>||</sup> School of Medicine, Indiana University.

**Table 1.** Selected Properties of the Databases<sup>a</sup>

	YEAST	YEAST_CORE	WORM	FLY	FLY_CONF	HUMAN
Interactions <sup>a</sup>	10741	6600	3992	20433	4733	25207
proteins <sup>b</sup>	4358	2640	2616	7003	4646	7560
proteins in IC_1 <sup>c</sup>	1578	793	1527	2282	2569	1997
sequences in genome <sup>d</sup>	6357	6357	19957	18484	18484	32035
no. of HUBS (cutoff) <sup>e</sup>	766 (5)	141 (5)	282 (5)	1910 (5)	213 (5)	2329 (5)
no. of HUBS (cutoff) <sup>f</sup>	398 (9)	228 (12)	229 (6)	651 (15)	360 (4)	721 (14)
maximum interaction <sup>g</sup>	288	111	187	175	40	188

<sup>a</sup> The table shows the total number of interactions (a), the total number interacting proteins (b), the number of proteins with one interaction (c), the number of sequences in the genomes (d), the number of HUB proteins (the cutoff value used to define HUB proteins) with fixed cutoff (e), the number of HUB proteins (the cutoff value used to define HUB proteins) with floating cutoff (f), and the maximum interaction of a single protein (g) for the various datasets used.

physicochemical feature(s).<sup>9</sup> Our goal in this study is to explore whether hub proteins are enriched in features such as intrinsic structural disorder and sequence repeats.

Recently, it has become clear that a large fraction of eukaryotic proteins lack a well-defined 3D structure, but manifest their functions in an intrinsically unstructured or disordered state.<sup>13–19</sup> Computational studies have shown that this feature increases with the increasing complexity of the organisms and prevails in regulatory and signal-transduction proteins.<sup>20,21</sup> This structural state confers important functional features, such as increased interaction capacity,<sup>22–24</sup> enhanced association rates,<sup>25–27</sup> and adaptability to different partners.<sup>28,29</sup> Apparently, these features are of potential benefit in functions realized in protein interaction networks, as suggested previously.<sup>30,31</sup> In fact, disorder has been noted to contribute to hub characteristics in several ways, where hubs may be mostly disordered, partially disordered, or ordered (i.e. highly structured), but interacting with disordered partners.<sup>31</sup> Disordered proteins or segments are often generated by the expansion of internal repeat regions<sup>32</sup> and often concomitantly exhibit low-complexity amino acid compositions.<sup>33,34</sup> Internal repeat regions can also encode for recurring structural elements in ordered proteins, whose presence could lead to generation of novel proteins or functional variants.<sup>35</sup> Whether disordered or ordered, sequence repeats might afford proteins enhanced evolutionary prospects due to an enlarged available surface area, which predisposes them for functioning via protein–protein interactions.<sup>35</sup> Since many disordered regions do not contain the salient sequence features of low complexity or repeats,<sup>36</sup> and low complexity does not represent an absolute discriminator for order and disorder,<sup>36</sup> disorder prediction is needed to indicate the lack of specific 3D structure in the absence of ligands and partners.<sup>16–19</sup> In fact, because of the significant difference in the various attributes of sequences encoding for disordered and ordered proteins, prediction of intrinsic disorder in proteins can play a crucial role in the comparison and analysis of different proteomes.<sup>21,37</sup>

Motivated by these implications, we have used bioinformatics methods to predict protein disorder, sequence repeats, and segments of low complexity in the interaction networks of *S. cerevisiae* (YEAST), *D. melanogaster* (FLY), *C. elegans* (WORM), and *H. sapiens* (HUMAN). In this study, we found that hub proteins tend to be larger and contain significantly longer and more frequent regions of these sequence features, which indicate such features as the structural basis for the large interaction capacity of hub proteins. Our extensive global analysis and findings provide strong evidence supporting the generalization of recent observations regarding the importance of disorder for a few hub proteins<sup>31</sup> and also for a highly restricted interactome dataset.<sup>38</sup> In more general terms, our

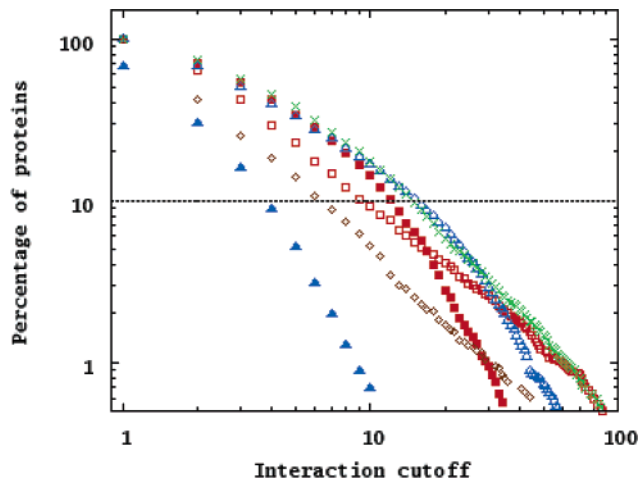
observations challenge the simplistic view that the evolution of scale-free behavior by random growth leads to preferential attachment.<sup>6</sup> We offer a more dynamic and realistic evolutionary perspective, in which network specialization is primarily accomplished via evolution of hub proteins through the accrual of “hub-friendly” sequence features.

## Materials and Methods

**Sequence Features.** Proteins were analyzed for four sequence features: length, regions of predicted protein disorder, low complexity, and sequence repeats. Sequences were downloaded from GenBank and were studied using the programs available via the Internet as follows: protein disorder, IUPred<sup>39,40</sup> available at <http://iupred.enzim.hu/>; low complexity, SEG<sup>33</sup> downloaded from the ftp site <ftp://ftp.ncbi.nlm.nih.gov/pub/seg/>; and sequence repeats, internal repeat finder<sup>41</sup> available at <http://nihserver.mbi.ucla.edu/Repeats/>. Disorder and low complexity could be assigned to residues, from which the percentage of the sequence covered could be calculated. In the case of repeats, however, the program returns only an estimation of the repeat number and repeat length, from which percentage coverage, instead of the actual location of the repeat region, could be calculated. Various combinations of these properties, calculated for each protein as the maximum percentage of the properties, were also analyzed.

**Datasets.** The list of all protein–protein interactions of yeast (YEAST) was downloaded from the BIND database (2004 November release). A “core dataset” of the interactome is also defined as the subset of interactions observed by several different large- or small-scale experiments, or confirmed by studies on paralogues. This YEAST\_CORE dataset was downloaded from the DIP database.<sup>42</sup> The *C. elegans* interactions (WORM) were also taken from BIND.<sup>3</sup> The *Drosophila* interactome (FLY) was downloaded from the CuraGen database at <http://www.curagen.com/>.<sup>2</sup> The interaction file also specifies a reliability score for each interaction. The subset of the “confident interactions” (FLY\_CONF) was also separately analyzed. The data for human interacting proteins (HUMAN) were downloaded from human protein reference database <http://www.hprd.org/>.<sup>4</sup> The major features of the datasets are shown in Table 1.

Hub proteins (HUBs) were defined in two different ways. On one hand, we applied a fixed cutoff, when proteins with five or more interactions were defined as candidate hubs, similarly to a previous work.<sup>38</sup> On the other hand, because hub function is a system property and it cannot be appropriately defined at the level of individual proteins, we also applied a floating cutoff definition, in which a unique cutoff was set for each interactome depending on the dataset. In this case, hub proteins were



**Figure 1.** Degree of connectivity in the interaction databases. The percentage of proteins with interacting partners above the given cutoff value is shown for YEAST (red squares), FLY (blue triangles), WORM (brown diamonds), and HUMAN (green x) databases on a log–log scale. The filled symbols correspond to confident/core interactions. A power-law distribution assumes a linear relationship.

defined as the top 10% of proteins with the highest number of interacting partners (cf. Table 1, Figure 1). The properties of hub proteins were compared to two reference datasets. IC\_1 contains proteins with exactly one interaction. The other reference dataset was RAN, a random sample of genome sequences. To generate RAN, the appropriate genome sequences were downloaded from the COGENT database at <http://cgg.ebi.ac.uk/services/cogent/info.html>. Due to computational considerations, a maximum of 10 000 proteins, representative of the whole genomes, were randomly selected and their sequence features were individually determined. We chose to compare hub proteins to these reference datasets primarily because of the above-mentioned uncertainty in hub definition. Due to this, the dataset complementary to hubs at any chosen cutoff value will also contain proteins that behave as hubs themselves, which will compromise the statistical difference between hubs and nonhubs. Comparing two distinct reference datasets, RAN, which statistically represents the whole interactome and thus contains hubs, and IC\_1, which by definition does not, are expected to provide the information necessary to resolve this dilemma.

**Statistical Approaches.** For each sequence, the total number or proportion of residues with a given feature was determined. Beside the average of these values, their distributions were also analyzed, to obtain a more detailed picture. The range of the values was divided into five bins so that roughly equal portions of RAN sequences fell into each bin. Since occasionally more than one-fifth of proteins lack disorder/low complexity/repeats, the first bin can cover more than 20% of proteins, whereby the remaining proteins could then be divided into four equal bins. Nevertheless, these unequal bins with similar number of data points give statistically more reliable results than bins of equal width, where the first bin would contain most data points and the last bin only a few.

The distribution of properties was compared between HUB, IC\_1, and RAN proteins (for definition and explanation, see Datasets). The statistical significance of the differences was also assessed. The basic assumption was that deviations arise because of limited sample size. This was estimated by scooping

into the reference dataset, selecting as many random proteins as HUB contained, and repeating this procedure 5000 times. These random samples were then used to calculate the standard deviation of the percentage of proteins falling into the given bin. The deviation was also approximated with the square root of the actual data points in the given bin. This estimation gave very similar results (not shown).

Given the distribution of the percentage of hub proteins and the reference proteins (IC\_1 or RAN) in each bin and their deviations, the likelihood that they derive from the same distribution can also be estimated using  $\chi^2$  statistics:  $\chi^2 = \sum (a_i - b_i)^2 / (2 \times \text{dev}_i^2)$ . For each  $i$  bin,  $a_i$  and  $b_i$  is the percentage of proteins falling in the given bin for HUB proteins and for proteins in the reference set, respectively, and  $\text{dev}_i$  is the standard deviation of the percentage of proteins in the bin, calculated from the 5000 random samples taken from the reference set, and used for both sets. By virtue of this value, we can test the hypothesis that the two distributions are not the same. If this value exceeds 13.3 (using 4 as the degrees of freedom), the two distributions are different at a confidence level of 99%.

## Results

**Connectivity of the Interactomes Studied.** The scale-free topology of interaction networks is primarily manifested in a power-law distribution of the degree of connectivity. Different interactomes, however, have been determined by different experimental techniques and vary in coverage, which may affect their global topological features.<sup>5</sup> For a comparison of the actual databases used herein (Table 1), Figure 1 shows the percentage of proteins versus their numbers of interactions on a log–log scale. The YEAST and WORM data follow most closely the linear relationship expected, but the other datasets deviate significantly from a straight line, as already noted.<sup>2,5</sup> Despite the deviation from strict linearity, however, the data are amenable to our proposed analyses, with a relatively small fraction of proteins having large numbers of interactions.

**Sequence Features in Hubs and Nonhubs, and the Effect of Cutoff Choice.** Four sequence features (sequence size and the number of residues with either of the three sequence attributes: structural disorder, sequence repeats, and low complexity) have been compared between hubs and nonhubs by calculating the differences between the averages for hubs proteins (HUB) and proteins with exactly one interaction (IC\_1) and a random sample of genome sequences (RAN), for all four species (Table 2). The mean and standard deviation values show that hubs are significantly longer, and contain more of, and a greater proportion of, structural disorder, sequence repeats, and low complexity than nonhubs. Hubs in this experiment have been defined by a fixed cutoff of five or more interaction partners. These comparisons show that the length of the protein and disorder are the strongest discriminators of hubs from reference proteins, with repeats and low complexity displaying smaller, but still significant, differences. Worth noting is the unusual behavior of the WORM database, where IC\_1 proteins also seem to be biased. A further point is that these data allow an insight into the evolution of hub function. Comparing the four species, hubs appear to have gained mostly in length and disorder, with a smaller increase in the other two features. Thus, these four features not only characterize hubs today, they also have contributed to the evolution of hub function and the increase in complexity of protein interactomes.

**Table 2.** Number of Residues and Proportion with Sequence Features of Hubs in Four Interactomes with Fixed Cutoff<sup>a</sup>

datasets	property	HUB average	IC_1		RAN	
			average	SD	average	SD
Average Number of Residues						
YEAST	disorder	143.69	87.38	5.09	92.11	5.45
YEAST	lc	45.20	32.86	1.79	34.35	1.93
YEAST	repeat	23.44	14.52	2.43	17.84	2.94
YEAST	length	532.61	430.23	12.01	469.72	13.61
WORM	disorder	140.34	144.09	12.82	77.94	9.54
WORM	lc	49.28	47.66	3.90	33.62	3.43
WORM	repeat	49.20	61.20	14.14	36.65	10.39
WORM	length	494.90	506.24	26.00	436.73	22.87
FLY	disorder	174.55	144.07	5.37	163.40	6.30
FLY	lc	68.90	57.95	2.20	65.65	2.55
FLY	repeat	43.22	38.43	3.35	42.75	3.58
FLY	length	482.66	507.85	10.06	541.08	11.45
HUMAN	disorder	203.34	168.29	5.91	119.14	4.22
HUMAN	lc	70.22	59.76	1.70	43.76	1.42
HUMAN	repeat	94.51	72.45	5.60	56.48	3.73
HUMAN	length	698.33	621.82	11.40	467.36	8.99
Proportion of Residues						
YEAST	disorder	0.2460	0.1532	0.0760	0.1730	0.0079
YEAST	lc	0.0818	0.0713	0.0034	0.0767	0.0036
YEAST	repeat	0.0331	0.0287	0.0041	0.0302	0.0040
WORM	disorder	0.2493	0.2124	0.0149	0.1636	0.0136
WORM	lc	0.1006	0.0906	0.0077	0.0808	0.0069
WORM	repeat	0.0873	0.0660	0.0099	0.0588	0.0095
FLY	disorder	0.3148	0.2280	0.0059	0.2594	0.0063
FLY	lc	0.1381	0.1010	0.0030	0.1140	0.0031
FLY	repeat	0.0875	0.0568	0.0034	0.0633	0.0034
HUMAN	disorder	0.2722	0.2292	0.0050	0.2389	0.0054
HUMAN	lc	0.1027	0.0999	0.0023	0.0955	0.0024
HUMAN	repeat	0.0967	0.0785	0.0039	0.0864	0.0042

<sup>a</sup> For each species (*S. cerevisiae*, *D. melanogaster*, *C. elegans*, and *H. sapiens*) and each sequence feature (disorder, low complexity (lc), repeats, and total length in the case of absolute numbers) the distribution of the number of residues or the proportion of the given feature in the sequences of hub proteins is compared to that of both the genome sequences (RAN) and the IC\_1 (proteins with exactly one interaction). HUB proteins were defined as proteins with at least five interactions. The averages were calculated for HUB proteins, for the RAN and IC\_1 datasets. The SD refers to the standard deviation calculated from the average properties over random samples of proteins, matching the number of HUB proteins but selected from RAN and IC\_1 datasets, respectively (see Materials and Methods).

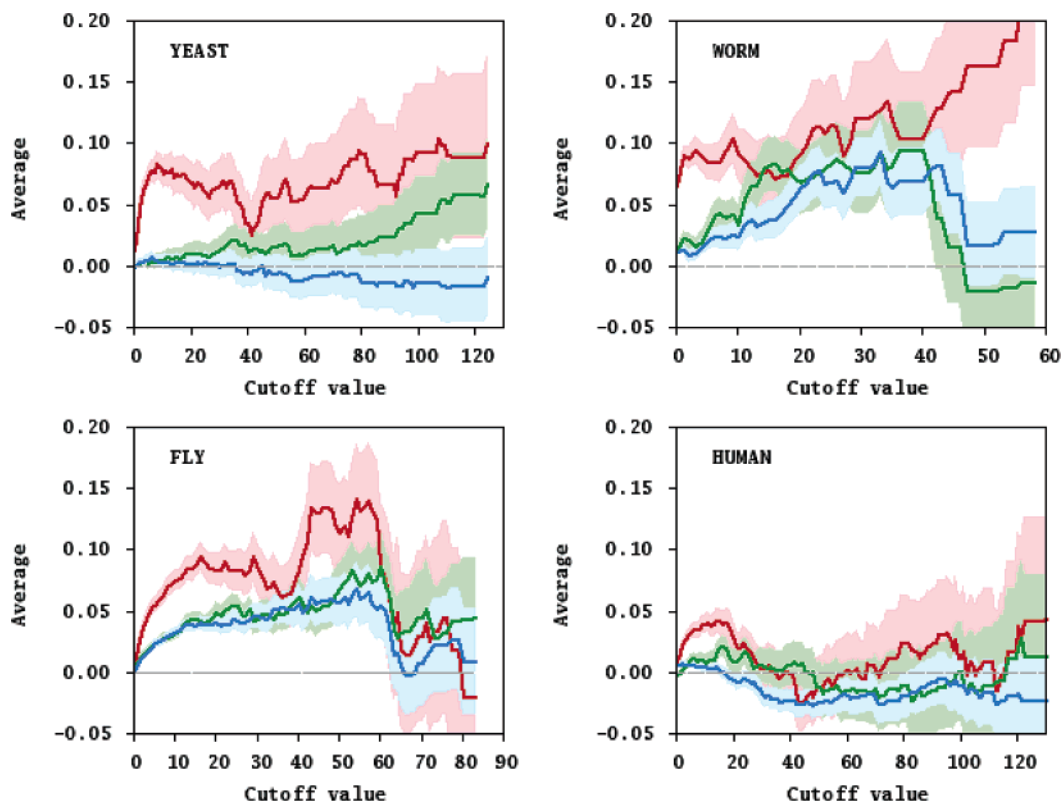
The literature offers no clue regarding the specification of the boundary between hubs and nonhubs; indeed, the concept of being a hub is qualitative rather than quantitative. Since the choice of the cutoff could affect the above results, an experiment was carried out to examine this uncertainty. Figure 2 shows the results of how the mean of the difference between hubs and random sample of proteins changes with the cutoff value. The mean for hub proteins is higher than the average for RAN and tends to increase with increasing cutoff values up to very large values with the exception of HUMAN, which shows a diminution of the difference with high values. The reason for this deviation is not apparent but may be related to the distinction that most of the data in the human interactome<sup>4</sup> have come from curated individual observations, whereas the other three datasets have been generated from high-throughput studies. Of course, as the boundary value increases, the number of hub proteins decreases, which increases the standard deviation and impairs the estimation of the significance of the difference. Nevertheless, these data support the overall conclusion that hubs have discernible structural characteristics that do not depend on hub definition, i.e., on the choice of the cutoff value.

**Characterizing Hubs by Applying a Floating Cutoff Definition.** Here we further address the point that hub function is a system property and cannot be exactly defined at the level of the individual proteins. Furthermore, various interactomes have been determined by different techniques, which are of different sensitivity and may not provide the same information even on

the very same protein. We therefore also adopted an alternative definition using a floating cutoff definition in which hub proteins were considered as the top 10% of proteins sorted according to the number of interaction partners. This defines hubs in terms of their relation to the entire interactome and thereby increases the statistical power of our analysis.

For the four species, the occurrence of the four attributes in hubs has been compared to that in IC\_1 and RAN, by dividing the range of values obtained into five bins so that roughly an equal portion of RAN sequences fell into each bin (for details, see Statistical Approaches). The results are shown in Figure 3. In practically all the cases, hub proteins are underrepresented in the first or first two bins and are overrepresented in the last or last two bins. In other words, hubs tend to have more disorder, more sequence repeats, or more low complexity regions than nonhubs. The trend is similar for the sizes of proteins: the polypeptide chains of hubs are significantly longer than those of nonhubs, with the possible exception of *Drosophila*, in which case there is an excess of hubs with medium lengths (centered at around 400 and 500 amino acids).

A concise description of all these comparisons is given in Table 3. Here, the difference between the distributions of sequence features and the total length of hubs and both random genome samples and proteins with exactly one interaction is determined. The difference is characterized by a  $\chi^2$  value and the corresponding probability that the two sets of data are of different distributions. Hubs and the reference datasets significantly differ in practically all features for all the



**Figure 2.** Effect of hub definition on the difference between hubs and random proteins. Three sequence features, disorder (red), repeat coverage (green), and low complexity (blue) are compared for YEAST, WORM, FLY, and HUMAN hubs (HUB) and a random selection of proteins (RAN). The difference between the average values for HUB and RAN is shown as a function of the number of interactions, above which a protein is considered a hub (cutoff). The light-colored stripe around the mean corresponds to the standard deviation.

species, with the possible exception of the low complexity in WORM and length in YEAST\_CORE. It is also to be noted that differences are more significant in some cases with RAN than IC\_1, which suggests IC\_1 as a class is not representative of the whole genome but is biased due to the methods used for high-throughput screening of protein–protein interactomes. In general, there are some minor differences between the datasets. Disorder is a strong distinguishing feature in the various interactomes, with length and repeats being also convincing in most cases. In several instances low complexity appears to be the least obvious but is still significant. Apart from these minor differences, however, it is safe to conclude that all three features are important, and thus conserved, in defining hub behavior.

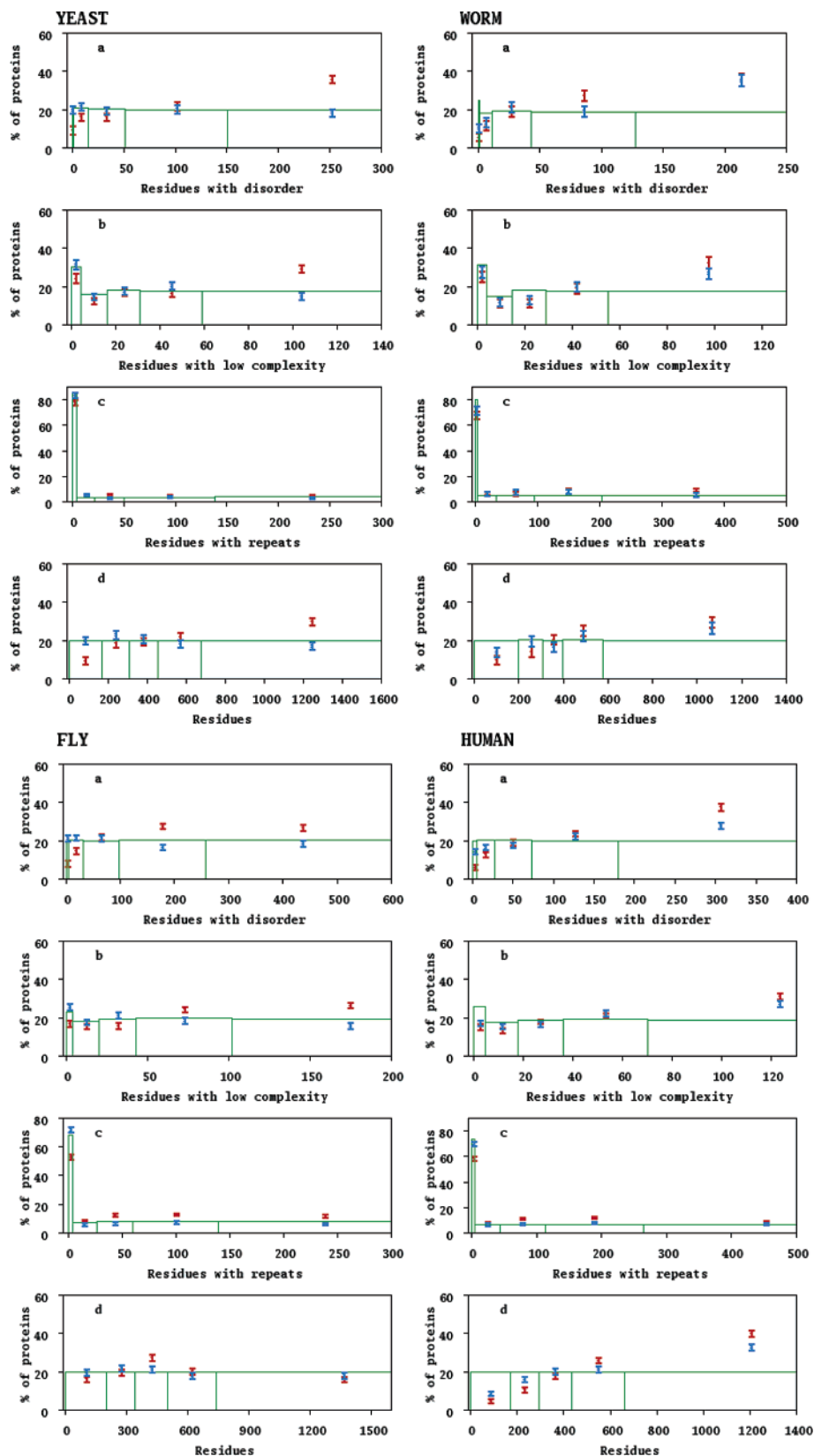
**Proportion of Various Sequence Features in Hubs.** In addition to the overall length of regions with a particular feature, it is also worth investigating how the proportion of regions with a given feature differs between hubs and nonhubs. Regions with a given feature were identified and normalized by the size of the protein for the four species, as shown in Figure 4. Again, the difference in favor of hubs is significant in most cases, with the exception of low complexity in YEAST\_CORE, as quantitatively rendered in Table 4. Compared to the total length of features, there are some differences in the order of the importance of features, which are probably of secondary importance. Overall, these data strongly suggest that not only lengthy regions of disorder/repeats/low complexity but also a high proportion of these features is likely to be important for conferring advantages in terms of hub behavior.

**Interdependence of Sequence Features.** Because the three features studied are not independent of each other, it is

important to determine the extent of their correlation. Although intrinsically unstructured or disordered proteins are often composed of repeats<sup>32</sup> and low sequence complexity correlates with the lack of a well-defined structure,<sup>36</sup> the three features are not perfectly correlated characteristics. Indeed their combinations are stronger indicators of hubs than any single one. To characterize their interdependence, the difference of the averages of individual features and their combinations between HUB and RAN proteins has been determined (Figure 5). Among the three properties, disorder exhibited the largest increase in all species, whereas their combination increased the difference even further. In contrast, low complexity in any combination led to only a minor increase and showed a relatively high correlation with both disorder and repeats (data not shown).

**All Interactions versus Confident Interactions.** A comparison of interactomes obtained in different studies has shown that individual studies may have provided a low coverage of the total interactome and contain a significant fraction of false positive interactions.<sup>4,43</sup> This suggests that the actual interaction databases may not be representative, which may cause artificial results in our studies. To minimize this possibility, we characterized specific subsets covering reliable interactions only. These are available for the *Drosophila* (FLY\_CONF) and yeast (YEAST\_CORE) interactome (for definitions, see Datasets).

Restricting our analysis to the subset of confident interactions did not alter the differences in any significant way (Tables 3 and 4), thus corroborating the prior major conclusions. The differences of the averages are significant in most of the cases by the measures  $\chi^2$  and probability of difference in the distributions of hubs and reference datasets.



**Figure 3.** Total length of sequence features of hubs in interactomes. Comparison of the total number of residues with disorder (a), low complexity (b), repeats (c), and the length of protein (d) for the yeast (YEAST), worm (WORM), *Drosophila* (FLY), and human (HUMAN) interactomes. Green columns represent the borders of bins that contain about the same number of proteins calculated for the random sample of genome sequences (RAN). Hub proteins are shown in red, and the random sample of proteins with exactly one interaction (IC\_1) are shown in blue. The height of the columns and the position of symbols represent the percent of proteins in the given database that fall into the given bin, i.e., range of feature. The horizontal position of symbols is arbitrary, because it represents all data within the given bin. The error bars correspond to the standard deviation calculated from IC\_1.

**Table 3.** Number of Residues with Sequence Features of Hubs in Four Interactomes with Floating Cutoff<sup>a</sup>

datasets	property	compared to RAN		compared to IC_1	
		$\chi^2$	probability	$\chi^2$	probability
YEAST	disorder	49.14	>0.999 99	58.81	>0.999 99
YEAST	lc	23.94	0.999 91	38.87	>0.999 99
YEAST	repeat	16.05	0.997 04	13.27	0.989 99
YEAST	length	27.53	0.999 98	42.27	>0.999 99
YEAST_CORE	disorder	24.89	0.999 94	8.79	0.933 46
YEAST_CORE	lc	13.43	0.990 64	4.74	0.685 45
YEAST_CORE	repeat	58.93	>0.999 99	36.29	>0.999 99
YEAST_CORE	length	21.65	0.999 76	3.36	0.500 88
WORM	disorder	53.13	>0.999 99	8.24	0.916 94
WORM	lc	25.53	0.999 96	2.71	0.392 13
WORM	repeat	20.43	0.999 58	4.31	0.633 74
WORM	length	17.55	0.998 48	5.45	0.755 59
FLY	disorder	53.67	>0.999 99	84.06	>0.999 99
FLY	lc	27	0.999 98	54.14	>0.999 99
FLY	repeat	63.67	>0.999 99	106.29	>0.999 99
FLY	length	17.84	0.998 67	12.28	0.984 60
FLY_CONF	disorder	29.62	0.999 99	28.75	0.999 99
FLY_CONF	lc	3.49	0.520 86	4.57	0.665 33
FLY_CONF	repeat	15.72	0.996 58	23.30	0.999 89
FLY_CONF	length	32.9	0.999 99	24.83	0.999 95
HUMAN	disorder	125.77	>0.999 99	40.93	>0.999 99
HUMAN	lc	64.37	>0.999 99	5.66	0.774 17
HUMAN	repeat	79.14	>0.999 99	41.68	>0.999 99
HUMAN	length	170.41	>0.999 99	28.93	0.999 99

<sup>a</sup> For each species (*S. cerevisiae*, *D. melanogaster*, *C. elegans*, and *H. sapiens*) and each sequence feature (disorder, low complexity (lc), repeats, and total length) the distribution of the number of residues with the given feature in the sequences of hub proteins is compared to that in both the genome sequences (RAN) and the IC\_1 (proteins with exactly one interaction). The difference is characterized by  $\chi^2$  values (see Materials and Methods) and the corresponding probability that the two sets of data are of different distributions.

## Discussion

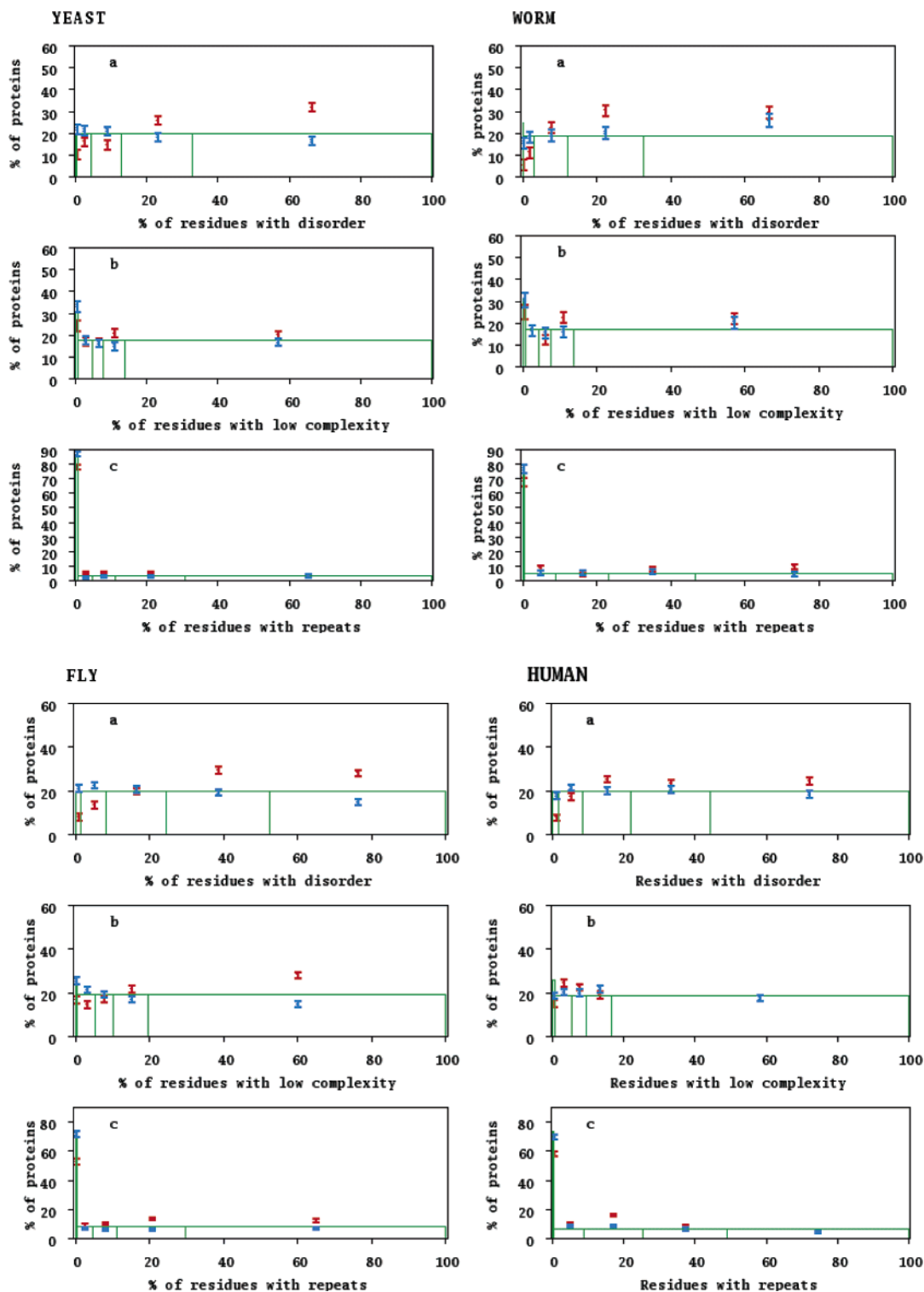
The interaction networks of different species show remarkable similarity in terms of the global feature of near scale-free topology. In the interactomes examined, we found significant deviations from a strict scale-free behavior (cf. Figure 1), with an increasing deficiency of hubs toward higher connectedness. In principle, this may be attributed to the suppression of hubs,<sup>2</sup> to an approximation of scale-free behavior due to limited sampling,<sup>5</sup> or to some other factors, such as the limited size of the network. Notwithstanding these reservations, the fundamental features of these networks derive primarily from the presence of highly connected hubs. Thus, the molecular basis of the function of hub proteins is key to understanding how interaction networks provide the bases for cell function. In this paper, we present evidence that hub proteins are significantly larger, have more predicted disorder, and contain more sequence repeats and/or low-complexity regions than nonhubs. Our studies into the interdependence of the features have shown that disordered segments and repeat regions are relatively independent, and their presence in hubs provides synergistic structural rationale for hub behavior. Low complexity, on the other hand, is much worse in distinguishing hubs from nonhubs and may actually compromise discrimination between these functional classes (cf. Figure 5). Interpretation of this sequence feature in terms of hub function, thus, may lead to misleading conclusions.

As noted in the Introduction, structural disorder confers many functional advantages, several of which provide the rationale for its prevalence in hubs,<sup>16,18,19</sup> as also suggested in previous works.<sup>24,31</sup> The observation that either the total length

of disordered segments or the proportion of disorder are equally good attributes of hub function may actually imply two alternative structural strategies. For example, the open structure of intrinsically unstructured or disordered proteins provides a large interaction surface, which enhances the capacity of the protein for interactions.<sup>23</sup> This is of clear benefit to proteins, which interact simultaneously with many partners such as the so-called party hubs.<sup>31,44</sup> The presence of long repeat regions may be rationalized on a similar ground.<sup>35</sup> For several actual examples of hubs, such as for caldesmon, BRCA1, and estrogen receptor  $\alpha$ , for example,<sup>31</sup> the advantage of a large amount of disorder has been demonstrated. Hub function may also benefit because disorder significantly increases the association rate of protein interactions,<sup>17,25</sup> as formulated in the fly casting<sup>26</sup> and protein fishing<sup>27</sup> models. In addition, multicomponent complexes do not assemble very well from rigid components due to steric clashes. In contrast, the use of coupled folding and binding by flexible subunits facilitates the formation of such multicomponent complexes by avoiding the steric problems encountered by rigid subunits.<sup>45</sup>

A somewhat different structural logic may apply to proteins that do not necessarily have long disordered regions but that have a high proportion of disorder. These hubs may rely on the malleability of their structures, which enable them to adapt to distinct partners. Such adaptability has been described for cyclin-dependent protein kinase inhibitors,<sup>28</sup> for glycogen synthase kinase  $3\beta$ ,<sup>46</sup> for  $\alpha$ -synuclein,<sup>47</sup> and for the hypoxia inducible factor  $1\alpha$ ,<sup>18</sup> and has been generalized as moonlighting<sup>29</sup> or polymorphism in the bound state.<sup>24</sup> A variation on this theme might be represented by ordered hubs, such as calmodulin,<sup>31</sup> or 14-3-3 proteins,<sup>48</sup> for which the partners utilize disorder in an adaptative process. While such a use of disorder by hub partners definitely occurs in some cases, we found no statistical difference between hub partners and other proteins (data not shown). To further investigate whether disorder may be important for some hub partners, it will be necessary to separately evaluate the partners of a large collection of highly ordered hubs. Overall, the excess of the features studied in hubs can be rationalized in terms of the functional specialization of these proteins. It would be interesting to test additionally whether the features are also correlated with the connectedness of hubs. Due to the relatively low number of hubs and the extremely wide range of the number of interaction partners (Table 1), the significance of this possible correlation could not be established given the current limited dataset (data not shown).

In addition, the excess of disorder and repeat regions in hubs also has general evolutionary implications. Hubs play a central role in defining the scale-free topology of interaction networks, but the prevailing model for the emergence of such a contextual arrangement fails to capture the basic capacity of proteins to undergo evolutionary changes. The most influential model of network evolution assumes that random growth and preferential attachment to already highly connected nodes explains the emergence of scale-free behavior.<sup>6,7</sup> The underlying evolutionary mechanism has been assumed to be gene duplication, which, due to mere chance, prefers nodes already connected to nodes with multiple links.<sup>9,10</sup> Although scale-free topology in principle may confer several selective advantages, such as error tolerance,<sup>8</sup> avoidance of jamming,<sup>49</sup> and hierarchical modularity,<sup>44</sup> upon which selective pressure may act, its suggested development oversimplifies the situation in which deletion of gene products, rewiring of physical contacts, and



**Figure 4.** Proportion of disorder, sequence repeats, and low complexity in hubs in interactomes. Comparison of the relative abundance of disorder (a), repeats (b), and low complexity (c) for the yeast (YEAST), worm (WORM), *Drosophila* (FLY), and human (HUMAN) interactomes. Hub proteins (red), a random sample of proteins with exactly one interaction (blue), and a random sample of genome sequences (green) are shown as in Figure 3.

critical differences between individual proteins all need to be taken into account.<sup>12</sup> In fact, it has been formally shown that this topology also arises simply if hub proteins attract novel partners due to their physicochemical nature that predisposes them for interactions.<sup>11</sup> It is of relevance here that IUPred, the algorithm used for assessing disorder,<sup>39,40</sup> relies on estimating the energy content that a given protein segment can realize.

This incorporates energy terms for both intramolecular and protein–solvent interactions. The importance of the latter in both network evolution and intrinsic disorder has been discussed recently.<sup>50</sup> A recent analysis of genetic regulatory networks has in fact shown that, for the evolution of a network with the observed global and local features, elements of both node copying (gene duplication) and link mutation (change in



**Table 4.** Proportion of Disorder, Sequence Repeats, and Low Complexity in Four Interactomes<sup>a</sup>

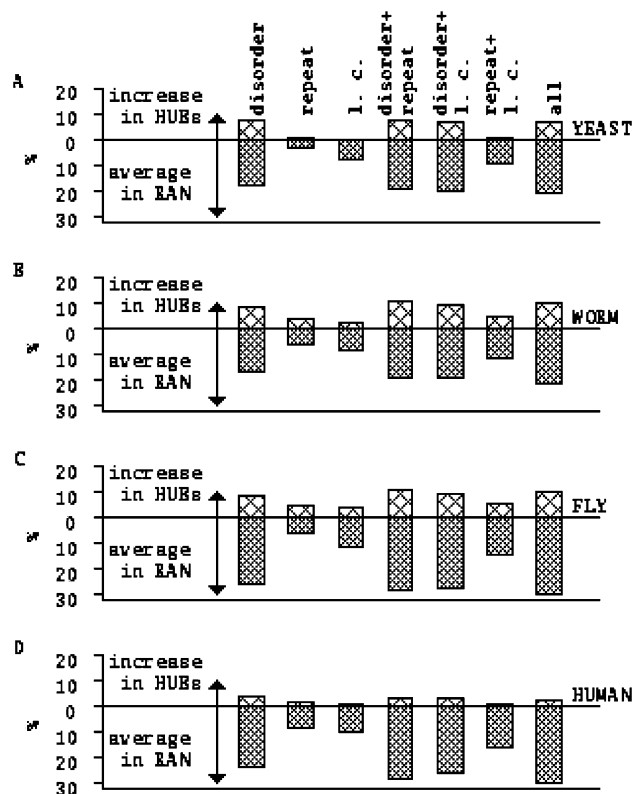
datasets	property	compared to RAN		compared to IC_1	
		$\chi^2$	probability	$\chi^2$	probability
YEAST	disorder	38.92	>0.999 99	67.88	>0.999 99
YEAST	lc	6.23	0.817 45	14.50	0.994 14
YEAST	repeat	17.05	0.998 11	32.66	>0.999 99
YEAST_CORE	disorder	31.28	>0.999 99	19.17	0.999 27
YEAST_CORE	lc	4.68	0.678 18	3.20	0.475 50
YEAST_CORE	repeat	59.17	>0.999 99	68.92	>0.999 99
WORM	disorder	48.10	>0.999 99	20.68	0.999 63
WORM	lc	8.29	0.918 54	6.09	0.807 22
WORM	repeat	23.25	0.999 89	13.61	0.991 35
FLY	disorder	69.66	>0.999 99	115.33	>0.999 99
FLY	lc	31.99	>0.999 99	72.41	>0.999 99
FLY	repeat	65.08	>0.999 99	112.49	>0.999 99
FLY_CONF	disorder	28.84	0.999 99	30.15	>0.999 99
FLY_CONF	lc	8.47	0.924 29	12.03	0.982 86
FLY_CONF	repeat	17.70	0.998 59	22.77	0.999 86
HUMAN	disorder	47.55	>0.999 99	43.79	>0.999 99
HUMAN	lc	34.03	>0.999 99	10.01	0.959 82
HUMAN	repeat	111.64	>0.999 99	52.17	>0.999 99

<sup>a</sup> For each species (*S. cerevisiae*, *D. melanogaster*, *C. elegans*, and *H. sapiens*) and each sequence feature (disorder, low complexity (lc), repeats) the distribution of the number of the proportion of the given feature for hub proteins is compared to that in both the genome sequences (RAN) and the IC\_1 (proteins with exactly one interaction). The difference is characterized by  $\chi^2$  values (see Materials and Methods) and the corresponding probability that the two sets of data are of different distributions.

interaction) events have to be invoked.<sup>51</sup> Although protein–protein interaction networks studied in our work and genetic regulatory networks differ in some basic aspects, the evolutionary complexity for one regulatory network supports our point that a more elaborate evolutionary model of biological networks is likely needed in general for biological networks.

By examining hubs of the four species, it is clear their length and disorder, and to a lesser degree their repeats and low complexity regions, tend to increase as the complexity of the organism and the underlying interactomes increases in complexity on the evolutionary tree. These observations support the idea that the evolution of protein interaction networks has involved an element of selection of certain proteins toward functioning in network organization, i.e., by becoming hubs. In this process, the generation and extension of internal repeat regions and the increase in disorder is proposed to have played an active role. This scenario fully conforms to the logic of specialization that enables biological entities to occupy more niches. Additionally, direct functional advantages could also derive from this specialization process. Disordered proteins are frequently involved in regulated interaction processes, due to their disposition for posttranslational modification.<sup>16,52</sup> This is of significant functional benefit, as interaction networks are very dynamic objects, prone to undergo profound reorganization events mostly conducted by “transient”<sup>53</sup> or “date”<sup>44</sup> hubs. An interesting possibility is that date hubs may also draw a functional advantage not from the disorder and ensuing adaptability of their own but that of their partners. In principle, this might alleviate the demand of structural adaptability of the hub and provide a simpler solution for the inclusion of the hub in distinct and functionally/structurally unrelated complexes. This possibility has been discussed previously.<sup>31</sup>

Another significant feature might derive from the observation that, since intrinsically disordered proteins are typically unfolded, they undergo little change upon treatment with heat or chemical denaturants. This resistance may provide protec-



**Figure 5.** Averages and correlations of the sequence features. The average percentage of the sequence properties and their various combinations in HUB proteins for the four species calculated. The region below the zero line corresponds to the average in the random genome subsets (RAN), whereas the region above the zero line shows the increase in HUB proteins for the three primary sequence properties (disorder, repeats, and low complexity, lc) and their various combinations. These latter ones were defined as the maximum of the two or three properties for each protein, averaged over the dataset.

tion against elimination of hub function, to which scale-free networks are very sensitive.<sup>8</sup> A further pertinent point is that intrinsically unstructured or disordered regions often bind with their partner(s) by virtue of an extended surface with interaction sites dispersed over the surface of the ordered protein partner.<sup>19</sup> A change in any of these sites might not entirely eliminate the interaction and may thus provide resistance against point mutations. This may be a good explanation why evolutionary variability shows very weak correlation with the number of interaction partners,<sup>54</sup> whereas removal of a hub is three times more probable to be lethal than other proteins.<sup>55</sup>

In summary, hub proteins are found to be enlarged and also to be enriched in predicted disorder, in sequence repeats, and in low complexity regions and in combinations involving two or more of these features. All of these characteristics and their combinations facilitate binding to multiple partners. The enrichment of these features over evolutionary time is probably necessary to explain these observations, suggesting a more complicated evolutionary history than the commonly accepted mechanisms based on simple, random gene duplication. Experimental studies to further test these proposed roles of intrinsic disorder in protein–protein interaction networks would be useful.

**Acknowledgment.** This work was supported by grants from the Economic Competitiveness Operative Program (GVOP),

Grant 3.1.1.-2004-05-0195/3.0; the Hungarian Scientific Research Fund (OTKA), Grants K60694, F043609, and T049073; the National R&D Program (NKFP), Grant MediChemBats21/A/005/2004; the NIH, Grant R01 LM007688; and the Indiana Genomics Initiative, funded in part by the Lilly Endowment. We also acknowledge the Bolyai János fellowships for Z.D. and P.T. and the Wellcome Trust International Senior Research Fellowship ISRF 067595 for P.T. Finally, A.K.D. would like to thank Zoran Obradovic, Vladimir Uversky, and Pedro Romero for their continuing collaborations that include discussions of importance to the current work.

## References

- Gavin, A. C.; Bosche, M.; Krause, R.; Grandi, P.; Marzioch, M.; Bauer, A.; Schultz, J.; Rick, J. M.; Michon, A. M.; Cruciat, C. M.; Remor, M.; Hofert, C.; Schelder, M.; Brajenovic, M.; Ruffner, H.; Merino, A.; Klein, K.; Hudak, M.; Dickson, D.; Rudi, T.; Gnau, V.; Bauch, A.; Bastuck, S.; Huhse, B.; Leutwein, C.; Heurtier, M. A.; Copley, R. R.; Edlmann, A.; Querfurth, E.; Rybin, V.; Drewes, G.; Raida, M.; Bouwmeester, T.; Bork, P.; Seraphin, B.; Kuster, B.; Neubauer, G.; Superti-Furga, G. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **2002**, *415*, 141–147.
- Giot, L.; Bader, J. S.; Brouwer, C.; Chaudhuri, A.; Kuang, B.; Li, Y.; Hao, Y. L.; Ooi, C. E.; Godwin, B.; Vitols, E.; Vijayadamar, G.; Pochart, P.; Machineni, H.; Welsh, M.; Kong, Y.; Zerhusen, B.; Malcolm, R.; Varrone, Z.; Collis, A.; Minto, M.; Burgess, S.; McDaniel, L.; Stimpson, E.; Spriggs, F.; Williams, J.; Neurath, K.; Ioime, N.; Agee, M.; Voss, E.; Furtak, K.; Ranzulli, R.; Aanensen, N.; Carrola, S.; Bickelhaupt, E.; Lazovatsky, Y.; DaSilva, A.; Zhong, J.; Stanyon, C. A.; Finley, R. L., Jr.; White, K. P.; Braverman, M.; Jarvie, T.; Gold, S.; Leach, M.; Knight, J.; Shimkets, R. A.; McKenna, M. P.; Chant, J.; Rothberg, J. M. A protein interaction map of *Drosophila melanogaster*. *Science* **2003**, *302*, 1727–1736.
- Li, S.; Armstrong, C. M.; Bertin, N.; Ge, H.; Milstein, S.; Boxem, M.; Vidalain, P. O.; Han, J. D.; Chesneau, A.; Hao, V.; Goldberg, D. S.; Li, N.; Martinez, M.; Rual, J. F.; Lamesch, P.; Xu, L.; Tewari, M.; Wong, S. L.; Zhang, L. V.; Berriz, G. F.; Jacotot, L.; Vaglio, P.; Reboul, J.; Hirozane-Kishikawa, T.; Li, Q.; Gabel, H. W.; Elewa, A.; Baumgartner, B.; Rose, D. J.; Yu, H.; Bosak, S.; Sequerra, R.; Fraser, A.; Mango, S. E.; Saxton, W. M.; Strome, S.; Van Den Heuvel, S.; Piano, F.; Vandenhaute, J.; Sardet, C.; Gerstein, M.; Doucette-Stamm, L.; Gunsalus, K. C.; Harper, J. W.; Cusick, M. E.; Roth, F. P.; Hill, D. E.; Vidal, M. A map of the interactome network of the metazoan *C. elegans*. *Science* **2004**, *303*, 540–543.
- Gandhi, T. K.; Zhong, J.; Mathivanan, S.; Karthick, L.; Chandrika, K. N.; Mohan, S. S.; Sharma, S.; Pinkert, S.; Nagaraju, S.; Periaswamy, B.; Mishra, G.; Nandakumar, K.; Shen, B.; Deshpande, N.; Nayak, R.; Sarker, M.; Boeke, J. D.; Parmigiani, G.; Schultz, J.; Bader, J. S.; Pandey, A. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat. Genet.* **2006**, *38*, 285–293.
- Han, J. D.; Dupuy, D.; Bertin, N.; Cusick, M. E.; Vidal, M. Effect of sampling on topology predictions of protein–protein interaction networks. *Nat. Biotechnol.* **2005**, *23*, 839–844.
- Barabasi, A. L.; Oltvai, Z. N. Network biology: Understanding the cell's functional organization. *Nat. Rev. Genet.* **2004**, *5*, 101–113.
- Babu, M. M.; Luscombe, N. M.; Aravind, L.; Gerstein, M.; Teichmann, S. A. Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol.* **2004**, *14*, 283–291.
- Albert, R.; Jeong, H.; Barabasi, A. L. Error and attack tolerance of complex networks. *Nature* **2000**, *406*, 378–382.
- Eisenberg, E.; Levanon, E. Y. Preferential attachment in the protein network evolution. *Phys. Rev. Lett.* **2003**, *91*, 138701.1–138701.4.
- Pastor-Satorras, R.; Smith, E.; Sole, R. V. Evolving protein interaction networks through gene duplication. *J. Theor. Biol.* **2003**, *222*, 199–210.
- Caldarelli, G.; Capocci, A.; De Los Rios, P.; Munoz, M. A. Scale-free networks from varying vertex intrinsic fitness. *Phys. Rev. Lett.* **2002**, *89*, 258702.1–258702.4.
- Aloy, P.; Russell, R. B. Taking the mystery out of biological networks. *EMBO Rep.* **2004**, *5*, 349–350.
- Wright, P. E.; Dyson, H. J. Intrinsically unstructured proteins: Re-assessing the protein structure–function paradigm. *J. Mol. Biol.* **1999**, *293*, 321–331.
- Uversky, V. N.; Gillespie, J. R.; Fink, A. L. Why are “natively unfolded” proteins unstructured under physiologic conditions? *Proteins* **2000**, *41*, 415–427.
- Dunker, A. K.; Lawson, J. D.; Brown, C. J.; Romero, P.; Oh, J. S.; Oldfield, C. J.; Campen, A. M.; Ratliff, C. M.; Hipps, K. W.; Ausio, J.; Nissen, M. S.; Reeves, R.; Kang, C.; Kissinger, C. R.; Bailey, R. W.; Griswold, M. D.; Chiu, W.; Garner, E. C.; Obradovic, Z. Intrinsically disordered protein. *J. Mol. Graphics Modell.* **2001**, *19*, 26–59.
- Dunker, A. K.; Brown, C. J.; Lawson, J. D.; Iakoucheva, L. M.; Obradovic, Z. Intrinsic disorder and protein function. *Biochemistry* **2002**, *41*, 6573–6582.
- Tompa, P. Intrinsically unstructured proteins. *Trends Biochem. Sci.* **2002**, *27*, 527–533.
- Dyson, H. J.; Wright, P. E. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* **2005**, *6*, 197–208.
- Tompa, P. The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett.* **2005**, *579*, 3346–3354.
- Iakoucheva, L.; Brown, C.; Lawson, J.; Obradovic, Z.; Dunker, A. Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.* **2002**, *323*, 573–584.
- Ward, J. J.; Sodhi, J. S.; McGuffin, L. J.; Buxton, B. F.; Jones, D. T. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* **2004**, *337*, 635–645.
- Fuxreiter, M.; Simon, I.; Friedrich, P.; Tompa, P. Prefolded structural elements feature in partner recognition by intrinsically unstructured proteins. *J. Mol. Biol.* **2004**, *338*, 1015–1026.
- Gunasekaran, K.; Tsai, C. J.; Kumar, S.; Zanuy, D.; Nussinov, R. Extended disordered proteins: Targeting function with less scaffold. *Trends Biochem. Sci.* **2003**, *28*, 81–85.
- Uversky, V. N.; Oldfield, C. J.; Dunker, A. K. Showing your ID: Intrinsic disorder as an ID for recognition, regulation and cell signaling. *J. Mol. Recognit.* **2005**, *18*, 343–384.
- Pontius, B. W. Close encounters: Why unstructured, polymeric domains can increase rates of specific macromolecular association. *Trends Biochem. Sci.* **1993**, *18*, 181–186.
- Shoemaker, B. A.; Portman, J. J.; Wolynes, P. G. Speeding molecular recognition by using the folding funnel: The fly casting mechanism. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 8868–8873.
- Evans, P. R.; Owen, D. J. Endocytosis and vesicle trafficking. *Curr. Opin. Struct. Biol.* **2002**, *12*, 814–821.
- Kriwacki, R. W.; Hengst, L.; Tennant, L.; Reed, S. I.; Wright, P. E. Structural studies of p21Waf1/Cip1/Sd1 in the free and Cdk2-bound state: Conformational disorder mediates binding diversity. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 11504–11509.
- Tompa, P.; Szasz, C.; Buday, L. Structural disorder throws new light on moonlighting. *Trends Biochem. Sci.* **2005**, *30*, 484–489.
- Dunker, A. K.; Obradovic, Z. The protein trinity-linking function and disorder. *Nat. Biotechnol.* **2001**, *19*, 805–806.
- Dunker, A. K.; Cortese, M. S.; Romero, P.; Iakoucheva, L. M.; Uversky, V. N. Flexible nets: The roles of intrinsic disorder in protein interaction networks. *FEBS J.* **2005**, *272*, 5129–5148.
- Tompa, P. Intrinsically unstructured proteins evolve by repeat expansion. *Bioessays* **2003**, *25*, 847–855.
- Wootton, J. C.; Federhen, S. Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem. (Oxford)* **1993**, *17*, 149–163.
- Wootton, J. C. Sequences with “unusual” amino acid compositions. *Curr. Opin. Struct. Biol.* **1994**, *4*, 413–421.
- Andrade, M. A.; Perez-Iratxeta, C.; Ponting, C. P. Protein repeats: Structures, functions, and evolution. *J. Struct. Biol.* **2001**, *134*, 117–131.
- Romero, P.; Obradovic, Z.; Li, X.; Garner, E. C.; Brown, C. J.; Dunker, A. K. Sequence complexity of disordered protein. *Proteins* **2001**, *42*, 38–48.
- Oldfield, C. J.; Cheng, Y.; Cortese, M. S.; Brown, C. J.; Uversky, V. N.; Dunker, A. K. Comparing and combining predictors of mostly disordered proteins. *Biochemistry* **2005**, *44*, 1989–2000.
- Patil, A.; Nakamura, H. Disordered domains and high surface charge confer hubs with the ability to interact with multiple proteins in interaction networks. *FEBS Lett.* **2006**, *580*, 2041–2045.
- Dosztányi, Z.; Csizmok, V.; Tompa, P.; Simon, I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.* **2005**, *347*, 827–839.
- Dosztányi, Z.; Csizmok, V.; Tompa, P.; Simon, I. IUPred: Web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **2005**, *21*, 3433–3434.

- (41) Pellegrini, M.; Marcotte, E. M.; Yeates, T. O. A fast algorithm for genome-wide analysis of proteins with repeated sequences. *Proteins* **1999**, *35*, 440–446.
- (42) Deane, C. M.; Salwinski, L.; Xenarios, I.; Eisenberg, D. Protein interactions: Two methods for assessment of the reliability of high throughput observations. *Mol. Cell Proteomics* **2002**, *1*, 349–356.
- (43) Bader, G. D.; Hogue, C. W. Analyzing yeast protein–protein interaction data obtained from different sources. *Nat. Biotechnol.* **2002**, *20*, 991–997.
- (44) Han, J. D.; Bertin, N.; Hao, T.; Goldberg, D. S.; Berriz, G. F.; Zhang, L. V.; Dupuy, D.; Walhout, A. J.; Cusick, M. E.; Roth, F. P.; Vidal, M. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature* **2004**, *430*, 88–93.
- (45) Namba, K. Roles of partly unfolded conformations in macromolecular self-assembly. *Genes Cells* **2001**, *6*, 1–12.
- (46) Dajani, R.; Fraser, E.; Roe, S. M.; Yeo, M.; Good, V. M.; Thompson, V.; Dale, T. C.; Pearl, L. H. Structural basis for recruitment of glycogen synthase kinase 3beta to the axin-APC scaffold complex. *Embo. J.* **2003**, *22*, 494–501.
- (47) Uversky, V. N. A protein-chameleon: conformational plasticity of alpha-synuclein, a disordered protein involved in neurodegenerative disorders. *J. Biomol. Struct. Dyn.* **2003**, *21*, 211–234.
- (48) Bustos, D. M.; Iglesias, A. A. Intrinsic disorder is a key characteristic in partners that bind 14-3-3 proteins. *Proteins: Struct., Funct., Bioinf.* **2006**, *63*, 35–42.
- (49) Toroczkai, Z.; Bassler, K. E. Network dynamics: Jamming is limited in scale-free systems. *Nature* **2004**, *428*, 716.
- (50) Fernandez, A.; Scott, R.; Berry, R. S. The nonconserved wrapping of conserved protein folds reveals a trend toward increasing connectivity in proteomic networks. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 2823–2827.
- (51) Louzoun, Y.; Muchnik, L.; Solomon, S. Copying nodes versus editing links: The source of the difference between genetic regulatory networks and the WWW. *Bioinformatics* **2006**, *22*, 581–588.
- (52) Iakoucheva, L. M.; Radivojac, P.; Brown, C. J.; O'Connor, T. R.; Sikes, J. G.; Obradovic, Z.; Dunker, A. K. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.* **2004**, *32*, 1037–1049.
- (53) Luscombe, N. M.; Babu, M. M.; Yu, H.; Snyder, M.; Teichmann, S. A.; Gerstein, M. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* **2004**, *431*, 308–312.
- (54) Fraser, H. B.; Hirsh, A. E.; Steinmetz, L. M.; Scharfe, C.; Feldman, M. W. Evolutionary rate in the protein interaction network. *Science* **2002**, *296*, 750–752.
- (55) Jeong, H.; Mason, S. P.; Barabasi, A. L.; Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **2001**, *411*, 41–42.

PR060171O

**JMB**Available online at [www.sciencedirect.com](http://www.sciencedirect.com) ScienceDirect

## Molecular Principles of the Interactions of Disordered Proteins

**Bálint Mészáros, Peter Tompa, István Simon and Zsuzsanna Dosztányi\***

*Institute of Enzymology  
Biological Research Center  
Hungarian Academy of  
Sciences, 1518 Budapest  
P.O. Box 7, Hungary*

Thorough knowledge of the molecular principles of protein–protein recognition is essential to our understanding of protein function at the cellular level. Whereas interactions of ordered proteins have been analyzed in great detail, complexes of intrinsically unstructured/disordered proteins (IUPs) have hardly been addressed so far. Here, we have collected a database of 39 complexes of experimentally verified IUPs, and compared their interfaces with those of 72 complexes of ordered, globular proteins. The characteristic differences found between the two types of complexes suggest that IUPs represent a distinct molecular implementation of the principles of protein–protein recognition. The interfaces do not differ in size, but those of IUPs cover a much larger part of the surface of the protein than for their ordered counterparts. Moreover, IUP interfaces are significantly more hydrophobic relative to their overall amino acid composition, but also in absolute terms. They rely more on hydrophobic–hydrophobic than on polar–polar interactions. Their amino acids in the interface realize more intermolecular contacts, which suggests a better fit with the partner due to induced folding upon binding that results in a better adaptation to the partner. The two modes of interaction also differ in that IUPs usually use only a single continuous segment for partner binding, whereas the binding sites of ordered proteins are more segmented. Probably, all these features contribute to the increased evolutionary conservation of IUP interface residues. These noted molecular differences are also manifested in the interaction energies of IUPs. Our approximation of these by low-resolution force-fields shows that IUPs gain much more stabilization energy from intermolecular contacts, than from folding, i.e. they use their binding energy for folding. Overall, our findings provide a structural rationale to the prior suggestions that many IUPs are specialized for functions realized by protein–protein interactions.

© 2007 Elsevier Ltd. All rights reserved.

*Keywords:* intrinsically unstructured proteins; protein–protein interactions; molecular recognition; disorder-to-order transition; protein–protein interface

\*Corresponding author

### Introduction

The recent success in high-throughput studies of protein structure, function and interactions provide solid evidence that for most proteins macromole-

cular interactions are indispensable for their functions.<sup>1,2</sup> Among these, protein–protein interactions (PPIs) are central for both function and control, and justify concerted efforts aimed at describing the complicated network of interacting proteins, i.e. the interactome.<sup>3,4</sup> At the level of individual proteins, most pertinent studies are directed towards solving structures of complexes to unveil the molecular principles of interactions that govern specific recognition.<sup>1</sup> The idea underlying all these efforts is that a better understanding of PPIs at both the individual and system levels will provide an improved atomistic picture of communication with-

Abbreviations used: PDB, Protein Data Bank; IUP, intrinsically unstructured protein; PPI, protein–protein interaction; MorEs, molecular recognition elements; MorFs, molecular recognition features.

E-mail address of the corresponding author:  
[zsuzsa@enzim.hu](mailto:zsuzsa@enzim.hu)

in the proteome, which is essential for our understanding of the living cell.

A common theme in these efforts is the detailed analysis of the molecular interfaces that proteins apply to recognize each other. Several studies have focused on dissecting these surfaces, and have shown that these usually are of the order of 1000 Å<sup>2</sup> in area, and they are distinguished from the average surface of proteins by an elevated hydrophobicity, evolutionary conservation of certain anchoring residues and characteristic shapes that differ for various classes of complexes, such as homodimers, heterodimers or enzyme–inhibitor complexes.<sup>5–8</sup> The insight gained from these analyzes enables the prediction of interfaces from structural information about the monomers and enables a structural interpretation of functionally relevant features of PPIs, such as strength, kinetics, specificity and evolution.

The issue of the molecular principles of PPIs, however, has so far been unduly neglected in the case of the newly recognized structural class of intrinsically disordered proteins. The discovery of such regions (IDRs) and full-length proteins (IUPs)<sup>9–11</sup> has been followed by the recognition that protein disorder is

widespread in eukaryotic proteomes, and correlates with signaling and regulatory,<sup>12–16</sup> and chaperone<sup>17</sup> functions. The relevance to our subject comes from the fact that all these functions rely on rapid and highly regulated PPIs and, in fact, several functional advantages ascribed to protein disorder are linked directly with its involvement in protein binding. When IUPs/IDRs are involved in PPIs, they undergo induced folding or disorder-to-order transition,<sup>18,19</sup> suggested to provide several advantages, such as specificity without excessive binding strength, increased speed of interaction, binding promiscuity or moonlighting, among others.<sup>14,20,21</sup> These advantages underscore the high frequency of disorder in proteins organizing the interactome, i.e. in hubs.<sup>22–25</sup> In terms of the molecular details of the recognition process, it has been suggested that IUPs often use short sequential recognition elements for binding, termed primary contact sites,<sup>26</sup> preformed structural elements,<sup>27</sup> or molecular recognition elements/features (MorEs/MorFs).<sup>28,29</sup> In direct connection with these concepts, a recent analysis has shown that short linear motifs in proteins often fall into locally disordered regions,<sup>30</sup> underlying the suggestion that interactions of IUPs/IDRs

**Table 1.** Disordered protein complexes

PDB ID	Chain ID	Resolution (Å)	Name
1axc	(B, A)	2.6	DNA-binding protein/DNA (human PCNA)
1cee	(B, A)	NMR	CDC42 – GTPase binding domain of WASP
1cmk	(I, E)	2.9	Phosphotransferase
1cqt	(I, A)	3.2	Gene regulation/DNA (ternary complex)
1dev	(B, A)	2.2	Signaling protein (SMAD2 MH2 – SARA)
1dpj	(B, A)	1.8	Hydrolase/hydrolase inhibitor (proteinase A – IA3 peptide inhibitor)
1f83	(BC, A)	2.0	Hydrolase/membrane protein (botulinum neurotoxin – synaptobrevin-II)
1fj	(C, ABDE)	2.02	Signaling protein
1fv1	(F, DE)	1.9	Immunodominant peptide from myelin basic protein
1g3j	(B, A)	2.1	XTCF3-CBD/β-catenin armadillo repeat complex
1i7x	(B, A)	3.0	β-Catenin/E-cadherin complex
1i8h	(A, B)	NMR	Membrane protein/isomerase
1iwq	(B, A)	2.0	Metal-binding protein
1j2j	(B, A)	1.6	GGA1 gat N-terminal region in complex with ARF1 GTP form
1jsu	(C, AB)	2.3	Transferase/cyclin/inhibitor
1kdx	(B, A)	NMR	Transcription regulation complex (KIX domain of CBP – PKID CREB)
1kil	(CD, ABE)	2.3	Membrane protein (complexin/snare complex)
1l8c	(B, A)	NMR	Gene regulation
1mv0	(A, B)	NMR	Tumor suppressor BIN1
1mxl	(I, C)	NMR	Ca-binding protein (cardiac troponin C – troponin I)
1o9a	(B, A)	NMR	Complex of 1f12f1 fibronectin with B3 from FNBB
1onv	(B, A)	NMR	RAP74 – MAP II CTD phosphatase FCP1
1p16	(D, B)	2.7	mRNA capping enzyme – RNA polymerase
1p4q	(A, B)	NMR	Transferase
1q68	(A, B)	NMR	T-cell surface glycoprotein – tyrosine-protein kinase
1rf8	(B, A)	NMR	Biosynthetic protein (EIF4E – M7GDP and EIF4GI)
1sb0	(B, A)	NMR	KIX domain of CBP – transactivation domain of c-MYB
1sc5	(B, A)	3.26	σ-28(FLIA)/FLGM complex
1sqk	(B, A)	2.5	Structural protein (ciboulot – skeletal actin)
1sqq	(I, A)	3.0	Oxidoreductase
1tba	(A, B)	NMR	Transcription factors (TBP-TAFII230 complex)
1th1	(C, A)	2.5	Cell adhesion/antitumor protein
1vit	(I, LH)	3.2	Serine protease/inhibitor
1wkw	(B, A)	2.1	Ternary complex of EIF4E-M7GPPA-4EBP1 peptide
1xtg	(B, A)	2.1	Neurotoxin BONT/A – synaptosomal-associated protein 25
1ycq	(B, A)	2.3	Oncogene protein (MDM2 – p53)
2auh	(B, A)	3.2	Transferase (grb14 bps region – receptor tyrosine kinase)
2b3g	(B, A)	1.6	P53N – RPA70N
2c1t	(D, B)	2.6	Nuclear transport complex (KAP60P:NUP2 complex)

**Table 2.** Ordered complexes

PDB ID	Chain ID	Resolution (Å)	Name
1a0o	(B, A)	2.95	CHEA – CHEY
1atn	(D, A)	2.8	Deoxyribonuclease I – actin
1avz	(C, AB)	3.0	Myristylation/transferase
1glb	(F, G)	2.6	Phosphotransferase
1hrp	(A, B)	3.0	Hormone (HCG)
1lpa	(A, B)	3.04	Hydrolase (carboxylic esterase)
1luc	(B, A)	1.5	Bacterial luciferase
1spb	(P, S)	2.0	Serine proteinase/prosegment
1ttq	(A, B)	2.0	Carbon-oxygen lyase (tryptophan synthase)
2btf	(P, A)	2.55	Acetylation and actin-binding ( $\beta$ -actin-profilin complex)
2pcb	(B, AC)	2.8	Cytochrome <i>c</i> peroxidase (ccp) – Cytochrome <i>c</i>
3hhr	(A, BC)	2.8	Human growth hormone complexed with its receptor
1aoz	(A, B)	1.9	Oxidoreductase (oxygen acceptor)
1cdt	(A, B)	2.5	Cytotoxin (cardiotoxin)
1fc1	(A, B)	2.9	Immunoglobulin (FC fragment)
1g6n	(A, B)	2.1	CAP-CAMP (DNA binding protein)
1glq	(A, B)	1.8	Glutathione transferase – glutathione
1hng	(A, B)	2.8	T lymphocyte adhesion glycoprotein
1il8	(A, B)	NMR	Interleukin 8
1msb	(A, B)	2.3	Hepatic lectin
1nl3	(A, B)	2.8	Seca protein translocation ATPase
1phh	(A, B)	2.3	Oxidoreductase (ternary complex)
1pp2	(R, L)	2.5	Hydrolase (Ca-free phospholipase)
1pyp	(A, B)	3.0	Acid anhydride hydrolase
1tar	(A, B)	2.2	Aspartate aminotransferase
1utg	(A, B)	1.34	Steroid binding (uteroglobin)
1vsg	(A, B)	2.9	Variant surface glycoprotein
1ypi	(A, B)	1.9	Isomerase (intramolecular oxidoreductase)
2ccy	(A, B)	1.67	Electron transport (heme protein)
2cts	(A, B)	2.0	Oxo-acid-lyase (citrate synthase)
2gn5	(A, B)	2.3	Gene 5/DNA binding protein
2or1	(L, R)	2.5	Gene regulating protein (434 repressor complex with operator)
2rhe	(A, B)	1.6	Immunoglobulin (Bence-Jones protein)
2rus	(A, B)	2.3	Rubisco complex with CO <sub>2</sub> and Mg
2rve	(A, B)	3.0	ECO RV endonuclease
2sod	(O, Y)	2.0	Oxidoreductase (Cu, Zn superoxide dismutase)
2ts1	(A, B)	2.3	Ligase (tyrosyl-transfer RNA synthetase)
2tsc	(A, B)	1.97	Methyltransferase (thymidylate synthase)
3enl	(A, B)	2.25	Carbon-oxygen lyase
3grs	(A, B)	1.54	Oxidoreductase (flavoenzyme)
3hvt	(B, A)	2.9	Nucleotidyltransferase (HIV virus reverse transcriptase)
3icd	(A, B)	2.5	Oxidoreductase (isocitrate dehydrogenase)
3sdh	(A, B)	1.4	Hemoglobin I (carbon-monoxo)
3sdp	(A, B)	2.1	Oxidoreductase (Fe superoxide dismutase)
3ssi	(A, B)	2.3	Serine protease inhibitor
4mdh	(A, B)	2.5	Oxidoreductase (cytoplasmic malate dehydrogenase)
5adh	(A, B)	2.9	Oxidoreductase (alcohol dehydrogenase-ADP-ribose)
1ahw	(C, AB)	3.0	Tissue factor – inhibitory FAB (5g9)
1fdl	(Y, LH)	2.5	IGG1 FAB fragment – lysozyme
1iai	(MI, LH)	2.9	Idiotypic – anti-idiotypic FAB complex
1jhl	(A, LH)	2.4	Lysozyme antibody D11.15 – lysozyme
1mel	(L, A)	2.5	VH antibody – lysozyme
1nca	(N, LH)	2.5	N9 neuraminidase-NC41-FAB
1yqv	(Y, LH)	1.7	FAB HYHEL5 – lysozyme
2jel	(P, LH)	2.5	JELL42 FAB/HPR complex
2vir	(C, AB)	3.25	Influenza virus hemagglutinin – neutralizing antibody
3hfm	(Y, LH)	3.0	IGG1 FAB fragment – lysozyme
1acb	(I, E)	2.0	Hydrolase (serine protease)
1avw	(B, A)	1.75	Trypsin – trypsin inhibitor
1brs	(D, A)	2.0	Barnase – barstar
1cho	(I, E)	1.8	Serine proteinase/inhibitor ( $\alpha$ -chymotrypsin-ovomucoid third domain)
1cse	(I, E)	1.2	Serine proteinase/inhibitor (subtilisin carlsberg – eglin-C)
1dfj	(E, I)	2.5	Ribonuclease inhibitor – ribonuclease A
1fss	(B, A)	3.0	Acetylcholinesterase – fasciculin-II
1mct	(I, A)	1.6	Trypsin – trypsin inhibitor
1stf	(I, E)	2.37	Hydrolase (papain – inhibitor stefin B)
1tab	(I, E)	2.3	Hydrolase (trypsin – Bowman-Birk inhibitor)
1tgs	(I, Z)	1.8	Proteinase/inhibitor (trypsinogen – trypsin inhibitor)
1ugh	(I, E)	1.9	Glycosylase (uracil-DNA glycosylase – protein inhibitor)
2ptc	(I, E)	1.9	$\beta$ -Trypsin – trypsin inhibitor
2sic	(I, E)	1.8	Subtilisin – streptomyces subtilisin inhibitor
4htc	(I, LH)	2.3	Hydrolase ( $\alpha$ -thrombin – recombinant hirudin)

represent a distinct mode of PPIs.<sup>31</sup> Whereas this notion has been often invoked in the literature, structural studies in support are rather scarce.

In two recent studies of the interfaces, several classes of protein complexes, including a limited number of IUPs<sup>32</sup> and complexes of short binding elements (molecular recognition elements/features), have been analyzed.<sup>33</sup> It was found that the surfaces and interfaces of two-state complexes and complexes of IUPs show special features, such as increased area per residue, or an elevated level of exposed hydrophobic residues buried only upon complex formation. These studies have limitations in terms of generalizations toward IUP behavior, either because they included only five *bona fide* IUP complexes, ten two-state complexes, and 44 ribosomal proteins,<sup>32</sup> or because selection of the 258 MoRF examples was based on length.<sup>33</sup> The resulting dataset contains complexes of small but stable proteins, like the trypsin inhibitor, beside many potentially disordered proteins. These results pertain to the behavior of IUPs, but the recent rapid growth in the number of experimentally verified IUPs,<sup>34</sup> and structures of complexes with an IUP as one partner,<sup>35–38</sup> enabled us to extend these studies and provide further statistically sound conclusions. On the basis of novel data, we identified 39 complexes, in which an IUP binds a globular protein partner, analyzed their interfaces and compared them with those of 72 complexes of globular proteins. We found that the chemical and physical features of interfaces of IUPs are distinct from those of globular proteins in many aspects, such as a much larger relative area, and preference for hydrophobic residues, which exceeds even the interior of the protein. Their much larger contact number per residue are indicative of structural adaptation driven by induced folding, and suggest that IUPs might preferentially gain stabilization energy from contacts with the partner. We provide evidence for this latter by estimating these energies *via* low-resolution force-fields, already exploited in predicting protein disorder by the IUPred algorithm.<sup>39,40</sup> These differences suggest that IUP binding represents a unique implementation of the principles of protein–protein binding, which provides the rationale for why IUPs fold only upon encountering their physiological partner, and why so often they function *via* molecular recognition.<sup>18,21,41</sup>

## Results

### Data for analysis

For this analysis, we have scanned the PDB for entries of complexes, one chain of which was proven to be disordered by experimental techniques. The database of such “disordered” complexes contains 39 complexes (Table 1). The interfaces of these have been analyzed and compared to those of 72 complexes, for which both components are ordered

in isolation (ordered complexes, Table 2). Figure 1 shows characteristic examples for the two classes, i.e. ordered and disordered, which already illustrate the differences that emerge as the quantitative study unfolds.

### Global characteristics of interfaces

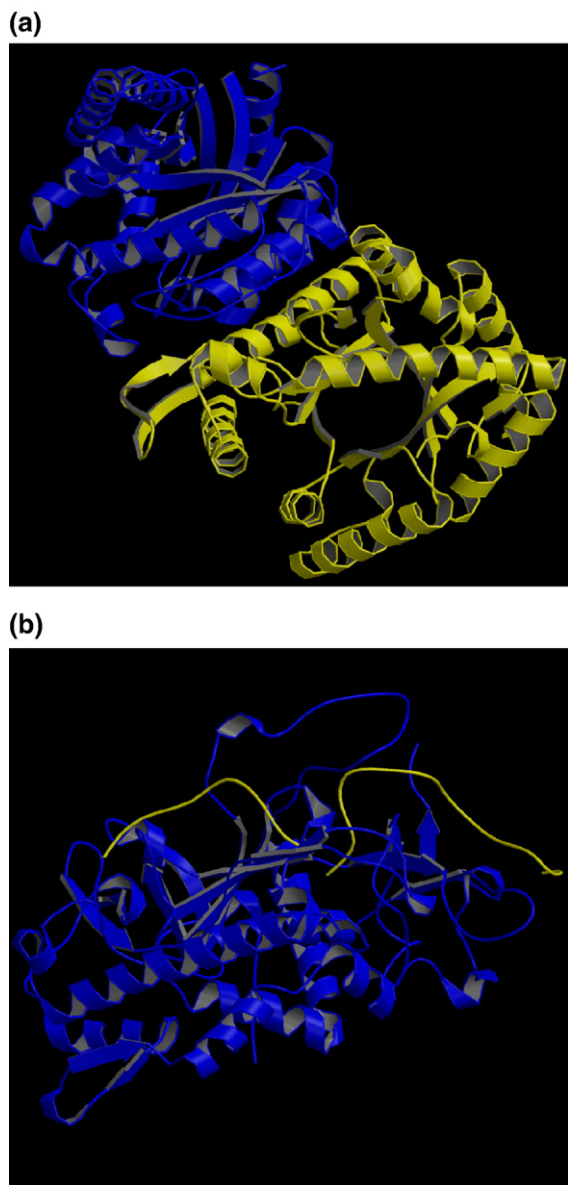
The first level of analysis is a rough comparison of the geometry of interfaces of the two classes of complexes. Figure 2(a) shows the distribution of the size of the interface area. The distributions are not significantly different, i.e. the size of IUP interfaces covers about the same range as those of ordered complexes, maybe with the lack of very large interfaces (>3000 Å<sup>2</sup>) for IUPs. Characteristic differences can be seen, however, if the interface area is plotted as a function of chain length (Figure 2(b)). It takes a much longer chain for globular proteins than for IUPs to create the same interface, which indicates that IUPs may have a much larger interaction potential, as already suggested.<sup>42</sup>

The explanation for this observation may come from the fact that IUPs have the same relative surface, but use a larger part of it for interaction, or they already have a larger surface per residue, or both. To find out which actually applies, we plotted the surface area per residue of proteins against their interaction areas per residue (Figure 3(a)). The two classes show a striking difference and clear separation, which validates previous observations made on a much smaller dataset.<sup>32</sup> IUPs have a much larger surface per residue, and they exceed globular proteins in their interface area per residue. When the ratio of the two values are calculated (Figure 3(b)), it is clear that IUPs have relatively larger surfaces, they also use a larger portion of their surface for interaction with their partner, sometimes 50% of the whole, as opposed to only 5%–15% for most ordered proteins.

Another way of comparing the interfaces is to count how many continuous segments (for definition, see Data and Methods) the binding surface is assembled from. Since folding of a globular protein brings distinct segments of the polypeptide chain in proximity, it is expected that their binding surfaces are more fragmented, i.e. they are assembled from more segments than those of IUPs. The picture that emerges (Figure 4) is in complete agreement with this expectation: in 70% of the cases the binding surface of the IUP represents a single sequentially continuous segment only, and with the exception of a single case (PDB 1sc5) they never contain more than three separate segments. On the other hand, ordered proteins hardly ever use a single segment for binding to their partner, and their segmentation number may occasionally even exceed the value of 10.

### Chemical nature of interfaces

As seen, global characteristics of the interfaces of IUPs and ordered proteins differ significantly. When

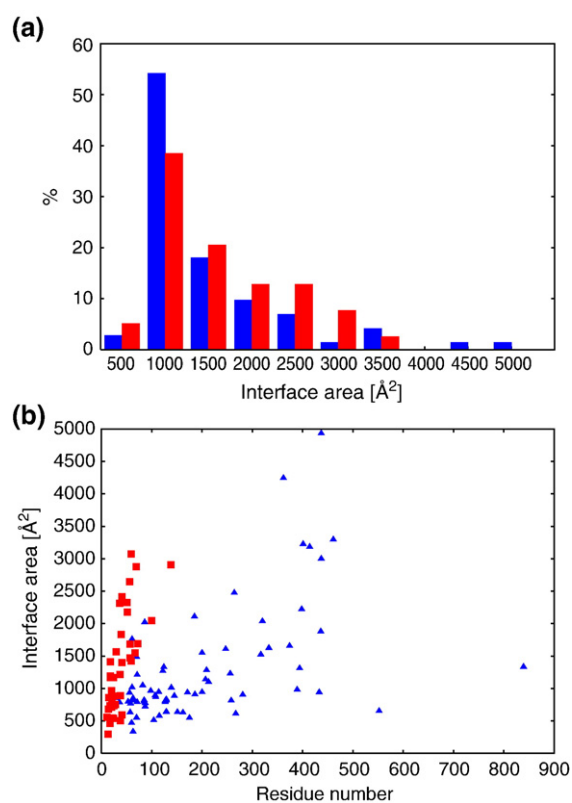


**Figure 1.** A typical example of ordered and disordered complexes. The Figure shows an example of the two basic types of complexes analyzed in this study. (a) Bacterial luciferase (PDB code 1luc) is an example for complexes between ordered proteins. (b) Botulinum neurotoxin (PDB code 1f83) shows a complex between a disordered and ordered protein, where the disordered protein chain wraps around an ordered protein in a largely extended conformation.

we ask about the physical and chemical nature of the interfaces, IUPs again stand out. By having much larger relative surfaces, it is natural their ratio of buried/exposed area is much smaller on average than that of globular proteins (Figure 5). As expected, this is true for polar residues, because proteins in general must not be able to bury polar residues upon folding. The surprise, however, comes from looking at hydrophobic residues, because they also are more exposed than buried in IUPs, which suggests that they are mostly used for

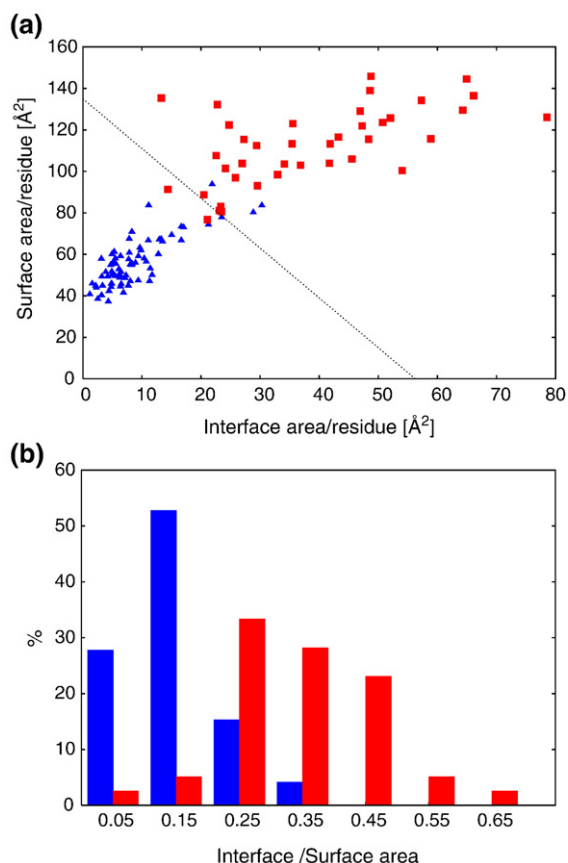
contact with the partner, and not for generating a hydrophobic core, as with globular proteins. In a way, it may be suggested that the hydrophobic core of IUPs is in the interface, and not within the polypeptide chain of the folded, partner-bound state. This observation suggests an unexpected structural strategy for function, in a sense that IUPs tend to expose their few hydrophobic residues for interaction with the partner. A similar observation has been made in the case of two-state complexes, the monomers of which may exist either as unfolded (disordered) in isolation or folded in the complex.<sup>32,43</sup>

The detailed analysis of the amino acid composition of the interfaces provides full evidence for this point, i.e. that IUPs preserve and expose their hydrophobic residues for partner binding. In agreement with previous studies,<sup>5,32</sup> the analysis of ordered proteins shows that their surface is enriched in polar/charged residues, and depleted in hydrophobic residues (Figure 6(a)). Relative to this distribution, their interfaces are more hydrophobic, and occupy a position intermediate between the hydrophobic interior and the polar surface of the

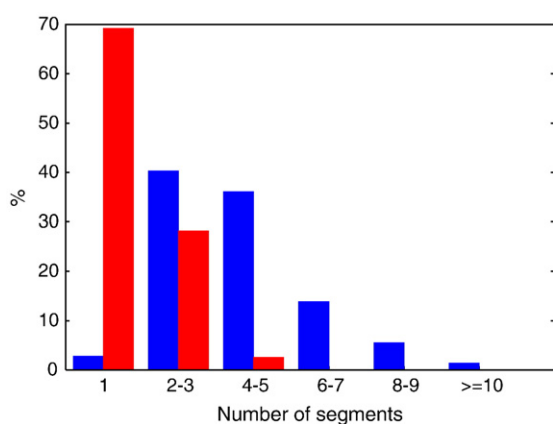


**Figure 2.** IUPs realize large contact surfaces in their complexes. (a) The distribution of the size of the interface area for the smaller chain of ordered complexes (blue bars), and for disordered proteins in complex with an ordered protein (red bars). (b) The interface area *versus* chain length for the smaller chain of the ordered (blue triangles) and disordered (red squares) type of complexes (the smaller chain is always the IUP in the disordered complexes) as a function of the length of the chain involved in the construct.

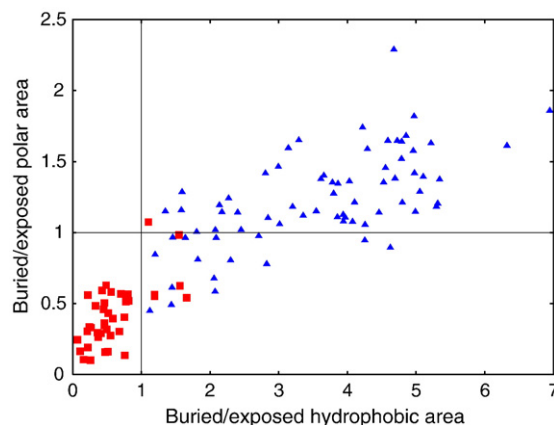




**Figure 3.** IUPs use a large fraction of their surfaces for binding. (a) Surface area per residue *versus* interface area per residue for the smaller chain of ordered complexes (blue triangles), and for disordered proteins in complex with an ordered protein (red squares). (b) The distribution of interfaces in terms of the interface area relative to the total surface area (blue for ordered and red for disordered complexes).



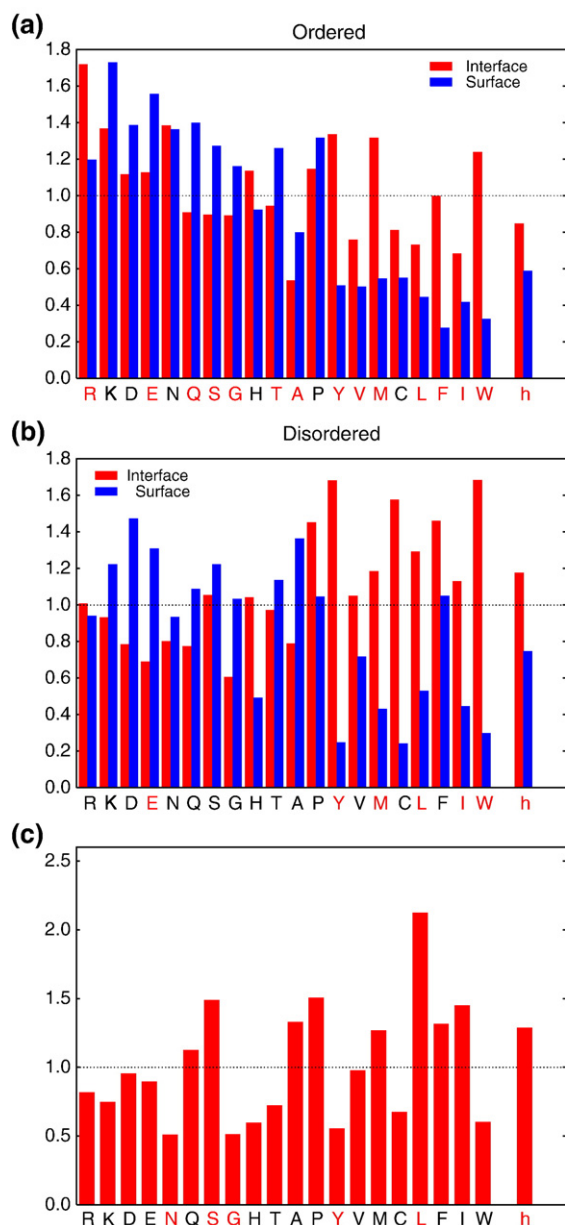
**Figure 4.** Lack of segmentation of the interfaces of IUPs. The distribution of the occurrence of interfaces with various numbers of non-continuous sequence segments (as defined in Data and Methods), given for the smaller chain of ordered complexes (blue), and for disordered proteins in complex with an ordered protein (red).



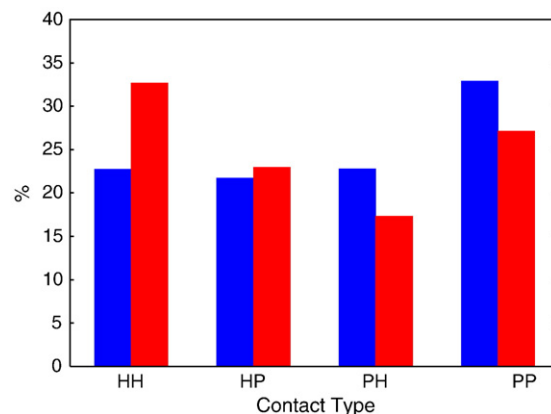
**Figure 5.** Buried and exposed surfaces of IUPs and ordered proteins in complexes. The ratio of the buried and exposed polar area *versus* the ratio of the buried and exposed hydrophobic area for the smaller chain of ordered complexes (blue triangles), and for disordered proteins in complex with an ordered protein (red squares).

protein, with the balance tilted towards polar residues. IUPs look completely different (Figure 6(b)). As already suggested by the analysis of the polarity/hydrophobicity of their surfaces (Figure 5), they keep a larger fraction of their hydrophobic residues exposed than ordered proteins do. Furthermore, unlike the case of ordered proteins, their interfaces are more hydrophobic than their surface in general, and more hydrophobic than the protein in general, and thus more hydrophobic than the buried regions of the protein. Thus, IUPs show a unique preference to expose and use their hydrophobic residues for interaction, whether they are aromatic (Trp, Tyr), or aliphatic (Leu, Ile). This difference comes from a greater relative hydrophobicity of IUP interfaces, and from the chemical differences between the two groups: interfaces of IUPs are significantly more hydrophobic than interfaces of ordered proteins (Figure 6(c)). It appears that IUPs rely much more on hydrophobic residues at their interface, which counteracts their unfavorable decrease in entropy upon folding.

When the types of contacts in the interface are counted (Figure 7), this distinction clearly shows. IUP interfaces rely much more on hydrophobic-hydrophobic contacts than ordered proteins do, which balance this by significantly more polar-polar contacts, probably due to being able to better shield these from hydration water. A further characteristic difference is that the interface residues of IUPs are engaged in larger numbers of contacts than those of ordered proteins (Figure 8). The reason of this difference is probably related directly to the different binding modes of the two structural classes, because IUPs undergo large-scale induced folding upon binding, and can better adapt to the structure of the partner, whereas structural adaptation of ordered proteins is limited due to their much lower level of conformational freedom.



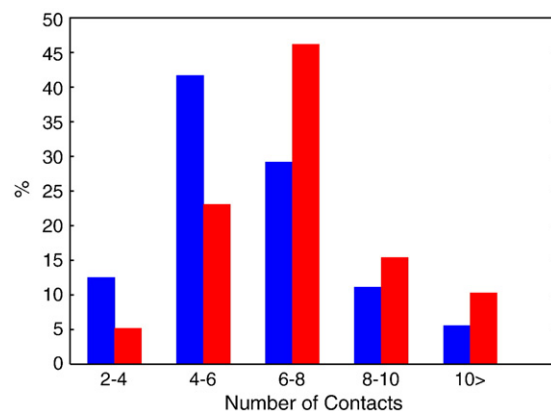
**Figure 6.** Amino acid composition of the surface and interface of IUPs and ordered proteins. (a) The amino acid composition of the surface (blue) and the interface area (red) relative to the total amino acid composition for the smaller chain of ordered complexes, and (b) for disordered proteins in complex with an ordered protein. The amino acids are sorted according to the hydrophobicity scale described by Fauchere and Pliska.<sup>50</sup> Additionally, the ratio of hydrophobic amino acids (Ala, Cys, Ile, Leu, Met, Prot, Val, Phe, Trp, Tyr) is shown for the surface and the interface residues normalized by the total amino acid composition (marked as h). Amino acids with statistically significant differences between the surface and interface compositions are marked by a red letter. Statistical significance was calculated by two-sided Student's *t*-test. (c) The ratio of the amino acid compositions of IUP interfaces and interfaces of ordered proteins. The ratio of hydrophobic amino acids in the two datasets is shown. Amino acids showing significant differences between the interface compositions of ordered and disordered complexes are marked by a red letter.



**Figure 7.** Contribution of hydrophobic and polar contacts to the interfaces of IUPs and ordered proteins. Histogram of the various types of contacts in the interfaces of the smaller chain of ordered complexes (blue), and of disordered proteins in complex with an ordered protein (red). For this analysis, amino acids were classified as hydrophobic (H: Ala, Cys, Ile, Leu, Met, Prot, Val, Phe, Trp, Tyr) or polar (Asp, Glu, Gly, His, Lys, Asn, Gln, Arg, Ser, and Thr). The first position corresponds to the smaller chain of the complex, which is always the IUP in the complex designated disordered, thus HP does not equal PH contacts.

### Interaction energies at the interface

A key question that arises from all the foregoing studies is whether all the characteristic differences between the interfaces of ordered and disordered proteins manifest themselves in differences in the interaction energies of the two types of complexes. We showed earlier that the concept of low-resolution force-fields can be used to recognize IUPs from the amino acid sequence because the estimated pair-wise inter-residue interaction energy of IUPs is less favorable



**Figure 8.** The histogram shows the distribution of the number of atom contacts per interacting residues-residues pairs at the interface (as defined in Data and Methods) observed for the smaller chain of the ordered complexes (blue), and for disordered proteins in complex with an ordered protein (red).

compared to globular proteins.<sup>39</sup> Here, we applied the same statistical potentials to characterize the energetic relationship of the complexes. The total interaction energy that arises from interactions within individual chains and from their interactions with the partners were calculated (Figure 9(a)). The picture shows a clear separation of the two classes of complexes. Considering the energy realized from intramolecular interactions, ordered proteins invariably fall within the stabilizing range, which underscores that they fold due to favorable interactions within the chain. Most of the IUPs, as already shown,<sup>39</sup> cannot form sufficient inter-residue interactions for a stable fold, which renders them disordered in isolation. The situation, however,

changes in the presence of the partner. IUPs tend to have more stabilizing interactions with the partner, which shifts the overall balance towards favoring the folded state. This explains why IUPs undergo induced folding in the presence of the partner and explains their observed preference for hydrophobic residues in the interface, since stabilization primarily comes from hydrophobic interactions.

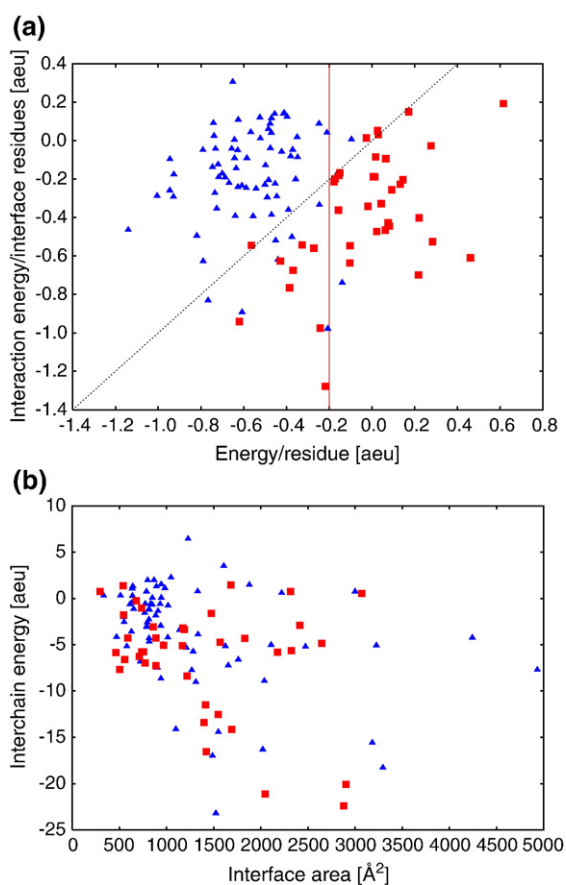
Total interface energies can also be calculated and plotted as a function of the interface area (Figure 9(b)). Overall, IUPs tend to have somewhat larger negative value of stabilization energies for an interface of the same size, most obvious within the range of smaller interfaces that range from 500 Å<sup>2</sup> to 1500 Å<sup>2</sup>. This suggests that a better fit and more hydrophobic contacts result in a somewhat greater binding energy. It should not be overlooked, however, that this enthalpic component is combined with a large unfavorable entropic term due to the large-scale induced folding of the IUPs, which probably makes binding of IUPs overall weaker.

### Conservation of the interface of IUPs

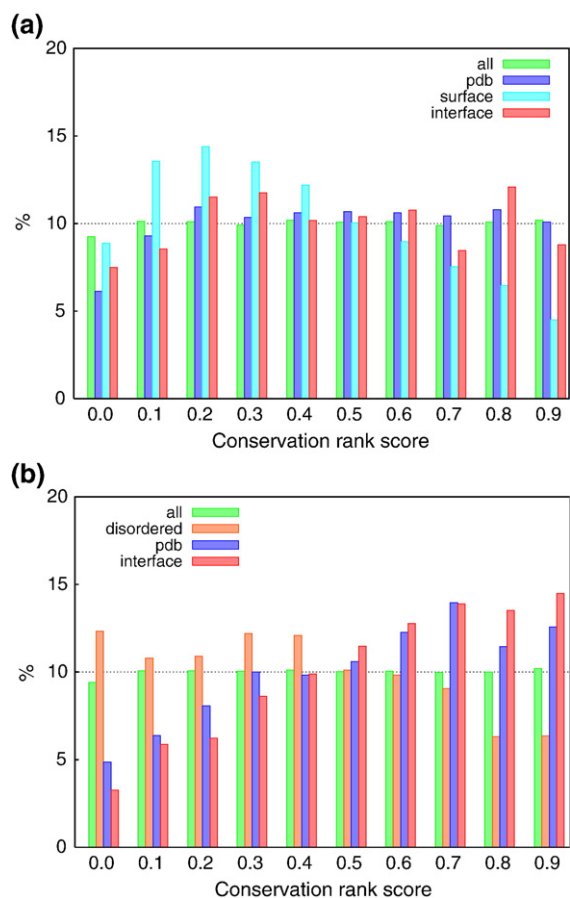
Since interfaces constitute the key structural element of PPIs, which, in turn, are intertwined with the function of proteins, we asked whether it has made its mark on the evolutionary conservation of proteins. To this end, we calculated a conservation rank score for various parts of the complexes, including the whole protein, the part seen in the PDB, and the interface itself (Figure 10). For both IUPs and ordered complexes, the distribution of conservation scores for the complete sequences is even, as expected. In other aspects, the two types of complexes differ significantly. In the case of globular proteins, positions corresponding to the PDB structure are close in distribution to the complete sequences, although they contain somewhat fewer variable residues. Surface residues not involved in the complexes are less conserved, whereas interface residues are significantly more conserved in comparison, although they do not show an overall significant conservation, probably due to the key contribution to interactions of only a few anchor residues.<sup>8,44</sup> IUPs show clearly distinguishable behavior. Disordered regions, on average, are the least conserved compared to all other types of regions. The regions that become ordered upon complex formation (structured part in PDB) are more conserved, even in comparison to the complete sequence. Interface positions show the most pronounced tendency to be conserved, in line with their importance in the function of IUPs, also apparent in their energetic contribution to binding-induced folding, as shown in the previous section.

### Discussion

The advance of intrinsic structural disorder in evolution is often associated with the advantages



**Figure 9.** Pairwise energies of intrachain *versus* interchain interactions. (a) Pairwise energies of intrachain *versus* interchain interactions normalized by the length of the chain and the number of residues at the interface, respectively. Pairwise energy was calculated by statistical force-fields, as given in Data and Methods. The energies are in arbitrary energy units (aeu). The vertical line at  $-0.2$  (arbitrary energy units) shows the borderline between ordered (blue triangles) and disordered proteins (red squares), used for predictions of protein disorder. Typically, disordered chains lie below the  $y=x$  line, indicating that in terms of statistical potentials they make more favourable interactions with their partner compared to the interactions made within the chain. (b) Total energies of interfaces as the function of the interface area of the smaller chain of ordered complexes (blue triangles), and of disordered proteins in complexes (red squares).



**Figure 10.** Conservation at the interfaces of ordered and disordered complexes. (a) The plot shows the fractional rank of conservation scores for ordered protein complexes. For each protein in the dataset, the alignments were generated at the level of the complete sequence of the SwissProt/Uniprot entry corresponding to the PDB sequence. The conservation scores were calculated as described,<sup>53</sup> and transformed into fractional rank score, which ranges between 0 and 1. The scores were divided into ten bins, and the percentage of positions falling into each bin was collected. Distribution of conservation rank scores are given for: the complete amino acid sequence (green), positions part of the structure (blue), surface positions (light blue) and interface positions (red). (b) Distribution of conservation rank scores for disordered proteins. The score is given for: the complete sequence (green), positions in disordered regions (as given in DisProt, orange), the region corresponding to the PDB file (blue), and the interface residues in the complex (red).

disorder imparts on protein–protein interactions. In accord, a great deal of relevant studies on IUPs have focused on the rules and principles of their interactions, such as the contribution of disorder to the function of hubs,<sup>23–25</sup> binding of IUPs by short recognition elements,<sup>26,29,30</sup> large-scale induced folding accompanying their partner binding<sup>18,19</sup> or the possibility of drug development by targeting IUP-binding sites.<sup>45</sup> Studies of key components designed to understand their interactions, i.e. atomistic analysis of their interaction sites, however, has so far lagged behind. Whereas the contact sur-

faces of ordered, globular proteins have been analyzed and characterized in great detail,<sup>5–8</sup> those of IUPs have scarcely been addressed. In two pertinent studies,<sup>32,33</sup> either a limited set of IUP complexes (five examples only) or complexes selected on the criterion of length (258 MoRFs) have been analyzed alongside two-state complexes and ribosomal proteins for possible generalizations on the interfaces of IUP PPIs. The recent rapid advance in the IUP field enabled us to extend these studies to the complexes of much more experimentally verified IUPs, and to provide a solid foundation to the principles of its involvement in PPIs. Our results reinforce or extend some previous observations, and suggest some useful novel generalizations.

The major finding of our work is that IUPs differ in how the chemical and physical principles of protein–protein interactions are implemented, as probably dictated by their disorder in the unbound state. IUPs tend to have much larger exposed surface per residue, of which they dedicate a much larger portion for contacting the partners. This has been suggested in earlier studies,<sup>32,33</sup> and is in excellent agreement with the proposition that IUPs have evolved to provide much larger relative intermolecular interfaces than globular proteins.<sup>42</sup> Our observations suggest that ordered and disordered proteins segregate in terms of the surface area per residue *versus* interface area per residue, i.e. the two types of proteins do not lie on a continuous scale;<sup>42</sup> rather, they represent separate and disparate solutions to similar evolutionary problems. These points provide compelling evidence for the often-invoked point that the involvement in PPIs is a key element of the functional repertoire of IUPs. The suggestion that IUPs carry out their functions by transient or permanent interactions in five out of six functional categories is in excellent agreement with these points.<sup>21,41</sup>

A further key observation of our studies is that in 70% of the cases the binding surface of the IUP represents a single continuous segment of the polypeptide chain. The possible explanation is that bringing together more segments to occupy adjacent positions would disproportionately increase the unfavorable entropic component of binding, and is avoided in most instances. This finding is a structural manifestation of the observation made by sequence analysis that short, isolated recognition segments of proteins tend to fall into locally disordered regions.<sup>30,31</sup> Our observation that interaction surfaces of IUPs tend to be more hydrophobic than the rest of the surface, or the entire chain, also conforms to earlier suggestions of entirely different origin. The concept of predicting short recognition segments of IUPs, i.e. MoREs/MoRFs,<sup>28,29</sup> from local anomalies in disorder scores can be associated with a local increase of hydrophobicity in an otherwise highly charged/polar disordered sequential environment. Interfaces of these motifs are characterized by an enrichment of usually buried residues, and depletion of otherwise exposed resi-

dues,<sup>33</sup> in agreement with our findings. Our analysis of the unique evolutionary design of short linear motifs (SLMs/ELMs) in proteins<sup>30</sup> is also in line with these observations. We found that ELMs have a basic design in that they have a few consensus residues of globular-like attributes, grafted on a carrier sequence typically disordered in nature.

This unexpected mode of IUP interactions suggests that they use their very few hydrophobic residues for intermolecular interactions rather than intramolecular stabilization of structure. In other words, they counter the tendency of hydrophobic amino acids to collapse into some structure by their special amino acid composition, keeping them exposed for interaction with the partner. In this sense, these proteins (at least their interaction segments) are really specialized for partner binding. In contrast to what has been suggested in a previous study,<sup>42</sup> our analysis of a larger number of complexes show that IUPs do not bury the majority of their hydrophobic residues, but keep most of them exposed even in the partner-induced folded state. This suggests a somewhat inside-out way of folding, in which interactions with the partner *via* primarily hydrophobic contacts promotes folding, which buries polar residues to a greater extent than with ordered, globular proteins.

The significance of these differences is underscored by our observation that the interfaces of IUPs differ from those of ordered proteins in terms of hydrophobicity. IUP interfaces are more hydrophobic with respect to the rest of the chain,<sup>32</sup> and in direct comparison with those of globular proteins, in line with results on MoRF binding motifs.<sup>33</sup> Further, they use more hydrophobic-hydrophobic than polar-polar contacts, and they realize more contacts per residue than globular proteins. Often, the interfaces of globular proteins are composed of conserved polar/charged residues that provide critical anchoring interactions, surrounded and sealed from hydration by more variable shielding residues.<sup>44</sup> The different geometry, segmentation and amino acid composition of IUP interfaces entails a different logic, since being unable to shield polar residues they must rely much more on hydrophobic-hydrophobic interactions. This also explains the relative paucity of charged residues, such as Arg, a noted anchor residue in the interactions of globular proteins,<sup>44</sup> in the interfaces of IUPs. All these features are in line with the newly observed evolutionary conservation of their interfaces, suggesting an increased number of anchoring residues compared to the very few such residues in the case of globular complexes.<sup>8,44</sup>

These molecular features are expected to manifest themselves in the energetics of interactions. Our estimation of the interaction energies by low-resolution force-fields shows that IUPs realize much more energy in their interaction than within the chain upon folding, and have a somewhat larger total energy at the interface than ordered proteins. Since our calculations estimate only the enthalpic component of binding energy, it is safe to conclude that the overall free energy of IUP binding is rather

small, due to a large part being spent on compensating the unfavorable entropic cost of folding from a disordered state. Thus, our results are in line with the general wisdom of the field, that the special binding mode of IUPs separate binding strength from specificity, an often-mentioned functional advantage of protein disorder.<sup>10,14,41</sup>

In conclusion, we might state that recent studies on the principles of the interactions of IUPs point to a coherent picture of how these proteins realize their function. All studies<sup>32,33</sup> agree that the interfaces of IUPs per residue are much larger, and IUPs, unlike globular proteins, use their hydrophobic residues for interaction rather than folding. The size of their interfaces is not particularly different from those of ordered proteins, but they tend to consist of a few, often only a single, segment. This is probably attributed to the fact that IUPs use the energy of binding to assist folding, which, according to our results, would be energetically too demanding if more segments had to fold and come in spatial proximity for binding. As a final word, we must note that in addition to all these intriguing theoretical insights gained from studying the interactions of IUPs, understanding their interaction principles has a much farther-reaching practical side. Recent studies have shown that protein disorder abounds in proteins involved in various diseases, such as cancer,<sup>12</sup> cardiovascular diseases,<sup>46</sup> and conformational diseases,<sup>47</sup> and it has been suggested that IUPs bind their partner in a special way, through a deep binding crevice amenable for interference with small-molecule inhibitors.<sup>45</sup> This latter feature has been demonstrated by developing inhibitors against the p53-MDM2 interaction.<sup>48</sup> It is not too far-fetched to suggest that a detailed understanding of the physical principles of the interactions of IUPs will open new possibilities to develop inhibitors against their PPIs, which hopefully will offer a wealth of opportunities for developing drugs to combat often lethal diseases.

## Data and Methods

### Databases

Disordered complexes were collected by identifying complexes in the Protein Data Bank (PDB) with experimental evidence of the disorder of one of the partners. The initial dataset was taken from the literature,<sup>27</sup> and was extended with further examples collected from the database of protein disorder, DisProt.<sup>34</sup> The disordered state was accepted if the protein showed at least 95% sequence identity with a protein found in DisProt, and at least 50% of its amino acids seen in the PDB structure were shown disordered in DisProt. In practical terms, disorder in these cases was almost always above 90%. Of the complexes that corresponded to these definitions, we discarded those in which the two interacting partners were actually part of the same SwissProt sequence, and any that contained a chimera protein. Further, we excluded ribosomal proteins, which were included in an earlier study.<sup>32</sup> Although parts of ribosomal proteins may lack a

well-defined structure in the absence of the partner, they interact with the highly charged RNA, and their interfaces must have rather special features. For the same reason, we excluded protein–DNA complexes. The database contains 39 independent complexes, presented in Table 1. Ordered complexes contained examples in which both partners had a well-defined structure when studied in isolation. These were taken from the literature.<sup>5,32,44</sup> Protein chimeras and fragments were discarded, which left 72 independent complexes, presented in Table 2. Different classes of complexes (homodimers, enzyme-inhibitor complexes, etc.) were not distinguished in this study.

### Calculation of surfaces

Accessible surfaces of proteins were calculated as described.<sup>49</sup> Polar and hydrophobic surfaces were defined as the accessible surfaces of polar and hydrophobic amino acids. Hydrophobicities were calculated by using the scale developed by Fauchere and Pliska.<sup>50</sup> On this basis, we considered N, Q, S, T, H, G, R, K, D and E as polar amino acids, and A, I, L, M, F, P, W, V, Y, and C as hydrophobic amino acids. Buried surface was defined as the difference between the standard surface and actual accessible surface of the protein. Standard surface was taken as the sum of the surfaces of its amino acids, as determined in an AXA sequential environment†. The interaction surface (interface) buried in a complex was defined as the difference between the surface area of the complex and the sum of the surface areas of the two separate protein subunits.

### Interactions, interaction energies and segmentation

Two atoms were considered in contact if the distances between their centers were less than the sum of their van der Waals radii plus 1 Å. Two amino acids were considered in contact if they had at least two heavy atoms in contact. Interaction energies within a folded chain and between interacting partners were approximated by low-resolution statistical amino acid contact potentials,<sup>39</sup> also used as the basis of the disorder prediction algorithm IUPred.<sup>40</sup> The potentials originate from the work of Thomas and Dill.<sup>51</sup> These energy-like quantities have been summed for all amino acid pairs in contact, either within a single chain or in the interface between the two interacting chains.

Part of a binding interface was considered to belong to a single segment of the parent protein if the respective amino acids were in contact with the partner and their distances in the polypeptide chain were not larger than five amino acid residues.<sup>5,30</sup> Segmentation was defined as the number of segments within the interface of a protein in the given complex.

### Evolutionary conservation

For each PDB entry in the database, the complete sequence of the protein was retrieved using the corresponding SwissProt/Uniprot entry.<sup>52</sup> In a few cases, the disordered part of the complex was composed of multiple chains and these were treated separately. Residue conservation was calculated at the level of the full protein sequence as described.<sup>53</sup> Homologous sequences were collected from the Uniref100 sequence database using Psi-

Blast,<sup>54</sup> with a cut-off value of 10e-30. From these sequences, a multiple alignment was created by CLUSTAL W.<sup>55</sup> For each position in the alignment, the conservation score was determined using the formula of Valdar01.<sup>53</sup> This method sums all possible pair-wise match scores between amino acids in an aligned column, weighted by the combination of the appropriate substitution matrix score and the sequence weight, which normalizes against the redundancy of sequences in the alignment. The scoring scheme penalizes gaps as well as mutations. The original algorithm was modified so that gaps longer than 50 residues were not taken into account, e.g. domain-sized deletions were not considered during the calculation of the conservation scores. The variation of conservation scores also depends on the number and diversity of the homologous sequences. To place the various sequence variations on the same platform, the conservation scores were transformed into the fractional rank in the alignment. This artificially stretches the distribution of conservation scores evenly between 0 and 1. The proteins that did not have a corresponding match in SwissProt, or the number of aligned sequences was below 20, were discarded. This reduced the number of proteins in the datasets to 31 for disordered and 39 for ordered protein complexes.

### Acknowledgements

This work was supported by grants GVOP-3.2.1.-2004-05-0195/3.0, Hungarian Scientific Research Fund (OTKA) T049073, K60694, ETT 245/2006 from the Hungarian Ministry of Health, NKFP MediChem2. This work was also supported by the scrIN-SILICO LSHP-CT-2005-012127 project within the EU FP6 program. The Wellcome Trust International Senior Research Fellowship ISRF 067595 for P. T. and Öveges grant from National Office for Research and Technology for I.S. are gratefully acknowledged.

### References

1. Aloy, P. & Russell, R. B. (2004). Ten thousand interactions for the molecular biologist. *Nature Biotechnol.* **22**, 1317–1321.
2. Aloy, P. & Russell, R. B. (2006). Structural systems biology: modelling protein interactions. *Nature Rev. Mol. Cell Biol.* **7**, 188–197.
3. Gavin, A. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M. *et al.* (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440**, 631–636.
4. Arifuzzaman, M., Maeda, M., Itoh, A., Nishikata, K., Takita, C., Saito, R. *et al.* (2006). Large-scale identification of protein-protein interaction of *Escherichia coli* K-12. *Genome Res.* **16**, 686–691.
5. Jones, S. & Thornton, J. M. (1996). Principles of protein-protein interactions. *Proc. Natl Acad. Sci. USA*, **93**, 13–20.
6. Lo Conte, L., Chothia, C. & Janin, J. (1999). The atomic structure of protein-protein recognition sites. *J. Mol. Biol.* **285**, 2177–2198.
7. Nooren, I. M. & Thornton, J. M. (2003). Diversity of protein-protein interactions. *EMBO J.* **22**, 3486–3492.

† <http://wolf.bms.umist.ac.uk/naccess/>

8. Keskin, O., Ma, B. & Nussinov, R. (2005). Hot regions in protein-protein interactions: the organization and contribution of structurally conserved hot spot residues. *J. Mol. Biol.* **345**, 1281–1294.
9. Dunker, A. K., Garner, E., Guillot, S., Romero, P., Albrecht, K., Hart, J. & Obradovic, Z. (1998). Protein disorder and the evolution of molecular recognition: theory, predictions and observations. *Pac. Symp. Biocomput.* **3**, 473–484.
10. Wright, P. E. & Dyson, H. J. (1999). Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* **293**, 321–331.
11. Uversky, V. N., Gillespie, J. R. & Fink, A. L. (2000). Why are “natively unfolded” proteins unstructured under physiologic conditions? *Proteins: Struct. Funct. Genet.* **41**, 415–427.
12. Iakoucheva, L., Brown, C., Lawson, J., Obradovic, Z. & Dunker, A. (2002). Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.* **323**, 573–584.
13. Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M. & Obradovic, Z. (2002). Intrinsic disorder and protein function. *Biochemistry*, **41**, 6573–6582.
14. Uversky, V. N., Oldfield, C. J. & Dunker, A. K. (2005). Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. *J. Mol. Recognit.* **18**, 343–384.
15. Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F. & Jones, D. T. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* **337**, 635–645.
16. Tompa, P., Dosztányi, Z. & Simon, I. (2006). Prevalent structural disorder in *E. coli* and *S. cerevisiae* proteomes. *J. Proteome Res.* **5**, 1996–2000.
17. Tompa, P. & Csermely, P. (2004). The role of structural disorder in the function of RNA and protein chaperones. *FASEB J.* **18**, 1169–1175.
18. Dyson, H. J. & Wright, P. E. (2002). Coupling of folding and binding for unstructured proteins. *Curr. Opin. Struct. Biol.* **12**, 54–60.
19. Domanski, M., Hertzog, M., Coutant, J., Gutsche-Perelroizen, I., Bontems, F., Carlier, M. F. *et al.* (2004). Coupling of folding and binding of thymosin  $\beta$ 4 upon interaction with monomeric actin monitored by nuclear magnetic resonance. *J. Biol. Chem.* **279**, 23637–23645.
20. Dyson, H. J. & Wright, P. E. (2005). Intrinsically unstructured proteins and their functions. *Nature Rev. Mol. Cell Biol.* **6**, 197–208.
21. Tompa, P. (2005). The interplay between structure and function in intrinsically unstructured proteins. *FEBS Letters*, **579**, 3346–3354.
22. Dunker, A. K., Cortese, M. S., Romero, P., Iakoucheva, L. M. & Uversky, V. N. (2005). Flexible nets. The roles of intrinsic disorder in protein interaction networks. *FEBS J.* **272**, 5129–5148.
23. Haynes, C., Oldfield, C. J., Ji, F., Klitgord, N., Cusick, M. E., Radivojac, P. *et al.* (2006). Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS Comput. Biol.* **2**, e100.
24. Patil, A. & Nakamura, H. (2006). Disordered domains and high surface charge confer hubs with the ability to interact with multiple proteins in interaction networks. *FEBS Letters*, **580**, 2041–2045.
25. Dosztányi, Z., Chen, J., Dunker, A. K., Simon, I. & Tompa, P. (2006). Disorder and sequence repeats in hub proteins and their implications for network evolution. *J. Proteome Res.* **5**, 2985–2995.
26. Csizsmok, V., Bokor, M., Banki, P., Klement, É., Medzihradzky, K. F., Friedrich, P. *et al.* (2005). Primary contact sites in intrinsically unstructured proteins: the case of calpastatin and microtubule-associated protein 2. *Biochemistry*, **44**, 3955–3964.
27. Fuxreiter, M., Simon, I., Friedrich, P. & Tompa, P. (2004). Prefolded structural elements feature in partner recognition by intrinsically unstructured proteins. *J. Mol. Biol.* **338**, 1015–1026.
28. Oldfield, C. J., Cheng, Y., Cortese, M. S., Romero, P., Uversky, V. N. & Dunker, A. K. (2005). Coupled folding and binding with  $\alpha$ -helix-forming molecular recognition elements. *Biochemistry*, **44**, 12454–12470.
29. Mohan, A., Oldfield, C. J., Radivojac, P., Vacic, V., Cortese, M. S., Dunker, A. K. & Uversky, V. N. (2006). Analysis of molecular recognition features (MoRFs). *J. Mol. Biol.* **362**, 1043–1059.
30. Fuxreiter, M., Tompa, P. & Simon, I. (2007). Local structural disorder imparts plasticity on linear motifs. *Bioinformatics*, **23**, 950–956.
31. Neduva, V. & Russell, R. B. (2005). Linear motifs: evolutionary interaction switches. *FEBS Letters*, **579**, 3342–3345.
32. Gunasekaran, K., Tsai, C. J. & Nussinov, R. (2004). Analysis of ordered and disordered protein complexes reveals structural features discriminating between stable and unstable monomers. *J. Mol. Biol.* **341**, 1327–1341.
33. Vacic, V., Oldfield, C. J., Mohan, A., Radivojac, P., Cortese, M. S., Uversky, V. N. & Dunker, A. K. (2007). Characterization of molecular recognition features, MoRFs, and their binding partners. *J. Proteome Res.* **6**, 2351–2366.
34. Sickmeier, M., Hamilton, J. A., LeGall, T., Vacic, V., Cortese, M. S., Tantos, A. *et al.* (2007). DisProt: the database of disordered proteins. *Nucl. Acids Res.* **35**, D786–D793.
35. Russo, A. A., Jeffrey, P. D., Patten, A. K., Massague, J. & Pavletich, N. P. (1996). Crystal structure of the p27<sup>Kip1</sup> cyclin-dependent-kinase inhibitor bound to the cyclin A-Cdk2 complex. *Nature*, **382**, 325–331.
36. Huber, A. H. & Weis, W. I. (2001). The structure of the  $\beta$ -catenin/E-cadherin complex and the molecular basis of diverse ligand recognition by  $\beta$ -catenin. *Cell*, **105**, 391–402.
37. Hertzog, M., van Heijenoort, C., Didry, D., Gaudier, M., Coutant, J., Gigant, B. *et al.* (2004). The  $\beta$ -thymosin/WH2 domain; structural basis for the switch from inhibition to promotion of actin assembly. *Cell*, **117**, 611–623.
38. Sorenson, M. K., Ray, S. S. & Darst, S. A. (2004). Crystal structure of the flagellar  $\sigma$ /anti- $\sigma$  complex  $\sigma^{28}$ /FlgM reveals an intact sigma factor in an inactive conformation. *Mol. Cell*, **14**, 127–138.
39. Dosztányi, Z., Csizsmok, V., Tompa, P. & Simon, I. (2005). The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.* **347**, 827–839.
40. Dosztányi, Z., Csizsmok, V., Tompa, P. & Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, **21**, 3433–3434.
41. Tompa, P. (2002). Intrinsically unstructured proteins. *Trends Biochem. Sci.* **27**, 527–533.
42. Gunasekaran, K., Tsai, C. J., Kumar, S., Zanuy, D. & Nussinov, R. (2003). Extended disordered proteins: targeting function with less scaffold. *Trends Biochem. Sci.* **28**, 81–85.

43. Tsai, C. J., Xu, D. & Nussinov, R. (1997). Structural motifs at protein-protein interfaces: protein cores versus two-state and three-state model complexes. *Protein. Sci.* **6**, 1793–1805.
44. Rajamani, D., Thiel, S., Vajda, S. & Camacho, C. J. (2004). Anchor residues in protein-protein interactions. *Proc. Natl Acad. Sci. USA*, **101**, 11287–11292.
45. Cheng, Y., Legall, T., Oldfield, C. J., Mueller, J. P., Van, Y. Y., Romero, P. *et al.* (2006). Rational drug design via intrinsically disordered protein. *Trends Biotechnol.* **24**, 435–442.
46. Cheng, Y., LeGall, T., Oldfield, C. J., Dunker, A. K. & Uversky, V. N. (2006). Abundance of intrinsic disorder in protein associated with cardiovascular disease. *Biochemistry*, **45**, 10448–10460.
47. Galzitskaya, O. V., Garbuzynskiy, S. O. & Lobanov, M. Y. (2006). Prediction of amyloidogenic and disordered regions in protein chains. *PLoS Comput. Biol.* **2**, e177.
48. Vassilev, L. T., Vu, B. T., Graves, B., Carvajal, D., Podlaski, F., Filipovic, Z. *et al.* (2004). In vivo activation of the p53 pathway by small-molecule antagonists of MDM2. *Science*, **303**, 844–848.
49. Lee, B. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379–400.
50. Fauchere, J. L. & Pliska, V. (1983). Hydrophobic parameters  $\pi$  of amino-acid side chains from the partitioning of N-acetyl-amino-acid amides. *Eur. J. Med. Chem.* **18**, 369–375.
51. Thomas, P. D. & Dill, K. A. (1996). An iterative method for extracting energy-like quantities from protein structures. *Proc. Natl Acad. Sci. USA*, **93**, 11628–11633.
52. Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, H. C. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, **23**, 1282–1288.
53. Valdar, W. S. & Thornton, J. M. (2001). Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Proteins: Struct. Funct. Genet.* **42**, 108–124.
54. Schaffer, A. A., Aravind, L., Madden, T. L., Shavirin, S., Spouge, J. L., Wolf, Y. I. *et al.* (2001). Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucl. Acids Res.* **29**, 2994–3005.
55. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids. Res.* **22**, 4673–4680.

*Edited by M. Sternberg*

(Received 16 May 2007; received in revised form 29 June 2007; accepted 2 July 2007)  
Available online 12 July 2007



# Prediction of Protein Binding Regions in Disordered Proteins

Bálint Mészáros, István Simon, Zsuzsanna Dosztányi\*

Institute of Enzymology, Biological Research Center, Hungarian Academy of Sciences, Budapest, Hungary

## Abstract

Many disordered proteins function via binding to a structured partner and undergo a disorder-to-order transition. The coupled folding and binding can confer several functional advantages such as the precise control of binding specificity without increased affinity. Additionally, the inherent flexibility allows the binding site to adopt various conformations and to bind to multiple partners. These features explain the prevalence of such binding elements in signaling and regulatory processes. In this work, we report ANCHOR, a method for the prediction of disordered binding regions. ANCHOR relies on the pairwise energy estimation approach that is the basis of IUPred, a previous general disorder prediction method. In order to predict disordered binding regions, we seek to identify segments that are in disordered regions, cannot form enough favorable intrachain interactions to fold on their own, and are likely to gain stabilizing energy by interacting with a globular protein partner. The performance of ANCHOR was found to be largely independent from the amino acid composition and adopted secondary structure. Longer binding sites generally were predicted to be segmented, in agreement with available experimentally characterized examples. Scanning several hundred proteomes showed that the occurrence of disordered binding sites increased with the complexity of the organisms even compared to disordered regions in general. Furthermore, the length distribution of binding sites was different from disordered protein regions in general and was dominated by shorter segments. These results underline the importance of disordered proteins and protein segments in establishing new binding regions. Due to their specific biophysical properties, disordered binding sites generally carry a robust sequence signal, and this signal is efficiently captured by our method. Through its generality, ANCHOR opens new ways to study the essential functional sites of disordered proteins.

**Citation:** Mészáros B, Simon I, Dosztányi Z (2009) Prediction of Protein Binding Regions in Disordered Proteins. *PLoS Comput Biol* 5(5): e1000376. doi:10.1371/journal.pcbi.1000376

**Editor:** Rita Casadio, University of Bologna, Italy

**Received:** December 11, 2008; **Accepted:** March 30, 2009; **Published:** May 1, 2009

**Copyright:** © 2009 Mészáros et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** This work was supported by grants GVOP-3.2.1.-2004-05-0195/3.0 and NKTH07a-TB\_INTER from the National Office for Research and Technology (NKTH) and K72569 from the Hungarian Scientific Research Fund (OTKA). The work was also supported by grant 2R01-LM07329-01 from the National Library of Medicine. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

\* E-mail: zsuzsa@enzim.hu

## Introduction

The classical point of view on protein function claims that the functionality of a protein requires the presence of a well-defined three dimensional structure. However, as the amount of experimental evidence against the generality of this concept grew, this paradigm had to be reassessed [1]. It has become evident that there is a large number of proteins that do not require a stable structure even under physiological conditions in order to fulfill their biological role [2–4]. These intrinsically unstructured/disordered proteins (IUPs/IDPs) lack a well defined tertiary structure and exhibit a multitude of conformations that dynamically change over time and population. The importance of protein disorder is underlined by the abundance of partially or fully disordered proteins encoded in higher eukaryotic genomes [5,6]. Disordered proteins are involved in many important biological functions [2,7], which complement the functional repertoire of globular proteins [7]. Recent characterization of IUPs based on their functions shows that disorder can help these proteins to fulfill their functions in various ways [8,9]. In the case of entropic chains, the biological function is directly mediated by disorder (e.g. MAP2 projection domain [10], titin's PEVK domain [11], NF-M and NF-H between neurofilaments [12,13], nucleoporin complex

[14]). Furthermore, disordered segments often act as flexible linkers between folded domains in multidomain proteins [2,15]. Alternatively, many disordered proteins function by binding specifically to other proteins, DNA or RNA. This process, termed coupled folding and binding involves a transition from disordered state to a more ordered state with stable secondary and tertiary structural elements [16,17].

The coupled folding and binding confers several functional advantages in certain types of molecular interactions. Since – at least partial – folding happens together with binding, the entropic penalty counterbalances the enthalpy gain coming from the binding [18,19]. This way disorder uncouples specificity from binding strength allowing for weak transient, still specific interactions that are essential for signaling processes. These properties enable disordered proteins to play an important role in molecular recognition including gene regulation, cell cycle control and other key cellular processes [20–23]. The kinetic and thermodynamic details of the binding are influenced by conformational preferences present prior to binding [24]. Although disordered proteins in general lack secondary and tertiary structure, some exhibit partial secondary structure at closer inspection. For example, CD analysis indicated that p21 and p27 possess  $\alpha$ -helical segments [19,25,26]. Detailed NMR

## Author Summary

Intrinsically unstructured/disordered proteins (IUPs/IDPs) do not adopt a stable structure in isolation but exist as a highly flexible ensemble of conformations. Despite the lack of a well-defined structure these proteins carry out important functions. Many IUPs/IDPs function via binding specifically to other macromolecules that involves a disorder-to-order transition. The molecular recognition functions of IUPs/IDPs include regulatory and signaling interactions where binding to multiple partners and high-specificity/low-affinity interactions play a crucial role. Due to their specific functional and structural properties, these binding regions have distinct properties compared to both globular proteins and disordered regions in general. Here, we present a general method to identify disordered binding regions from the amino acid sequence. Our method targets the essential feature of these regions: they behave in a characteristically different manner in isolation than bound to their partner protein. This prediction method allows us to compare the binding properties of short and long binding sites. The evolutionary relationship between the amount of disordered binding regions and general disordered regions in various organisms was also analyzed. Our results suggest that disordered binding regions can be recognized even without taking into account their adopted secondary structure or their specific binding partner.

characterization of p27 and other proteins showed that several segments can have a pronounced tendency to adopt  $\alpha$ -helical, or even  $\beta$  strand conformations [9]. Upon binding, these inherent structural preferences can either be solidified or overwritten by the partner molecule [27]. Some regions can preserve flexibility even within the complex, mitigating the unfavorable entropy term [28]. This allows the fine-tuning of the affinity of interactions over a wide range. As a general rule, however, these interactions are driven largely enthalpically by the favorable interactions formed with the partner molecule [18,19,29].

The inherent flexibility of disordered proteins offers further advantages in binding. It results in a malleable interface that can allow binding to several partners or to adopt different conformations, manifested in increased binding capability [8,20]. In accordance, several analyses of protein interaction networks revealed that disordered proteins are abundant among hub proteins, proteins with a large number of interacting partners [30,31]. In a different scenario, the binding partners of an ordered protein are disordered, as shown for binding of 14-3-3 proteins, thus allowing a single protein to bind multiple partners [32]. Beside their involvement in protein-protein interactions, these proteins are also subjects of various post-translational modifications that control their functions, localization and turnover [33]. In this way, these proteins can integrate and mediate multiple signals of various sources, and act as the central elements in signaling or regulatory networks. The centrality of these proteins, however, is also their weakness. It has been suggested that the targeted attack of hubs can cause serious disruption in protein interaction networks [34]. Furthermore, disordered proteins are often associated with various diseases [35]. For example, the primary importance of p53 originates from its involvement of 50% of cancers [36]. In general, 79% of human cancer associated proteins have been classified as IUPs, compared to 47% of all eukaryotic proteins in SwissProt database [22]. Disordered proteins were also suggested to be common in diabetes and cardiovascular diseases [35,37]. Several disordered proteins - such as A $\beta$ ,  $\tau$ ,  $\alpha$  synuclein,

and prion protein - are involved in neurodegenerative diseases and are also prone to amyloid formation [38–40]. On the other hand, due to their specific way of interactions, disordered proteins can also be attractive targets for drug discovery. A novel strategy for drug discovery exploiting binding sites within disordered regions has already been suggested [41]. This adds further support to the importance of finding specific functional sites in proteins that undergo disorder-to-order transition upon binding or *disordered binding regions* in short.

Despite their importance, the number of well characterized examples of disordered proteins undergoing disorder-to-order transition is very small. The PDB also offers only a limited sample of proteins adopting a well defined conformation as part of a complex. However, recent comparisons of these structures with complexes formed between ordered proteins pointed out several differences [42–44]. In general, disordered proteins adopted a largely extended conformation in the complex exposing the majority of their residues for interacting with their partner. The interface of disordered proteins was enriched in hydrophobic residues compared to the interface of ordered proteins, but also to disordered regions in general. The higher number of interchain contacts was suggested to be a sign of better adaptation of disordered proteins to the surface of their partner. In general, the regions that become ordered were shorter as compared to globular domains, usually less than 30–40 residues. While the interface of globular proteins was most often formed by distant segments of the amino acid sequence brought together by folding, disordered binding sites were much more localized in the primary structure. These features demonstrate that the underlying principles of molecular recognition of disordered binding regions are different from the complex formation of globular proteins [43].

Disordered binding sites are also expected to be distinguishable from general disordered sites that are not directly involved in binding. A common notion is that protein disorder comes in many flavors, and these should be targeted by specific prediction methods [45,46]. However, training specific methods would require significantly larger datasets than those that are available today. Nevertheless, existing general protein disorder prediction methods might already be equipped for this problem. It has been suggested that specific patterns of disorder prediction profiles can be associated with regions undergoing disorder-to-order transitions [47]. Since these regions can be ordered as well as disordered, there is no clear recipe whether these regions should be predicted ordered, disordered, or as borderline cases. A recent analysis compared several methods to recognize short protein-protein interaction motifs containing  $\alpha$ -helical elements in their bound state, the so-called  $\alpha$ -MoRFs [48]. As expected, the various methods showed large variations in predicted order/disorder tendency corresponding to binding regions. One of the earliest prediction method PONDR VL-XT [49–51] was quite consistent in predicting these regions as ordered within a broader disordered region, giving them the characteristic appearance of dips in the prediction output. Based on this specific prediction output, a method was developed to recognize  $\alpha$ -MoRFs from the amino acid sequence [48,52]. First, regions predicted with dips in the output of VL-XT were selected and were filtered further by a neural network using several additional properties. This prediction method is restricted to recognize short,  $\alpha$ -helical binding regions within disordered proteins.

Here we present a general method to identify specific binding regions undergoing disorder-to-order transition. Our method relies on the general disorder prediction method IUPred [53,54]. IUPred is based on the assumption that disordered proteins have a specific amino acid composition that does not allow the formation

of a stable well-defined structure. The method utilizes statistical potentials that can be used to calculate the pairwise interaction energy from known coordinates. Using a dataset of globular proteins only, a method was developed to estimate the pairwise interaction energy of proteins directly from the amino acid sequence. By virtue of this algorithm, disordered residues can be predicted by having unfavorable estimated pairwise energies. The estimation of the energy for each residue is based on its amino acid type, and the amino acid composition of its sequential neighborhood. Through the amino acid composition of the sequential environment, IUPred can take into account that the disorder tendency of residues can be modulated by their environment [53]. This property of IUPred is exploited in order to recognize regions that are most likely to undergo a disorder-to-order transition based on their estimated pairwise energies in different contexts. The prediction of binding sites is based on estimating the energy content in free and in the bound states, and identifying segments that are potentially sensitive to these changes. In a previous work, the ability to predict specific contacts was emphasized in order to recognize disordered regions that are involved in binding externally rather than internally [46]. In our model, however, there was no attempt made to model specific interactions. Instead, the environment is taken into account simply at the level of amino acid composition. Here we show that this simple model captures the essential property of disordered binding regions and allows their robust prediction. We termed our disordered binding site prediction method ANCHOR, to reflect the primary importance of short segments driving the complex formation between a disordered protein and its partner.

## Results

### The outline of the algorithm

The goal of the present work was to recognize a special class of disordered segments from the amino acid sequence, namely those that are capable of undergoing a disorder-to-order transition upon binding to a globular protein partner. The essential feature of such binding regions is that they behave in a characteristically different manner in isolation than bound to their partner protein. In their free state, they behave as disordered proteins, existing as a highly flexible structural ensemble. In their bound state they usually adopt a rigid conformation, similar to regions within globular structures. This capability to behave in drastically different ways in different environments is targeted by our approach. We seek to identify segments in a generally disordered region that cannot form enough favorable intrachain interactions, however they have the capability to energetically gain by interacting with a globular partner protein. Our prediction is based on three properties.

1. The first criterion ensures that a given residue belongs to a long disordered region, and filters out globular domains.
2. The second criterion corresponds to the isolated state and it ensures that a residue is not able to form enough favorable contacts with its own local sequential neighbors to fold, otherwise it would be prone to adopt a well defined structure on its own.
3. The third criterion tests the feasibility that a given residue can form enough favorable interactions with globular proteins upon binding. This basically ensures that there is an energy gain by interacting with globular regions.

These properties are estimated individually and are combined into a single predictor via optimized weights.

In more detail, the prediction of these three properties relies on the energy estimation framework implemented in IUPred, a general disorder prediction method. The core element of IUPred

is the energy predictor matrix  $\mathbf{P}$ . The parameters in  $P_{ij}$  were trained on globular proteins with known structures only, without relying on any kind of disordered dataset. These parameters were determined to minimize the difference between the estimated energies and the energies calculated from the known structures on the dataset of globular proteins. Using the energy predictor matrix IUPred predicts the  $E$  interaction energy for each residue based on the following formula in default:

$$E_i^k = \sum_{j=1}^{20} P_{ij} f_j^k(w_0) \quad (1)$$

where  $i$  denotes the type of the  $k$ -th amino acid,  $P_{ij}$  is the element of the energy predictor matrix that estimates the pairwise energy of residue of type  $i$  in the presence of residue type  $j$ ,  $f_j^k(w_0)$  is the fraction of residue type  $j$  in the sequential environment within  $w_0$  residues from residue  $k$ . The size of neighborhood considered ( $w_0$ ) equals 100 residues in both directions and the result is smoothed over a window size of 10 (also in both directions from the  $k$ -th residue so in fact 21 residues are considered in total). For the final prediction output, the energies are transformed into probability values, denoted as  $s_k$ . For more details see Dosztányi et al. [53].

The disordered binding site prediction is based on three different scores that are calculated with a slight modification of the original energy estimation scheme. The parameters of  $P_{ij}$  were taken directly from IUPred. The following three scores are assigned to each residue in a protein according to the above described criteria (1–3):

1, To measure the tendency of the neighborhood of an amino acid for being disordered we use the IUPred algorithm and assign an  $S_k$  score to the  $k$ -th residue of the chain by averaging the IUPred scores in the  $w_l$  neighborhood of the residue in question:

$$S_k = \frac{1}{N} \sum_{k \neq j = b_{lower}}^{b_{upper}} s_j \quad (2)$$

where  $s_j$  is the IUPred score of the  $j$ -th residue of the chain,  $N$  is the number of amino acids in the averaging and  $b_{lower}$  and  $b_{upper}$  are the lower and upper boundaries of the neighborhood of the  $i$ -th residue, that is  $b_{lower} = \max(k - w_l; 1)$  and  $b_{upper} = \min(k + w_l; l)$ , where  $l$  is the chain length.

2, We estimate the pairwise interaction energy the given residue may gain by forming intrachain contacts. This is done the exact same way as in IUPred using (1), only here the size of the considered neighborhood ( $w_2$ ) is left as a parameter and is set during the training of the predictor:

$$E_i^{int,k} = \sum_{j=1}^{20} P_{ij} f_j^k(w_2) \quad (3)$$

The smaller window size corresponds to more local behavior.

3, The pairwise energy that the residue may gain by interacting with a globular protein is approximated using the average amino acid composition of globular proteins:

$$E_i^{glob} = \sum_{j=1}^{20} P_{ij} \bar{f}_{glob,j} \quad (4)$$

where  $\bar{f}_{glob,j}$  is the fraction of residue type  $j$  in the averaged reference amino acid composition of globular proteins shown in

**Table 1.** Reference amino acid composition of globular proteins.

AA	F %
R	3.68
K	6.37
D	4.92
E	5.43
N	4.69
Q	3.86
S	8.05
G	8.46
H	2.00
T	6.35
A	7.67
P	4.89
Y	3.86
V	7.13
M	1.84
C	2.43
L	8.22
F	3.19
I	5.20
W	1.76

Amino acid composition of the reference globular protein dataset comprised of all the amino acids in the longer chains of the ordered complexes dataset. Amino acids are sorted by increasing hydrophobicity based on the Fauchere-Pliska hydrophobicity scale [94]. AA denotes amino acid codes and  $f$  denotes the fraction of the respective amino acid expressed as a percentage.  
doi:10.1371/journal.pcbi.1000376.t001

Table 1. By subtracting this energy from  $E_i^{\text{int},k}$  one can estimate the energy that the residue may gain by interacting with a hypothetical globular protein compared to forming intrachain contacts ( $E_i^{\text{gain},k} = E_i^{\text{int},k} - E_i^{\text{glob}}$ ).

The final prediction score of the residue is given by the linear combination of the above three terms:

$$I_k = p_1 S_k + p_2 E_i^{\text{int},k} + p_3 E_i^{\text{gain},k} \quad (5)$$

where the  $p_1$ ,  $p_2$  and  $p_3$  coefficients are determined during the training of the predictor together with the optimal values of  $w_1$  and  $w_2$  window sizes.  $I_k$  is then converted into a  $p$  value that expresses the probability of that residue being in a disordered binding site. For a binary classification residues with scores above 0.5 are predicted to be in a disordered binding site. Since the second and third terms of (5) may vary heavily between neighboring residues, the final score is smoothed in a window of 4 residues.

The optimal values for the three weights ( $p_1$ ,  $p_2$  and  $p_3$ ) and the two window sizes ( $w_1$  and  $w_2$ ) are determined using a dataset of disordered protein complexes and ordered monomeric proteins by three-fold cross validation (See Methods and Figure S1 for a schematic representation and outline of this procedure). The small dataset of known disordered proteins bound to ordered proteins represent a serious bottleneck during optimization. Therefore, it is a clear advantage of our approach that it greatly reduces the dependence on the existing dataset of disordered complexes, and leaves us with only 5 parameters to be optimized on this small dataset.

The behavior of various scores is shown for an example, the N terminal domain (residues 1–100) of human p53 tumor suppressor protein that plays an important regulatory role [55]. Its N terminal region is completely disordered [56] and is known to be able to bind to (at least) three different globular proteins as shown in Figure 1. The segment between residues 17–27 binds to MDM2 [57], the other two binding sites overlap with residues 33–56 binding to RPA 70N [58] and residues 45–58 binding to the B subunit of RNA polymerase II [59]. The three calculated quantities for this domain are also shown in Figure 1. It is worth noting that the MDM2 binding site in the N-terminal region of p53 appears to be on the border of being disordered. Although the disordered prediction is part of ANCHOR, the output of this prediction ( $E_{\text{int}}$ , described in Theory) is linearly combined with two other quantities meaning that predicted disorder is not strictly a prerequisite of a successful disordered binding site prediction.

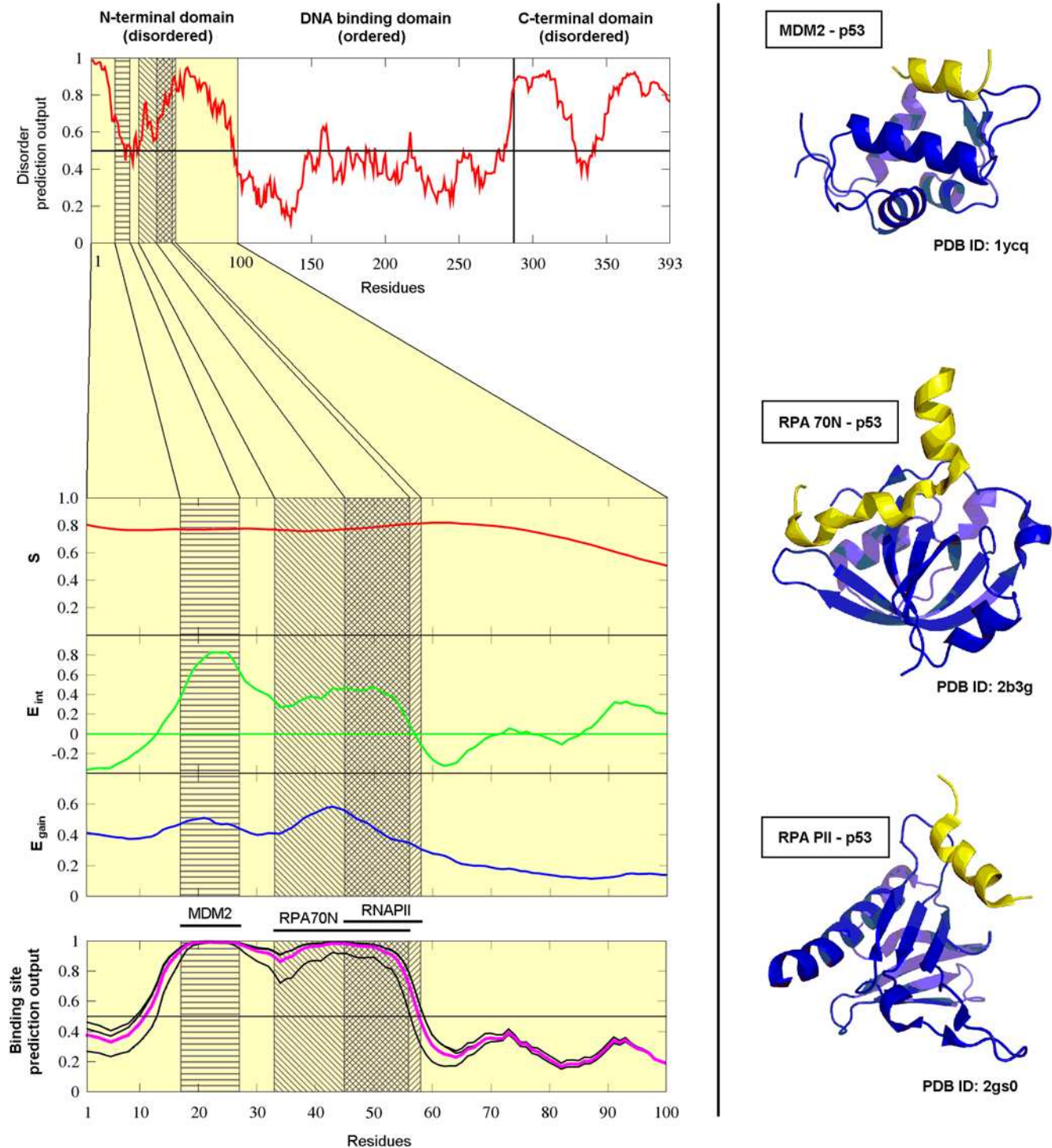
### Testing of the algorithm

Testing of the predictor was done by dividing both our negative and positive datasets (*Globular proteins* and *Short disordered complexes*) into three subsets, training the predictor on two of these and evaluating it on the remaining third one. This was done in all three possible combinations yielding three optimal parameter sets. The parameters calculated on the training sets are shown in Table 2 together with the respective True Positive Rates (TPR) and the fraction of the amino acids in disordered regions of the Disprot dataset predicted to be in disordered binding sites (F values). The optimal parameters were chosen to maximize the amount of correctly predicted disordered binding sites (TPR) while minimizing predicted binding sites in globular proteins (FPR) and also restricting predicted binding sites within disordered regions in general (F). The fact that the three parameter sets do not differ significantly implies that our method is robust.

The output of the predictor with all three parameter sets and the combined final predictor (the average of these three) are shown for the example of the N terminal region of p53 in Figure 1. A few additional well characterized examples are shown in the Supporting Information (Figure S2, Figure S3, Figure S4, Figure S5, and Figure S6).

The results obtained on the three independent testing subsets as well as their average are given in Table 3. Since the cutoffs are given by the training process such that we achieve exactly 5% False Positive Rate (FPR) on the respective training sets (ie. the part of the original Globular proteins dataset that was used in the training of the respective subpredictor), the FPR's are also quoted (they can differ slightly from 5%). Besides the overall TPR calculated on a residue basis (marked  $TPR_{AA}$ ), we also calculated the percentage of binding sites identified, termed  $TPR_{SEG}$ . A binding site was considered to be found if at least five of its amino acids are correctly classified. The results show that ANCHOR performs at 62%  $TPR_{AA}$  with a slightly higher  $TPR_{SEG}$  of 68% on average, while maintaining a 5% FPR. ANCHOR is also specific to disordered binding sites as opposed to disorder to general. If all disordered proteins had approximately equal capability of binding then the fraction of correctly identified disordered binding sites (TPR) could not be significantly different from the fraction of disordered regions predicted to be binding sites (F value). As this is not the case (TPR = 62% vs. F = 42%) we can conclude that common features of known disordered binding sites that distinguish them from general disordered protein regions are successfully recognized.

Another standard way of describing prediction algorithms is by Receiver Operating Characteristic (ROC) curves [60], that is the TPR versus the FPR of the algorithm. This relationship is mapped



**Figure 1. The construction of the ANCHOR prediction method demonstrated on the N-terminal domain of human p53.** *Left:* IUPred prediction score for the full length human p53 (top) and  $S$ ,  $E_{int}$  and  $E_{gain}$  calculated for the disordered N terminal domain of human p53 (middle). Grey boxes show the three binding sites with the overlap of the RPA70N and RNAPII binding sites shown in dark grey. The outputs of the three individually optimized predictors are shown in black and their average, the final prediction score is shown in purple (bottom). *Right:* PDB structures of the binding sites in the N-terminal region of p53 (yellow) complexed with the respective partners (blue): MDM2 (top, PDB ID: 1ycq [57]), RPA 70N (middle, PDB ID: 2b3g [58]) and RNA PII (bottom, PDB ID: 2gs0 [59]). doi:10.1371/journal.pcbi.1000376.g001

by scanning the interval between 0 and 1 with the score cutoff. The three ROC curves of the predictor with the three different parameter sets evaluated on the respective testing sets are shown in Figure 2. A single number measure to characterize the

performance is the area under the curve (AUC) with random predictors scoring  $AUC = 0.5$  and perfect predictors scoring  $AUC = 1$ . The AUC values of the predictors trained and tested on the respective subsets are 0.8675, 0.8781 and 0.8993.

**Table 2.** Parameter and prediction accuracy values obtained during the optimization of ANCHOR.

	$w_1$	$w_2$	$p_1$	$p_2$	$p_3$	$F$ (%)	$TPR$ (%)	$FPR$ (%)
Training set 1	25	60	0.4630	0.3847	0.7985	46.0	69.8	5.0
Training set 2	27	60	0.6075	0.4149	0.6773	47.4	67.7	5.0
Training set 3	29	90	0.6990	0.4585	0.5488	43.4	64.8	5.0

Optimal parameters of the predictor determined during training.  $w_1$ ,  $w_2$ ,  $p_1$ ,  $p_2$  and  $p_3$  are the optimized parameters,  $F$  is the fraction of the residues in the disordered regions in the Disprot database that are predicted to be in binding sites,  $TPR$  and  $FPR$  are the True- and False Positive Rates, respectively.

doi:10.1371/journal.pcbi.1000376.t002

Since the interacting regions of a disordered and an ordered protein are inherently different we expect that the predictor will only recognize binding sites in disordered proteins that interact with globular proteins but are not part of globular proteins themselves. In order to verify this hypothesis we tested the combined final predictor on a dataset of complexes containing only ordered chains (that is three-state complexes – see Methods). The prediction was done on the short interacting chain of the complexes. This gave a false positive rate of only 3.7% that is even lower than the value obtained on our testing set, although this might be only a consequence of the relatively small size of our ordered complex set (72 complexes). Overall, we could ensure that our predictor makes very few mistakes on both globular proteins and complexes of globular proteins, while it can still recognize the majority of disordered binding regions. This implies that our algorithm is specific to disordered binding sites as opposed to globular proteins, the interface between globular proteins or disordered proteins in general.

Our predictor was also tested on a completely independent dataset of  $\alpha$ -MoRFs, short disordered complexes that was assembled by Cheng et al. [48] and composed of 40 proteins containing binding regions that adopt mostly  $\alpha$ -helical structure upon binding. The results of the prediction on this dataset can be seen in Table 4. Although the residue based TPR is somewhat lower than that calculated on our testing set (57.0% instead of 61.8%), the segment based TPR is almost the same for the two sets (67.5% and 68.3%). Overall these results are comparable to the ones calculated on our training set.

### Amino acid based evaluation of the predictor

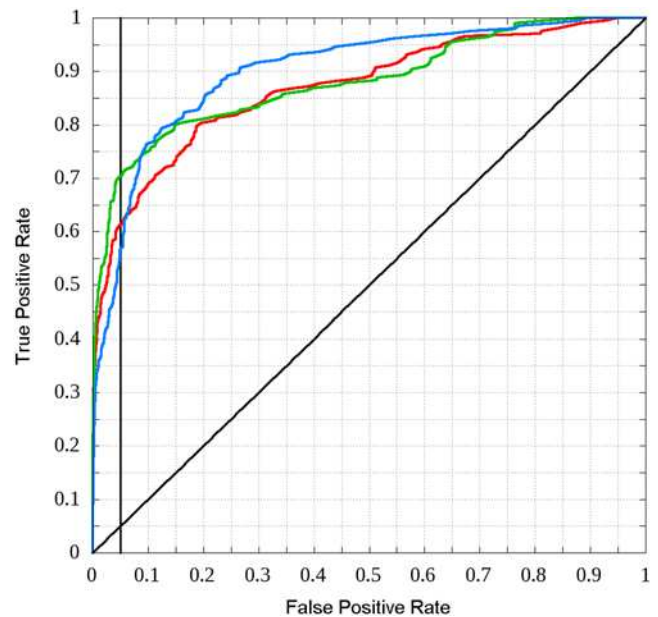
The specific construction of the algorithm for the prediction of interaction energy implies that the method will be sensitive to amino acid compositions. The differences between the composi-

**Table 3.** Prediction efficiency of ANCHOR evaluated on the testing datasets.

	$TPR_{AA}$ (%)	$TPR_{SEG}$ (%)	$FPR$ (%)
Testing set 1	61.1	62.5	5.7
Testing set 2	69.5	80.0	4.4
Testing set 3	54.7	62.5	5.1
Average	61.8	68.3	5.1

Results of the testing of ANCHOR on the three testing datasets.  $TPR_{AA}$  denotes the ratio of correctly identified amino acids belonging to binding sites.  $TPR_{SEG}$  denotes the ratio of binding sites found by the algorithm.

doi:10.1371/journal.pcbi.1000376.t003

**Figure 2. ROC curves obtained during the testing of ANCHOR.** ROC curves of the predictor with parameter sets optimized on each of the three training subsets and evaluated on the respective testing subsets are shown with red, green and blue lines. The line with unity slope corresponding to random prediction is also shown. The vertical line corresponds to  $FPR = 0.05$ , where the final predictor (the average of these three) is used.

doi:10.1371/journal.pcbi.1000376.g002

tion of disordered binding sites and the amino acid composition of any of the negative sets (globular proteins, ordered interfaces and disordered proteins in general) are shown in Figure 3A, 3B, and 3C, respectively. The amino acid compositions of all three datasets are significantly different from that of disordered binding segments (data not shown).

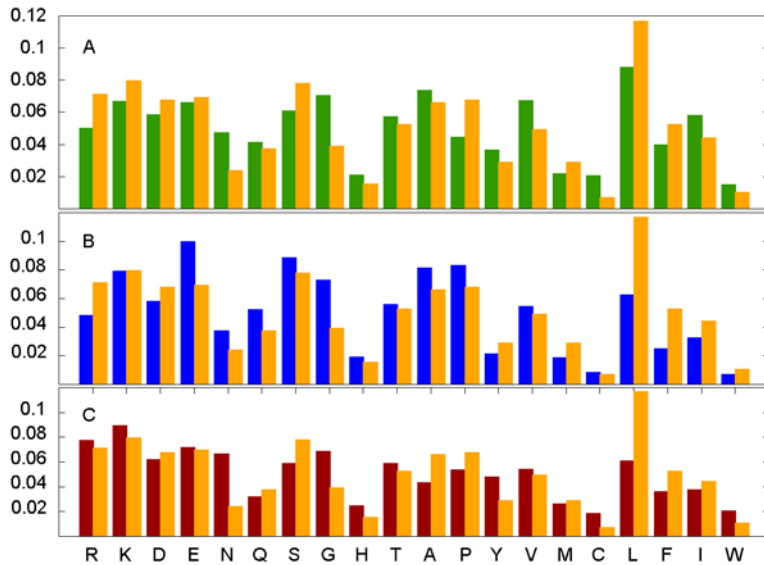
The final prediction is based on three different scores that combine local and global disorder tendency with sensitivity to the structural environment. Although the individual quantities that are combined for the final score can work selectively better or worse for various types of residues, the effect of these differences on the efficiency of the final prediction is not trivial. This effect was tested by comparing the amount of the different amino acids in the short disordered binding sites to the amount recovered from these by the predictor. These data are shown in Table 5 together with the calculated  $p$  values quantifying their differences. As all of the  $p$  values are fairly large, these differences are likely to occur by chance alone. For example, proline rich binding sites are found

**Table 4.** Prediction efficiency of ANCHOR evaluated on an independent dataset ( $\alpha$ -MoRFs dataset).

	H	E	C	Total	SEG
In dataset	263	8	210	479	40
Found	147	5	121	273	27
Ratio (TPR)	55.9%	62.5%	57.6%	57.0%	67.5%

Prediction results for the  $\alpha$ -MoRFs dataset. SEG denotes segment based results where each binding site is considered one segment and one such segment is considered found if at least five of its amino acids are correctly identified.

doi:10.1371/journal.pcbi.1000376.t004



**Figure 3. The distinct amino acid composition of short disordered binding sites.** The average amino acid composition of the interacting parts of the short disordered binding sites compared to the average amino acid composition of (A) the globular proteins dataset, (B) the disordered proteins dataset and (C) the interacting parts of the shorter chains of the ordered complexes. Amino acids are arranged according to increasing hydrophobicity.

doi:10.1371/journal.pcbi.1000376.g003

**Table 5.** The independence of the efficiency of ANCHOR from the amino acid composition of the binding sites.

AA	$N_{int}$	$N_{found}$	$p$
R	42	21	0.122
K	47	36	0.362
D	40	27	1.000
E	41	20	0.116
N	14	6	0.252
Q	22	11	0.358
S	46	34	0.497
G	23	14	0.758
H	9	7	1.000
T	31	20	1.000
A	39	33	0.068
P	40	19	0.113
Y	17	11	1.000
V	29	20	1.000
M	17	16	0.085
C	4	2	1.000
L	69	47	0.857
F	26	19	0.764
I	31	26	0.146
W	6	5	1.000

$N_{int}$  shows the number of interacting residues in the short disordered binding sites,  $N_{found}$  shows the amount of these that are correctly found by the predictor. As there are types of amino acids that are rare, Fisher's exact test was used to calculate (two-tailed)  $p$  values to determine if the predictor works significantly better or worse for certain amino acid types with high  $p$  values corresponding to no significant difference.

doi:10.1371/journal.pcbi.1000376.t005

with similar accuracy as binding sites enriched in hydrophobic amino acids. Therefore, one may conclude that there is no statistical evidence based on the available dataset that the efficiency of the predictor depends significantly on the amino acid composition of the disordered binding site in question.

### Secondary structures and the efficiency of ANCHOR

The relationship between the efficiency of the prediction and the secondary structure types was also assessed, by considering the three types of secondary structural elements: helix (H, including  $\alpha$ - and  $3_{10}$  helices), extended (E) and coil (C, including everything else) as defined by DSSP [61]. The number of amino acids in different conformations that can be found in the PDB structures of our positive training set (short disordered complexes), in the interacting residues of these structures and the interacting residues that are correctly identified by the predictor are shown in Table 6. These data are represented graphically as distributions in Figure 4. The secondary structure content in this type of interactions is heavily biased towards coil conformation. It can also be seen on Figure 4 that the predictor seems to work slightly better for H and E conformations. However assessing the difference of the distributions of secondary structures in interacting residues and in the subset identified correctly by ANCHOR shows that this difference is not statistically significant at a 5% level ( $\chi^2 = 5.32$ ,  $p = 0.070$ ).

Furthermore, a similar result holds true if binding sites are categorized based on their dominant secondary structure type - that is there is no significant correlation between the secondary structure type the binding regions adopt upon binding and the efficiency of the predictor. (Dataset S1 shows the secondary structure types determined for the short disordered chains in the disordered complexes as described in Protocol S1.) Overall, this means that there is no significant difference in the efficiency of the prediction on different secondary structural elements.

### Testing on long disordered regions

Since the predictor was trained on the short disordered dataset it is informative to see how it performs on long disordered binding

**Table 6.** Secondary structure distributions in the short disordered binding site dataset.

	Total in PDB		Interacting residues		Correctly identified	
	Number	Fraction (%)	Number	Fraction (%)	Number	Fraction (%)
H	297	35.7	200	33.6	144	36.7
E	25	3.0	25	4.2	23	5.9
C	510	61.3	371	62.2	225	57.4
<b>Total</b>	<b>832</b>		<b>596</b>		<b>392</b>	

The number and fraction of amino acids in different secondary structures in the disordered chains of the complexes. The three groups show these data for all the amino acids in the PDB structures, the ones in interaction and the ones that are correctly identified as part of binding site by ANCHOR.

doi:10.1371/journal.pcbi.1000376.t006

sites. There is experimental evidence that at least some long disordered chains are not uniform concerning binding strength but contain short stretches of strongly interacting residues separated by segments that interact with the partner only weakly if at all [19]. In these cases, it is expected that the predictor will be unable to identify the weakly interacting parts since – though these parts may also form interchain contacts – they would not be able to bind to the partner in the absence of their sequential neighbors. The distribution of predicted binding regions for the short and long disordered chains in Figure 5A shows a strong preference for predicting multiple interacting regions for longer chains. This inevitably yields lower residue based TPR but the segment based TPR is not expected to drop. Testing the predictor on the long disordered data confirms this assumption with a decreased residue based TPR of 47.7% (as opposed to 65.8% obtained on running the final predictor on the whole set of short disordered complexes) but with a basically unchanged segment based TPR of 78.6% (compared to the 76.1% calculated on short disordered complex-

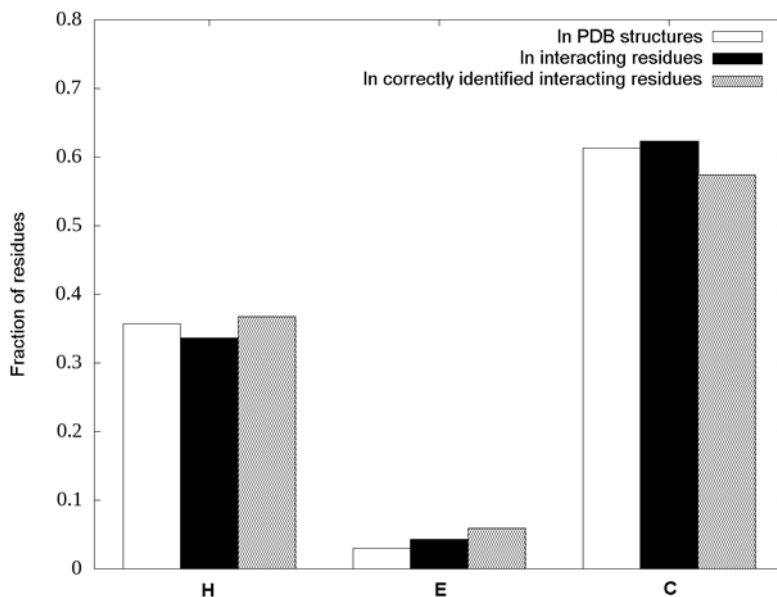
es). These data suggest that the method either finds short disordered binding sites as a whole or completely misses it. However, this may not be true for long binding regions. Figure 5B shows the distribution of the fraction of amino acids successfully identified during prediction in the two types of binding sites. The effect can clearly be seen as about 59% of short binding regions are either fully recovered or are completely missed (the sum of the rightmost and leftmost columns) whereas this ratio is only about 29% for long binding sites.

This type of behavior is illustrated on the disordered human p27. This protein is involved in controlling eukaryotic cell division through interactions with cyclin-dependent kinases. Its kinase inhibitory domain binds both subunits of the CDK2-cyclin A complex in an extended conformation (PDB ID: 1jsu [62]). It is known from kinetic measurements that the binding of p27 is hierarchical through its three domains: first, the D1 domain (residues 25–36) binds to cyclinA which anchors the neighboring LH domain (residues 38–60) that exhibits transient helical structure in monomer state as well [63]. After the binding of D1 this transient structure is stabilized and positions the rest of the chain (D2 domain, residues 62–90) in the correct position to bind to CDK2.

Figure 6 shows the prediction output for p27. Four interacting regions are identified with the first one (27–37) clearly corresponding to D1. The gap between the first two regions (38–58) coincides with the weakly interacting LH domain. The last three regions (59–67, 74–77 and 79–90) cover the strongly interacting D2. Figure 6 also shows the number of atomic contacts/residue for p27 (averaged in a window of size 3). This contact number profile exhibits well pronounced peaks that line up with the regions that are predicted by our algorithm. The figure also shows the four predicted regions mapped to the crystal structure of the complex.

#### Wiskott-Aldrich Syndrome protein (WASp)

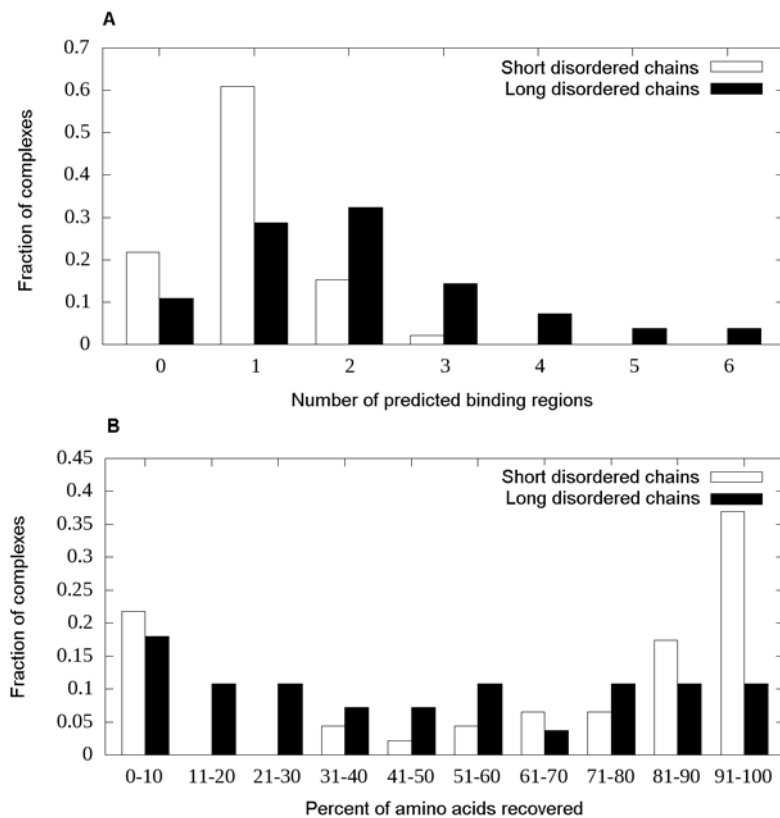
The examples discussed so far represent various fragments of proteins. Here we present an additional case showing the prediction output for a complete protein sequence.



**Figure 4.** Secondary structure distributions in the short disordered binding site dataset. Fraction of amino acids in different secondary structures in the disordered chains of the complexes. The three groups denote the fractions calculated on all the residues in the PDB structures, only the interacting ones and the ones correctly identified by the predictor.

doi:10.1371/journal.pcbi.1000376.g004





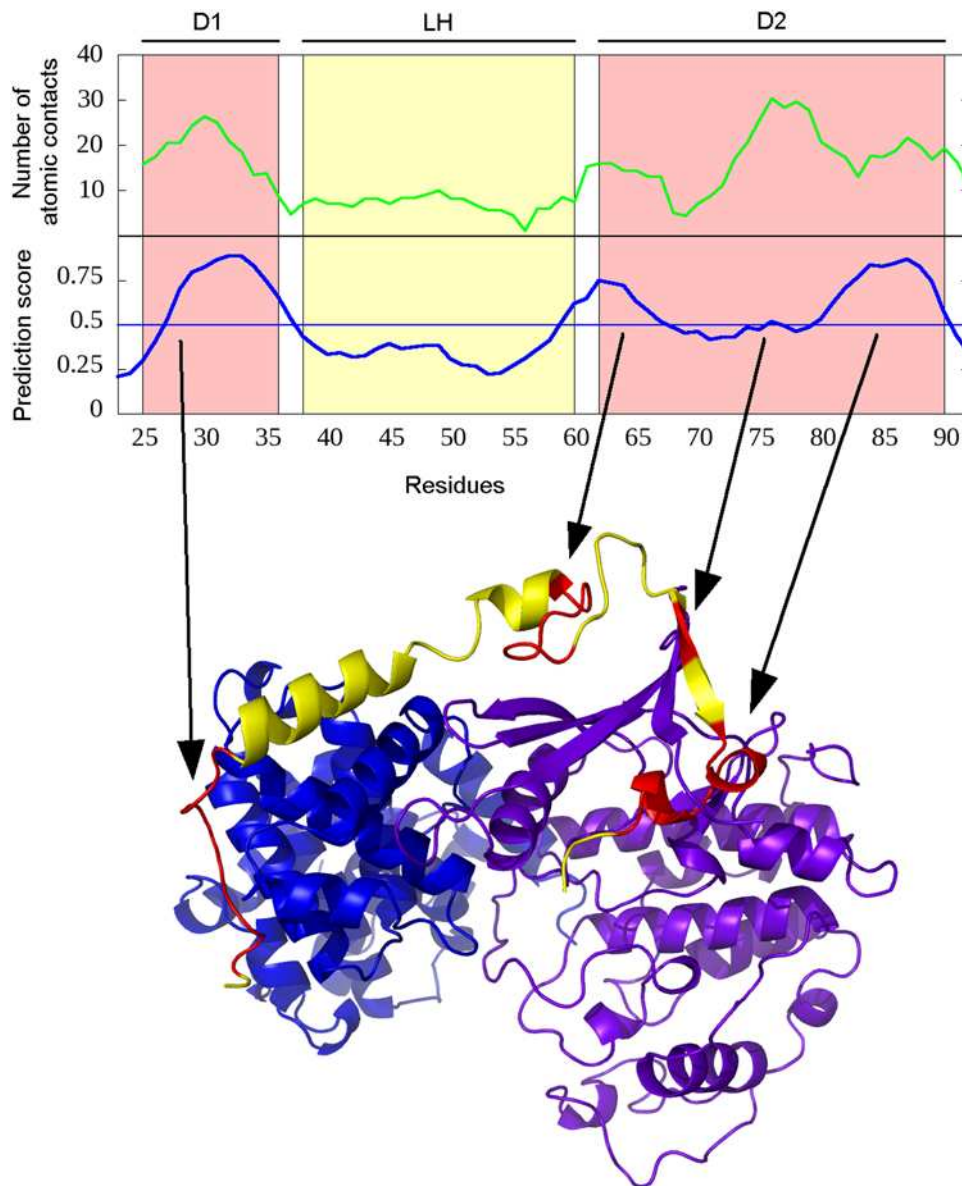
**Figure 5. Prediction accuracies and segmentation for the short and long disordered binding sites.** (A) The distribution of the number of binding segments predicted in short (white bars) and long (black bars) binding sites. It shows the segmented nature of longer binding sites. (B) The distribution of the fraction of correctly recovered interacting residues in both the short (white bars) and long (black bars) disordered binding sites. doi:10.1371/journal.pcbi.1000376.g005

The human Wiskott-Aldrich Syndrome protein (WASP) is a 502 residue long protein that is expressed in the cells of the hematopoietic system [64]. Its mutations can be linked to the Wiskott-Aldrich Syndrome (WAS), a disease characterized by actin cytoskeleton defects leading to deficiencies in blood clotting and immune response. The protein is composed of various functional domains. It contains the WH1 domain near the N terminus (residues 39–148), the GTPase-binding domain (GBD, 230–310), a polyproline-rich region and a C-terminal verpulin homology/central region/acidic region (VCA, 430–502) domain [65] that also contains the WH2 domain (430–447). Apart from the structured WH1 domain, it is predicted to be largely disordered and contains several low complexity regions (enriched in P, G and acidic amino acids). There is experimental evidence that the activated WASp hubs a number of interactions with partners including CDC42, RAC, NCK, FYN, SRC kinase FGR, BTK, ABL, PSTPIP1, WIP, and the p85 subunit of PLC-gamma as well as the Arp2/3 complex. However, the location of many of these binding regions is not known. The domain structure of WASp is shown in Figure 7 together with the known binding regions.

In its inactive state WASp exists in an autoinhibited form with the GBD domain bound to the VCA domain. When WASp is activated, the GBD domain is bound to CDC42 and this interaction disrupts the GBD-VCA interaction. This initiates a conformational change where WASp opens up and becomes able to bind to the Arp2/3 complex leading to its activation and actin nucleation. Both GBD and VCA regions were shown to be disordered in their free state [65,66], with GBD adopting a loosely

packed, compact conformation. However, the structure of both complexes could be determined using NMR, by covalently linking GBD to CDC42 or the VCA region, respectively [65,67]. In these two structures WASp GBD adopts related but distinct folds. The plasticity that can be seen by comparing these two complexes is enabled by the absence of discrete tertiary structure in isolation. As it can be seen on Figure 7, ANCHOR captures these disordered binding sites correctly.

It is known that WASp is able to bind to SRC Homology 3 (SH3) domains through one of its proline rich regions although the exact binding site is not known. The interaction with SH3 domains is usually mediated by a short, linear sequence motif that is present in the interaction partner. In the collection of Eukaryotic Linear Motifs (ELM) database (<http://elm.eu.org/> [68]) there are five different motifs annotated as SH3 recognition sites. Multiple instances of the following three can be found in human WASp: LIG\_SH3\_1, LIG\_SH3\_2 and LIG\_SH3\_3 represented by the following consensus sequences: [RKY]..P..P, P..P.[KR] and ...[PV]..P, for interaction with Class I/ClassII SH3 domains and those SH3 domains with a non-canonical Class I recognition specificity, respectively. The found motifs are clustered in two separate regions mainly falling into the proline-rich regions of WASp (Figure 7). Although there is no direct evidence for the location of interaction with SH3 domains on human WASp, the interaction sites have been identified for Las17 [69], the yeast homologue of this protein. In total, four distinct regions containing multiple binding sites were identified experimentally in Las17 that interact with various SH3 domains. These sites correspond to the proline rich regions in WASp



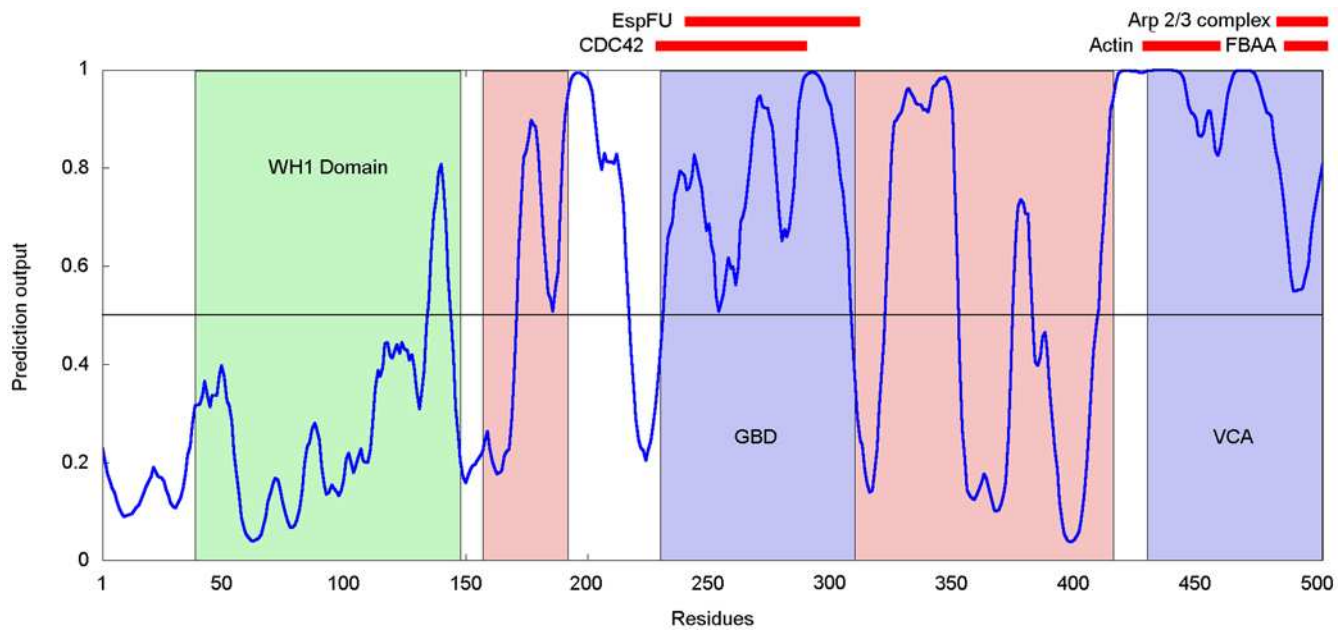
**Figure 6. ANCHOR prediction for human p27.** *Top:* Number of atomic contacts (green) and prediction output (blue) and for the N-terminal binding region of human p27. “D1” and “D2” denote the two strongly interacting domains (red boxes) and “LH” denotes the weakly interacting linker domain between them (yellow box). *Bottom:* Crystal structure of human p27 (red and yellow) complexed with CDK2 (magenta) and Cyclin A (blue) (PDB ID: 1jsu [62]). Red parts denote regions that are predicted to bind by the predictor. These regions correspond to the experimentally verified strongly binding regions of p27. The figure was generated by PyMOL. doi:10.1371/journal.pcbi.1000376.g006

(155–194 and 306–427) that also match with several SH3 binding motifs. As linear motifs were shown to have a preference to reside in disordered regions [70], it is plausible to expect ANCHOR to be able to recognize the SH3 binding region of WASp. In accordance with this, both regions containing putative SH3 binding sites contain binding sites predicted by ANCHOR. This prediction can restrict the candidate sequence regions for SH3 binding and can guide experimental studies to localize true binding sites.

### Complete proteome scans

In order to gain some evolutionary insight concerning disordered binding sites, the predictor was run on the 736 complete proteomes (53 archaea, 639 bacteria and 44 eukaryota,

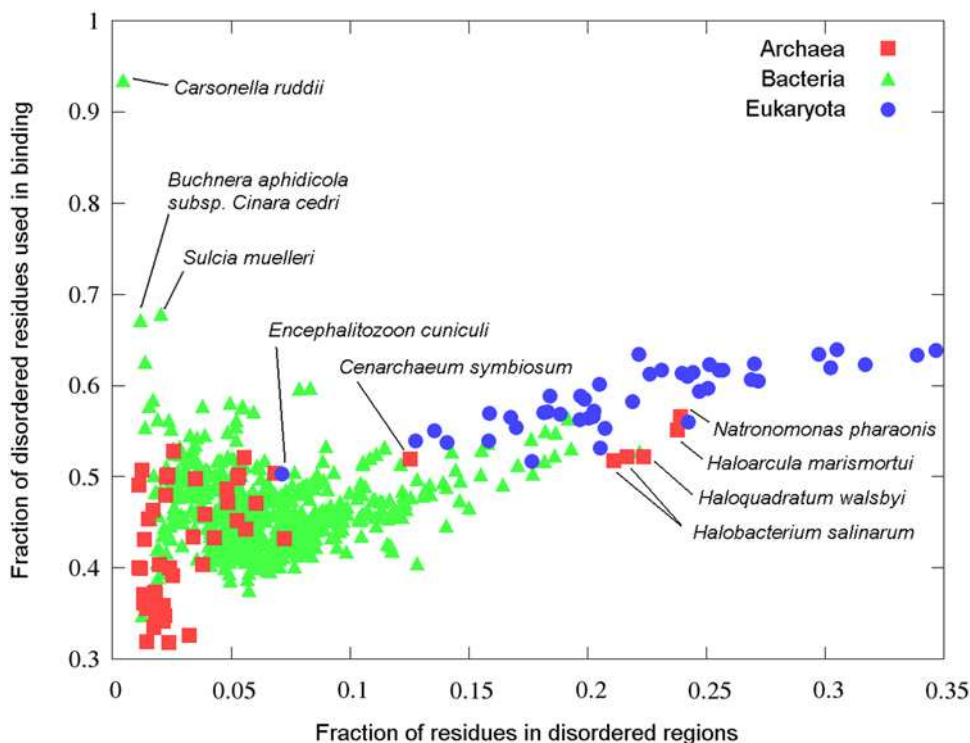
see Dataset S5, Dataset S6, and Dataset S7, respectively) that are currently available from the SwissProt database (<ftp://ftp.expasy.org/>). In agreement of previous analyses [5,6] there is a clear trend of increasing amount of protein disorder as the complexity of the organism increases (see Figure 8). However, Figure 8 also shows that the fraction of disordered amino acids predicted to be in disordered binding sites increases even compared to fraction of disordered residues, as the complexity of organisms grows. Generally, archaea have the least amount of both disorder and binding sites. On the other hand, eukaryota have generally the largest ratio of disordered and binding amino acids with bacteria being between these two groups on average. However there are a few exceptions to these general trends, marked separately on Figure 8.



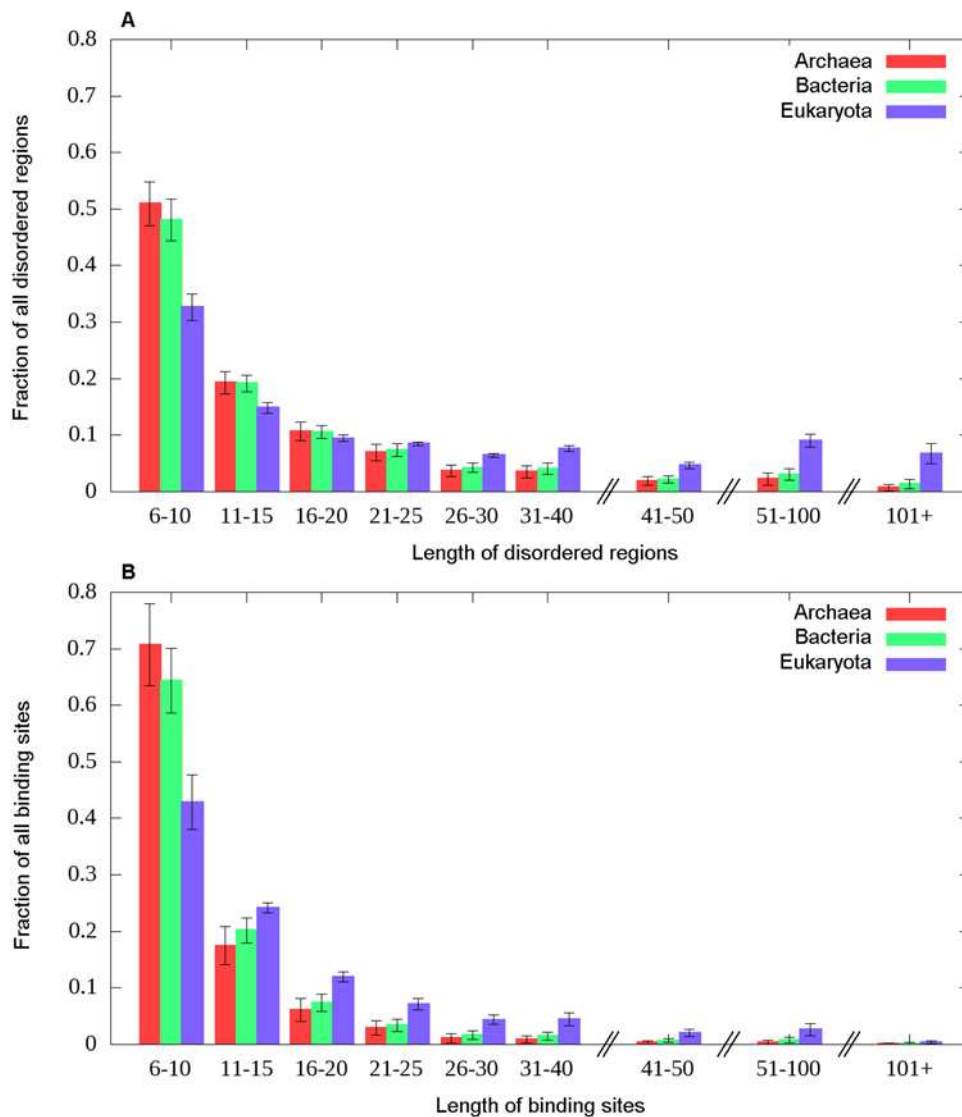
**Figure 7. ANCHOR prediction for human WASp.** Red bars mark known interaction sites, green box marks the globular WH1 domain, blue boxes mark the GBD and VCA domains. Light red boxes indicate the regions with putative SH3 domain interaction sites.  
doi:10.1371/journal.pcbi.1000376.g007

Considering archaea, mesophiles generally contain a larger amount of disorder and a larger fraction of disordered binding sites than most extremophiles (thermophiles, cryophiles and acidiphiles). However the group of halophile archaea (archaea

that favor high saline concentration) is a distinct exception with fraction of disordered amino acids ranging from 0.2 to 0.25 as opposed to other extremophiles' values not exceeding 0.07. This group includes all the halophile archaea in our study, namely



**Figure 8. Fraction of disordered and disordered binding site residues in complete proteomes.** The number of amino acids in disordered binding sites divided by the number of amino acids in disordered regions plotted as a function of the number of amino acids in disordered regions divided by the total number of residues in the proteome of the organism for the 736 complete proteomes deposited in the SwissProt database, colored according to the three kingdoms of life. The outlying points are marked with the name of the corresponding organism.  
doi:10.1371/journal.pcbi.1000376.g008



**Figure 9. Length distribution of disordered and disordered binding sites in complete proteomes.** The length distribution of (A) the disordered protein segments determined by IUPred and (B) predicted disordered binding sites determined by ANCHOR for the 736 complete proteomes available, grouped according to the three kingdoms of life. doi:10.1371/journal.pcbi.1000376.g009

*Natronomonas pharaonis*, *Haloarcula marismortui*, *Haloquadratum walsbyi* and two types of *Halobacterium salinarum*. *Cenarchaeum symbiosum*, the only example of obligate endosymbiont among archaea also has an unusually large amount of disordered protein segments in its proteome (0.12). While *Cenarchaeum symbiosum* is closely related to thermophile archaeas, it is adapted to the much lower living temperature of its host [71]. This adaptation could explain the relatively large amount protein disorder and disordered binding sites. In general, these clear differences in the predicted disorder between various archaea organisms points to different strategies to adapt to various extreme environmental conditions resulting in biased amino acid compositions. However, we cannot rule out the possibility that under such extreme conditions, as high salt concentration or high temperature, the amount of disorder can be over- or underpredicted depending how these conditions affect the presence of protein disorder.

Among bacterial proteomes, there are a few examples of organisms that seem to utilize a surprisingly large fraction of their

disordered amino acids in binding. The three most extreme cases (*Carsonella ruddii*, *Sulcia muelleri* and *Buchnera aphidicola subsp. Cinara cedri*) are marked separately on Figure 8. These are the three smallest complete bacterial proteomes, none of them reaching the size of the smallest archaea proteome. These organisms present extreme cases of streamlined genomes as a result of endosymbiosis [72–74]. As these proteomes are very small, the predicted amount of disorder and disordered binding sites are within the false positive range, and should be treated more cautiously.

Eukaryotes tend to appear more consistent both in using larger amount of disordered residues and larger fraction of disordered residues for binding compared to the other two kingdoms (Figure 8). The only notable outlier both in terms of extremely low amount disordered proteins and disordered binding sites is *Encephalitozoon cuniculi*. This organism is the only microsporidian parasite in our dataset and has an extremely small proteome. This lack of complexity and dependence on a eukaryotic host to function might explain the lack of disordered proteins.

The length distributions of the predicted disordered regions and binding sites in the three kingdoms of life was also analyzed and are shown in Figure 9A and 9B, respectively. As complexity increases, longer disordered segments are preferred, and the difference between eukaryota and lower complexity organisms becomes even more apparent for longer regions (over 30 residues). A similar trend can be observed in the length distribution of disordered binding sites. While in archaea and bacteria predicted binding regions are generally below 30 residues, longer binding sites in eukaryota organisms are much more common. There are at least three different effects that can contribute to this phenomenon. First, as the number of binding sites rise there is also an increasing possibility of these binding sites becoming very close to each other or even overlapping with each other. This scenario was demonstrated in the case of the N-terminal domain of p53, as shown in Figure 1. Second, extremely large disordered binding regions may be needed for special functions. Some members of the mucin protein family provide an example for this. Human MUC1 contains a large repeat region (20–120 repeats, one repeat being 20 amino acids long) that enables it to aggregate and to perform its function [75]. As each repeat is correctly identified as a disordered binding site, the whole repeat region is predicted as one large binding region. This mechanism can create binding sites up to the length of several hundreds of residues in extreme cases. Third, we cannot exclude the possibility that longer binding sites are not always segmented by weakly interacting regions like in the case of p27, thus forming long, continuous binding regions. Nevertheless, the majority of predicted binding sites is shorter than 30 residues, although such restriction on the length of disordered binding sites was not enforced.

## Discussion

Regions undergoing disorder-to-order transitions upon binding are essential elements in the molecular recognition process involving disordered proteins. The main property of these binding regions is that they can exist in a disordered state as well as in bound state, adopting at least partially a well-defined conformation. The presence of these two separate states discriminates them from monomeric globular proteins as well as from complexes formed between globular proteins and from disordered proteins in general. They are also expected to differ from dual personality fragments [76], which occur within globular proteins, however, mostly as a result of perturbations of environmental conditions. In this work we aimed to recognize such disordered binding regions from the amino acid sequence. So far, the limited number of well characterized examples hindered the development of general prediction methods. Nevertheless, biophysical considerations suggest that in most cases there is a strong signal in the amino acid sequence highlighting regions involved in coupled folding and binding. These regions are linear in sequence, unlike in the case of globular proteins, where distinct sites in the amino acid sequence are brought together to form the interface for interaction [43]. An additional difference is that binding of disordered proteins is driven by a large enthalpic component to compensate for the entropy penalty due to the loss of conformational freedom [9]. These features result in a relatively short sequence segment containing residues with a pronounced tendency to make interactions, leading to a characteristic sequence signal.

Our approach relies on a basic physical model of disordered binding sites and it is based on modeling the interaction capacity in the free disordered state and in the bound ordered state. Previously, it was shown that ordered proteins can be discriminated from disordered proteins based on estimated pairwise

energy content and this approach was implemented in IUPred, a general disorder prediction method [53]. This method takes into account that disorder/order tendency can be modulated by the sequential neighborhood simply at the level of amino acid composition, without attempting to model the specific interactions. Taking it one step further, the same energy estimation calculations were used to identify disordered binding regions in proteins. Our model assumes that the specific properties of disordered binding sites are dictated by the combination of preferences to bind to an ordered protein on the one hand, and the ability to remain in a disordered state in isolation, on the other. Based on this simple model, ANCHOR achieved approximately 67% accuracy at predicting 5% false positive rate (Tables 2–4). Furthermore, this approach was validated by the ability to reproduce the specific amino acid composition of disordered binding sites, that is distinct from that of ordered proteins as well as disordered proteins in general (Table 5).

During binding, the formation of intermolecular contacts is accompanied by the formation or the stabilization of secondary structure elements. The secondary structure composition of the binding sites is highly unequal (Table 6 and Figure 4). The most dominant secondary structure element adopted in the bound conformation is coil, while  $\beta$  strand conformation is rare. Helical conformations are observed as frequently in disordered complexes as in globular proteins [27]. It was found that the adopted secondary structure can be predicted from the amino acid sequence with similar accuracy as in the case of globular proteins, suggesting that the adopted secondary structure can be imprinted into the sequence of the binding motif [27]. The secondary structure observed in the complex can also be dictated by the template structure. An extreme example of this is the C-terminal region of p53 (see Supporting Information), observed in all three secondary structure classes [32]. It is clear that not all of these conformations can be the result of inherent preferences. Interestingly, our prediction method does not seem to be sensitive to the adopted secondary structure conformation and it works with the same accuracy for all secondary structure conformations (Table 6 and Figure 4). This independence of secondary structure elements underlines the generality of ANCHOR. These results also suggest that disordered binding sites can be recognized without taking into account of the adopted secondary structure in the majority of cases. Nevertheless, the details of conformational preferences can be still crucial in selecting the specific binding partner, or determining the kinetic and thermodynamic properties of the associations.

Beside our algorithm, a previously published method called  $\alpha$ -MoRF predictor also exploited a general disorder prediction method to recognize short binding elements [48,52]. Although the direct comparison between the two methods was not possible, because the  $\alpha$ -MoRF predictor is not yet publicly available, some basic differences between the two methods should be noted. First, the  $\alpha$ -MoRF predictor directly relies on the prediction output of PONDR VLXT, which essentially predicts binding regions as ordered structural elements, and a subsequent neural network is applied to filter out valid disordered binding sites. Although very high accuracies were reported for the performance of the neural network based filtering, the complete method is limited by finding dips based on PONDR VLXT [49–51]. Therefore it should be taken into account that this program is a first generation prediction method that was trained on only 15 proteins. In the case of IUPred, dips corresponding to certain binding sites were also observed, although to a smaller extent [48,53]. This observation, however, is not directly exploited in our prediction method. Instead, the core parameters of the energy prediction of

IUPred are used to create three separate scores characterizing three important attributes of disordered binding regions. The second main difference is that ANCHOR is not restricted to a single secondary structure class like the  $\alpha$ -MoRF predictor that was trained to recognize only  $\alpha$ -helical segments. The example of the C-terminal region of p53 (Figure S2), where four short overlapping regions were shown to bind in different conformations representing all three secondary structure classes, indicates that such restriction can be a serious disadvantage for recognizing some extremely adaptable disordered binding motifs.

An alternative approach for binding site identification is based on the observation that protein-protein interactions are often mediated through short linear motifs (approximately three to eight residues) [77]. Such motifs are defined by a consensus pattern, which captures the key residues involved in function or binding. Prominent examples include the nuclear receptor box motif, MDM2 binding sites, SH2/SH3 domain recognition patterns or 14-3-3 domain binding sites [68]. Although there are known examples of motifs that reside within globular domains, many of them are required to be in a disordered region to function properly and it was suggested that such motifs share many similarities with disordered binding regions [70]. Our preliminary results support previous observations of the partial overlap between short linear motifs and disordered binding segments. Nevertheless, short disordered binding sites and sequence specific linear motifs capture different aspects of certain binding regions. Linear motifs are defined on the basis of a per residue binding strength, and they are specific to a certain partner or to a group of partner molecules. However, such short linear motifs can also occur purely by chance, with no biological significance. Also, sequence patterns alone cannot ensure the accessibility of the site and the potential flexibility of the binding region that could be necessary for the complex formation. Complementary to sequence motifs, ANCHOR aims to capture a broader structural context. Based on their specific structural properties, it can recognize disordered binding regions that are capable of undergoing disorder-to-order transition. The predictions are made without taking into account the partner molecules and are expected to be less sensitive to sequence details. For certain motifs, this molecular environment can be a prerequisite of functionality and could help to identify biologically significant binding motifs.

In our work we assumed, that short binding regions undergoing disorder-to-order transition can be viewed as elementary binding units that are necessary for the molecular recognition. Therefore, such examples were used for the optimization of our method. In accordance with their elementary unit picture, ANCHOR recognized them generally as a single continuous binding site (Figure 5). Regions undergoing disorder-to-order transition, however, are not limited to such short segments as there are several examples of longer disordered segment becoming ordered upon complex formation. Such segments can be as long as 100 residues. However, these longer regions can contain segments which bind only weakly or might not become ordered at all [63,78,79]. This segmentation of longer binding regions can occur for structural reasons. The segmentation can prevent the accumulation of the critical amount of residues that would lead to the formation a collapsed structure or non-specific aggregates. The possible functional advantages of the segmented nature of a binding site were demonstrated for the well characterized example of p27. The kinase inhibitory domain of p27 can be divided into several subdomains which dock and fold in a stepwise manner on the surface of the Cdk2-cyclin A complex [19]. These segments can also evolve independently, increasing the repertoire for specificity for different cellular location or species. Intervening

segments of higher flexibility are accessible for modifications such as phosphorylations and ubiquitinations. This way p27 can integrate and process various signals to regulate cell proliferation, in which the flexibility and modularity of p27 is essential [63]. The segmented nature of binding is reflected in the prediction output, with predicted binding sites corresponding to the strongly interacting regions (Figure 6 for p27, and Figure S4 for a similar example, calpastatin). In the dataset of longer disordered binding segments, we found this segmentation to be quite general. In these cases, the predicted sites generally give only partial coverage of the PDB structure, and multiple binding sites are predicted in the majority of cases (Figure 5). This suggests that our prediction method is likely to find those sites that interact more strongly, anchoring the disordered segments to their partner protein. While the segmented nature of binding is prominent in the case of long binding regions, to a smaller extent, it can also affect shorter binding regions. Indeed, around 20% of short disordered binding regions are predicted as 2 or 3 segments (Figure 5). This could also account for the significantly lower per residue efficiency compared to the segment based efficiency.

By looking at further individual examples, one can already see remarkable variations in the details of disorder-to-order transitions even within the limited collection that is available today. The adopted conformation in these complexes can be quite different, both in terms of secondary or tertiary structure. Furthermore, the transition to an ordered structure might not be complete [28]. This could leave terminal residues or linker regions flexible and inaccessible to structure determination. It was also suggested that specific binding can be possible even without adopting a well-defined conformation as in the case of the  $\zeta$ -chain of T-cell receptor [80] (see Figure S6). Differences are also present at the level of the sequence. Some binding regions rely largely on hydrophobic or aromatic residues (MDM2 binding regions, Figure 1), others use proline rich regions (WASp SH3 binding regions, Figure 7). Disordered binding regions can contain conserved linear motifs, while large divergence in sequence was noted in other cases (C terminal domain of histones [81]). These examples represent multiple ways disordered regions can be utilized for binding. A single protein sequence can contain several distinct binding regions, however, a single region can be involved in binding to multiple partners, or use these regions in combination to hub several interactions (p53 – see Figure 1 and Figure S2, WASp – see Figure 7). In an alternative scenario, disorder present in the partner molecules allows to bind a well-folded protein by a large number of proteins ( $\beta$ -catenin [82], Figure S3). Even further variations are expected as the number of examples will grow in the future. Nevertheless, the success of ANCHOR confirms our hypothesis, that despite these differences disordered binding regions have a common property that predispose them for coupled folding and binding.

The occurrence of disordered binding sites is clearly tied to the presence of disordered protein regions. Their relationship was further analyzed at the level of complete proteomes. Previous studies have shown that the amount of predicted disordered regions increases with the complexity of organisms throughout evolution and reaches a high level in multicellular organisms [5,6]. This increase can be mostly attributed to the appearance of long, domain-sized segments of protein disorder or fully disordered proteins (Figure 9A). Our analysis showed that the amount of disordered binding segments increases in eukaryotes in a similar way, however, their fraction is elevated even compared to disordered regions in general (Figure 8). The observed trend is valid through a wide range of organisms, and occasional exceptions occur either due to adaptation to extreme habitat

conditions, or as a result of endosymbiosis. These findings imply that the newly introduced disordered proteins and protein segments mainly serve as a carrier for new binding regions in eukaryotic organisms. The importance of disordered regions in protein-protein interactions is also supported by the increased ratio of disordered proteins among hub proteins [30,31]. Disordered segments are often involved for complex signaling and regulatory processes [20] such as cell cycle control, gene regulation or signal transduction in the intracellular region of transmembrane proteins [83]. These processes rely on interactions involving multiple partners and high specificity/low affinity interactions, that disordered binding segments can provide by their very nature. The disordered segments can harbor multiple binding sites which can act relatively independently. In other cases segmented binding sites can be involved in simultaneous binding to larger complexes. Overlapping binding sites (such in the case of p53 N and C terminal regions) suggest competition between binding partners. We are only beginning to comprehend how disordered binding regions are exploited to provide versatile interaction sites in proteins.

In conclusion, disordered binding regions represent a specific subclass of disordered proteins that can undergo a disorder-to-order transition upon binding. These binding sites generally have distinct properties both structurally and functionally. Due to the inherent flexibility, these regions are difficult to study experimentally [84], making specific prediction methods even more valuable. While there are several methods available for prediction of disordered regions [85,86], recognizing disordered binding sites was regarded as a more challenging problem [9] due to the limited number of well-characterized examples. In this work we report a general method to recognize disordered binding sites based on a basic biophysical model. Our method relies on a simple energy estimation procedure that was developed earlier for the IUPred disorder prediction method. This way, the problem of small datasets can be largely avoided. We showed that these regions can be characterized by highly disordered sequential neighborhood, unfavorable intrachain energies and more favorable interaction energies with a globular partner. The combination of these properties allowed the recognition of disordered binding sites independent of their secondary structure or amino acid composition, underlining the generality of the method. As such binding sites are essential functional elements of disordered proteins, their prediction directly provides information about functionally important residues in these proteins. In this way, ANCHOR broadens the repertoire of prediction methods for functional sites in proteins aiming to decrease the large number of unannotated sequences [87]. Generally, the complete understanding of protein-protein interactions involving disordered binding sites requires the knowledge of their partners as well as possible post-translational modifications that can influence their binding. While predictions can be made even without taking the partner molecule into account, certain cases might require incorporating the specific feature of the partner. Nevertheless, our method can provide the starting point for such scientific explorations, by finding potential regions involved in such binding.

## Methods

### Databases

The primary source of data for the present analysis is a carefully assembled dataset of binding regions undergoing disorder-to-order transition. The strict requirement of the experimental verification of both the disordered status in isolation and the formation of an ordered structure in complex distinguishes our dataset from a

previously collected dataset for disordered binding regions [88]. The length of disordered regions involved in the binding can vary on a large scale. In the case of longer regions it is not guaranteed that each residue is equally important for binding, therefore complexes of short disordered regions were treated separately, and only these were used for tuning the method.

**Short disordered complexes.** Complexes from the PDB [89] were collected by scanning the chains in the PDB entries against the Disprot database [90]. A complex was accepted if it consisted of a chain with length between 10 and 30 residues that was found in the Disprot database as part of an annotated disordered segment and at least one interacting partner that was at least 40 residues long. Furthermore, complexes containing transmembrane proteins, RNA or DNA, chimeras, disulfide bonds between the disordered and ordered chains or a large number of unknown residues (marked with an X) were excluded. A few experimentally verified disordered complexes missing from Disprot were added to this set [42,43,62,91–93]. A sequence similarity filter of 50% has also been applied to remove closely related proteins or protein segments. This procedure yielded a set of 46 complexes that are listed in Dataset S1.

**Long disordered complexes.** Complexes containing long disordered chains were collected in the same fashion as short ones but with different criteria for the length of the interacting partners. Here the length of the disordered chains was required to be at least 30 residues and they had to have an interacting partner of 70 residues or more. The resulting set of 28 complexes is listed in Dataset S2.

**$\alpha$ -MoRFs dataset.** This dataset originally consisted of 53 complexes [48]. Complexes that were contained in our Short disordered complexes dataset as well were excluded in order to get a truly independent set. Three complexes were further removed from the remainder since one of them is part of the ribosome subunit S23 and the other two can be found in the PDB with structures containing only the disordered chain – that is they are presumably capable of folding on their own. The rationale behind this exclusion is that our predictor is neither trained to recognize RNA/DNA-protein interactions nor to identify globular-globular interfaces. This left 40 complexes in total.

**Globular proteins.** Globular proteins were collected from PDB entries that had only one chain of at least 30 residues [53]. Also transmembrane proteins and complexes with RNA/DNA were filtered out. This dataset contains 553 proteins and is presented in Dataset S3.

**Ordered complexes.** This set contains protein complexes that consist of two partners both of which are ordered. These data were taken from the literature [43]. The dataset does not include cases of crystal packing dimers, chimeras and fragments and consists of 72 complexes (Dataset S4).

**Disordered proteins.** For the analysis of disordered proteins and protein segments the 3.7 version of Disprot database was used (<http://www.disprot.org/>) [90], considering only annotated disordered segments of 10 residues or longer.

### Parameter optimization

The optimal parameters were determined by a three fold cross-validation, by dividing both our negative and positive datasets (Globular proteins and Short disordered complexes, respectively) into three parts. In each turn we used two parts for training and the remaining part for testing. To avoid any bias, the different subsets were chosen such that the distribution of chain lengths in both the positive and negative sets and the distribution of secondary structure types in the positive set were approximately the same. Our approach relies on IUPred, a general disorder

prediction method, and its energy predictor matrix. These parameters (ie. the elements of the energy predictor matrix) have been determined earlier, independently of disordered binding regions. Only five additional parameters,  $w_1$ ,  $w_2$ ,  $p_1$ ,  $p_2$  and  $p_3$  were optimized for this specific problem and were selected by a grid search procedure. Specifically,  $w_1$  was varied in the range of 20 to 100 in steps of 10 (giving 9 possible values),  $w_2$  was varied in the range of 5 to 35 in steps of 2 (giving 16 possible values), and  $p_1$ ,  $p_2$  and  $p_3$  was selected from 1000 sets of randomly generated values. Taking into account that the prediction performance is insensitive to the norm and the sign of the vector corresponding to the  $p_1$ ,  $p_2$  and  $p_3$  values, the search was restricted to 1000 random sets that were evenly distributed on the surface of the upper half of the unit sphere. This means that  $p_1$  and  $p_2$  were randomly selected from the interval  $[-1;1]$  and  $p_3$  was selected from the interval  $[0;1]$  in a way that the sum of their squares is always equal to 1. This yielded 1000 different ( $p_1, p_2, p_3$ ) combinations. These, combined with all possible values of  $w_1$  and  $w_2$  gave 144,000 different parameter sets in total. These were considered in order to select the optimal one, containing the five optimal parameters for each round of the cross-validation.

To quantify the performance of the predictor given a set of parameters we calculated the True Positive Rate (TPR) at False Positive Rates (FPR) fixed at 5% calculated on globular proteins as the negative set. However, a full characterization of the performance of the algorithm would also require a set of disordered proteins that are known *not* to bind to globular proteins. Unfortunately, such dataset cannot be constructed since there is hardly any way to give evidence for a protein that it does not contain binding sites. This problem was addressed by calculating the fraction of amino acids that are predicted as binding sites in general disordered regions of Disprot database that are correctly recognized as disordered by IUPred. This fraction was denoted as  $F$ . Optimal parameters should combine high TPR with low  $F$  at the expense of very low FPR.

During optimization of the algorithm, the performance on three different datasets needed to be monitored at the same time (set of globular proteins, set of disordered binding sites and Disprot). The best parameter set was chosen manually, by reducing the parameter set in a step-wise manner based on the following steps:

- 1, Calculate TPR (at fixed FPR = 5%) and F for each of the 144,000 candidate sets of parameters
- 2, Discard all for which  $F > 50\%$
- 3, Discard all for which  $TPR < 60\%$
- 4, From the remainder choose the 20 for which the difference between TPR and F is the largest
- 5, Choose the one for which TPR is maximal (the TPR-F difference among these 20 sets vary only within a range of less than 0.02 so that is not a good measure to choose the best one)

The negative and positive sets were divided into three parts, resulting in three different optimal parameter sets. The final predictor algorithm is constructed by averaging these three outputs. As the training sets only contained binding regions of at least 10 amino acids and we aim to identify at least 5 residues of each region, all predicted binding sites were removed that did not exceed 5 consecutive residues. A schematic figure of the training procedure is given in Figure S1.

### Availability

ANCHOR is available upon request from the authors.

### Supporting Information

**Dataset S1** 46 complexes of short disordered and long globular proteins. Column 4 contains the secondary structure type of the

bound disordered chains based on the structure found in the PDB record as defined in Data and Methods. Thick lines separate the three groups used during parameter optimization.

Found at: doi:10.1371/journal.pcbi.1000376.s001 (0.07 MB DOC)

**Dataset S2** 28 complexes of long disordered and long globular proteins. Column 4 contains the secondary structure type of the bound disordered chains based on the structure found in the PDB record as defined in Data and Methods.

Found at: doi:10.1371/journal.pcbi.1000376.s002 (0.05 MB DOC)

**Dataset S3** 553 monomeric globular proteins that were used as a negative dataset [2]. Columns correspond to the grouping used during parameter optimization.

Found at: doi:10.1371/journal.pcbi.1000376.s003 (0.20 MB DOC)

**Dataset S4** 72 complexes of ordered proteins [3]. The interaction is considered between the shortest chains and its interaction partners.

Found at: doi:10.1371/journal.pcbi.1000376.s004 (0.08 MB DOC)

**Dataset S5** The 53 complete archaea proteomes available from SwissProt (ftp://ftp.expasy.org/) used for full proteome scans. The fraction of total amino acids in disordered regions and the fraction of disordered amino acids in disordered binding sites are indicated together for each organism.

Found at: doi:10.1371/journal.pcbi.1000376.s005 (0.09 MB DOC)

**Dataset S6** The 639 complete bacteria proteomes available from SwissProt (ftp://ftp.expasy.org/) used for full proteome scans. The fraction of total amino acids in disordered regions and the fraction of disordered amino acids in disordered binding sites are indicated together for each organism.

Found at: doi:10.1371/journal.pcbi.1000376.s006 (0.86 MB DOC)

**Dataset S7** The 44 complete eukaryota proteomes available from SwissProt (ftp://ftp.expasy.org/) used for full proteome scans. The fraction of total amino acids in disordered regions and the fraction of disordered amino acids in disordered binding sites are indicated together for each organism.

Found at: doi:10.1371/journal.pcbi.1000376.s007 (0.08 MB DOC)

**Figure S1** Development of ANCHOR. In the first step, our Short Disordered Binding Sites dataset and Globular Proteins dataset (positive and negative datasets) are split up and only 2/3 is used in the subsequential steps. Then a parameter set ( $w_1$ ,  $w_2$ ,  $p_1$ ,  $p_2$ ,  $p_3$ ) is selected from the 144,000 random ones. This parameter set is used to calculate  $S$ ,  $E_{int}$  and  $E_{gain}$  for every position in every sequence in the three input datasets using the fixed energy predictor matrix  $P$  (see Theory). Based on this calculations the evaluating measures are calculated: TPR is calculated on Short Disordered Binding Sites, FPR is calculated on Globular Proteins and  $F$  is calculated on Disordered Proteins. Based on these measures, the best parameter set out of 144,000 is chosen (see Data and Methods). Then this parameter set is evaluated on the remaining one third of the datasets. These results are reported in Table 3. This procedure is repeated for all three subsets of Short Disordered Binding Sites and Globular Proteins. The output of the three optimized predictors are combined into one final predictor by averaging their output.

Found at: doi:10.1371/journal.pcbi.1000376.s008 (0.05 MB PPT)



**Figure S2** ANCHOR prediction output for the C-terminal domain of human p53. Prediction for the C-terminal disordered domain of human p53. The regulatory binding site around residues 375–390 is able to adopt all three secondary structural elements upon binding to globular partners [4].  
Found at: doi:10.1371/journal.pcbi.1000376.s009 (0.04 MB TIF)

**Figure S3** ANCHOR prediction output for Tcf4. Prediction output for transcription factor Tcf4 (blue) together with the number of atomic contacts (green) determined in the complexed form with Beta-catenin (PDB ID: 2gl7 [5]). Beta-catenin is known to bind several disordered binding regions.  
Found at: doi:10.1371/journal.pcbi.1000376.s010 (0.03 MB TIF)

**Figure S4** ANCHOR prediction output for human calpastatin. Prediction output for the I. domain of human calpastatin. Subdomains A, B and C (grey boxes) are known to bind to calpain and inhibit it. Subdomains A and C bind via a preformed alpha-helix, while subdomain B does not exhibit strong structural preference in solution [6].  
Found at: doi:10.1371/journal.pcbi.1000376.s011 (0.04 MB TIF)

**Figure S5** ANCHOR prediction output for the KID domain of CREB. Prediction output for the KID domain of CREB. The region marked with a grey box interacts with the KIX domain of CBP via two preformed alpha-helices [7].  
Found at: doi:10.1371/journal.pcbi.1000376.s012 (0.03 MB TIF)

## References

- Wright PE, Dyson HJ (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol* 293: 321–331.
- Dyson HJ, Wright PE (2005) Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* 6: 197–208.
- Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, et al. (2001) Intrinsically disordered protein. *J Mol Graph Model* 19: 26–59.
- Tompa P (2002) Intrinsically unstructured proteins. *Trends Biochem Sci* 27: 527–533.
- Dunker AK, Obradovic Z, Romero P, Garner EC, Brown CJ (2000) Intrinsic protein disorder in complete genomes. *Genome Inform Ser Workshop Genome Inform* 11: 161–171.
- Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* 337: 635–645.
- Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, et al. (2007) Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. *J Proteome Res* 6: 1882–1898.
- Tompa P (2005) The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett* 579: 3346–3354.
- Galea CA, Wang Y, Sivakolundu SG, Kriwacki RW (2008) Regulation of cell division by intrinsically unstructured proteins: intrinsic flexibility, modularity, and signaling conduits. *Biochemistry* 47: 7598–7609.
- Chen J, Kanai Y, Cowan NJ, Hirokawa N (1992) Projection domains of MAP2 and tau determine spacings between microtubules in dendrites and axons. *Nature* 360: 674–677.
- Linke WA, Kulke M, Li H, Fujita-Becker S, Neagoe C, et al. (2002) PEVK domain of titin: an entropic spring with actin-binding properties. *J Struct Biol* 137: 194–205.
- Mukhopadhyay R, Kumar S, Hoh JH (2004) Molecular mechanisms for organizing the neuronal cytoskeleton. *Bioessays* 26: 1017–1025.
- Hoh JH (1998) Functional protein domains from the thermally driven motion of polypeptide chains: a proposal. *Proteins* 32: 223–228.
- Alber F, Dokudovskaya S, Veenhoff LM, Zhang W, Kipper J, et al. (2007) The molecular architecture of the nuclear pore complex. *Nature* 450: 695–701.
- Bruschweiler R, Liao X, Wright PE (1995) Long-range motional restrictions in a multidomain zinc-finger protein from anisotropic tumbling. *Science* 268: 886–889.
- Dyson HJ, Wright PE (2002) Coupling of folding and binding for unstructured proteins. *Curr Opin Struct Biol* 12: 54–60.
- Uversky VN (2002) Natively unfolded proteins: a point where biology waits for physics. *Protein Sci* 11: 739–756.
- Demarest SJ, Martinez-Yamout M, Chung J, Chen H, Xu W, et al. (2002) Mutual synergistic folding in recruitment of CBP/p300 by p160 nuclear receptor coactivators. *Nature* 415: 549–553.
- Lacy ER, Filippov I, Lewis WS, Otieno S, Xiao L, et al. (2004) p27 binds cyclin-CDK complexes through a sequential mechanism involving binding-induced protein folding. *Nat Struct Mol Biol* 11: 358–364.

**Figure S6** ANCHOR prediction output for the  $\zeta$ -chain of T-cell receptor. Prediction output for the zeta-chain of the T-cell receptor. The transmembrane region is marked with red box and the three intracellular ITAM regions are marked with blue boxes.

Found at: doi:10.1371/journal.pcbi.1000376.s013 (0.12 MB TIF)

**Protocol S1** Protocol including references for the Supporting Information.

Found at: doi:10.1371/journal.pcbi.1000376.s014 (0.04 MB DOC)

## Acknowledgments

The fruitful discussions with Dr. Monika Fuxreiter and Prof. Burkhard Rost are gratefully acknowledged. We would like to thank Gabor E. Tusnady and Petr Kulhánek for their continuous support in computational problems. We are grateful to Christopher J. Oldfield and A. Keith Dunker for providing us the  $\alpha$ -MoRF dataset. We would also like to thank to Mark Adamsbaum for his critical comments on the manuscript.

## Author Contributions

Conceived and designed the experiments: BM IS ZD. Performed the experiments: BM. Analyzed the data: BM IS ZD. Wrote the paper: BM IS ZD.

- Uversky VN, Oldfield CJ, Dunker AK (2005) Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. *J Mol Recognit* 18: 343–384.
- Liu J, Perumal NB, Oldfield CJ, Su EW, Uversky VN, et al. (2006) Intrinsic disorder in transcription factors. *Biochemistry* 45: 6873–6888.
- Iakoucheva LM, Brown CJ, Lawson JD, Obradovic Z, Dunker AK (2002) Intrinsic disorder in cell-signaling and cancer-associated proteins. *J Mol Biol* 323: 573–584.
- Fuxreiter M, Tompa P, Simon I, Uversky VN, Hansen JC, et al. (2008) Malleable machines take shape in eukaryotic transcriptional regulation. *Nat Chem Biol* 4: 728–737.
- Dunker AK, Garner E, Guillot S, Romero P, Albrecht K, et al. (1998) Protein disorder and the evolution of molecular recognition: theory, predictions and observations. *Pac Symp Biocomput*. pp 473–484.
- Kriwacki RW, Hengst L, Tennant L, Reed SI, Wright PE (1996) Structural studies of p21Waf1/Cip1/Sd1 in the free and Cdk2-bound state: conformational disorder mediates binding diversity. *Proc Natl Acad Sci U S A* 93: 11504–11509.
- Bienkiewicz EA, Adkins JN, Lumb KJ (2002) Functional consequences of preorganized helical structure in the intrinsically disordered cell-cycle inhibitor p27(Kip1). *Biochemistry* 41: 752–759.
- Fuxreiter M, Simon I, Friedrich P, Tompa P (2004) Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. *J Mol Biol* 338: 1015–1026.
- Tompa P, Fuxreiter M (2008) Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem Sci* 33: 2–8.
- Spolar RS, Record MT Jr (1994) Coupling of local folding to site-specific binding of proteins to DNA. *Science* 263: 777–784.
- Dosztanyi Z, Chen J, Dunker AK, Simon I, Tompa P (2006) Disorder and sequence repeats in hub proteins and their implications for network evolution. *J Proteome Res* 5: 2985–2995.
- Haynes C, Oldfield CJ, Ji F, Klitgord N, Cusick ME, et al. (2006) Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS Comput Biol* 2: e100. doi:10.1371/journal.pcbi.0020100.
- Oldfield CJ, Meng J, Yang JY, Yang MQ, Uversky VN, et al. (2008) Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. *BMC Genomics* 9(Suppl 1): S1.
- Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, et al. (2004) The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res* 32: 1037–1049.
- Albert R, Jeong H, Barabási AL (2000) Error and attack tolerance of complex networks. *Nature* 406: 378–382.
- Uversky VN, Oldfield CJ, Dunker AK (2008) Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu Rev Biophys* 37: 215–246.
- Vogelstein B, Lane D, Levine AJ (2000) Surfing the p53 network. *Nature* 408: 307–310.
- Cheng Y, LeGall T, Oldfield CJ, Dunker AK, Uversky VN (2006) Abundance of intrinsic disorder in protein associated with cardiovascular disease. *Biochemistry* 45: 10448–10460.

38. Frankfort SV, Tulner LR, van Campen JP, Verbeek MM, Jansen RW, et al. (2008) Amyloid beta protein and tau in cerebrospinal fluid and plasma as biomarkers for dementia: a review of recent literature. *Curr Clin Pharmacol* 3: 123–131.
39. Waxman EA, Giasson BI (2008) Molecular mechanisms of alpha-synuclein neurodegeneration. *Biochim Biophys Acta*;In press.
40. Marc D, Mercery R, Lantier F (2007) Scavenger, transducer, RNA chaperone? What ligands of the prion protein teach us about its function. *Cell Mol Life Sci* 64: 815–829.
41. Cheng Y, LeGall T, Oldfield CJ, Mueller JP, Van YY, et al. (2006) Rational drug design via intrinsically disordered protein. *Trends Biotechnol* 24: 435–442.
42. Gunasekaran K, Tsai CJ, Nussinov R (2004) Analysis of ordered and disordered protein complexes reveals structural features discriminating between stable and unstable monomers. *J Mol Biol* 341: 1327–1341.
43. Meszaros B, Tompa P, Simon I, Dosztanyi Z (2007) Molecular principles of the interactions of disordered proteins. *J Mol Biol* 372: 549–561.
44. Vacic V, Oldfield CJ, Mohan A, Radivojac P, Cortese MS, et al. (2007) Characterization of molecular recognition features, MoRFs, and their binding partners. *J Proteome Res* 6: 2351–2366.
45. Vucetic S, Brown CJ, Dunker AK, Obradovic Z (2003) Flavors of protein disorder. *Proteins* 52: 573–584.
46. Schlessinger A, Punta M, Rost B (2007) Natively unstructured regions in proteins identified from contact predictions. *Bioinformatics* 23: 2376–2384.
47. Garner E, Romero P, Dunker AK, Brown C, Obradovic Z (1999) Predicting binding regions within disordered proteins. *Genome Inform Ser Workshop Genome Inform* 10: 41–50.
48. Cheng Y, Oldfield CJ, Meng J, Romero P, Uversky VN, et al. (2007) Mining alpha-helix-forming molecular recognition features with cross species sequence alignments. *Biochemistry* 46: 13468–13477.
49. Romero, Obradovic, Dunker K (1997) Sequence data analysis for long disordered regions prediction in the calcineurin family. *Genome Inform Ser Workshop Genome Inform* 8: 110–124.
50. Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, et al. (2001) Sequence complexity of disordered protein. *Proteins* 42: 38–48.
51. Li X, Romero P, Rani M, Dunker AK, Obradovic Z (1999) Predicting protein disorder for N-, C-, and internal regions. *Genome Inform Ser Workshop Genome Inform* 10: 30–40.
52. Oldfield CJ, Cheng Y, Cortese MS, Romero P, Uversky VN, et al. (2005) Coupled folding and binding with alpha-helix-forming molecular recognition elements. *Biochemistry* 44: 12454–12470.
53. Dosztanyi Z, Csizmok V, Tompa P, Simon I (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol* 347: 827–839.
54. Dosztanyi Z, Csizmok V, Tompa P, Simon I (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* 21: 3433–3434.
55. Chumakov PM (2007) Versatile functions of p53 protein in multicellular organisms. *Biochemistry (Mosc)* 72: 1399–1421.
56. Dawson R, Muller L, Dehner A, Klein C, Kessler H, et al. (2003) The N-terminal domain of p53 is natively unfolded. *J Mol Biol* 332: 1131–1141.
57. Kussie PH, Gorina S, Marechal V, Elenbaas B, Moreau J, et al. (1996) Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain. *Science* 274: 948–953.
58. Bochkareva E, Kaustov L, Ayed A, Yi GS, Lu Y, et al. (2005) Single-stranded DNA mimicry in the p53 transactivation domain interaction with replication protein A. *Proc Natl Acad Sci U S A* 102: 15412–15417.
59. Di Lello P, Jenkins LM, Jones TN, Nguyen BD, Hara T, et al. (2006) Structure of the Tfb1/p53 complex: Insights into the interaction between the p62/Tfb1 subunit of TFIIF and the activation domain of p53. *Mol Cell* 22: 731–740.
60. Fawcett T (2006) An introduction to ROC analysis. *Pattern Recognit Lett* 27: 861–874.
61. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22: 2577–2637.
62. Russo AA, Jeffrey PD, Patten AK, Massague J, Pavletich NP (1996) Crystal structure of the p27Kip1 cyclin-dependent-kinase inhibitor bound to the cyclin A-Cdk2 complex. *Nature* 382: 325–331.
63. Galea CA, Nourse A, Wang Y, Sivakolundu SG, Heller WT, et al. (2008) Role of intrinsic flexibility in signal transduction mediated by the cell cycle regulator, p27 Kip1. *J Mol Biol* 376: 827–838.
64. Ochs HD, Notarangelo LD (2005) Structure and function of the Wiskott-Aldrich syndrome protein. *Curr Opin Hematol* 12: 284–291.
65. Kim AS, Kakalis LT, Abdul-Manan N, Liu GA, Rosen MK (2000) Autoinhibition and activation mechanisms of the Wiskott-Aldrich syndrome protein. *Nature* 404: 151–158.
66. Marchand JB, Kaiser DA, Pollard TD, Higgs HN (2001) Interaction of WASP/Scar proteins with actin and vertebrate Arp2/3 complex. *Nat Cell Biol* 3: 76–82.
67. Abdul-Manan N, Aghazadeh B, Liu GA, Majumdar A, Ouerfelli O, et al. (1999) Structure of Cdc42 in complex with the GTPase-binding domain of the ‘Wiskott-Aldrich syndrome’ protein. *Nature* 399: 379–383.
68. Puntrevoll P, Linding R, Gemund C, Chabanis-Davidson S, Mattingsdal M, et al. (2003) ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res* 31: 3625–3630.
69. Tong AH, Drees B, Nardelli G, Bader GD, Brannetti B, et al. (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science* 295: 321–324.
70. Fuxreiter M, Tompa P, Simon I (2007) Local structural disorder imparts plasticity on linear motifs. *Bioinformatics* 23: 950–956.
71. Preston CM, Wu KY, Molinski TF, DeLong EF (1996) A psychrophilic crenarchaeon inhabits a marine sponge: *Cenarchaeum symbiosum* gen. nov., sp. nov. *Proc Natl Acad Sci U S A* 93: 6241–6246.
72. Perez-Brocail V, Gil R, Ramos S, Lamelas A, Postigo M, et al. (2006) A small microbial genome: the end of a long symbiotic relationship? *Science* 314: 312–313.
73. Nakabachi A, Yamashita A, Toh H, Ishikawa H, Dunbar HE, et al. (2006) The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. *Science* 314: 267.
74. Wu D, Daugherty SC, Van Aken SE, Pai GH, Watkins KL, et al. (2006) Metabolic complementarity and genomics of the dual bacterial symbiosis of sharpshooters. *PLoS Biol* 4: e188. doi:10.1371/journal.pbio.0040188.
75. Hanisch FG, Muller S (2000) MUC1: the polymorphic appearance of a human mucin. *Glycobiology* 10: 439–449.
76. Zhang Y, Stec B, Godzik A (2007) Between order and disorder in protein structures: analysis of “dual personality” fragments in proteins. *Structure* 15: 1141–1147.
77. Neduva V, Russell RB (2005) Linear motifs: evolutionary interaction switches. *FEBS Lett* 579: 3342–3345.
78. Kiss R, Kovács D, Tompa P, Perczel A (2008) Local structural preferences of calpastatin, the intrinsically unstructured protein inhibitor of calpain. *Biochemistry* 47: 6936–6945.
79. Hurley TD, Yang J, Zhang L, Goodwin KD, Zou Q, et al. (2007) Structural basis for regulation of protein phosphatase 1 by inhibitor-2. *J Biol Chem* 282: 28874–28883.
80. Sigalov A, Aivazian D, Stern L (2004) Homo-oligomerization of the cytoplasmic domain of the T cell receptor zeta chain and of other proteins containing the immunoreceptor tyrosine-based activation motif. *Biochemistry* 43: 2049–2061.
81. Hansen JC, Lu X, Ross ED, Woody RW (2006) Intrinsic protein disorder, amino acid composition, and histone terminal domains. *J Biol Chem* 281: 1853–1856.
82. Sampietro J, Dahlberg CL, Cho US, Hinds TR, Kimelman D, et al. (2006) Crystal structure of a beta-catenin/BCL9/Tcf4 complex. *Mol Cell* 24: 293–300.
83. De Biasio A, Guarnaccia C, Popovic M, Uversky VN, Pintar A, et al. (2008) Prevalence of intrinsic disorder in the intracellular region of human single-pass type I proteins: the case of the notch ligand delta-4. *J Proteome Res* 7: 2496–2506.
84. Dosztanyi Z, Sandor M, Tompa P, Simon I (2007) Prediction of protein disorder at the domain level. *Curr Protein Pept Sci* 8: 161–171.
85. Dosztanyi Z, Tompa P (2008) Prediction of protein disorder. *Methods Mol Biol* 426: 103–115.
86. Ferron F, Longhi S, Canard B, Karlin D (2006) A practical overview of protein disorder prediction methods. *Proteins* 65: 1–14.
87. Rost B, Liu J, Nair R, Wrzeszczynski KO, Ofra Y (2003) Automatic prediction of protein function. *Cell Mol Life Sci* 60: 2637–2650.
88. Mohan A, Oldfield CJ, Radivojac P, Vacic V, Cortese MS, et al. (2006) Analysis of molecular recognition features (MoRFs). *J Mol Biol* 362: 1043–1059.
89. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. *Nucleic Acids Res* 28: 235–242.
90. Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, et al. (2007) DisProt: the Database of Disordered Proteins. *Nucleic Acids Res* 35: D786–D793.
91. Huber AH, Weis WI (2001) The structure of the beta-catenin/E-cadherin complex and the molecular basis of diverse ligand recognition by beta-catenin. *Cell* 105: 391–402.
92. Hertzog M, van Heijenoort C, Didry D, Gaudier M, Coutant J, et al. (2004) The beta-thyosin/WH2 domain; structural basis for the switch from inhibition to promotion of actin assembly. *Cell* 117: 611–623.
93. Sorenson MK, Ray SS, Darst SA (2004) Crystal structure of the flagellar sigma/anti-sigma complex sigma(28)/FlgM reveals an intact sigma factor in an inactive conformation. *Mol Cell* 14: 127–138.
94. Fauchere JL, Charton M, Kier LB, Verloop A, Pliska V (1988) Amino acid side chain parameters for correlation studies in biology and pharmacology. *Int J Pept Protein Res* 32: 269–278.

# IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding

Bálint Mészáros, Gábor Erdős and Zsuzsanna Dosztányi\*

MTA-ELTE Momentum Bioinformatics Research Group, Department of Biochemistry, Eötvös Loránd University, Budapest H-1117, Hungary

Received February 26, 2018; Revised April 17, 2018; Editorial Decision April 28, 2018; Accepted May 11, 2018

## ABSTRACT

The structural states of proteins include ordered globular domains as well as intrinsically disordered protein regions that exist as highly flexible conformational ensembles in isolation. Various computational tools have been developed to discriminate ordered and disordered segments based on the amino acid sequence. However, properties of IDRs can also depend on various conditions, including binding to globular protein partners or environmental factors, such as redox potential. These cases provide further challenges for the computational characterization of disordered segments. In this work we present IUPred2A, a combined web interface that allows to generate energy estimation based predictions for ordered and disordered residues by IUPred2 and for disordered binding regions by ANCHOR2. The updated web server retains the robustness of the original programs but offers several new features. While only minor bug fixes are implemented for IUPred, the next version of ANCHOR is significantly improved through a new architecture and parameters optimized on novel datasets. In addition, redox-sensitive regions can also be highlighted through a novel experimental feature. The web server offers graphical and text outputs, a RESTful interface, access to software download and extensive help, and can be accessed at a new location: <http://iupred2a.elte.hu>.

## INTRODUCTION

Intrinsically disordered proteins and protein regions (IDPs/IDRs) carry out important biological functions without relying on a single well-defined conformation, defying the traditional structure-function paradigm (1). Such regions are best characterized as ensembles of highly fluctuating conformations in isolation but their

detailed properties are delicately tailored for their specific function (2). The activities of IDPs can directly emerge from their flexible nature, exhibiting entropic chain functions or serving as linkers between ordered domains. Disordered proteins can also mediate protein-protein interactions by recognizing specific partners and undergo a disorder-to-order transition by adopting a more structured conformation. Such disordered binding regions or MoRFs (molecular recognition features) commonly occur in modular proteins involved in signaling and regulation (3,4). The specific properties of these compact functional modules, such as their plasticity and flexibility, enable their regulation depending on cellular cues through various mechanisms including post-translational modifications (PTMs) or competitive binding (5). While the majority of known disordered binding regions lose their flexibility upon interaction (with the exception of fuzzy complexes (6,7)), an order-to-disorder transition is the key for the function of another group of proteins. These conditionally disordered proteins are folded in isolation but their functional state requires a local or global unfolding to a more disordered state. The transition can be induced by interactions with other macromolecules or changes in environmental factors, such as pH, temperature or redox potential (8). One example for such conditional disorder is presented by Hsp33 from *Escherichia coli*. This redox-sensing chaperone becomes active upon oxidative stress, which induces a transition to a more disordered state exposing the substrate binding surface of the protein (9).

The growing number of examples of experimentally verified disordered segments are collected into dedicated databases, such as the DisProt database, which currently holds 2,167 such disordered regions from 803 proteins (10). However, these entries only provide a small sample of IDPs/IDRs that are widespread in all domains of life but are most prevalent in eukaryotic organisms (11–14). At this scale, protein disorder can only be studied through computational approaches. The distinct sequence properties of IDPs compared to that of globular proteins enable the discrimination of these two groups at the amino acid se-

\*To whom correspondence should be addressed. Tel: +36 1 372 2500; Fax: +36 1 372 8537; Email: dosztanyi@caesar.elte.hu

quence level at reasonable accuracies. So far, over 50 prediction methods have been developed using a wide arsenal of approaches, including simple amino acid propensity scales, simplified biophysical models, machine learning techniques and meta-servers (15–17). IUPred is one of the commonly used methods for predicting protein disorder and it is based on capturing the basic biophysical properties of IDPs (18,19). The basic assumption of this method is that intrinsically disordered proteins have a specific amino acid composition that does not allow the formation of enough favorable inter-residue interactions to stabilize a well-defined structural state (20,21). In IUPred, the interaction capacity of each residue is captured by an energy estimation scheme. While there are other methods that can achieve higher accuracies on particular datasets, IUPred still provides robust predictions with a favorable trade-off between speed and accuracy (22–24). As a result, IUPred is frequently used in itself or in combination with other tools to provide information about disorder (25).

The next challenge following the prediction of protein disorder is the characterization of the functional properties of IDPs/IDRs. Towards this end, most efforts focused on predicting regions of disordered proteins that are involved in protein-protein interactions, although methods that aim to predict regions binding to DNA and RNA, or to recognize linker regions have also been developed (26,27). The first publicly available method developed to recognize disordered binding regions was ANCHOR (28,29). Similarly to IUPred, this method relies on the energy estimation approach to characterize the disordered tendency and binding capacity of protein segments. Apart from ANCHOR, machine learning methods, in particular support vector machines (SVM) have also been developed for the prediction of disordered binding regions. MoRFpred and fMoRFpred utilize SVM models in their predictions incorporating sequence conservation data and amino-acid physicochemical properties, in addition to predictions of intrinsic disorder, relative solvent accessibility and residue flexibility (30,31). MFSPSSMpred and DISOPRED3 predict MoRFs based on an SVM with a radial basis function kernel, and using sequence-derived features and evolutionary profiles as inputs (32,33). MoRFchibi also employs SVMs, but uses a dual architecture to efficiently discriminate short MoRF regions from their flanking regions and to recognize similarity to already known instances (34).

The precision of the computational identification of disordered binding regions is usually evaluated against predicting such regions within globular proteins. However, these prediction methods should also have a discriminatory power against disordered regions in general. The main challenge is that currently we do not have a clear idea about the prevalence of disordered binding regions in proteins in general. One well-characterized example, p53 shows a nearly complete coverage by overlapping binding regions within its N- and C-terminal disordered segments (35). Other examples suggest that this could be a common scenario for many IDPs/IDRs, however, methods are often evaluated on proteins with a single known disordered binding site. A further limitation for accurate method development originates from a limited set of well-characterized examples used for training and testing. As a result, larger datasets were re-

sorted to PDB complexes formed between short and longer segments, assuming that the short segments are usually associated with disorder (30). However, this approach resulted in noisy datasets without experimental verification. In this regard, a major new development was the launch of the DIBS database, which collects protein complexes where one partner was shown experimentally to be both disordered in isolation and being involved in disorder-to-order transition (36). This database currently contains 773 entries, providing a reliable platform for further method development for recognizing disordered binding regions.

Conditionally disordered regions provide further computational challenges for the characterization of IDPs (8). An important category in this class corresponds to redox potential regulated proteins that play important roles in oxidant signalling and protein biogenesis events (37). Fascinating examples, such as Hsp33(9), COX17(38) or CP12 (39) indicate that redox sensing can be coupled to disorder-to-order or order-to-disorder transitions. While the limited number of such cases currently prevents systematic analyses, we found that the biophysical model of IUPred is already equipped to highlight redox-sensitive regions in proteins.

Recently, we have relocated our web-server IUPred to a new location (25). This gave us access to further improvements. Here, we describe the IUPred2A web server, which provides a combined interface to collect predictions for disordered regions via an improved version of IUPred, disordered binding segments via a new version of ANCHOR, and can highlight redox-sensitive regions in proteins based on the energy estimation method. These predictions can be accessed through an HTML server, a RESTful web server and as a downloadable software.

## METHODS

### IUPred2

IUPred uses an energy estimation method at its core. This approach utilizes a low-resolution statistical potential to characterize the tendencies of amino acid pairs to form contacts, observed in a collection of globular protein structures (40). When the structure is known, the statistical potential allows the calculation of the energy for each residue based on its interactions with other contacting residues in the structure. The sum of these residue-level energy terms can be used to quantify the total stabilizing energy contribution of intrachain interactions in a given protein structure. To open up a way to estimate these energies directly from the amino acid sequence without a known structure, a novel method was developed (18). In this model, the energy of each residue in the amino acid sequence is estimated based on the following formula:

$$e_i^k = \sum_{j=1}^{20} P_{ij} c_j^k,$$

where  $e_i^k$  is the energy of the residue in position  $k$  of type  $i$ ,  $P_{ij}$  is the  $ij$ th element of the energy predictor matrix, and  $c_j$  is the  $j$ th element of amino acid composition vector, specifying the ratio of amino acid type  $j$  in the sequence neighbourhood of position  $k$ .  $\mathbf{P}$  is a  $20 \times 20$  energy predictor matrix that connects the amino acid composition vector to

the energy of the given residue. Its parameters were optimized on a set of globular proteins to minimize the difference between the energy calculated from the known structures using the statistical potential and the energy estimated from the amino acid sequence. Based on the energy estimation, residues that have favorable energies are predicted as ordered and residues with unfavorable energies are predicted as disordered. The energies calculated for each residue in the amino acid sequence are smoothed with the window size ( $w^0$ ) and are transformed into a score between 0 and 1, so they can be interpreted as quasi-probabilities of a given residue being disordered.

The resulting method, IUPred (19) is able to recognize regions of proteins that are not compatible with ordered regions based on their inability to form enough favorable intrachain interactions. As the method relies on a low-resolution biophysical model of protein folding, its parameters are easily interpretable. Furthermore, calculations involve only simple arithmetics and as a result IUPred not only makes reliable and highly robust predictions, but is currently one of the fastest available disorder prediction algorithms, making it especially suited for large-scale studies.

In the current version, IUPred2, the force field and the architecture of the method were left unchanged. However, integration into several resources, such as MobiDBlite (41), MobiDB 3.0 (42) and InterPro (43) made it necessary to implement several minor bug-fixes. IUPred2 was tested on both the original testing sets of disordered and globular structures (18), and the newest version of DisProt (10) as a positive testing set, and a custom-built negative testing set of single domain ordered proteins with known structures (see Supplementary material). The efficiencies of IUPred2 and the original IUPred are consistent with earlier independent testing results (22,24), and are virtually the same. This is evidenced by the high similarities between the two receiver operating characteristic (ROC) curves of the two algorithms on both pairs of testing datasets (see Supplementary material for the ROC curves), with the areas under the curves being nearly identical (AUC = 0.855 and 0.856 for IUPred2 and IUPred on the new testing sets, and AUC = 0.924 and 0.926 on the original testing sets). From a practical point of view, these efficiencies correspond to true positive rates of 59.6% and 68.72% when using IUPred2 with 5% and 10% false positive rates, respectively, on the new testing sets.

## ANCHOR2

Similarly to IUPred, ANCHOR also utilizes the energy estimation approach, for the identification of disordered binding sites. Besides the general disorder tendency, two additional terms were also incorporated into the method that estimate the energy associated with interaction with a globular protein and with the local disordered sequence environment (28). These tendencies were combined using a linear combination and were transformed to yield a normalized score between 0 and 1 representing the probability of a given residue being part of a disordered binding region. In the presented IUPred2A server, ANCHOR was substantially reworked to give better predictions.

*Concept and architecture of ANCHOR2.* Retaining the original idea behind ANCHOR, the new ANCHOR2 methods also employs a simple biophysics-based model to describe disordered binding regions. In this framework, residues belonging to disordered binding sites have to fulfill two distinct criteria: (i) they have to be able to form favourable interactions with the binding surface of an ordered protein and (ii) they should be embedded in a generally disordered sequence environment. These two criteria are formulated as follows:

$$S_k = (E_{\text{gain},k}(w_1) - E_{\text{gain},0})(I_k(w_2) - I_0),$$

where  $S_k$  is the score assigned to residue  $k$ ;  $E_{\text{gain},k}(w_1) = E_{\text{loc},k}(w_1) - E_{\text{int},k}$  is the energy the residue gains by making interactions with an averaged ordered interacting surface (represented by the composition vector  $E_{\text{int}}$ ) instead of its own sequential environment (represented by the composition vector  $E_{\text{loc},k}(w_1)$ , calculated in a  $w_1$  half-window sequential neighborhood of residue  $k$ );  $I_k(w_2)$  is the averaged IUPred score in the  $w_2$  half-window sequential neighborhood of residue  $k$ ;  $E_{\text{gain},0}$  and  $I_0$  are parameters that determine the minimum energy gain and minimum average disorder tendency a residue has to possess in order to become a disordered binding site. The sign of  $E_{\text{gain}}$  is chosen in a way that high positive values mark true binding residues (as usually expected from prediction methods), which is different from the standard choice for true free energy. Keeping this in mind, the architecture of ANCHOR2 has a clear biophysical meaning and contains only four parameters ( $w_1$ ,  $w_2$ ,  $E_{\text{gain},0}$  and  $I_0$ ) that need to be optimized during training.

*Training and benchmarking.* ANCHOR2 was trained and tested using the disordered binding regions in the DIBS database (36) filtered for 30% sequence identity as the positive set, using only short binding regions below 30 residues yielding a total of 374 protein regions. Four distinctively different datasets were used as negative (see Supplementary material). The first negative dataset (*ordered monomers*) comprises sequence regions (also filtered for 30% sequence identity) that encode single structural domains with determined monomeric structures in the PDB (4,549 protein regions). The second dataset contains 389 *flexible linker* regions, used previously in the assessment of DISOPRED3(33). These two datasets can be considered as verified in a sense that they are unlikely to contain currently unknown disordered binding regions. The third dataset (*decoy sequences*) were collected as  $\sim 15,000$  protein segments taken randomly from the human proteome, excluding extracellular proteins, transmembrane regions and known structural Pfam domains to increase the expected ratio of disordered regions. The fourth negative dataset contains 1,042 known disordered protein regions from the *DisProt* database (10) that do not overlap with entries in DIBS. These two datasets cannot be assumed to be devoid of currently unknown disordered binding regions (unverified datasets). However, for parameter optimization and testing, the positive dataset, the ordered monomer set and the decoy set were split, and two thirds of all three were used in training and the remaining one third was used in testing.

During training the four adjustable parameters  $w_1$ ,  $w_2$ ,  $E_{\text{gain},0}$  and  $I_0$  were tuned to their optimal values. The  $E_{\text{gain}}$  term of the score basically describes the distinction between disordered binding regions and other sequence regions in general (non-binding disordered segments in particular). In accord,  $w_1$  was set to achieve the highest information gain (similarly to the protocol employed in (44)) calculated in the separation of the positive and decoy training sets (see Supplementary material). While the decoy set can in theory contain any number of disordered binding regions, due to the random assignment we expect their numbers to be fairly low. In contrast to the energy gain term, the  $I$  term of the score primarily describes the separation between disordered binding regions and ordered proteins. Thus,  $w_2$  was set to achieve the highest information gain in the distinction between the positive and the ordered monomer training sets (see Supplementary material). As a final step,  $E_{\text{gain},0}$  and  $I_0$  were also set to best discriminate the elements of the positive and the two negative training sets.

Testing of ANCHOR2 was done by calculating residue-based ROC curves evaluating the ability of the method to separate the testing positive dataset from any of the four negative testing datasets. To better gain insights into the strengths and weaknesses of ANCHOR2, three other methods capable of predicting disordered binding regions: the original ANCHOR, DISOPRED3 and MoRFchibi, were also evaluated on the same datasets. The obtained ROC curves for all four negative testing sets are shown in Figure 1, while the calculated AUC values for all methods for all datasets are shown in Table 1. The obtained efficiencies of the four methods outline the clear differences between their applicability. Both DISOPRED3 and MoRFchibi are machine learning based methods and were trained to have very low false positive prediction rates in both ordered and disordered protein regions. However, this comes at the expense of recognizing disordered binding regions that are not similar to currently known ones. ANCHOR and ANCHOR2 on the other hand incorporate a direct description of protein disorder in their predictions and thus excel at giving an extremely low false positive rate on ordered protein regions. They are also remarkable at distinguishing flexible linkers, but predict a higher ratio of disordered binding sites in generic disordered protein datasets, such as DisProt. While this may involve over-prediction, it is worth noting that the exact number of true disordered binding regions in DisProt sequences are not known and thus it is hard to determine the optimal behaviour of disorder binding site predictions on these data.

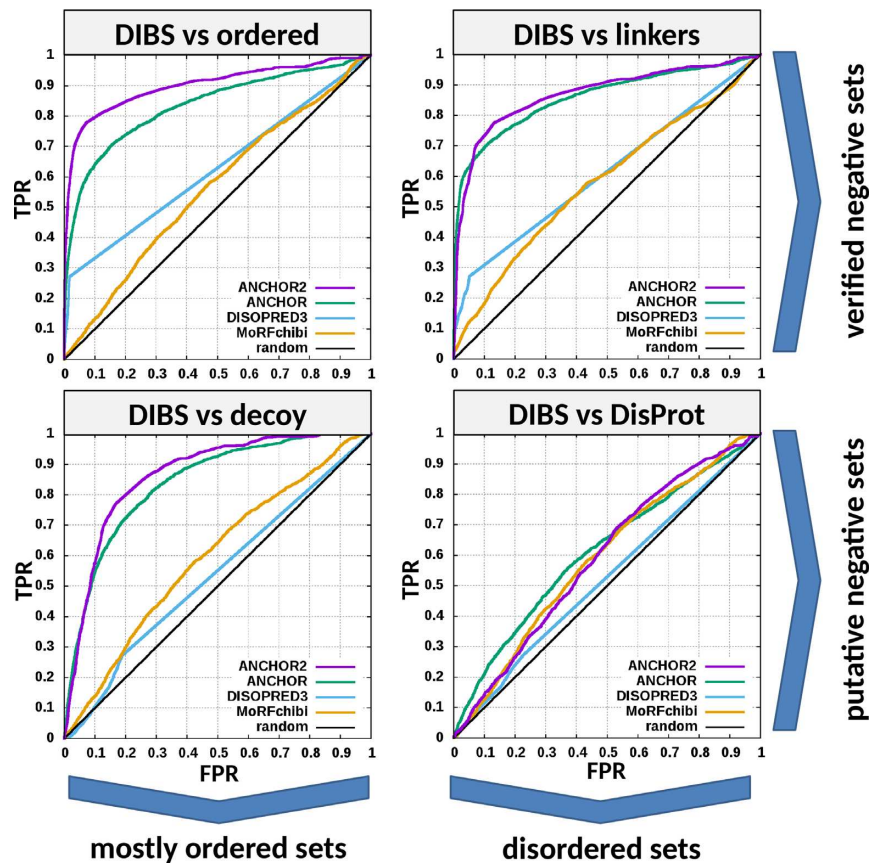
As a final step, the prediction score of ANCHOR was normalized to fall between 0 and 1 in such a way that the ratio of binding residues stayed below 50% even in the DisProt database, where it was the highest among the negative datasets. Using this threshold, the ratio of residues predicted to be binding in the positive and negative datasets is shown in Table 2. While this reduces the apparent efficiency of ANCHOR2 as compared to the scaling used in the original ANCHOR (the 0.5 cutoff corresponded to 5% false positive prediction on ordered protein segments), ANCHOR2 is still able to correctly predict nearly 64% of residues in known binding regions (true positive rate), with over 72%

of known binding regions harboring at least one correctly predicted residue (segment-level true positive rate).

### Redox-state dependent prediction of protein disorder

In another group of conditionally disordered proteins, changes of the oxidation status are coupled to disorder-to-order or order-to-disorder transitions (37). One example for this behaviour is provided by the human small copper chaperone Cox17. This protein can be viewed as a prototype for proteins that are synthesized on cytosolic ribosomes and diffuse as intrinsically disordered proteins to the mitochondrial intermembrane space, where they become oxidized and fold into their functional conformations (38). The activity of Hsp33 also depends on oxidative conditions, however, for this protein the functional state is disordered. Under non-stress conditions, Hsp33 is a compactly folded zinc-binding protein with negligible activity. Oxidative stress causes the formation of two intramolecular disulfide bonds and the release of  $\text{Zn}^{2+}$  ions. This leads to the unfolding of the zinc-binding domain, exposing the substrate binding surface of the chaperone that is necessary for its activity (45).

The key sensors built into these redox-regulated proteins are cysteine residues which can undergo reversible thiol oxidation in response to the oxidation status of the molecular environment. Under reducing conditions cysteine residues can behave as polar amino acids, most similar to serine, without contributing much to protein stability. However, they can also play essential roles in stabilizing the folded conformation by coordinating  $\text{Zn}^{2+}$  ions under reducing conditions, or by forming disulfide bonds that are commonly used by extracellular proteins that experience oxidative conditions (46). In our energy estimation scheme, the strong stabilizing feature of cysteine residues can be adequately captured, with the most extreme energy terms corresponding to interactions mediated by cysteine residues. In order to capture the other end of the spectrum, cysteine residues can be changed to serine in the amino acid sequence. Thus, we generate two disorder prediction profiles, one corresponding to the state that is achieved through cysteine stabilization (redox-plus) and one without cysteine stabilization (redox-minus), modeled by a cysteine/serine swap. In many cases the two profiles would not differ significantly. However, our assumption is that in the case of conditionally disordered redox proteins the two profiles would be separated and would highlight redox-sensitive regions based on their different disorder tendencies. These regions are defined when the redox-minus line predicts disorder for a minimal region of 10 residues, while no disorder is predicted for the same region by the redox-plus profile. This core region is then extended in both directions to the point where the separation in the disorder score between the two lines falls below 0.15. Thus identified redox-sensitive regions are merged if their sequence separation is less than 10 residues (for details see Supplementary material). While this feature of IUPred2A cannot be tested rigorously, examples provided in later sections and on the server help pages indicate that the prediction of redox-sensitive regions can be used to explore this phenomenon at the large-scale. Our preliminary data suggests that redox sensitive regions can be



**Figure 1.** ROC curves of four methods predicting disordered binding regions/MoRFs on the four different negative testing datasets. The upper row shows testing on verified negative data containing virtually no disordered binding regions. Negative sets of the bottom row might contain an unknown number of disordered binding regions, albeit with a significantly lower frequency compared to the positive set.

**Table 1.** Area under the curve (AUC) values calculated from the ROC curves in Figure 1

		Methods			
		ANCHOR2	ANCHOR	DISOPRED3	MoRFchibi
<b>Datasets</b>	<b>ordered monomers</b>	<b>0.901</b>	0.835	0.627	0.561
	<b>linkers</b>	<b>0.870</b>	0.859	0.612	0.581
	<b>decoy</b>	<b>0.865</b>	0.840	0.536	0.595
	<b>DisProt</b>	0.590	<b>0.610</b>	0.522	0.588

AUC values can range from 0.5 for random predictions to 1 for perfect predictions. The highest AUC values for each negative dataset are highlighted in bold.

**Table 2.** Prediction rates of ANCHOR2 on training and testing datasets

Dataset name	Dataset type	Fraction of residues predicted to be disordered binding regions by ANCHOR2
<b>DIBS training</b>	Verified positive	57.31% (66.40% at segment level)
<b>DIBS testing</b>	Verified positive	63.83% (72.58% at segment level)
<b>Ordered monomers training</b>	Verified negative	2.38%
<b>Ordered monomers testing</b>	Verified negative	2.44%
<b>Linker regions</b>	Verified negative	6.03%
<b>Decoy training</b>	Putative negative	10.69%
<b>Decoy testing</b>	Putative negative	11.55%
<b>DisProt</b>	Putative negative	50%

Datasets were evaluated using 0.5 cutoff to discriminate between disordered binding regions and non-binding residues.

quite common in the human proteome: the few experimentally characterized examples indicate that how this redox sensitivity is used in biological context can be more complex and can be fully understood only based on further experiments.

## SERVER DESCRIPTION

### Input

To ease the transition of users of the original IUPred server, the user interface of IUPred2A inherits a lot from its predecessor, enabling fast and straightforward usage. The main page features entry boxes, which accept a FASTA formatted or plain protein sequence, or any valid UniProt accession/ID. The sequences of corresponding UniProt entries are accessed through an SQL database containing information about the specified input, or extract the information directly from UniProt, in case of an SQL database failure. In addition, a multi-FASTA formatted file with a maximum size of 200 megabytes can also be uploaded. The new web-server also incorporates RESTful services using custom links for searches. For IUPred2 predictions, three types of predictions can be chosen depending on the type of structural regions the user wants to analyse: short stretches of disorder (such as flexible loops or linkers), long disordered regions (such as disordered domains), or structured domains. These options are directly inherited from the previous IUPred implementation (19,25). In addition, IUPred2A features optional context dependent prediction options, using either ANCHOR2 for the identification of disordered binding sites, or the redox-sensitive feature to uncover redox potential dependent disorder. Once the proper inputs are filled, the server calculates the results on a Django 2.0 based back-end. Each prediction is calculated on-the-fly server side, utilizing the latest MPI technology for maximum efficiency. To ease the load on the server, multi-FASTA uploads are treated separately and are queued until the server has enough free capacity.

### Output

The latest version of Bokeh (0.12.14) is responsible for the visualization of the results that is directly integrated into the Django framework. The graphical output presents the requested predictions. By default, it contains disorder predictions from IUPred2 and binding site predictions from ANCHOR2, but the individual predictions can be turned on and off on the plot. Alternatively, the redox-sensitive regions are highlighted. Integration with the UniProt resource enables the display of various additional information about the requested protein (when available), such as PFAM annotations (47), low-throughput post-translational modifications (including phosphorylation, methylation and acetylation sites) from PhosphoSitePlus (48), related structures from the PDB (49) and experimentally verified disordered regions from three different databases: generic disorder from DisProt (10) and disordered binding regions from DIBS (36) and MFIB (50). Besides the visual output, both text based and JSON formatted outputs are downloadable for each prediction. Despite the intensive use of cutting-

edge web technologies, IUPred2A supports all HTML5 and WebP compatible browsers.

### Supporting features

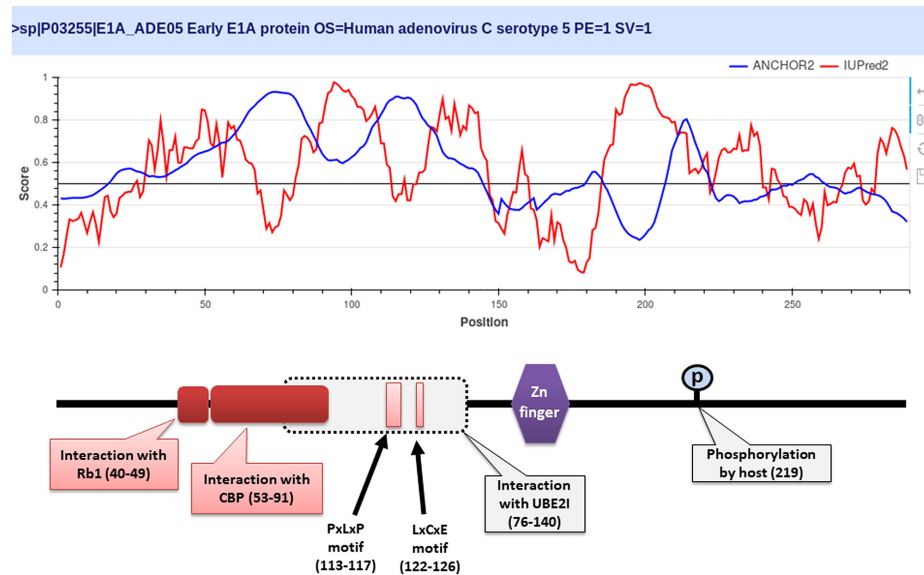
To further enhance the usability of IUPred2A, the site features the description of the method, together with various examples that highlight its functionality and aid the correct interpretation of the results. Furthermore, IUPred2A also supports the local use of IUPred2 and ANCHOR2, as both methods are available for download as Python3 codes.

## EXAMPLES

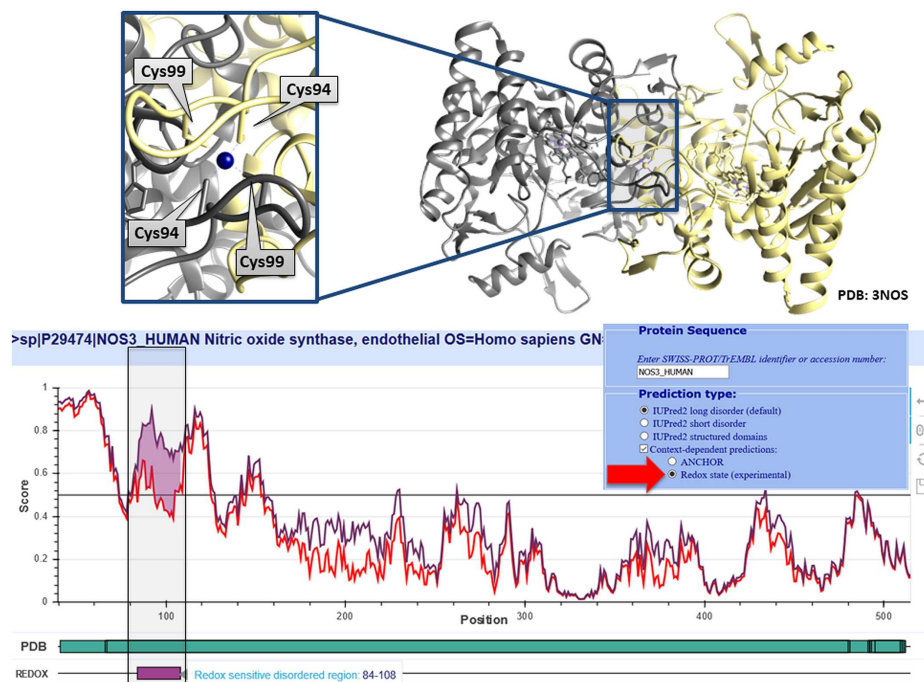
ANCHOR2 can correctly recognize many disordered binding regions that machine learning methods are likely to overlook due to their very conservative estimates of the occurrence of these functional modules. This is demonstrated through the example of the oncogenic Human adenovirus C early E1A protein (Figure 2). E1A is a largely disordered protein (51), which is essential for forcing the host cell into S phase via modulation of the Rb1/E2F1 pathway (52) and the inhibition of apoptosis via modulation of p53 degradation (53). These host-pathogen interactions are mediated by several binding events. Rb1 and CBP are targeted by two N-terminal tandem binding sites with determined complex structures deposited in the PDB. These known disordered binding regions are identified by ANCHOR2 as two distinct neighbouring peaks in the output score. While no other E1A-human protein complexes are currently known in structural detail, E1A harbors two additional known motifs capable of forming host-specific interactions. Both motifs, together with the putative binding site for the deubiquitinase UBE2I are correctly recognized by ANCHOR2 as a separate peak in the prediction score. A distinct peak C-terminal of the structured zinc-finger has no known binding partners; however it entails a serine residue that was shown to be phosphorylated by host kinases (54), hinting at an additional important binding region with currently limited characterization.

In the case of disordered binding regions, the transition between the disordered and the folded state is induced by the presence of a protein partner. However, in certain cases both the structural state and molecular interactions can be influenced by redox potential. A prime example of such behaviour is presented by the endothelial nitric oxide synthase (NOS3). Dimerization of this protein is essential for its oxidoreductase activity. The dimer interface is formed through a Zn<sup>2+</sup>-cysteine complex, where Cys94 and Cys99 from each subunit coordinate the Zn<sup>2+</sup>. These cysteines appeared susceptible to redox modifications which promote a disulfide bond formation within each monomer and subsequent release of Zn<sup>2+</sup>. This results in the disruption of the dimer and a transition to the monomeric state, paralleled by the disruption of the enzyme activity (46,55). Figure 3 shows the prediction for NOS3 generated using the experimental redox-state option of IUPred2, correctly capturing the redox-sensitive region involved in this structural transition.





**Figure 2.** The output of IUPred2 and ANCHOR2 for the oncogenic Human adenovirus C early E1A protein. Top: IUPred2 and ANCHOR2 scores are shown in red and blue. Bottom: schematic architecture of E1A. Disordered binding regions with known complex structure are shown in deep red boxes. Light red boxes correspond to known linear motifs. Grey box marks the region sufficient for interaction with UBE2I.



**Figure 3.** The output of the redox-state dependent IUPred2 predictor for the N-terminal region of NOS3. Top: the coordination of  $Zn^{2+}$  by cysteines 94 and 99 from both chains in the dimeric NOS3 structure. Bottom: the output of IUPred2 using the redox state modeling option, where the estimated sensitivity of the disorder tendency is marked in purple. The plot is zoomed into the N-terminal region that can be seen in the dimeric complex (PDB: 3NOS).

## CONCLUSION

The current paper presents the new IUPred2A server that serves as a unified platform for both generic and context-dependent prediction of protein disorder. IUPred2A combines and supersedes our general disorder prediction method IUPred and disordered binding region prediction

method ANCHOR. While IUPred2 features only slight improvements over its predecessor, ANCHOR2 was completely re-trained and re-tested built on a new architecture, bringing a significant improvement over the original version. In addition, IUPred2A also incorporates a new experimental feature that targets the identification of protein regions capable of redox-state dependent transition between

disordered and ordered states. These methods are available through a completely rewritten server at a new location. The IUPred2A server retains all options for data input from previous versions, but also significantly expands its functionality by introducing RESTful services, and automated data integration from a range of databases with information about protein structure. Furthermore, completely rewritten codes for IUPred2 and ANCHOR2 are available for download to aid local large-scale analyses.

Concurrent machine learning algorithms typically excel at correctly predicting protein regions with a substantial similarity to training examples. However, owing to their biophysics-based models, IUPred2 and ANCHOR2 are expected to be able to correctly recognize protein regions that share limited to no resemblance to currently known disordered regions or binding sites. This, together with the fact that both IUPred and ANCHOR present virtually the fastest methods with high accuracies in their respective fields (56), make them outstandingly suited for de novo identification of binding- and non-binding disordered protein regions in large-scale studies.

While the computational identification of protein disorder in general has already been targeted by several methods, the possible context dependence of structural features has been generally overlooked from a prediction standpoint. IUPred2A presents the first attempt at the unified description of the context-dependence of protein disorder by being able to describe the lack of structure and its dependence on the presence of a partner protein or a change in redox environment. As IUPred2A is rooted in a biophysical model of molecular interactions, it holds the potential for the future extension of its architecture to successfully incorporate the effects of other structure-modifying environmental factors, such as pH or post-translational modifications.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

The authors are grateful to Domenico Cozzetto and David T. Jones for kindly providing the flexible linker dataset used for benchmarking DISOPRED3. The constructive remarks of László Dobson concerning IUPred2A functionality are gratefully acknowledged.

## FUNDING

Hungarian Academy of Sciences [LP2014-18 'Lendület']; Országos Tudományos Kutatási Alapprogramok [K108798]. Funding for open access charge: OTKA K108798.

*Conflict of interest statement.* None declared.

## REFERENCES

1. Wright, P.E. and Dyson, H.J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.*, **293**, 321–331.
2. van der Lee, R., Buljan, M., Lang, B., Weatheritt, R.J., Daughdrill, G.W., Dunker, A.K., Fuxreiter, M., Gough, J., Gsponer, J., Jones, D.T. *et al.* (2014) Classification of intrinsically disordered regions and proteins. *Chem. Rev.*, **114**, 6589–6631.
3. Mészáros, B., Tompa, P., Simon, I. and Dosztányi, Z. (2007) Molecular principles of the interactions of disordered proteins. *J. Mol. Biol.*, **372**, 549–561.
4. Vacic, V., Oldfield, C.J., Mohan, A., Radivojac, P., Cortese, M.S., Uversky, V.N. and Dunker, A.K. (2007) Characterization of molecular recognition features, MoRFs, and their binding partners. *J. Proteome Res.*, **6**, 2351–2366.
5. Van Roey, K., Gibson, T.J. and Davey, N.E. (2012) Motif switches: decision-making in cell regulation. *Curr. Opin. Struct. Biol.*, **22**, 378–385.
6. Borgia, A., Borgia, M.B., Bugge, K., Kissling, V.M., Heidarsson, P.O., Fernandes, C.B., Sottini, A., Soranno, A., Buholzer, K.J., Nettels, D. *et al.* (2018) Extreme disorder in an ultrahigh-affinity protein complex. *Nature*, **555**, 61–66.
7. Miskei, M., Antal, C. and Fuxreiter, M. (2017) FuzDB: database of fuzzy complexes, a tool to develop stochastic structure-function relationships for protein complexes and higher-order assemblies. *Nucleic Acids Res.*, **45**, D228–D235.
8. Jakob, U., Kriwacki, R. and Uversky, V.N. (2014) Conditionally and transiently disordered proteins: awakening cryptic disorder to regulate protein function. *Chem. Rev.*, **114**, 6779–6805.
9. Reichmann, D., Xu, Y., Cremers, C.M., Ilbert, M., Mittelman, R., Fitzgerald, M.C. and Jakob, U. (2012) Order out of disorder: working cycle of an intrinsically unfolded chaperone. *Cell*, **148**, 947–957.
10. Piovesan, D., Tabaro, F., Mičetić, I., Necci, M., Quaglia, F., Oldfield, C.J., Aspromonte, M.C., Davey, N.E., Davidović, R., Dosztányi, Z. *et al.* (2017) DisProt 7.0: a major update of the database of disordered proteins. *Nucleic Acids Res.*, **45**, D219–D227.
11. Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F. and Jones, D.T. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.*, **337**, 635–645.
12. Lobanov, M.Y. and Galzitskaya, O.V. (2015) How common is disorder? Occurrence of disordered residues in four domains of life. *Int. J. Mol. Sci.*, **16**, 19490–19507.
13. Peng, Z., Yan, J., Fan, X., Mizianty, M.J., Xue, B., Wang, K., Hu, G., Uversky, V.N. and Kurgan, L. (2015) Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. *Cell. Mol. Life Sci.*, **72**, 137–151.
14. Xue, B., Dunker, A.K. and Uversky, V.N. (2012) Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. *J. Biomol. Struct. Dyn.*, **30**, 137–149.
15. He, B., Wang, K., Liu, Y., Xue, B., Uversky, V.N. and Dunker, A.K. (2009) Predicting intrinsic disorder in proteins: an overview. *Cell Res.*, **19**, 929–949.
16. Dosztányi, Z., Mészáros, B. and Simon, I. (2010) Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins. *Brief. Bioinform.*, **11**, 225–243.
17. Meng, F., Uversky, V.N. and Kurgan, L. (2017) Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions. *Cell. Mol. Life Sci.*, **74**, 3069–3090.
18. Dosztányi, Z., Csizmók, V., Tompa, P. and Simon, I. (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.*, **347**, 827–839.
19. Dosztányi, Z., Csizmók, V., Tompa, P. and Simon, I. (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, **21**, 3433–3434.
20. Garbuzynskiy, S.O., Lobanov, M.Y. and Galzitskaya, O.V. (2004) To be folded or to be unfolded? *Protein Sci.*, **13**, 2871–2877.
21. Dosztányi, Z., Csizmók, V., Tompa, P. and Simon, I. (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.*, **347**, 827–839.
22. Peng, Z.-L. and Kurgan, L. (2012) Comprehensive comparative assessment of in-silico predictors of disordered regions. *Curr. Protein Pept. Sci.*, **13**, 6–18.
23. Walsh, I., Giollo, M., Di Domenico, T., Ferrari, C., Zimmermann, O. and Tosatto, S.C.E. (2015) Comprehensive large-scale assessment of intrinsic protein disorder. *Bioinformatics*, **31**, 201–208.

24. Necci, M., Piovesan, D., Dosztányi, Z., Tompa, P. and Tosatto, S.C.E. (2018) A comprehensive assessment of long intrinsic protein disorder from the DisProt database. *Bioinformatics*, **34**, 445–452.
25. Dosztányi, Z. (2018) Prediction of protein disorder based on IUPred. *Protein Sci.*, **27**, 331–340.
26. Peng, Z., Wang, C., Uversky, V.N. and Kurgan, L. (2017) Prediction of disordered RNA, DNA, and protein binding regions using DisoRDPbind. *Methods Mol. Biol.*, **1484**, 187–203.
27. Meng, F. and Kurgan, L. (2016) DFLpred: High-throughput prediction of disordered flexible linker regions in protein sequences. *Bioinformatics*, **32**, i341–i350.
28. Mészáros, B., Simon, I. and Dosztányi, Z. (2009) Prediction of protein binding regions in disordered proteins. *PLoS Comput. Biol.*, **5**, e1000376.
29. Dosztányi, Z., Mészáros, B. and Simon, I. (2009) ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics*, **25**, 2745–2746.
30. Disfani, F.M., Hsu, W.-L., Mizianty, M.J., Oldfield, C.J., Xue, B., Dunker, A.K., Uversky, V.N. and Kurgan, L. (2012) MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. *Bioinformatics*, **28**, i75–i83.
31. Yan, J., Dunker, A.K., Uversky, V.N. and Kurgan, L. (2016) Molecular recognition features (MoRFs) in three domains of life. *Mol. Biosyst.*, **12**, 697–710.
32. Fang, C., Noguchi, T., Tominaga, D. and Yamana, H. (2013) MFSPSSMpred: identifying short disorder-to-order binding regions in disordered proteins based on contextual local evolutionary conservation. *BMC Bioinformatics*, **14**, 300.
33. Jones, D.T. and Cozzetto, D. (2015) DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics*, **31**, 857–863.
34. Malhis, N. and Gsponer, J. (2015) Computational identification of MoRFs in protein sequences. *Bioinformatics*, **31**, 1738–1744.
35. Gibson, T.J. (2009) Cell regulation: determined to signal discrete cooperation. *Trends Biochem. Sci.*, **34**, 471–482.
36. Schad, E., Fichó, E., Pancsa, R., Simon, I., Dosztányi, Z. and Mészáros, B. (2018) DIBS: a repository of disordered binding sites mediating interactions with ordered proteins. *Bioinformatics*, **34**, 535–537.
37. Reichmann, D. and Jakob, U. (2013) The roles of conditional disorder in redox proteins. *Curr. Opin. Struct. Biol.*, **23**, 436–442.
38. Fraga, H., Pujols, J., Gil-Garcia, M., Roque, A., Bernardo-Seisdedos, G., Santambrogio, C., Bech-Serra, J.-J., Canals, F., Bernadó, P., Grandori, R. *et al.* (2017) Disulfide driven folding for a conditionally disordered protein. *Sci. Rep.*, **7**, 16994.
39. Gontero, B. and Maberly, S.C. (2012) An intrinsically disordered protein, CP12: jack of all trades and master of the Calvin cycle. *Biochem. Soc. Trans.*, **40**, 995–999.
40. Thomas, P.D. and Dill, K.A. (1996) An iterative method for extracting energy-like quantities from protein structures. *Proc. Natl. Acad. Sci. U.S.A.*, **93**, 11628–11633.
41. Necci, M., Piovesan, D., Dosztányi, Z. and Tosatto, S.C.E. (2017) MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins. *Bioinformatics*, **33**, 1402–1404.
42. Piovesan, D., Tabaro, F., Paladin, L., Necci, M., Micetic, I., Camilloni, C., Davey, N., Dosztányi, Z., Mészáros, B., Monzon, A.M. *et al.* (2018) MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins. *Nucleic Acids Res.*, **46**, D471–D476.
43. Finn, R.D., Attwood, T.K., Babbitt, P.C., Bateman, A., Bork, P., Bridge, A.J., Chang, H.-Y., Dosztányi, Z., El-Gebali, S., Fraser, M. *et al.* (2017) InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res.*, **45**, D190–D199.
44. Erdős, G., Szaniszló, T., Pajkos, M., Hajdu-Soltész, B., Kiss, B., Pál, G., Nyitray, L. and Dosztányi, Z. (2017) Novel linear motif filtering protocol reveals the role of the LC8 dynein light chain in the Hippo pathway. *PLoS Comput. Biol.*, **13**, e1005885.
45. Reichmann, D., Xu, Y., Cremers, C.M., Ilbert, M., Mittelman, R., Fitzgerald, M.C. and Jakob, U. (2012) Order out of disorder: working cycle of an intrinsically unfolded chaperone. *Cell*, **148**, 947–957.
46. Pace, N.J. and Weerapana, E. (2014) Zinc-binding cysteines: diverse functions and structural motifs. *Biomolecules*, **4**, 419–434.
47. Finn, R.D., Coghill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.
48. Hornbeck, P.V., Zhang, B., Murray, B., Kornhauser, J.M., Latham, V. and Skrzypek, E. (2015) PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.*, **43**, D512–D520.
49. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
50. Fichó, E., Reményi, I., Simon, I. and Mészáros, B. (2017) MFIB: a repository of protein complexes with mutual folding induced by binding. *Bioinformatics*, **33**, 3682–3684.
51. Ferreon, A.C.M., Ferreon, J.C., Wright, P.E. and Deniz, A.A. (2013) Modulation of allostery by protein intrinsic disorder. *Nature*, **498**, 390–394.
52. Egan, C., Bayley, S.T. and Branton, P.E. (1989) Binding of the Rb1 protein to E1A products is required for adenovirus transformation. *Oncogene*, **4**, 383–388.
53. Lowe, S.W. and Ruley, H.E. (1993) Stabilization of the p53 tumor suppressor is induced by adenovirus 5 E1A and accompanies apoptosis. *Genes Dev.*, **7**, 535–545.
54. Tremblay, M.L., McGlade, C.J., Gerber, G.E. and Branton, P.E. (1988) Identification of the phosphorylation sites in early region 1A proteins of adenovirus type 5 by amino acid sequencing of peptide fragments. *J. Biol. Chem.*, **263**, 6375–6383.
55. Zou, M.-H., Shi, C. and Cohen, R.A. (2002) Oxidation of the zinc-thiolate complex and uncoupling of endothelial nitric oxide synthase by peroxynitrite. *J. Clin. Invest.*, **109**, 817–826.
56. Malhis, N., Wong, E.T.C., Nassar, R. and Gsponer, J. (2015) Computational identification of MoRFs in protein sequences using hierarchical application of Bayes rule. *PLoS One*, **10**, e0141603.

Structural bioinformatics

## ANCHOR: web server for predicting protein binding regions in disordered proteins

Zsuzsanna Dosztányi\*, Bálint Mészáros and István Simon

Institute of Enzymology, Biological Research Center, Hungarian Academy of Sciences, PO Box 7, H-1518 Budapest, Hungary

Received on June 25, 2009; revised on August 7, 2009; accepted on August 23, 2009

Advance Access publication August 28, 2009

Associate Editor: Anna Tramontano

### ABSTRACT

**Summary:** ANCHOR is a web-based implementation of an original method that takes a single amino acid sequence as an input and predicts protein binding regions that are disordered in isolation but can undergo disorder-to-order transition upon binding. The server incorporates the result of a general disorder prediction method, IUPred and can carry out simple motif searches as well.

**Availability:** The web server is available at <http://anchor.enzim.hu>. The program package is freely available for academic users.

**Contact:** zsuzsa@enzim.hu

### 1 INTRODUCTION

Many disordered proteins contain important functional elements involved in protein–protein interactions. Disordered binding regions play a critical role in various biological processes, involving regulation and signaling (Dyson and Wright, 2002). These segments differ from protein interaction sites of globular proteins due to their distinct structural properties (Mészáros *et al.*, 2007). Such regions exist as a highly flexible structural ensemble in isolation and adopt a well-defined conformation only upon binding to their specific partner molecules. It was suggested that certain disorder prediction methods can be indicative of disordered binding regions (Garner *et al.*, 1999). Specialized methods have been developed to regions adopting  $\alpha$ -helical conformation in their bound state (Cheng *et al.*, 2007) or for the binding partners of calmodulin (Radivojac *et al.*, 2006). In contrast, ANCHOR is a general method for recognizing disordered binding regions.

ANCHOR aims to capture the basic biophysical properties of disordered binding regions using estimated energy calculations (Mészáros *et al.*, 2009). Estimated energies can be assigned to each residue in a sequence and were shown to well-approximate the corresponding energies calculated from known structures of globular proteins (Dosztányi *et al.*, 2005b). Generally, disordered regions can be discriminated from ordered proteins by unfavorable estimated energies. This concept is utilized in the IUPred server for the prediction of protein disorder (Dosztányi *et al.*, 2005a). The estimated energies can also detect regions that are likely to gain energetically by interacting with globular proteins. Predictions in ANCHOR combine the general disorder tendency with the sensitivity to the structural environment (Mészáros *et al.*, 2009).

\*To whom correspondence should be addressed.

Because of this additional property, ANCHOR scores are relatively independent from IUPred scores.

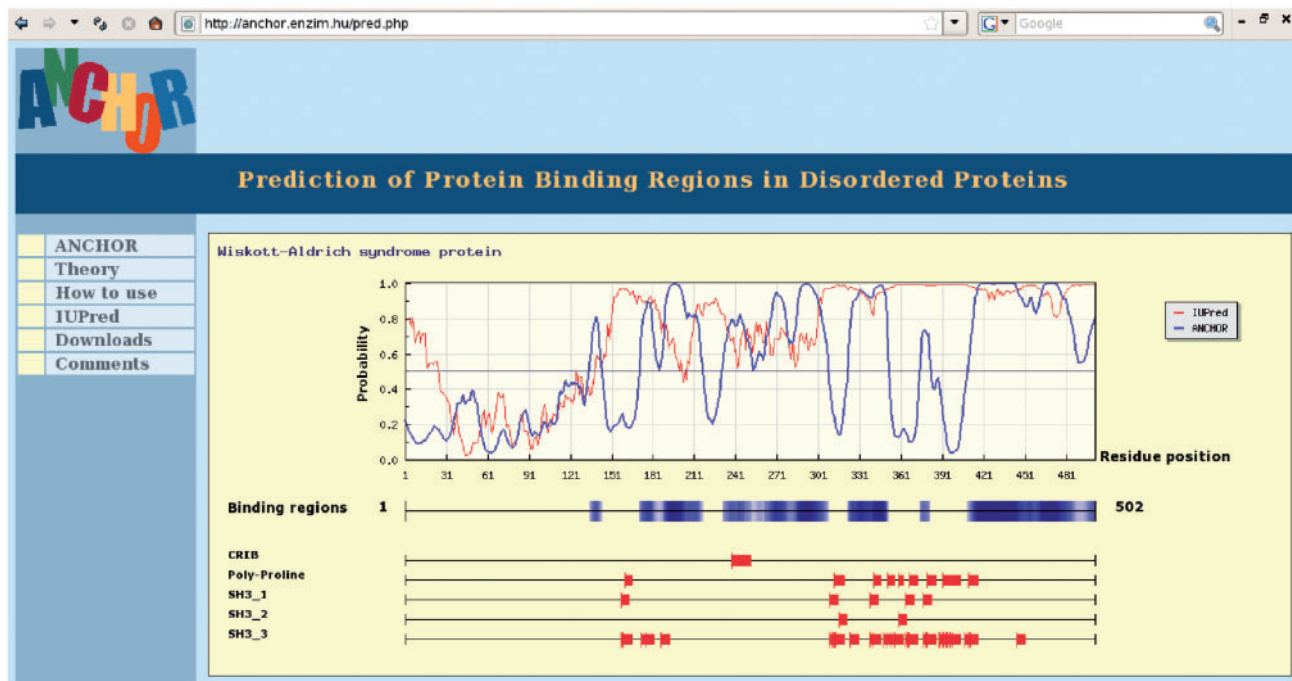
The developed method was able to recognize disordered binding regions with almost 70% accuracy at the segment level on various datasets. We also ensured that disordered binding regions could be discriminated from generally disordered regions and that the false positive rate on a dataset of globular proteins was <5%. Since the publication of the original paper (Mészáros *et al.*, 2009), we have found that the false positive rate can be further reduced by eliminating segments with IUPred scores too low to be compatible with disordered binding regions. Additionally, short predicted segments of length less than six residues are also filtered out.

ANCHOR predicts disordered binding regions without any information about the partner protein(s). A complementary approach identifies protein binding regions using motif searches. It was suggested that interaction with certain proteins or protein families are mediated through specific linear motifs that capture key residues responsible for binding. A growing number of such linear motifs are now being categorized in the ELM server (Puntervoll *et al.*, 2003). The presence of sequence motifs reduces the complex task of finding putative protein binding sites to a simple pattern matching problem. However, such matches can contain many false positives, suggesting that the definition of the binding motif should include information about the specific structural context. Since several instances of linear motifs occur within disordered regions, disordered binding regions could help to filter out false positive matches. Therefore, complementing the prediction of disordered binding regions with specific motif searches can prove useful in many cases and help to explore other motifs.

### 2 THE ANCHOR SERVER

The minimum input of the web server is a single amino acid sequence. Sequences can also be specified by their corresponding UniProt IDs or ACs. A list of motifs can also be submitted, specified as regular expressions with or without their names. A few examples, including known eukaryotic linear motifs are given in the help to guide the user with the format. The motif search, however, is not restricted to known linear motifs, any kind of regular expression can be specified.

The basic output of our prediction method is a probability score, indicating the likelihood of the residue to be part of a



**Fig. 1.** An example of ANCHOR graphical output for the Wiskott–Aldrich Syndrome protein (WASP) with various motif searches. The N-terminal of the protein contains an ordered domain, otherwise it is largely disordered. Multiple disordered binding regions were predicted, and several of these can be confirmed experimentally [see Mészáros *et al.* (2009) for more details]. The results of the motif searches show regions containing various SH3 binding sites as specified in the ELM database. Additionally, proline rich regions and the CRIB motif implicated in binding to Cdc42 can also be located.

disordered binding region along each position in the sequence. Regions that have a score  $>0.5$  and pass the filtering criteria are predicted as disordered binding regions. The returned plot shows the prediction profile calculated by ANCHOR, the disordered binding region prediction method, together with IUPred, a general disorder prediction method. Predicted disordered binding regions and matched motifs are also indicated underneath the profile as horizontal bars. The graphical output is followed by a simple text output, summarizing the predicted and filtered binding regions, the location of the found motifs and the returned prediction profile. An example for the graphical output is presented in Figure 1. The core program of ANCHOR is written C, while motif searches are carried out by a Perl wrapper. This Perl program is called by the web server written in PHP. The graphical output is generated by the JpGraph software (JpGraph, 2005; <http://www.aditus.nu/jpgraph/>). The default option for graphical/text output is automatically determined by the browser type, but it can be changed by user. Additionally, list of sequences can also be submitted to generate simple text output on a larger scale.

**Funding:** Hungarian Scientific Research Fund (OTKA-K72569); the National Office for Research and Technology, Hungary (NKTH07a-TB\_INTER).

**Conflict of Interest:** none declared.

## REFERENCES

- Cheng, Y. *et al.* (2007) Mining alpha-helix-forming molecular recognition features with cross species sequence alignments. *Biochemistry*, **46**, 13468–13477.
- Dosztányi, Z. *et al.* (2005a) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, **21**, 3433–3434.
- Dosztányi, Z. *et al.* (2005b) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.*, **347**, 827–839.
- Dyson, H.J. and Wright, P.E. (2002) Coupling of folding and binding for unstructured proteins. *Curr. Opin. Struct. Biol.*, **12**, 54–60.
- Garner, E. *et al.* (1999) Predicting binding regions within disordered proteins. *Genome Inform. Ser. Workshop Genome Inform.*, **10**, 41–50.
- Mészáros, B. *et al.* (2007) Molecular principles of the interactions of disordered proteins. *J. Mol. Biol.*, **372**, 549–561.
- Mészáros, B. *et al.* (2009) Prediction of protein binding regions in disordered proteins. *PLoS Comput. Biol.*, **5**, e1000376.
- Puntervoll, P. *et al.* (2003) ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res.*, **31**, 3625–3630.
- Radivojac, P. *et al.* (2006) Calmodulin signaling: analysis and prediction of a disorder-dependent molecular recognition. *Proteins*, **63**, 398–410.

# IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation

Gábor Erdős, Mátyás Pajkos and Zsuzsanna Dosztányi <sup>\*</sup>

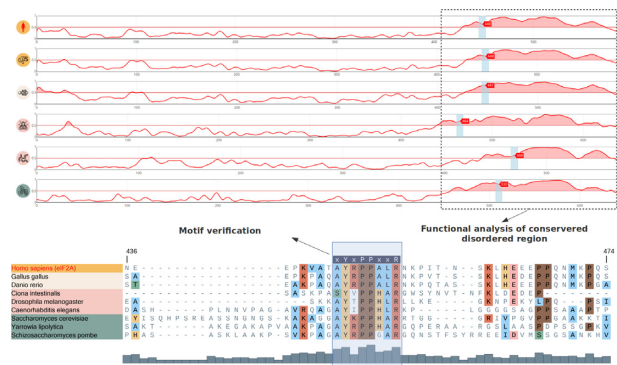
Department of Biochemistry, Eötvös Loránd University, Pázmány Péter stny 1/c, Budapest H-1117, Hungary

Received March 15, 2021; Revised April 21, 2021; Editorial Decision April 30, 2021; Accepted May 14, 2021

## ABSTRACT

Intrinsically disordered proteins and protein regions (IDPs/IDRs) exist without a single well-defined conformation. They carry out important biological functions with multifaceted roles which is also reflected in their evolutionary behavior. Computational methods play important roles in the characterization of IDRs. One of the commonly used disorder prediction methods is IUPred, which relies on an energy estimation approach. The IUPred web server takes an amino acid sequence or a Uniprot ID/accession as an input and predicts the tendency for each amino acid to be in a disordered region with an option to also predict context-dependent disordered regions. In this new iteration of IUPred, we added multiple novel features to enhance the prediction capabilities of the server. First, learning from the latest evaluation of disorder prediction methods we introduced multiple new smoothing functions to the prediction that decreases noise and increases the performance of the predictions. We constructed a dataset consisting of experimentally verified ordered/disordered regions with unambiguous annotations which were added to the prediction. We also introduced a novel tool that enables the exploration of the evolutionary conservation of protein disorder coupled to sequence conservation in model organisms. The web server is freely available to users and accessible at <https://iupred3.elte.hu>.

## GRAPHICAL ABSTRACT



## INTRODUCTION

A significant part of the genome of various organisms encode protein segments that do not form a well-defined structure in isolation even under physiological condition (1–3). These regions are called intrinsically disordered regions (IDRs) and their presence defines intrinsically disordered proteins (IDPs). IDRs are best characterized as fluctuating conformational ensembles whose behaviour is often context dependent (4,5). Despite the lack of well-defined structure, disordered regions play important functional roles in many cellular processes and are associated with various diseases (4,6). IDRs are multifaceted in terms of their function and can serve as entropic chains (serving e.g. as a spring) or flexible linkers between domains, mediate interactions through short linear motifs (SLiMs) or through sequentially longer and evolutionary conserved functional units called intrinsically disordered domains (IDDs) (4). A growing number of examples highlights the important role of IDPs driving or regulating the formation of membraneless organelles through liquid-liquid phase separation as well (7). Recognizing the importance of IDRs motivated efforts to develop various computational resources to facilitate the identification of biological relevant disordered regions and their functional characterization.

<sup>\*</sup>To whom correspondence should be addressed. Tel: +36 1372 2500/8537; Email: [zsuzsanna.dosztanyi@ttk.elte.hu](mailto:zsuzsanna.dosztanyi@ttk.elte.hu)

The central resource of experimentally verified disordered segments is the DisProt database (8). Through a community effort, entries are identified on the basis of various types of experimental data, collected from the literature by manual annotation. The IDEAL database has a similar focus (9), while DIBS and MFIB collect specific subsets of IDRs that undergo a disorder-to-order transition upon binding to globular protein or other disordered protein partners, respectively (10,11). Ordered structures are usually collected from the Protein Data Bank (PDB) (12). However, the PDB also contains regions that only adopt a well-defined structure in complex but would be disordered in isolation. The presence of missing residues in X-ray structures or high variations between conformers which satisfy the NMR constraints are usually taken as an indication of disorder. Despite existing experimental information, there can be ambiguity in the structural status of protein regions when order and disorder annotations overlap.

In recent years, the number of regions annotated as disordered has been growing steadily, with >1700 entries currently collected in the DisProt database (8). However, the overwhelming majority of IDRs is still uncovered. The large-scale analysis of ordered and disordered regions in proteins is available only through prediction tools which can recognize these segments from the amino acid sequence. More than 50 prediction methods have been developed over the last 20 years relying on different principles, including simple amino acid scales and biophysical models or various machine learning techniques, including deep learning techniques (13). The performance of these tools had been evaluated in the CASP experiments (14). However, these evaluations only included a small number of disordered regions with usually short length and provided limited insights into the usability of disorder prediction methods. Recently, a Critical Assessment of protein Intrinsic Disorder (CAID) prediction experiment was launched as a community-based blind test to determine the state of the art in predicting intrinsically disordered regions (15). Based on the first round, top performing methods were using machine learning approaches incorporating multiple sequence derived inputs, but the performance also varied depending on the evaluation criteria.

An intriguing aspect of prediction of protein disorder is what is the best way to incorporate evolutionary information. Several prediction algorithms include information derived from sequence profiles or raw multiple sequence alignment at the expense of significantly slower running time (15). Although these types of inputs can increase prediction accuracy, the gain is generally smaller relative to other problems like secondary structure prediction. In general, disordered regions are evolutionary less conserved compared to ordered regions, due to the lack of structural constraints in the case of IDRs (16–18). However, sequence based analyzes of functional IDRs showed that these modules can be as conserved in evolutionary terms as globular domains (17,19,20). The strict conservation is often limited to a few key amino acids, which could be surrounded by less conserved positions (18). However, the appearance of this island-like conservation pattern corresponding to these functional motifs is often compromised due to difficulties finding the optimal sequence alignment. In some

cases, larger disordered segments can also show strong evolutionary conservation, and can also be used to define sequence families (21). On the other hand, disordered characteristics can also be preserved over evolution without any sequential constraints. This type of conservation can occur in case of entropic chains, such as the projection domain of microtubule-associated protein 2 (MAP2), which serves as a spacer in the cytoskeleton by repealing molecules that approach microtubules (4). It was suggested that based on the relationship between the conservation of disorder and the conservation of sequence, three basic scenarios can occur. While the strict categorization largely depends on cut-off values (22), the simultaneous inspection of the disorder profile linked to sequence alignment can provide important insights for the evolutionary analysis of IDPs.

In this paper, we present the updated version of the IUPred (IUPred3) disorder prediction method. IUPred is based on a unique energy estimation approach that provides fast and robust prediction of disordered tendency. In addition, the same approach can also be used to highlight context dependent disordered regions, which can undergo a disorder-to-order transition as a result of binding (i.e. ANCHOR) or changes in redox conditions (23). IUPred is also incorporated into databases such as ELM (24) and Mobidb (25) or meta-tools (26,27). IUPred is also used to predict disordered binding regions (28–30), or to aid the identification of linear motif sites, both for *de-novo* discovery and to filter out false positive instances (31,32). A recent application of this method explored cancer associated mutations within IDRs (33). In the new implementation of IUPred we focus on features that can help the identification of biologically relevant disordered regions. We directly incorporated experimentally verified unambiguous ordered and disordered segments in the prediction profiles. We also developed a novel visualization tool which can highlight evolutionary conserved features of disordered regions by linking conservation protein disorder to sequence alignments of model organisms.

## MATERIALS AND METHODS

### IUPRED3: the algorithm

IUPred is based on an energy estimation method (34). For this, a pairwise statistical potential is generated from a library of known structures. Using this empirical force field, an energy-like quantity can be assigned to each residue based on the contacts it makes with other residues in the structure. These energies are estimated from the amino acid sequence using a  $20 \times 20$  energy estimation matrix. The parameters in this matrix are calculated by least square fitting to minimize the difference between the energies calculated from known structures and the energies calculated from the sequence. The energy estimation depends only on the amino acid type of each residue and the composition of its sequential neighborhood. The basic assumption of this approach is that residues with favorable energies are ordered while residues with unfavorable energies are disordered (34). The energies of neighboring residues are smoothed with a moving average using a window size of 10. Then, as a final step the energies are converted into a score between 0 and 1.

**Table 1.** Influence of an additional layer of smoothing for the performance of IUPred on the CAID dataset using the previous method (no second layer smoothing), using the Savitzky-Golay filter with parameters (19, 5) and using moving average smoothing with window size 11 compared to other state-of-the-art disorder prediction tools

		AUC	F1 score
No smoothing (IUPred2A)	-	0.736	0.417
Medium	Sav.Gol (19,5)	0.738	0.421
Strong	MA (11)	0.744	0.428
IUPred3 + experimental data	MA(11)	0.798	0.472
DisoMine		0.765	0.43
Predisorder		0.747	0.44

In our experience, the resulting IUPred profile can be still quite noisy. Therefore, here we introduce additional options to apply another layer of smoothing for the web server as well as for the downloadable IUPred3 package. As a first option, we apply a medium level smoothing using the Savitzky-Golay filter with parameters 19 and 5. This type of smoothing follows the ups and downs of the original prediction profile, but still eliminates significant parts of local noise. The second option uses a moving average with window size 11. Both options, but especially the strong smoothing options, improved the overall performance of the prediction when tested on the CAID DisProt dataset (Table 1). Nevertheless, the medium level smoothing can also be useful, because it can indicate local tendencies better, which could correspond to disordered binding sites within disordered regions, or flexible loops within ordered domains.

### Experimentally verified information

To collect experimentally verified disordered regions, we downloaded consensus disordered regions from the DisProt database (version 8.1). We also collected 54972 monomeric structures from the PDB using the Protein Interfaces, Surfaces, and Assemblies (PISA) service of the European Bioinformatics Institute (EBI). These structures were filtered for missing residues and served as a basis of our ordered dataset. However, the two types of annotations can overlap. To resolve these issues, we used a strict definition of order and disorder. Basically we eliminated DisProt annotations which overlapped with a monomeric structure or with a Pfam family which mapped to a monomeric structure. Altogether we identified 3160 ordered domains and 462 disordered regions. In addition, filtering based on experimentally characterized domains significantly boosts the performance of the method (Table 1).

### Disorder conservation tool

In IUPred3, we introduce a novel viewer of evolutionary conservation that enables the user to inspect disorder conservation along with sequence conservation. It is based on a precalculated dataset of orthologous sequences and multiple sequence alignments. First, orthologs were obtained by applying all-against-all GOPHER algorithm based prediction using protein sequences of the latest QFO (release 2020.04) reference dataset as the searching database (35,36). The orthology calculations were carried out for

48 eukaryota species with a total number of 876 605 proteins. Multiple sequence alignments of orthologs were constructed for each protein using the MAFFT algorithm (default parameters) (37). The orthologs were classified into the most specific term using 6 main evolutionary levels (Mammalia, Vertebrata, Eumetazoa, Opisthokonta, Eukarya and Plant). Users can also upload their own alignments which extends the application of this tool beyond eukaryotic species.

## SERVER DESCRIPTIONS

### Version control

To enable the smooth transition between different versions of the prediction methods related to IUPred, we restructured the website. The URL of the current version points to <https://iupred3.elte.hu> and also to <https://iupred.elte.hu>, which from now on will always be the latest version of IUPred. The previous iterations were moved to other domains, the original (34) version was renamed to IUPred1 and relocated to <https://iupred1.elte.hu>, the previous version (23) is available at <https://iupred2a.elte.hu>, as before. Many features of the web server were transferred from this earlier implementation, including download options.

### Submission page

The main page features entry boxes which accept a FASTA formatted or plain protein sequence, or any valid UniProt accession or an ID. The sequences of corresponding UniProt entries are accessed through an SQL database containing information about the specified input, or extract the information directly from UniProt, in case of an SQL database failure. In addition, a multi-FASTA formatted file with a maximum size of 200MB can also be uploaded. The web-server also incorporates RESTful services using custom links for searches.

IUPred3 offers multiple types of prediction options from which the user can choose. These include the default long disorder option, the short disorder option, which is tailored to recognize missing residues from X-ray structures, and the structural domains option. Additional options enable the prediction of context-dependent disordered regions such as disordered binding regions (ANCHOR method) or redox regulated disordered regions. In the current version we implemented novel features for the most commonly used long disorder prediction option, and introduced a new tool to explore disorder conservation. Alongside with the novel methods we also introduced an option for the users to be able to choose from different smoothing functions.

To further generalise the usability of the novel feature of IUPred3 to visualize disorder conservation a new submission option has been added, where users can upload FASTA formatted multiple sequence alignments containing up to 50 sequences. If such an alignment is supplied, IUPred3 will automatically use the first six sequences to calculate the disorder conservation and presents the results similarly to a standard 'Disorder conservation' analysis.



### Disorder prediction output

Once the proper inputs are selected and submitted, the server calculates the results on the latest Django based back-end. Each prediction is calculated on-the-fly server side, utilizing the latest MPI technology for maximum efficiency. Multi-FASTA uploads are treated separately and are queued until the server has enough free capacity.

The output of the requested prediction is presented in a graphical output visualized using a combination of Bokeh (1.4.0) and PlotlyJS (1.58.4) integrated into the Django frontend template framework. Integration with the UniProt resource enables the display of various additional information about the requested protein (when available). In case of a sequence input, IUPred tries to match the given sequence to a UniProt entry based on hashes generated from the sequence. If a unique matching entry is found, IUPred will map the input to the found entry in UniProt. Additional annotations include information on experimentally verified disordered regions from three different databases: generic disorder from DisProt (8) and disordered binding regions from DIBS (10) and MFIB (11), together with known motifs from the ELM database (24). Low-throughput post-translational modifications (including Ser, Thr and His phosphorylations, methylation, ubiquitylation and acetylation sites) from PhosphoSitePlus (38) are also indicated. In addition, PFAM annotations (39) with the different types of sequence families (domain, families, repeats, motifs, disordered) highlighted with different colors. Regions that have structural information based on known structures in the PDB (40) are also mapped to the selected entry. By selecting the 'Show structures' option, the mapped PDB regions are shown individually for each structure.

Besides the graphical output, both text based and JSON formatted outputs are downloadable for each prediction.

All functions of IUPred support all modern HTML5 and WebP compatible browsers.

### Disorder conservation output

Here we introduce a novel feature of IUPred3 that outputs both disorder and sequence conservation information of a given query protein relying on orthologous sequences of model organisms. The disorder conservation visualization is available directly from the submission page, but can also be accessed from the disorder output. The 'IUPred3 disorder conservation' tool uses the latest PlotlyJS library alongside with msaJS (1.0.0) (41).

Disordered profiles and multiple sequence alignments of orthologs are visualized in two separate viewers which are linked to each other. Disorder predictions are shown for six species. The disorder profiles are linked with a custom built hover function that maps the corresponding positions in each sequence. If there is no corresponding ortholog sequence in the given species, this bar is left empty. Alongside with the mapping of disorder profiles, the disorder conservation tool displays the multiple sequence alignment of 48 orthologs of the query protein (if found) using the msaJS library (41). Orthologs of model organisms are classified into six main evolutionary levels from unicellular eukaryotes to

mammalian in a nested way instead of listing sequences without any order. Each level is indicated with different colors to orient the users. The alignment is also mapped to the hover function of the prediction plots marking the central residue selected by the user. To ease the analysis of the multiple sequence alignment, pressing the Ctrl button locks the alignment in its current position, and the prediction plots can be reset to their default state. Disordered regions in the prediction plots are highlighted, however the cut-off value (default is 0.5) can be adjusted at the top of the page. Users might also search for interesting regions in the sequences of the model organisms using the respective input field above the plot. The field accepts regions in the format of start-end as well as standard regular expressions, for example '15-45' or '[RK].TQT', respectively.

### Supporting features

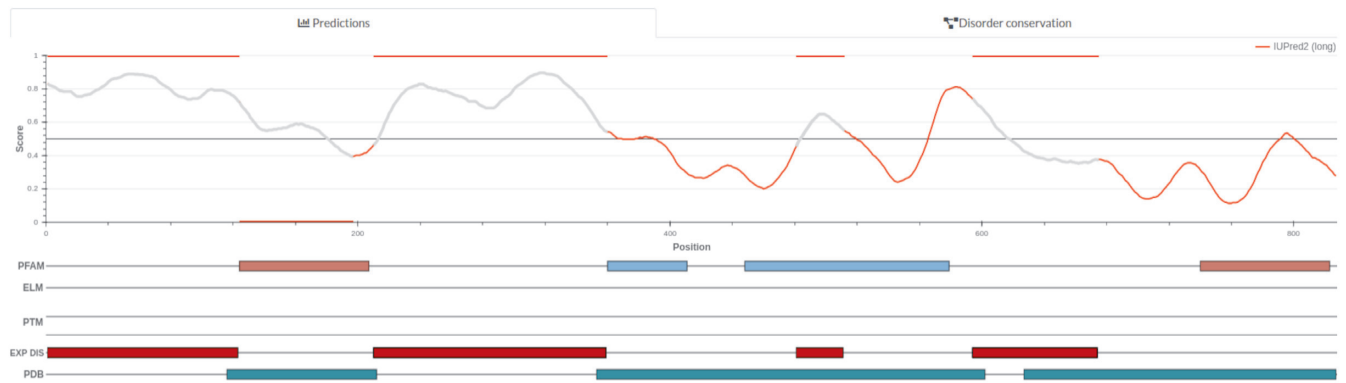
IUPred3 also features the description of the method on the website, as well as various examples that highlight its functionality. Furthermore, IUPred3 is also available as a standalone downloadable package alongside ANCHOR2 and the experimental redox sensitive conditional disorder prediction. Besides the standard executable we supply the package with an importable python library to further ease the use of the software (42).

### USE CASES

#### Example 1. Combining prediction with experimentally verified information

Many different annotations can exist with different reliabilities even for a single protein. These include experimental disorder, structural information or mapped sequence families. The complexity of these different levels of annotations can be demonstrated in the case of yeast Rap1.

Rap1 (repressor-activator protein 1) in yeast is a multifunctional protein that controls telomere silencing and the activation of glycolytic and ribosomal genes (43). Yeast Rap1 contains multiple regions matching DisProt entries and PDB structures (Figure 1). In this case, we accept disorder annotations, because there is no overlapping monomeric annotation neither in this protein, nor in other DisProt entries with the same domains. The BRCT domain which is located near the N-terminal is considered as a true ordered region. The solution structure of this domain was determined previously which reveals there is no disordered part of the core domain (44). Indeed, we identified nine fully resolved monomeric PDB structures in total corresponding to the BRCT Pfam family. Furthermore, the corresponding HMM profile did not match any of the experimentally verified disordered regions. However, there was no monomer based evidence identified for the other three structured regions and currently these are not considered as true ordered regions despite structural and domain annotations. Supporting this, the central region forms a complex with DNA, and probably has no or limited stability without it. Furthermore, the second DNA-binding domain overlaps with an experimentally verified disordered region.



**Figure 1.** The output of IUPred3 for the repressor-activator protein 1 of *Saccharomyces cerevisiae*. The strong smoothing option was used to generate this plot. At the upper part of the figure the disordered and ordered unambiguous experimentally verified protein regions are marked by red lines at the top and bottom of the plot, respectively. According to the manual curation of experimental data, the part of protein that has unambiguous verified order/disorder profile is coloured grey. The bottom part shows the various annotations for Rap1. Disordered regions from DisProt are shown in deep red boxes. Light red and blue boxes correspond to Pfam families and domains, respectively. Green boxes mark mapped consensus regions of PDB structures.

To highlight more reliable annotations, regions considered as true ordered or true disordered are indicated by a line at 0 and 1, and the prediction line faded to grey in the corresponding region. Additional annotations which were not accepted as true order or disorder due to some inconsistencies, are only highlighted below the plot. For these regions, the prediction is shown by a red line. Altogether, the visualization of the unambiguous experimental dataset of disordered and ordered protein regions helps the users to have a more clear view of the structural state information on the query.

### Example 2. Combined view of sequence and disorder conservation in model organisms

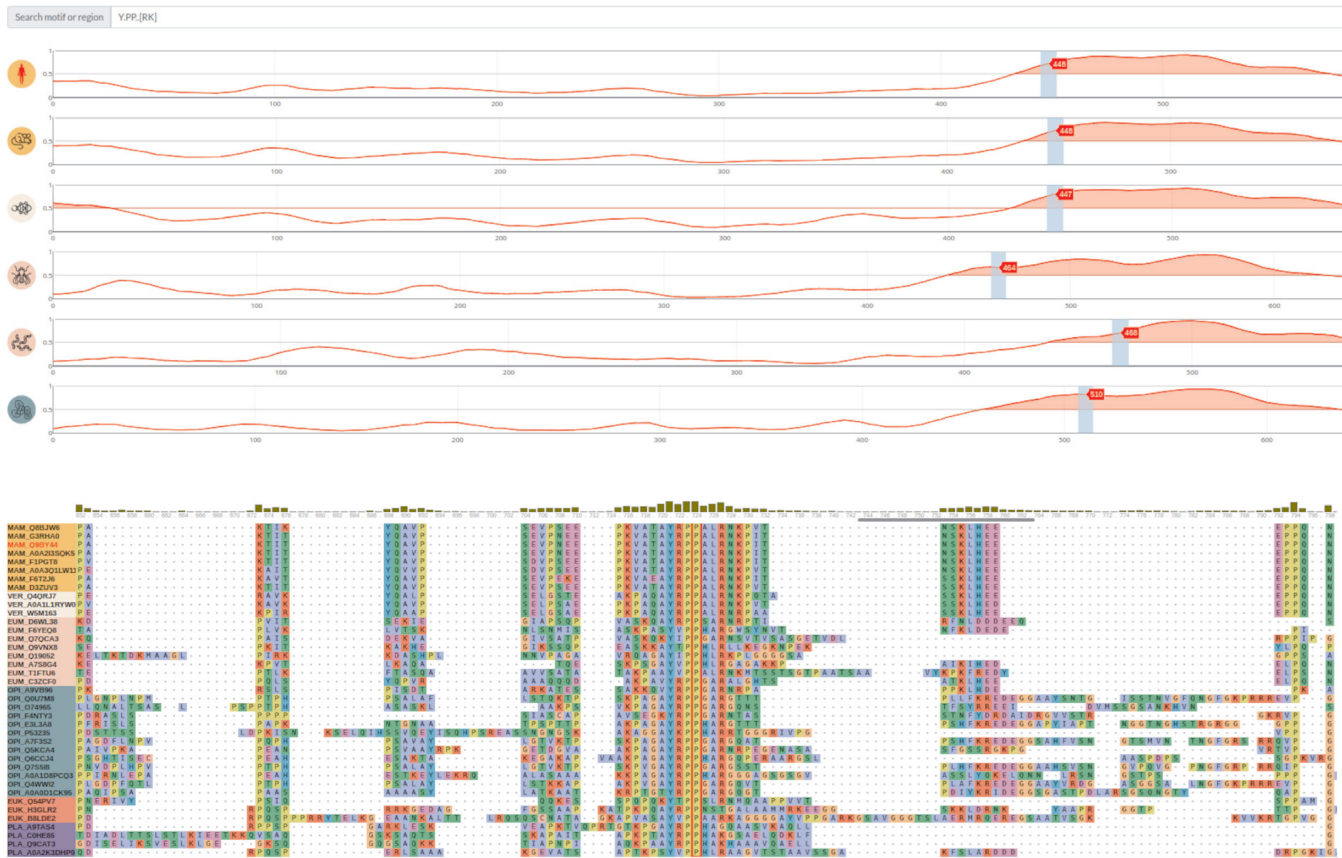
Previous analyses highlighted that the relationship between disorder and evolutionary conservation can be quite complex and include cases when both disorder and sequence is conserved, when only the disorder profile is conserved or when the sequence conservation is limited to few positions that can be indicative of putative linear motif sites. The two-level based visualization approach introduced here can be used to identify the different scenarios. Tools, such as Jalview (45) or ProViz (46) can visualize sequence alignment and also show disorder information for a single protein, but they cannot provide information on the evolutionary conservation of IDPs. Altogether, this new visualization tool of IUPred3 offers a simple way to inspect the disorder conservation based evolutionary history of IDPs.

One potential application of this tool is to locate putative linear motif sites within conserved disordered regions. An example for an evolutionary conserved disordered region is the human Eukaryotic translation initiation factor 2A (eIF2A) (Figure 2). The eIF2A protein is thought to participate in translation initiation during the translation of the first few amino acids (47). Orthologs of human eIF2A protein can be predicted not only in vertebrates but also in eumetazoa and unicellular eukaryotic organisms. This is supported by previous results in which yeast homolog

of eIF2A was identified based on homology searches (47). These proteins contain a conserved disordered region in their C-terminal. While the overall sequence conservation is low, it contains likely linear motif sites. For example, the YxPPxΦR motif in eIF2A is preserved over the evolution which is clearly observable in the multiple sequence alignment of orthologs (Figure 2). This corresponds to a consensus translation initiation factor (eIF4E) binding motif (YxPPxΦR) that was originally identified based on the interaction of eIF4E and DDX3X RNA helicase (48). However, in a previous study it was shown that the interaction between eIF2A and eIF4E is not dependent on the YxPPxΦR motif. This suggests that eIF2A might have a second binding region and the motif is involved in regulation of eIF4E activity (32). Although the YxPPxΦR motif in eIF2A is not well characterized, its evolutionary conservation indicates an ancient functional relevance.

### CONCLUSION

Disordered prediction tools can be used for multiple problems, including identifying regions suitable for structure determination, and are important starting points in the quest to characterize the function of non-globular regions. IUPred is one of the commonly used disordered prediction methods that is also often used in different contexts, to characterize individual proteins as well as for large-scale analysis (49). Here, we describe novel features introduced into the IUPred web server. We provide a way to filter out known ordered regions and to leverage experimentally verified disordered segments. We offer smoothing options which can help to eliminate noise in the prediction profile. These options can make it easier for the user to identify biologically relevant disordered regions. We also introduce a novel visualization tool which can be used to understand how the conservation of disorder is linked to the conservation of sequence. As the patterns of evolutionary conservation of disordered regions covers a wide range of behaviours, we expect this tool to be useful to understand the complex relationship between protein disorder and evolutionary history.



**Figure 2.** The output of disorder conservation for the human eIF2A protein. At the top of the figure, IUPred3 profiles of the human eIF2A and its orthologs from five generally known model organisms are depicted, and predicted disordered regions are highlighted by red. The bottom of the figure represents the multiple sequence alignment of orthologs identified from an extended set of eukaryotic model organisms. The human eIF2A as the query protein is highlighted by red in both parts of the figure. Model organisms are classified by taxonomic levels which are indicated with different colours. Using the regular expression based motif search box, the YxPPxΦR motif of eIF2A is highlighted by blue rectangles in each profile.

## FUNDING

ELIXIR Hungary ([www.elixir-hungary.org](http://www.elixir-hungary.org)) and ELIXIR Implementations Studies (IDP Community Implementation Study, Improving IDP tools interoperability and integration into ELIXIR and Integration and standardization of intrinsically disordered protein data). Funding for open access charge: ELIXIR Implementation Study.

*Conflict of interest statement.* None declared.

## REFERENCES

- Dunker, A.K., Garner, E., Guillot, S., Romero, P., Albrecht, K., Hart, J., Obradovic, Z., Kissinger, C. and Villafranca, J.E. (1998) Protein disorder and the evolution of molecular recognition: theory, predictions and observations. *Pac. Symp. Biocomput.*, 473–484.
- Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F. and Jones, D.T. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.*, **337**, 635–645.
- Tompa, P., Dosztanyi, Z. and Simon, I. (2006) Prevalent structural disorder in *E. coli* and *S. cerevisiae* proteomes. *J. Proteome Res.*, **5**, 1996–2000.
- van der Lee, R., Buljan, M., Lang, B., Weatheritt, R.J., Daughdrill, G.W., Dunker, A.K., Fuxreiter, M., Gough, J., Gspomer, J., Jones, D.T. *et al.* (2014) Classification of intrinsically disordered regions and proteins. *Chem. Rev.*, **114**, 6589–6631.
- Jakob, U., Kriwacki, R. and Uversky, V.N. (2014) Conditionally and transiently disordered proteins: awakening cryptic disorder to regulate protein function. *Chem. Rev.*, **114**, 6779–6805.
- Dyson, H.J. and Wright, P.E. (2005) Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.*, **6**, 197–208.
- Li, P., Banjade, S., Cheng, H.-C., Kim, S., Chen, B., Guo, L., Llaguno, M., Hollingsworth, J.V., King, D.S., Banani, S.F. *et al.* (2012) Phase transitions in the assembly of multivalent signalling proteins. *Nature*, **483**, 336–340.
- Hatos, A., Hajdu-Soltész, B., Monzon, A.M., Palopoli, N., Álvarez, L., Aykac-Fas, B., Bassot, C., Benítez, G.I., Bevilacqua, M., Chasapi, A. *et al.* (2020) DisProt: intrinsic protein disorder annotation in 2020. *Nucleic Acids Res.*, **48**, D269–D276.
- Fukuchi, S., Amemiya, T., Sakamoto, S., Nobe, Y., Hosoda, K., Kado, Y., Murakami, S.D., Koike, R., Hiroaki, H. and Ota, M. (2014) IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners. *Nucleic Acids Res.*, **42**, D320–D325.
- Schad, E., Fichó, E., Pancsa, R., Simon, I., Dosztányi, Z. and Mészáros, B. (2018) DIBS: a repository of disordered binding sites mediating interactions with ordered proteins. *Bioinformatics*, **34**, 535–537.
- Fichó, E., Reményi, I., Simon, I. and Mészáros, B. (2017) MFIB: a repository of protein complexes with mutual folding induced by binding. *Bioinformatics*, **33**, 3682–3684.
- Burley, S.K., Bhikadiya, C., Bi, C., Bittrich, S., Chen, L., Crichlow, G.V., Christie, C.H., Dalenberg, K., Di Costanzo, L., Duarte, J.M. *et al.* (2021) RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied

- research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.*, **49**, D437–D451.
13. Liu, Y., Wang, X. and Liu, B. (2019) A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. *Brief. Bioinform.*, **20**, 330–346.
  14. Monastyrskyy, B., Kryshchafovich, A., Moul, J., Tramontano, A. and Fidelis, K. (2014) Assessment of protein disorder region predictions in CASP10. *Proteins*, **82**, 127–137.
  15. Necci, M., Piovesan, D., Predictors, C., Curators, D. and Tosatto, S.C.E. (2020) Critical assessment of protein intrinsic disorder prediction. *Nat. Methods*, **18**, 472–481.
  16. Brown, C.J., Takayama, S., Campen, A.M., Vise, P., Marshall, T.W., Oldfield, C.J., Williams, C.J. and Dunker, A.K. (2002) Evolutionary rate heterogeneity in proteins with long disordered regions. *J. Mol. Evol.*, **55**, 104–110.
  17. Brown, C.J., Johnson, A.K. and Daughdrill, G.W. (2010) Comparing models of evolution for ordered and disordered proteins. *Mol. Biol. Evol.*, **27**, 609–621.
  18. Davey, N.E., Shields, D.C. and Edwards, R.J. (2009) Masking residues using context-specific evolutionary conservation significantly improves short linear motif discovery. *Bioinformatics*, **25**, 443–450.
  19. Dunker, A.K., Oldfield, C.J., Meng, J., Romero, P., Yang, J.Y., Chen, J.W., Vacic, V., Obradovic, Z. and Uversky, V.N. (2008) The unfoldomics decade: an update on intrinsically disordered proteins. *BMC Genomics*, **9**(Suppl. 2), S1.
  20. Pajkos, M., Zeke, A. and Dosztányi, Z. (2020) Ancient evolutionary origin of intrinsically disordered cancer risk regions. *Biomolecules*, **10**, 1115.
  21. Tompa, P., Fuxreiter, M., Oldfield, C.J., Simon, I., Dunker, A.K. and Uversky, V.N. (2009) Close encounters of the third kind: disordered domains and the interactions of proteins. *Bioessays*, **31**, 328–335.
  22. Ahrens, J.B., Nunez-Castilla, J. and Siltberg-Liberles, J. (2017) Evolution of intrinsic disorder in eukaryotic proteins. *Cell. Mol. Life Sci.*, **74**, 3163–3174.
  23. Mészáros, B., Erdős, G. and Dosztányi, Z. (2018) IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.*, **46**, W329–W337.
  24. Kumar, M., Gouw, M., Michael, S., Sámano-Sánchez, H., Pancsa, R., Glavina, J., Diakogianni, A., Valverde, J.A., Bukirova, D., Čalyševa, J. et al. (2020) ELM-the eukaryotic linear motif resource in 2020. *Nucleic Acids Res.*, **48**, D296–D306.
  25. Piovesan, D., Necci, M., Escobedo, N., Monzon, A.M., Hatos, A., Mičetić, I., Quaglia, F., Paladin, L., Ramasamy, P., Dosztányi, Z. et al. (2021) MobiDB: intrinsically disordered proteins in 2021. *Nucleic Acids Res.*, **49**, D361–D367.
  26. Necci, M., Piovesan, D., Clementel, D., Dosztányi, Z. and Tosatto, S.C.E. (2020) MobiDB-lite 3.0: fast consensus annotation of intrinsic disorder flavours in proteins. *Bioinformatics*, **36**, 5533–5534.
  27. Barik, A., Katuwawala, A., Hanson, J., Paliwal, K., Zhou, Y. and Kurgan, L. (2020) DEPICTER: intrinsic disorder and disorder function prediction server. *J. Mol. Biol.*, **432**, 3379–3387.
  28. Varadi, M., Guharoy, M., Zsolyomi, F. and Tompa, P. (2015) DisCons: a novel tool to quantify and classify evolutionary conservation of intrinsic protein disorder. *BMC Bioinformatics*, **16**, 153.
  29. Peng, Z. and Kurgan, L. (2015) High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. *Nucleic Acids Res.*, **43**, e121.
  30. Disfani, F.M., Hsu, W.-L., Mizianty, M.J., Oldfield, C.J., Xue, B., Dunker, A.K., Uversky, V.N. and Kurgan, L. (2012) MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. *Bioinformatics*, **28**, i75–i83.
  31. Erdős, G., Szaniszló, T., Pajkos, M., Hajdu-Soltész, B., Kiss, B., Pál, G., Nyitray, L. and Dosztányi, Z. (2017) Novel linear motif filtering protocol reveals the role of the LC8 dynein light chain in the Hippo pathway. *PLoS Comput. Biol.*, **13**, e1005885.
  32. Davey, N.E., Cowan, J.L., Shields, D.C., Gibson, T.J., Coldwell, M.J. and Edwards, R.J. (2012) SLiMPrints: conservation-based discovery of functional motif fingerprints in intrinsically disordered protein regions. *Nucleic Acids Res.*, **40**, 10628–10641.
  33. Mészáros, B., Hajdu-Soltész, B., Zeke, A. and Dosztányi, Z. (2020) In: *How Mutations of Intrinsically Disordered Protein Regions Can Drive Cancer*. Cold Spring Harbor Laboratory.
  34. Dosztányi, Z., Cizmók, V., Tompa, P. and Simon, I. (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.*, **347**, 827–839.
  35. Davey, N.E., Edwards, R.J. and Shields, D.C. (2007) The SLiMDisc server: short, linear motif discovery in proteins. *Nucleic Acids Res.*, **35**, W455–W459.
  36. Altenhoff, A.M., Boeckmann, B., Capella-Gutierrez, S., Dalquen, D.A., DeLuca, T., Forslund, K., Huerta-Cepas, J., Linard, B., Pereira, C., Przytycki, L.P. et al. (2016) Standardized benchmarking in the quest for orthologs. *Nat. Methods*, **13**, 425–430.
  37. Katoh, K., Misawa, K., Kuma, K.-I. and Miyata, T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**, 3059–3066.
  38. Hornbeck, P.V., Zhang, B., Murray, B., Kornhauser, J.M., Latham, V. and Skrzypek, E. (2015) PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.*, **43**, D512–D520.
  39. Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J. et al. (2021) Pfam: The protein families database in 2021. *Nucleic Acids Res.*, **49**, D412–D419.
  40. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
  41. Yachdav, G., Wilzbach, S., Rauscher, B., Sheridan, R., Sillitoe, I., Procter, J., Lewis, S.E., Rost, B. and Goldberg, T. (2016) MSAViewer: interactive JavaScript visualization of multiple sequence alignments. *Bioinformatics*, **32**, 3501–3503.
  42. Erdős, G. and Dosztányi, Z. (2020) Analyzing protein disorder with IUPred2A. *Curr. Protoc. Bioinformatics*, **70**, e99.
  43. Tomar, R.S., Zheng, S., Brunke-Reese, D., Wolcott, H.N. and Reese, J.C. (2008) Yeast Rap1 contributes to genomic integrity by activating DNA damage repair genes. *EMBO J.*, **27**, 1575–1584.
  44. Zhang, W., Zhang, J., Zhang, X., Xu, C. and Tu, X. (2011) Solution structure of Rap1 BRCT domain from *Saccharomyces cerevisiae* reveals a novel fold. *Biochem. Biophys. Res. Commun.*, **404**, 1055–1059.
  45. Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M. and Barton, G.J. (2009) Jalview Version 2 – a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.
  46. Jehl, P., Manguy, J., Shields, D.C., Higgins, D.G. and Davey, N.E. (2016) ProViz—a web-based visualization tool to investigate the functional and evolutionary features of protein sequences. *Nucleic Acids Res.*, **44**, W11–W15.
  47. Zoll, W.L., Horton, L.E., Komar, A.A., Hensold, J.O. and Merrick, W.C. (2002) Characterization of mammalian eIF2A and identification of the yeast homolog. *J. Biol. Chem.*, **277**, 37079–37087.
  48. Shih, J.-W., Tsai, T.-Y., Chao, C.-H. and Wu Lee, Y.-H. (2008) Candidate tumor suppressor DDX3 RNA helicase specifically represses cap-dependent translation by acting as an eIF4E inhibitory protein. *Oncogene*, **27**, 700–714.
  49. Dosztányi, Z. (2018) Prediction of protein disorder based on IUPred. *Protein Sci.*, **27**, 331–340.

Cite this: *Mol. BioSyst.*, 2012, **8**, 296–307

www.rsc.org/molecularbiosystems

PAPER

## Is there a biological cost of protein disorder? Analysis of cancer-associated mutations†‡

Mátyás Pajkos, Bálint Mészáros, István Simon and Zsuzsanna Dosztányi\*

Received 17th June 2011, Accepted 21st August 2011

DOI: 10.1039/c1mb05246b

As many diseases can be traced back to altered protein function, studying the effect of genetic variations at the level of proteins can provide a clue to understand how changes at the DNA level lead to various diseases. Cellular processes rely not only on proteins with well-defined structure but can also involve intrinsically disordered proteins (IDPs) that exist as highly flexible ensembles of conformations. Disordered proteins are mostly involved in signaling and regulatory processes, and their functional repertoire largely complements that of globular proteins. However, it was also suggested that protein disorder entails an increased biological cost. This notion was supported by a set of individual IDPs involved in various diseases, especially in cancer, and the increased amount of disorder observed among disease-associated proteins. In this work, we tested if there is any biological risk associated with protein disorder at the level of single nucleotide mutations. Specifically, we analyzed the distribution of mutations within ordered and disordered segments. Our results demonstrated that while neutral polymorphisms were more likely to occur within disordered segments, cancer-associated mutations had a preference for ordered regions. Additionally, we proposed an alternative explanation for the association of protein disorder and the involvement in cancer with the consideration of functional annotations. Individual examples also suggested that although disordered segments are fundamental functional elements, their presence is not necessarily accompanied with an increased mutation rate in cancer. The presented study can help to understand how the different structural properties of proteins influence the consequences of genetic mutations.

### Introduction

For several decades, molecular biology studies largely concentrated on globular proteins, based on the assumption that a well-defined structure is necessary for the proper function of proteins, and the loss of structure leads to the loss of function. In exploring the genetic background of various diseases similar biases were also present, by focusing on mutations that could be placed into structural context. With the increase of available genome sequences, it has become evident that a large number of naturally occurring proteins do not require a well-folded structure to fulfill their biological role.<sup>1–3</sup> These intrinsically unstructured/disordered proteins (IUPs/IDPs) exist as highly flexible ensembles of rapidly interconverting conformations, even under physiological conditions.<sup>1–3</sup> IDPs are surprisingly common, especially in higher eukaryotes,<sup>4,5</sup> and are involved in many vital cellular functions. These include regulation,

transcription and translation, signal transduction, protein phosphorylation, storage of small molecules, chaperone action, transport, and assembly of large multiprotein complexes.<sup>6</sup> The increased flexibility of these proteins is pertinent for their specific functions and offers several functional advantages. IDPs provide a larger interaction surface area than globular proteins of similar length.<sup>7,8</sup> They generally interact with their partners with relatively high specificity and low affinity and can bind to multiple partners.<sup>9,10</sup> The plasticity of these proteins also enables them to adapt to the surface of their partners.<sup>11</sup> They are often subject to various post-translational modifications that facilitate the regulation of their function in the cell.<sup>12,13</sup> Consequently, disordered proteins can capture and integrate various signals in a complex way through their disordered segments and participate in a large number of interactions.<sup>13</sup> These properties explain their prevalence in signaling and regulatory functions,<sup>5,14</sup> as well as serving as hubs of interaction networks.<sup>15,16</sup>

Given the functional importance of disordered protein regions, their malfunction is expected to have serious biological consequences. IDPs indeed are often associated with various diseases, especially with cancer.<sup>17</sup> This observation is supported by the list of IDPs, such as BRCA1, p27, p21 and CBP, that are

*Institute of Enzymology, Hungarian Academy of Sciences, PO Box 7, H-1518 Budapest, Hungary. E-mail: zsuzsa@enzim.hu*

† Published as part of a Molecular BioSystems themed issue on Intrinsically Disordered Proteins: Guest Editor M. Madan Babu.

‡ Electronic supplementary information (ESI) available. See DOI: 10.1039/c1mb05246b

involved in various forms of cancer. One of best characterized disordered proteins, p53, is directly inactivated in more than 50% of cancers.<sup>18</sup> At a more general level, the higher proportion of disordered proteins among cancer associated proteins was also observed. According to the analysis of the SwissProt database, 79% of human cancer associated proteins have been classified as IDPs, compared to 47% of all eukaryotic proteins.<sup>19</sup> The correlation between protein disorder and cancer was further underscored in the case of two common forms of generic alterations, chromosomal rearrangements<sup>20</sup> and copy number variations.<sup>21</sup> In addition to cancer, disordered proteins were also suggested to be common in diabetes and cardiovascular diseases.<sup>17,22</sup> Several disordered proteins—such as A $\beta$ ,  $\tau$ ,  $\alpha$  synuclein, and prion proteins—are involved in neurodegenerative diseases and are also prone to amyloid formation.<sup>23,24</sup> Altogether, these results lead to the conclusion that protein disorder comes with a “biological cost” that is reflected in an increased risk of cancer and other diseases.<sup>2,17</sup> This calls for the understanding of the role of protein disorder in various diseases.

Large scale sequencing efforts now enable us to explore the relationship between protein disorder and disease-causing genetic mutations at a more detailed level. The completion of the Human Genome Project is being followed by concerted efforts to categorize commonly occurring sequence alterations.<sup>25,26</sup> As a result, the dbSNP database already contains more than 13 million sequence variations. Recently, dbSNP started to include personal genomics data by incorporating the results from the pilot study of 1000 Genomes Project.<sup>27</sup> The rapid accumulation of DNA variation data enabled the evaluation of evolutionary constraints at the level of single nucleotide polymorphisms (SNPs).<sup>28</sup> Furthermore, advances in sequencing technologies also opened new ways to explore how genetic changes lead to diseases. Before the Human Genome Project, the identification of potential cancer-causing genes often relied on prior assumptions about the approximate location of mutated regions in the genome or some information about their biological function.<sup>29</sup> Consequently, traditional approaches could indirectly favor better-characterized ordered proteins and introduce a bias against disordered segments. Cancer genome projects can decipher the genetic background of cancer without such biases by directly analyzing the differences between cancer and normal cells at the DNA level.<sup>30</sup> From the currently available studies of breast, colorectal,<sup>31,32</sup> pancreatic cancers<sup>33</sup> and glioblastoma,<sup>34</sup> an unexpectedly complex landscape of cancer emerged. According to this, cancer is a result of the accumulation of a relatively large number of mutations each of which carries a small fitness advantage towards tumor progression. While there are a few frequently occurring mutations, the distribution of mutations is dominated by a much larger number of infrequently mutated genes.<sup>32</sup>

With the rapid explosion of data on sequence variations and the expanding catalogue of cancer-associated mutations, we can have a fresh look on how the structural properties of proteins determine the distribution of neutral and cancer-associated mutations. In this work we tested the hypothesis regarding the biological cost of protein disorder in terms of single point mutations. We considered cancer-associated proteins identified by traditional biochemical essays as well as

**Table 1** The number of proteins and residues for the 12 cancer-associated mutation databases and the polymorphism database. The number of mutations and/or polymorphisms are also shown where applicable

Datasets	Number of			
	Proteins	Residues	Mutations	Polymorphisms
SP_cancer	1403	1 250 776	5246	—
SP_cancer_annotated	113	91 683	1555	—
SP_poly	11 510	7 776 050	—	36 583
CGP_br/col_1	924	795 543	1239	3536
CGP_br/col_2	1335	1 332 469	1739	6098
CGP_pan	711	769 634	790	3848
CGP_glio	1089	1 074 168	1195	5794
CGP_CAN_br/col_1	174	203 731	395	908
CGP_CAN_br/col_2	243	298 114	513	1372
CGP_CAN_pan	64	72 317	130	289
CGP_CAN_glio	36	43 031	77	210
COSMIC	8957	6 898 559	22 708	26 435
COSMIC_census	261	238 130	5375	673

by the various cancer genome projects. Using these datasets, the distributions of commonly occurring polymorphisms and cancer-associated mutations within ordered and disordered regions of proteins were investigated. A functionally relevant subclass of disordered segments corresponding to disordered binding regions was also studied in a similar manner. In order to explore indirect relationships between cancer and the structural state of proteins, we also considered functional categories of cancer-associated proteins. A closer look at interesting examples can give further insights into the role of protein disorder in cancer-associated proteins.

## Results

We have compiled 12 datasets of cancer-associated proteins from various resources (see Data and methods and Table 1). The datasets differed in their size and the primary way the specific proteins were identified. It is worth noting that in this study, genetic variations were restricted to single amino acid substitutions, therefore proteins that were associated with cancer *via* chromosomal translocations or copy number variations were not considered.

The first type of dataset was collected from the SwissProt database,<sup>35</sup> primarily from literature searches (SP\_cancer). A subset of this dataset with specific annotation in the OMIM database was also considered (SP\_cancer\_annotated). These two datasets, especially the annotated subset, are expected to be dominated by the cancer mutations identified in more traditional ways. The second type of datasets was compiled from four cancer genome projects. Two of these corresponded to breast and colorectal cancers (CGP\_br/col\_1 and CGP\_br/col\_2),<sup>31,32</sup> one to pancreatic cancer (CGP\_pan)<sup>33</sup> and another one to glioblastoma (CGP\_glio).<sup>34</sup> In each case, a subset of genes were selected that were more likely to contain driver mutations. These mutations are expected to actively contribute to the tumorigenesis as opposed to passenger mutations which occur purely by chance. These CAN sets were also analyzed separately (CGP\_CAN). The largest dataset was compiled from the COSMIC database (COSMIC).<sup>36</sup> It included cancer mutation data collected both from the literature and the outcomes of large-scale cancer genome projects. An additional

dataset corresponded to a more restricted subset of proteins in COSMIC that were part of cancer census genes.<sup>37</sup> These proteins could be casually linked to oncogenesis (COSMIC\_census). The number of proteins, amino acids and mutations in each dataset are given in Table 1.

### Protein disorder in cancer-associated proteins

We evaluated the disorder content in our datasets to confirm that protein disorder is common in human cancer-associated proteins.<sup>19</sup> The length and average disorder content were analyzed in these datasets. As a reference, we used the complete human proteome downloaded from the SwissProt database.<sup>38</sup> The disorder content was calculated using the IUPred disorder prediction method.<sup>39,40</sup> The results were confirmed with two other popular disorder prediction methods, DISOPRED2<sup>5</sup> and VSL2.<sup>41</sup>

Fig. 1 shows the disorder content and the percentage of proteins with disordered regions over 30 residues, as well as the average length of proteins in the various datasets as compared to the average values of the human proteome obtained with IUPred. In contrast to earlier results,<sup>19</sup> the percentage of disordered residues in these datasets was not significantly different compared to the background (Fig. 1 and Table S1 (ESI†)). Significant differences were only observed in the case of two breast–colorectal datasets (CGP\_br/col\_2 and CGP\_CAN\_br/col\_2) and the COSMIC census dataset. In the case of SP\_cancer\_annotated data, the disorder content actually decreased compared to the average disorder content in the human proteome, although this difference was not statistically significant. These results did not depend on the choice of the disorder prediction software, as DISOPRED2 and VSL2, two other fundamentally different methods produced remarkably similar outputs (see Fig. S1, ESI†). It should be noted that Iakoucheva *et al.*<sup>19</sup> compared the disorder

content of cancer proteins to those of all eukaryotic proteins in the SwissProt database. This could explain why the differences in their work were much larger compared to our work.

There was, however, a significant increase in the proportion of proteins containing long disordered segments among cancer-associated proteins compared to the human proteome. With the exception of SP\_cancer\_annotated and the CGP\_CAN\_glio datasets, all differences were significant. The results calculated with IUPred (Fig. 1B) were again confirmed by the two other prediction methods (Fig. S1, ESI†). In agreement with earlier results,<sup>42</sup> cancer-associated proteins were also significantly longer. The increase in length and in fraction of proteins with long disordered segments points to the increased modularity and complexity of cancer-associated proteins.

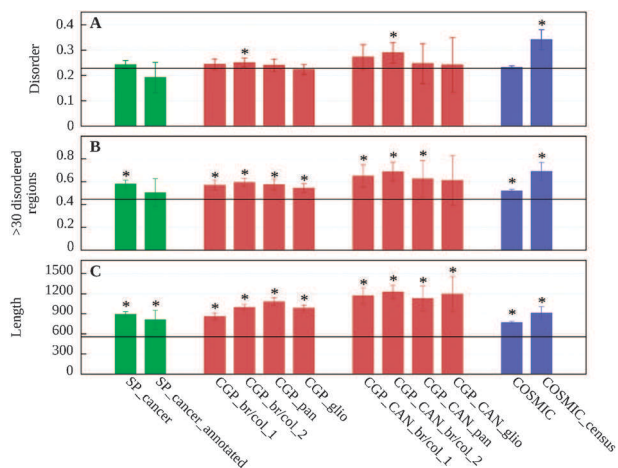
### Polymorphisms in ordered and disordered regions

The rates of evolution are largely governed by the stringency of functional and structural constraints. As ordered and disordered segments in proteins have distinct properties in this regard, these characteristic differences are expected to be reflected in the distribution of genetic variations in these regions. To test this assumption, we analyzed the differences in the distribution of SNPs within disordered and ordered segments of cancer-associated proteins. Polymorphism data were collected from the SwissProt resource and the dbSNP database (release 132). On average, we observed around five polymorphisms per thousand amino acid positions, although this number varied slightly among the various datasets (see Table 1).

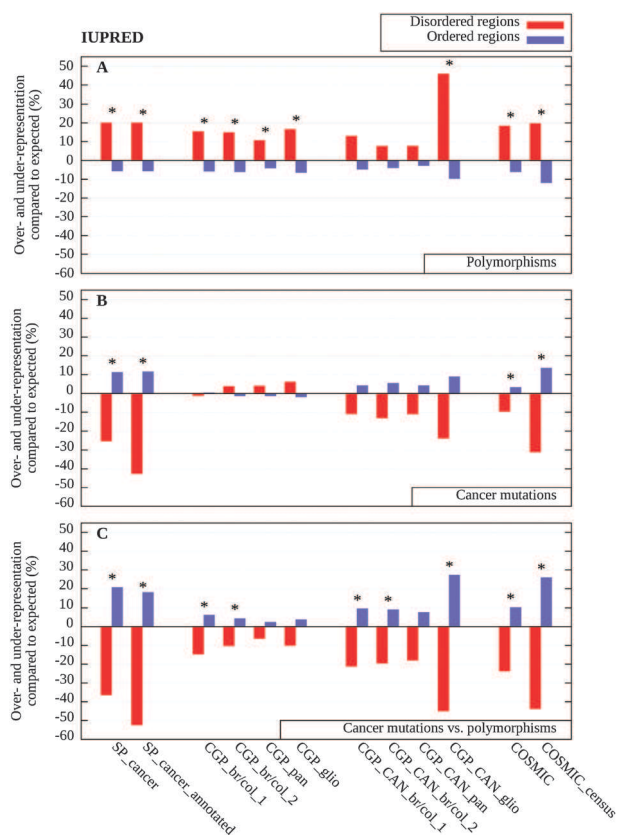
For each protein in our datasets, we tallied the number of observed polymorphisms in ordered and disordered segments. These numbers were compared to the expected number of polymorphisms based on the assumption that the mutations are distributed evenly in the sequence. The results presented in Fig. 2A show the relative difference between the observed and expected number of polymorphisms within both disordered and ordered segments predicted with IUPred. A more detailed account of the results for each dataset including the *p*-values showing the statistical significance is presented in Table S2 (ESI†). The results indicate that a significantly larger number of polymorphisms fell within disordered segments compared to ordered regions. The enrichments ranged from 7.7% (CGP\_CAN\_pan) to 45.9% (CGP\_CAN\_glio) in the various datasets, with an average of 15.0% (see Table S2, ESI†). With the exception of some of the CAN gene sets, the differences were statistically significant in all datasets and largely agreed for all three disorder prediction methods (see also Fig. S2A and S3A, ESI†). These data indicate that in cancer-associated proteins, disordered regions generally are more tolerant to mutations compared to ordered proteins. This trend is in agreement with the lower evolutionary conservation of disordered proteins, observed at various levels.<sup>28,43–45</sup>

### Cancer-associated mutations

As a next step we investigated if there is any preference of cancer-associated mutations towards order or disorder in proteins. The cancer-associated mutations collected from various sources were projected onto positions in the protein sequence, and the order/disorder status of the corresponding



**Fig. 1** Average ratio of disordered residues (A), ratio of proteins containing > 30 residue long disordered regions (B) and length (C) in the 12 datasets analyzed. Black horizontal lines represent the average values obtained for the proteins of the human proteome taken from SwissProt. Flags show the confidence interval of  $\alpha = 0.01$  calculated from the standard error of the mean of randomly selected samples from the human proteome (see Data and methods). Significant differences are marked with asterisks (see Table S1, ESI†).



**Fig. 2** Over- and under-representation of mutations in disordered (red) and ordered regions (blue) calculated with IUPred, as compared to background distributions (see Data and methods). (A) The distribution of polymorphisms as compared to the uniform random distribution; (B) the distribution of cancer-associated mutations as compared to the uniform random distribution and (C) the distribution of cancer-associated mutations as compared to the expected values weighted by the distribution of polymorphisms shown in (A). Significant differences are marked with asterisks (see Table S2, ESI†).

residues was determined by the IUPred disorder prediction algorithm. Similarly to polymorphisms, the observed number of mutations within ordered and disordered segments was compared to the expected number of mutations based on the assumption that the mutations are distributed evenly in the sequence.

Compared to polymorphisms, cancer-associated mutations followed a reversed trend and were more likely to appear within ordered regions (Fig. 2B). This tendency was strongest in the SwissProt datasets, but was also present in the four CGP\_CAN, as well for the complete COSMIC dataset and its subset of cancer census proteins. The SwissProt and COSMIC datasets produced statistically significant differences (see Table S2, ESI†). Results obtained with IUPred again were in agreement with results of the two other disordered prediction methods (Fig. S2 and S3, ESI†). The complete dataset of cancer genomes showed a slightly different trend. In these cases, cancer-associated mutations were slightly tilted towards disordered segments. The weak preference of these sequence variations for disordered segments can be due to the higher number of randomly occurring passenger mutations present in these datasets. Indeed, the normalization which takes into

account the uneven distribution of polymorphisms, compensated for this behavior. As a result, the underrepresentation of cancer-associated mutations within disordered regions became even more apparent and unequivocal within all datasets analyzed (Fig. 2C). The normalization also increased the statistical significance of the results (Table S2, ESI†). The reversed trend was statistically significant in the manually curated datasets (SP\_cancer, SP\_cancer\_annotated, COSMIC, and COSMIC\_census). Some of the cancer genomes project also produced significant differences after the normalization, despite the increased noise present in these datasets due to the higher content of passenger mutations. Altogether, these results clearly contradicted the original hypothesis about the increased risk of cancer associated with protein disorder, at least in terms of single nucleotide mutations.

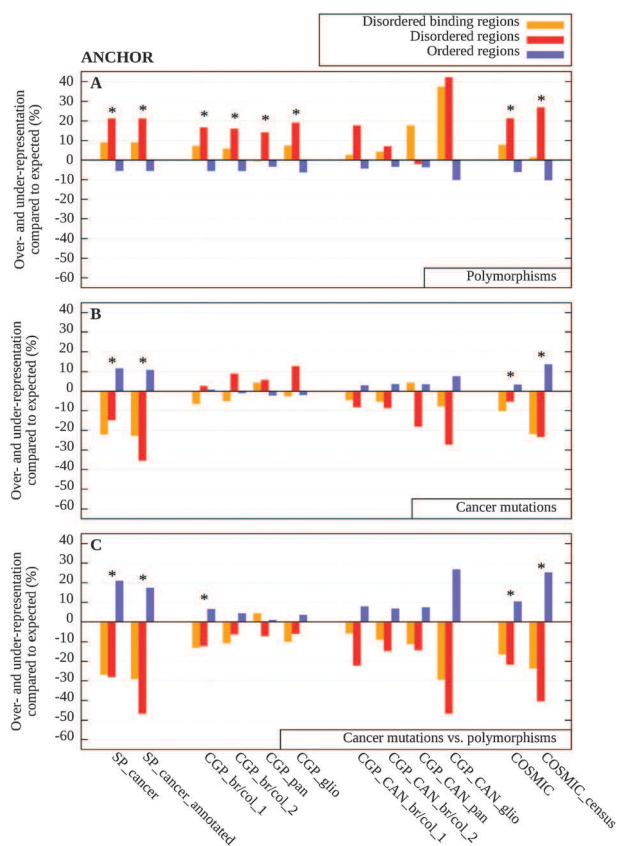
### Disordered binding regions

As disordered proteins are quite heterogeneous both in terms of their structural and functional properties, deviations from the general behavior can occur in certain cases. We specifically analyzed predicted disordered binding regions that are expected to be enriched in functionally relevant sites. Disordered proteins often function *via* binding to other macromolecules that involves a disorder-to-order transition.<sup>2,10</sup> Although binding to other macromolecules can induce a transition to a fully or at least partially ordered structure in the case of many IDPs, their complexes have distinct properties compared to complexes formed by ordered proteins.<sup>46,47</sup> The actual binding regions often correspond to short, localized elements in the sequence and have unique sequence properties compared to both ordered and disordered segments in general.<sup>48</sup> Using a sequence based prediction method, called ANCHOR,<sup>48,49</sup> we examined the distribution of polymorphisms and cancer-associated mutations within disordered binding regions.

Based on the predictions, a distinct group was formed from the residues of disordered binding regions. Residues not predicted as disordered binding sites were divided into two separate groups depending on whether they were predicted as disordered or as ordered. Disordered binding residues are usually part of a disordered segment, however, in some cases they can also correspond to local dips in the prediction profile in which case they are predicted as ordered.<sup>50</sup> Therefore, both disordered and ordered datasets contained fewer residues compared to the previous analysis. The results for the three groups are presented in Fig. 3A–C and Table S3 (ESI†). There are significant differences among the three sets in the distributions of observed SNPs (Fig. 3A), with the exception of the small CAN gene sets. While SNPs were clearly over-represented in disordered segments and underrepresented in ordered regions, disordered binding regions fell between these two categories, but their behavior was still closer to disordered segments. One of the CAN gene sets (pancreatic cancer) differed slightly from this trend, in this case more SNPs were observed in disordered binding regions than in disordered segments in general.

The distribution of cancer-associated mutations within disordered binding regions was largely similar to that of disordered regions in general, with some differences in the case of the cancer





**Fig. 3** Over- and under-representation of mutations in disordered binding regions (orange), disordered (red) and ordered regions (blue) calculated with ANCHOR, as compared to background distributions (see Data and methods). (A) The distribution of polymorphisms as compared to the uniform random distribution; (B) the distribution of cancer-associated mutations as compared to the uniform random distribution and (C) the distribution of cancer-associated mutations as compared to the expected values weighted by the distribution of polymorphisms shown in (A). Significant differences are marked with asterisks (see Table S3, ESI†).

genome datasets (Fig. 3B). These deviations can also be attributed to the increased number of passenger mutations within disordered segments and disappeared when the uneven distribution of polymorphisms was taken into account. In this normalized data, disordered binding regions had a smaller depletion of cancer-associated mutations in most cases compared to disordered regions in general (Fig. 3C). This behavior was expected for regions with increased functional importance.

### Functional correlations

We also analyzed cancer-associated proteins in terms of their functional categories and their number of protein–protein interactions. First, we assessed which functional groups were overrepresented within cancer-associated proteins. For this analysis, the GeneOntology functional categories were used (see Data and methods). The occurrence of each of the considered 50 biological processes and 41 molecular functions in the COSMIC\_census dataset was compared to the expected occurrence of these functions in the human proteome. The list of biological processes and molecular functions that exhibited

statistically significant differences is shown in Table 2. The significantly enriched processes among cancer-associated proteins included signal transduction, involvement in cell-cycle and proliferation, DNA- and protein binding, phosphorylation and regulation of transcription. These proteins on the other hand were significantly depleted in transport processes in general and particularly in ion transport. In other cases, the differences were not significant at the  $\alpha = 0.01$  level. In general, our results are in complete agreement with an earlier study,<sup>42</sup> and correlate well with the functional enrichments of disordered proteins.<sup>5,14</sup>

Cancer-associated proteins represent a specific group of proteins that are enriched in certain functions, contain more disordered regions, generally are longer and involved in a larger number of interactions (25.5 per protein as compared to 5.5 per protein in the human proteome). However, all these features also correlate with each other. To untangle these complicated relationships, we studied the association between these distinct features. Specifically, we considered the length of the protein, the ratio of its residues residing in disordered segments or disordered binding regions, the number of cancer-associated mutations taken from the COSMIC census database and the number of protein–protein interactions as well as the above identified significant functional classes (see Data and methods). The mutual information and the Jaccard distances were calculated between all pairs of features. The obtained distances between the different features are shown in Table 3. These distances were also subject to multidimensional scaling to reduce the dimensionality to two. The resulting scaled location of each feature is presented in Fig. 4.

It can be seen that the association between the ratios of residues in disordered regions and disordered binding sites is the highest indicating the relatively constant ratio of disordered residues that are involved in binding. Apart from this strong association, the functional features shared the most information with all the other features. This indicated the central role of function that largely determines the disorder content together with the amount of disordered binding regions, the number of protein–protein interactions, the required length for a given protein and its involvement in cancer. These data suggest that the association between increased amount of protein disorder and cancer in terms of single nucleotide mutations is indirect.

### Examples

Besides analyzing the general features of cancer-associated proteins, a few examples are also presented here to gain further insights into how disordered regions and their binding sites contribute to the function of these proteins. The examples were selected from the COSMIC dataset and stand out with the largest number of mutations falling into ordered (p53, PTEN) or disordered regions ( $\beta$ -catenin, ACP). The domain structure (according to PFAM<sup>51</sup>), the predicted disordered regions and disordered binding regions and the distribution of cancer-associated mutations are shown in Fig. 5. Interestingly, these proteins basically contained no neutral polymorphisms.

**p53.** The largest number of mutations occurred within p53 (TP53). It is a transcription factor that regulates a large number of genes (> 100 genes) and controls a number of key tumor suppressing functions such as cell cycle arrest, DNA repair,

**Table 2** List of GO biological processes and molecular functions that are significantly over- or under-represented in the COSMIC census database as compared to the human proteome. *p*-values were obtained using the exact Fisher test (see Data and methods)

	GO ID	Description	Number of COSMIC census proteins with the given term	Expected number of proteins with the given term	<i>p</i> -value	Over- or under-representation
Biological processes	GO:0007165	Signal transduction	51	26	$1.418 \times 10^{-3}$	0.96
	GO:0008283	Cell proliferation	17	4	$3.055 \times 10^{-3}$	3.25
	GO:0006811	Ion transport	0	8	$3.696 \times 10^{-3}$	-1.00
	GO:0006810	Transport	9	24	$5.370 \times 10^{-3}$	-0.63
	GO:0007049	Cell cycle	20	7	$8.084 \times 10^{-3}$	1.86
Molecular functions	GO:0005515	Protein binding	184	65	$1.305 \times 10^{-26}$	1.83
	GO:0003677	DNA binding	84	27	$4.907 \times 10^{-10}$	2.11
	GO:0000166	Nucleotide binding	72	25	$6.844 \times 10^{-8}$	1.88
	GO:0004672	Protein kinase activity	36	6	$5.573 \times 10^{-7}$	5.00
	GO:0003700	Transcription factor activity	44	12	$3.463 \times 10^{-6}$	2.67
	GO:0016301	Kinase activity	37	8	$3.192 \times 10^{-6}$	3.63
	GO:0016740	Transferase activity	48	18	$5.276 \times 10^{-5}$	1.67
	GO:0030528	Transcription regulator activity	17	5	$7.340 \times 10^{-3}$	2.40

**Table 3** Jaccard distances of the 6 features calculated on the COSMIC census database as compared to the human proteome (see Data and methods)

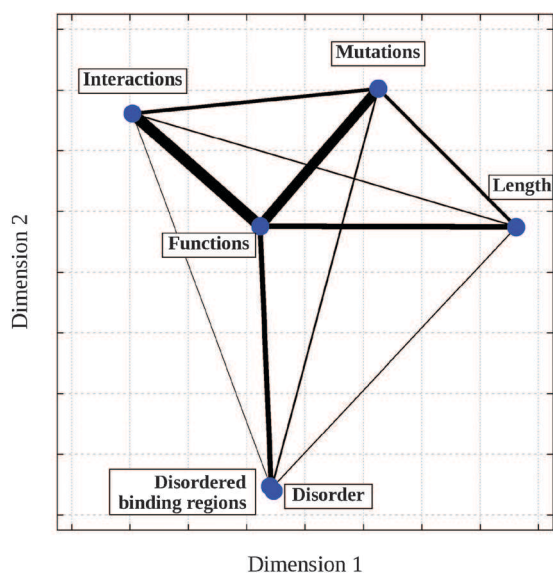
	Length	Disorder %	Binding regions %	COSMIC census mutations	Interactions	Functions
Length	0.0000	0.9871	0.9860	0.9597	0.9776	0.9157
Disorder %		0.0000	0.5170	0.9753	0.9896	0.9208
Binding regions %			0.0000	0.9732	0.9860	0.9162
COSMIC census mutations				0.0000	0.9444	0.8808
Interactions					0.0000	0.8670
Functions						0.0000

senescence and apoptosis.<sup>52,53</sup> p53 protein is expressed at a low level in normal cells and at a high level in response to DNA damage and oncogenic transformation. Whilst the activation of p53 often leads to apoptosis, p53 inactivation facilitates tumor progression. From structural point of view, it is composed of the central DNA binding domain that is largely ordered, and the disordered N- and C-termini (Fig. 5A).<sup>54,55</sup> These disordered regions harbor several binding sites. Specifically, binding partners for the N-terminal regions include MDM2, RPA 70N and the B subunit of RNA polymerase II.<sup>48</sup> The C-terminal contains the tetramerization domain that becomes ordered only upon forming a tetramer. The dynamic nature of this complex, however, is underlined by the presence of the nuclear localization signal hidden within this structure.<sup>56</sup> A remarkable example for the plasticity of disordered binding regions is presented by a short segment near the end of the sequence that was observed to bind to several partners, such as S100 $\beta$ , CBP, Cyclin A2 and sirtuin, in different local conformations.<sup>11</sup> Although p53 contains a significant amount of disorder that is essential for its central role, cancer-associated mutations are concentrated within the ordered DNA binding domain.<sup>54</sup>

**PTEN.** PTEN is also among the most frequently inactivated tumor suppressor genes in various cancers, with the second largest number of mutations collected in COSMIC. The PTEN gene encodes a dual specificity phosphatase that can act on both proteins and phosphoinositide substrates.<sup>57,58</sup> It negatively regulates the intracellular level of phosphatidylinositol-3,4,5-triphosphate in cells and functions as a tumor suppressor by

negatively regulating Akt/PKB signaling pathways. PTEN contains two key domains, the phosphatase (catalytic) domain, and the C2 (lipid membrane-binding) domain (Fig. 5B).<sup>59</sup> The C-terminal region is disordered, and the very end of the sequence contains the disordered binding region that can form a complex with the PDZ domain.<sup>60</sup> The observed cancer-associated mutations occur throughout the length of PTEN, but they are enriched in the C2 and in especially the phosphatase domains (Fig. 5B). Therefore, this protein is another example where cancer-associated mutations clearly prefer ordered regions. Although p53 and PTEN were thought to act independently as tumor suppressors, with an interesting twist, it turned out that they can interact both directly and indirectly.<sup>61</sup> The sites of the physical interaction were localized within the C2 domain of PTEN and the C-terminal region of p53, which is involved in multiple interactions.<sup>62</sup> The complex crosstalk between these two proteins is also supported by the recent finding demonstrating that PTEN and p53 somatic mutations are mutually exclusive in the case of human breast cancers.<sup>63</sup>

**$\beta$ -catenin.**  $\beta$ -catenin (CTNNB1) goes against the general trend with a significantly higher number of cancer-associated mutations falling into disordered segments.  $\beta$ -catenin is an essential structural component of the cadherin-based cell adhesion complex, and it is also involved in the Wnt/Wingless growth factor signaling pathway.<sup>64</sup> In cell adhesion,  $\beta$ -catenin helps link cadherin adhesion molecules to cytoskeletal actin filaments. In its signal transduction role,  $\beta$ -catenin functions as a transcriptional co-activator of target genes involved in cell

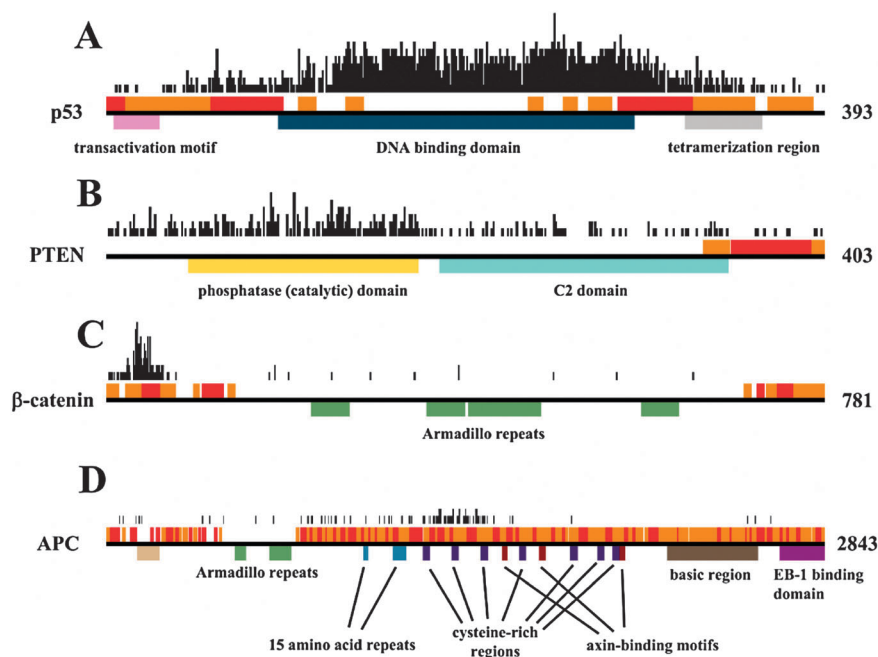


**Fig. 4** Two-dimensional mapping of various features based on the distances calculated on the COSMIC census database relative to the human proteome. Coordinates were obtained using multidimensional scaling (see Data and methods) by projecting the original Jaccard distances into two dimensions. The widths of the connecting lines are inversely proportional to the original Jaccard distances (see Table 3).

differentiation and proliferation.<sup>65</sup> The core region of  $\beta$ -catenin is composed of 12 copies of a 42 amino acid sequence motif known as an armadillo repeat (Fig. 5C). These repeats form a superhelix of helices that features a long, positively charged groove.<sup>66</sup> This groove mediates the interaction of  $\beta$ -catenin with several unrelated partners, largely based on charge

complementarity. Besides the central repeat region, the protein also contains short disordered segments on both termini. Nearly all cancer-associated mutations are located in the N-terminal disordered region (Fig. 5C). Of particular importance is the second predicted disordered binding site that also contains a short linear motif, the DSGxxS diphosphodegron. This region is recognized by the SCF- $\beta$ -TrCP E3 ligase with the binding site located at the top face of the  $\beta$ TrCP1 WD40  $\beta$ -propeller.<sup>67</sup> The complex formation targets  $\beta$ -catenin for proteasome destruction depending on the phosphorylation state of its degron.<sup>68</sup> Mutations in this region can lead to malignant transformation by increasing the cytoplasmic pool of  $\beta$ -catenin. This prompts its translocation to the nucleus, where it activates downstream elements of the Wnt pathway leading to cell overgrowth.<sup>64</sup>

**APC.** A large number of cancer-associated mutations within disordered regions are also present in another key element of the Wnt signalling pathway, the adenomatous polyposis coli (APC) protein.<sup>69</sup> Mutations of this protein frequently occur in colorectal tumors.<sup>70</sup> APC is a large (2843 residues) protein with several putative functions in cell cycle control, differentiation, migration, apoptosis, and the maintenance of chromosomal stability. It acts as a tumor suppressor based on its ability to bind to  $\beta$ -catenin and to promote its rapid degradation.<sup>71,72</sup> By downregulating CTNNB1, APC acts as a negative regulator in Wnt signaling. The central region of APC contains multiple  $\beta$ -catenin interaction motifs, including three 15 amino acid repeats and seven 20 amino acid cysteine-rich repeats.<sup>73,74</sup> The large majority of cancer-associated mutations are located within the first three 20 amino acid repeats (Fig. 5D). The protein contains several additional domains or motifs, such as



**Fig. 5** Domain structure, location of disordered binding regions and disordered segments and the number of cancer-associated mutations per position shown for (A) p53, (B) PTEN, (C)  $\beta$ -catenin and (D) APC. Black horizontal lines mark the full length proteins, colored boxes below show the various Pfam domains, red and orange boxes above show the disordered and disordered binding regions, respectively. The black boxes above the structural descriptions show the number of known cancer-associated mutations for each residue.

the oligomerization domain, armadillo repeats, axin binding repeats, basic region, and EB-1 interaction domain. With the exception of the armadillo repeats, these regions are largely disordered and contain several binding regions. Nevertheless, the cancer-associated mutations are significantly less frequent within these regions. This also indicates that it is not disorder in itself, but it is rather the specific function that can pose an increased biological risk in this case.

## Discussion

IDP regions are important elements of cancer-associated proteins.<sup>12</sup> In general, disordered proteins are fundamentally different from globular proteins both in their structural and functional properties. This necessitates the understanding of how these regions contribute to the development of cancer.

Ordered and disordered proteins are expected to differ in terms of their tolerance to mutations. The basic assumption is that neutral polymorphisms are less likely to occur in positions with stronger structural and functional constraints. In globular proteins, functionally relevant sites are often restricted to a few residues that form the active site, but nearly all residues contribute to the formation of the 3D structure at some level.<sup>75,76</sup> This represents a large evolutionary constraint for globular proteins. Functionally important residues of IDPs, such as residues directly involved in binding or undergoing post-translational modifications, can experience constraints similar to the active sites of globular proteins. In terms of structural constraints, however, mutations generally are expected to have a smaller impact on the structural properties of disordered segments, due to the lack of a well-defined structure. The increased evolutionary constraints of ordered residues compared to disordered ones have been observed at various evolutionary distances, ranging from human polymorphisms,<sup>28</sup> to the divergence between mouse and human.<sup>43</sup> Similar conclusions were drawn from the comparison of evolutionarily related sequences from different organisms that indicated that disordered segments were generally less conserved.<sup>44,45</sup> Deviations from this trend were observed in only a few cases and were mostly attributed to the involvement in protein–protein interactions.<sup>44</sup>

In complete agreement with this view, the larger tolerance to mutations of disordered segments was also present in cancer-associated proteins. Our results showed that a significantly fewer number of SNPs were observed in ordered regions compared to disordered regions in cancer-associated proteins. In contrast, cancer-associated mutations were more likely to occur within ordered segments. This effect was even larger, when the uneven distribution of polymorphisms was taken into account. These results suggest that disordered residues are more tolerant to mutations at two levels. Firstly, disordered regions can allow a larger number of genetic variations without affecting the function. Secondly, if a mutation occurs, it is more likely to cause cancer if the affected residue is located within an ordered region. The lower sensitivity of disordered regions to genetic variations is likely to originate from the specific structural properties of these regions. The analysis of disordered binding regions showed that functionally relevant sites within disordered regions can slightly deviate from this

behavior. Disordered binding regions could be placed between disordered regions in general and ordered regions, both in terms of the appearance of polymorphisms and cancer associated mutations. These suggest stronger evolutionary constraints within disordered binding regions, in accordance with their functional importance. Nevertheless, within the broader context of binding regions, only a few residues might be directly responsible for the specificity of the binding.<sup>77</sup> These residues could present even higher evolutionary constraints.

While results obtained on the various datasets agreed quite well, there were some variations. These differences can be associated with potential biases of the datasets. For example, since cancer genome projects rely on identifying nucleotide changes between normal and cancer cell lines at the level of genome, differences can also occur by random sites that are not actively involved in tumorigenesis. Mutations that occur randomly throughout the sequence do not bias our results, although they could decrease the statistical significance of the observed differences. However, we observed that neutral SNPs were not distributed randomly, but were more likely to occur within disordered regions. We accounted for this by using a different type of normalization. This leads to a more consistent picture with more pronounced differences, showing that cancer-associated mutations are more likely to occur within ordered regions. The normalization had the largest effect on the pure data of cancer genomes projects, where a higher number of non-disease causing mutations were expected. In the other cases, the results did not change much. Nevertheless, passenger mutations can also be present in the other databases. This is supported by the fact that only a few neutral polymorphisms were described in the case of our examples while they had the highest number of cancer-associated mutations. Due to the potential problems of passenger mutations, we used the term “cancer-associated mutation” throughout the manuscript. To weed out these mutations, further studies are needed. One of the important conclusions of our work is that such random mutations are not distributed evenly and affect disordered regions even more. This phenomenon should be taken into account in selecting driver mutations.

Other databases may suffer from different types of biases. For example, the SP\_cancer\_annotated dataset had a smaller percentage of disordered residues, in contrast to the increase of protein disorder in all other datasets. The preference of cancer associated mutations for ordered residues was also unusually high in this case. We suspect that the slightly different behavior in this case originates from the experimental biases of traditional approaches that could have favored ordered proteins. We could observe some differences within the various cancer genome projects as well, for example in the distribution of disordered binding regions (Fig. 3). The results obtained for breast and colorectal cancers agreed well in the two cases, but there were some differences when the CAN gene sets of glioblastoma and pancreatic cancer were considered. Although larger statistical variations can be expected in these cases due to the small size of these datasets, the results caution us that different types of cancer might be associated with different molecular and functional properties. Nevertheless, the 12 datasets analyzed in this work presented quite a consistent picture altogether, despite their different sizes, origins, and potential

biases. The consistency of these results lends confidence to our findings, showing that while in cancer genes neutral polymorphisms are more likely to occur within disordered regions, cancer-associated mutations are more common in ordered regions.

Our general finding is in contrast with the results obtained in the analyses of another major form of genetic aberrations leading to cancer, chromosomal translocations. In this case, a direct link between disorder and cancer was found.<sup>20</sup> This was rationalized based on that ordered proteins are more likely to be misfolded and degraded as a result of translocation, while disordered proteins could survive with an aberrant function.<sup>2,20</sup> A third form of commonly occurring genetic variations is copy number variation (CNV), which corresponds to the enrichment or depletion of certain genomic regions. CNVs are frequently observed in cancer and other diseases. In a recent study, a strong correlation between dosage sensitive gene products and protein disorder was found, and it was related to the interaction promiscuity of IDPs.<sup>21</sup> Interestingly, in two of the analyzed examples mutations affected disordered regions that regulated the level of  $\beta$ -catenin, a central element of the Wnt signalling pathway. These examples are in agreement with the observation that disordered proteins are generally under tight cellular control.<sup>78,79</sup> In contrast, the level of p53 is regulated by MDM2.<sup>52</sup> The specific binding site, however, did not show an increased rate of cancer-associated mutations (Fig. 5A). In order to resolve these seemingly contradictory results, cancer-associated mutations have to be placed into a network context. The network view was also suggested to be crucial in order to reduce the complexity of the landscape of cancer genomes.<sup>33</sup>

In conclusion, our results clearly show that protein disorder in itself is not responsible for the increased biological risk in terms of cancer-associated mutations. It seems plausible that the functional involvement of a protein determines both its disorder content and its involvement in cancer, thus presenting a correlation between these two features, without an existing casual link between them. Our study was restricted to single amino acid changes, however, other type of genetic alterations can also lead to cancer. A strong association between protein disorder and cancer was suggested in copy number variations or chromosomal translocations. The exploration of the role of protein disorder in these cases necessitates many further studies and taking into account the specific functions of these proteins and the way they are regulated. The present work, nevertheless, demonstrated that genetic mutations affect ordered and disordered regions in different ways, in accordance with the distinct structural and functional properties of these segments. In order to understand the background of various diseases, these differences have to be taken into account.

## Data and methods

### Datasets

**SwissProt cancer datasets.** We used three different resources to collect various cancer-associated genetic variations. The first dataset was downloaded from the UniProt/SwissProt Knowledgebase<sup>35</sup> and was derived primarily from literature

reports using strict inclusion criteria. This dataset contains polymorphisms with no clinical relevance, disease related amino acid mutations and some unclassified variants. Cancer-associated mutations were collected from the pre-compiled database available at <http://www.uniprot.org/docs/humsavar>. In the full dataset (SP\_cancer) those entries were kept, where the 'Disease name' field either matched one of the selection keywords ('cancer', 'tumor', 'lymphoma', 'leukemia', 'carcinoma', 'glioma', 'glioblastoma', 'melanoma' and 'sarcoma') or had an OMIM reference to a type of cancer (checked on the <http://www.omim.org/> site). A smaller list was also created by selecting the mutations from SP\_cancer that had 'Disease' annotations in the database omitting ones with 'Unclassified' tags (SP\_cancer\_annotated).

**Cancer genome project datasets.** The second type of datasets corresponded to four cancer genome projects collecting the result of comprehensive genome-wide analyses. Two of these studies described the mutations of breast and colorectal cancer (CGP\_br/col\_1 and CGP\_br/col\_2 datasets<sup>31,32</sup>), one focused on pancreatic cancer (CGP\_pan dataset<sup>33</sup>) and one on glioblastoma (CGP\_glio dataset<sup>34</sup>). In these studies, somatic mutations in cancer were determined by sequencing the major fraction of human genes and identifying nucleotide changes. Any alterations that were also present in normal samples or could be found in single-nucleotide polymorphism (SNP) databases were removed. The list of somatic mutations could still contain nonfunctional "passenger" alterations. To distinguish genes likely to contribute to tumorigenesis from those in which passenger mutations occurred by chance, a list of candidate cancer genes (CAN genes) was established based on the probability that the number of mutations in a given gene was greater than expected from the background mutation rate. The mutations described in these selected genes were used to compile the four datasets CGP\_CAN\_br/col\_1, CGP\_CAN\_br/col\_2, CGP\_CAN\_pan and CGP\_CAN\_glio. The list of gene identifiers and the nucleotide changes were downloaded from the supplementary materials of the original publications.

**COSMIC.** The third dataset was collected from the COSMIC database.<sup>36</sup> This is currently the most comprehensive catalogue of somatic mutations in cancer. Data are gathered from two sources, publications in the scientific literature (v52 contains 11 437 curated articles) and the full output of the genome-wide screens from the Cancer Genome Project (CGP) at the Sanger Institute, UK. This dataset also incorporated the outcome of cancer genome projects. A small subset of the COSMIC database was also part of the cancer census datasets that were casually linked to oncogenesis.<sup>37</sup> These genes constituted the COSMIC\_census dataset.

Although there could be some overlap between the three datasets, we opted to keep them separately in order to be able to observe any potential biases. Our analysis was restricted to single missense substitutions. Altogether we analyzed 12 different datasets. The number of proteins and mutations in each dataset are listed in Table 1.

**Polymorphisms.** In the case of SP datasets, the polymorphisms present in the SwissProt resource were also collected in the SP\_poly dataset and were used as reference.<sup>35</sup> In all other

cases, polymorphisms were collected using the UCSC Genome Browser.<sup>80</sup> Single genes were mapped to the genomic location corresponding to the UCSC Santa Cruz hg19/GRCh37 build. Those sequences, that could not be mapped, were changed or retracted, were discarded from further analyses. The polymorphism data were obtained by mapping the SNPs of dbSNP (release 132)<sup>25</sup> to the genomic coordinates. This release contained over 13 million SNPs. It also incorporated the results of the 1000 Genomes pilot projects that collected variations *via* whole genome shotgun sequencing from two families with high coverage and 179 individuals with low coverage.<sup>27</sup> We used the Common SNPs corresponding to uniquely mapped variants that appear in at least 1% of the population. The commonness of these variations suggests that these are likely to be neutral polymorphisms with no clinical relevance. To ensure the quality of the polymorphisms data, we only used validated SNPs.

All cancer associated mutations and polymorphisms were transformed into a common format specifying the used identifier of the sequence, the sequence position of the mutation and the original and mutated amino acids. This format enabled a simple selection at the level of unique mutations, therefore identical polymorphisms were only counted once. The numbers of polymorphisms are also listed in Table 1.

**Human proteome.** The proteins of the human proteome were downloaded from the “complete proteome” page of the UniProt database. Only reviewed entries were kept, resulting in a dataset of 20 232 proteins.

### Functional annotations

Functional classifications were based on GeneOntology (GO)<sup>81</sup> terms assigned to human proteins in UniProt. We retrieved all GO terms for all proteins in the human proteome and mapped them to high level GO terms described in the Generic GOSlim subset of GO. This subset contained 127 terms covering all three parts of GO annotations: biological processes (50 terms), cellular components (36 terms) and molecular functions (41 terms). All proteins from COSMIC, where possible, were mapped to UniProt sequences and were assigned the relevant GOSlim terms.

### Interactions

Protein–protein interactions were taken from the current release of the IntAct database ([www.ebi.ac.uk/intact/](http://www.ebi.ac.uk/intact/)).<sup>82</sup>

### Statistical analysis

**Comparison of the 12 cancer databases and the human proteome.** The average ratio of disordered residues, ratio of proteins containing >30 residue long disordered regions and average length were calculated in the 12 cancer datasets analyzed (Fig. 1). These averages were compared to the average values calculated in the human proteome. For each of the 12 datasets, standard errors of the mean were calculated by selecting 10 000 random samples from the human proteome of the same size as the respective dataset. In each of the 10 000 random selections, the means were calculated. From these means the standard error of the mean was established and used to test the difference between the random samples and the database average. The mean values, standard errors and the

appropriate *p*-values are shown in Table S1 (ESI†). Fig. 1 shows the confidence intervals of  $\alpha = 0.01$  (corresponding to 2.576 standard errors) in each case.

**Over- and under-representation of polymorphisms and cancer-associated mutations.** For each protein in our dataset, the sequences were downloaded from the Uniprot database or the UCSC Genome Browser. Using the sequence, the IUPred method<sup>39,40</sup> was used to assess which residues were part of disordered regions. These results were also calculated with two other disorder prediction methods, DISOPRED2<sup>5</sup> and VSL2.<sup>41</sup> The ANCHOR method<sup>48,49</sup> was used to predict regions involved in disordered binding regions. While there are several methods to predict disordered residues, ANCHOR is the only publicly available method for the prediction of disordered binding regions. For each protein, the number of polymorphisms and cancer-associated mutations within these regions were calculated. These numbers were compared to the expected number of mutations calculated in the following way: to calculate the expected number of mutations for ordered and disordered regions, the number of observed mutations was divided according to the ratio of ordered and disordered residues in the given sequence. This model takes into account that the number of mutations can change from one protein to another. The number of expected and observed mutations was summed up separately for ordered and disordered segments. Using these numbers, the statistical significance of the differences in the two distributions was assessed by the  $\chi^2$  test.

In this null model we assumed that the selection pressure on disordered and ordered regions is the same, and the probability that a mutation occurs in ordered or disordered regions is equal. We expect that the observed differences are mainly the result of selection acting at the protein level. It should be noted that other factors can also contribute to the selections, for example, by affecting the stability of DNA, mRNA, or interactions with regulatory factors. We checked that taking into account the different codon usage, or differences in transition–transversion rates does not affect our results.

In the case of cancer-associated mutations, an additional model was used to calculate the expected number of mutations. This took into account the uneven distribution of polymorphisms between ordered and disordered regions. The model was based on a normalization factor calculated from the ratio of the observed number of SNPs relative to their expected number. The normalization factor was calculated for disordered and ordered residues, in each dataset. The expected number of mutations was recalculated by weighting them according to the normalization factor for disordered and ordered residues within each dataset. Using these references, the statistical significance could be calculated similarly to the previous case. Unfortunately, current data do not enable us to calculate this factor for proteins individually. However, when datasets were divided into subgroups, for example based on the number of mutations, the results did not change.

**Distributions of functional categories.** The distribution of each GO term was analyzed using the COSMIC\_census dataset. To determine significantly over- or under-represented terms, the distribution of these terms in the human proteome

was used as a reference. A random subset was selected from the human proteome dataset and was parsed for occurrence numbers of each term. This was repeated 100 times and then the average occurrence of each term was calculated. These occurrence numbers were compared to the occurrence numbers in the COSMIC\_census dataset using left and right sided Fisher tests to assign significance values to the under- and over-representation of terms.

**Features.** The calculated length, ratio of disordered residues and disordered binding residues, interaction numbers and the number of COSMIC census mutations for COSMIC census proteins and the randomly selected reference human proteins were categorized into 5 bins to provide a coarse-grained description. The boundaries of the bins for each feature are shown in Table S4 (ESI<sup>†</sup>). The sixth feature describing the functional involvement of the proteins was represented by 'functional profiles'. These profiles were calculated based on the significantly over- and under-represented GO terms shown in Table 3. For each protein, a 13 element binary vector was assigned that showed which of the 13 considered GO terms the protein was annotated with.

**Mutual information and Jaccard distance.** The association between different features calculated on proteins was measured by calculating the mutual information ( $I(X, Y)$ ) between all  $X$  and  $Y$  pairs of features using the standard formula:

$$I(X, Y) = \sum_x \sum_y p(x, y) \log_2 \left( \frac{p(x, y)}{p'(x)p''(y)} \right)$$

where  $p'(x)$  and  $p''(y)$  are the probability distributions of the features  $X$  and  $Y$  respectively and  $p(x, y)$  is their joint probability distribution. As the maximal information of different features can vary (and hence their maximal mutual information can also vary), to be able to compare the association of different parameter pairs directly, the mutual information was scaled:

$$D(X, Y) = 1 - \frac{I(X, Y)}{H(X, Y)}$$

where  $H(X, Y)$  is the joint entropy of  $X$  and  $Y$ :

$$H(X, Y) = - \sum_x \sum_y p(x, y) \log_2 p(x, y)$$

The resulting  $D(X, Y)$  Jaccard distance is a universal metric with  $D(X, Y) = 1$  if  $X$  and  $Y$  are completely independent and  $D(X, Y) = 0$  if  $X$  and  $Y$  are identical.

The multidimensional scaling of the obtained distances was calculated using the R package.

## Acknowledgements

This work was sponsored by the Hungarian Scientific Research Fund (OTKA) [grant number K 72569]. The Bolyai Janos fellowship for Z.D. and Charles Simonyi fellowship for I.S. are also gratefully acknowledged. We would like to thank Lajos Kalmár for his critical comments on the project.

## References

- 1 A. K. Dunker, J. D. Lawson, C. J. Brown, R. M. Williams, P. Romero, J. S. Oh, C. J. Oldfield, A. M. Campen, C. M. Ratliff, K. W. Hipps, J. Ausio, M. S. Nissen, R. Reeves, C. Kang, C. R. Kissinger, R. W. Bailey, M. D. Griswold, W. Chiu, E. C. Garner and Z. Obradovic, *J. Mol. Graphics Modell.*, 2001, **19**, 26–59.
- 2 H. J. Dyson and P. E. Wright, *Nat. Rev. Mol. Cell Biol.*, 2005, **6**, 197–208.
- 3 P. Tompa, *Trends Biochem. Sci.*, 2002, **27**, 527–533.
- 4 A. K. Dunker, Z. Obradovic, P. Romero, E. C. Garner and C. J. Brown, *Genome Inf. Ser. Workshop Genome Inf.*, 2000, **11**, 161–171.
- 5 J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton and D. T. Jones, *J. Mol. Biol.*, 2004, **337**, 635–645.
- 6 P. Tompa, *FEBS Lett.*, 2005, **579**, 3346–3354.
- 7 K. Gunasekaran, C. J. Tsai, S. Kumar, D. Zanuy and R. Nussinov, *Trends Biochem. Sci.*, 2003, **28**, 81–85.
- 8 B. Mészáros, P. Tompa, I. Simon and Z. Dosztányi, *J. Mol. Biol.*, 2007, **372**, 549–561.
- 9 A. K. Dunker, E. Garner, S. Guillot, P. Romero, K. Albrecht, J. Hart, Z. Obradovic, C. Kissinger and J. E. Villafranca, *Pac. Symp. Biocomput.*, 1998, 473–484.
- 10 H. J. Dyson and P. E. Wright, *Curr. Opin. Struct. Biol.*, 2002, **12**, 54–60.
- 11 C. J. Oldfield, J. Meng, J. Y. Yang, M. Q. Yang, V. N. Uversky and A. K. Dunker, *BMC Genomics*, 2008, **9**(suppl 1), S1.
- 12 L. M. Iakoucheva, P. Radivojac, C. J. Brown, T. R. O'Connor, J. G. Sikes, Z. Obradovic and A. K. Dunker, *Nucleic Acids Res.*, 2004, **32**, 1037–1049.
- 13 C. A. Galea, Y. Wang, S. G. Sivakolundu and R. W. Kriwacki, *Biochemistry*, 2008, **47**, 7598–7609.
- 14 H. Xie, S. Vucetic, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, V. N. Uversky and Z. Obradovic, *J. Proteome Res.*, 2007, **6**, 1882–1898.
- 15 C. Haynes, C. J. Oldfield, F. Ji, N. Klitgord, M. E. Cusick, P. Radivojac, V. N. Uversky, M. Vidal and L. M. Iakoucheva, *PLoS Comput. Biol.*, 2006, **2**, e100.
- 16 Z. Dosztányi, J. Chen, A. K. Dunker, I. Simon and P. Tompa, *J. Proteome Res.*, 2006, **5**, 2985–2995.
- 17 V. N. Uversky, C. J. Oldfield and A. K. Dunker, *Annu. Rev. Biophys.*, 2008, **37**, 215–246.
- 18 B. Vogelstein, D. Lane and A. J. Levine, *Nature*, 2000, **408**, 307–310.
- 19 L. M. Iakoucheva, C. J. Brown, J. D. Lawson, Z. Obradovic and A. K. Dunker, *J. Mol. Biol.*, 2002, **323**, 573–584.
- 20 H. Hegyi, L. Buday and P. Tompa, *PLoS Comput. Biol.*, 2009, **5**, e1000552.
- 21 T. Vavouri, J. I. Semple, R. Garcia-Verdugo and B. Lehner, *Cell (Cambridge, Mass.)*, 2009, **138**, 198–208.
- 22 Y. Cheng, T. LeGall, C. J. Oldfield, A. K. Dunker and V. N. Uversky, *Biochemistry*, 2006, **45**, 10448–10460.
- 23 V. N. Uversky, *Front. Biosci.*, 2009, **14**, 5188–5238.
- 24 P. Tompa, *FEBS J.*, 2009, **276**, 5406–5415.
- 25 S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski and K. Sirotkin, *Nucleic Acids Res.*, 2001, **29**, 308–311.
- 26 D. F. Conrad, J. E. Keebler, M. A. Depristo, S. J. Lindsay, Y. Zhang, F. Casals, Y. Idaghdour, C. L. Hartl, C. Torroja, K. V. Garimella, M. Zilversmit, R. Cartwright, G. A. Rouleau, M. Daly, E. A. Stone, M. E. Hurles and P. Awadalla, *Nat. Genet.*, 2011, **43**, 712–714.
- 27 The 1000 Genomes Project Consortium, *Nature*, 2010, **467**, 1061–1073.
- 28 J. Liu, Y. Zhang, X. Lei and Z. Zhang, *Genome Biol.*, 2008, **9**, R69.
- 29 A. Bardelli and V. E. Velculescu, *Curr. Opin. Genet. Dev.*, 2005, **15**, 5–12.
- 30 H. Ledford, *Nature*, 2010, **464**, 972–974.
- 31 T. Sjöblom, S. Jones, L. D. Wood, D. W. Parsons, J. Lin, T. D. Barber, D. Mandelker, R. J. Leary, J. Ptak, N. Silliman, S. Szabo, P. Buckhaults, C. Farrell, P. Meeh, S. D. Markowitz, J. Willis, D. Dawson, J. K. Willson, A. F. Gazdar, J. Hartigan, L. Wu, C. Liu, G. Parmigiani, B. H. Park, K. E. Bachman,

- N. Papadopoulos, B. Vogelstein, K. W. Kinzler and V. E. Velculescu, *Science*, 2006, **314**, 268–274.
- 32 L. D. Wood, D. W. Parsons, S. Jones, J. Lin, T. Sjoblom, R. J. Leary, D. Shen, S. M. Boca, T. Barber, J. Ptak, N. Silliman, S. Szabo, Z. Dezso, V. Ustyanksky, T. Nikolskaya, Y. Nikolsky, R. Karchin, P. A. Wilson, J. S. Kaminker, Z. Zhang, R. Croshaw, J. Willis, D. Dawson, M. Shipitsin, J. K. Willson, S. Sukumar, K. Polyak, B. H. Park, C. L. Pethiyagoda, P. V. Pant, D. G. Ballinger, A. B. Sparks, J. Hartigan, D. R. Smith, E. Suh, N. Papadopoulos, P. Buckhaults, S. D. Markowitz, G. Parmigiani, K. W. Kinzler, V. E. Velculescu and B. Vogelstein, *Science*, 2007, **318**, 1108–1113.
- 33 S. Jones, X. Zhang, D. W. Parsons, J. C. Lin, R. J. Leary, P. Angenendt, P. Mankoo, H. Carter, H. Kamiyama, A. Jimeno, S. M. Hong, B. Fu, M. T. Lin, E. S. Calhoun, M. Kamiyama, K. Walter, T. Nikolskaya, Y. Nikolsky, J. Hartigan, D. R. Smith, M. Hidalgo, S. D. Leach, A. P. Klein, E. M. Jaffee, M. Goggins, A. Maitra, C. Iacobuzio-Donahue, J. R. Eshleman, S. E. Kern, R. H. Hruban, R. Karchin, N. Papadopoulos, G. Parmigiani, B. Vogelstein, V. E. Velculescu and K. W. Kinzler, *Science*, 2008, **321**, 1801–1806.
- 34 D. W. Parsons, S. Jones, X. Zhang, J. C. Lin, R. J. Leary, P. Angenendt, P. Mankoo, H. Carter, I. M. Siu, G. L. Gallia, A. Olivi, R. McLendon, B. A. Rasheed, S. Keir, T. Nikolskaya, Y. Nikolsky, D. A. Busam, H. Tekleab, L. A. Diaz, Jr., J. Hartigan, D. R. Smith, R. L. Strausberg, S. K. Marie, S. M. Shinjo, H. Yan, G. J. Riggins, D. D. Bigner, R. Karchin, N. Papadopoulos, G. Parmigiani, B. Vogelstein, V. E. Velculescu and K. W. Kinzler, *Science*, 2008, **321**, 1807–1812.
- 35 Y. L. Yip, M. Famiglietti, A. Gos, P. D. Duek, F. P. David, A. Gateau and A. Bairoch, *Hum. Mutat.*, 2008, **29**, 361–366.
- 36 S. A. Forbes, N. Bindal, S. Bamford, C. Cole, C. Y. Kok, D. Beare, M. Jia, R. Shepherd, K. Leung, A. Menzies, J. W. Teague, P. J. Campbell, M. R. Stratton and P. A. Futreal, *Nucleic Acids Res.*, 2011, **39**, D945–D950.
- 37 P. A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman and M. R. Stratton, *Nat. Rev. Cancer*, 2004, **4**, 177–183.
- 38 UniProt Consortium, *Nucleic Acids Res.*, 2011, **39**, D214–D219.
- 39 Z. Dosztányi, V. Csizmók, P. Tompa and I. Simon, *Bioinformatics*, 2005, **21**, 3433–3434.
- 40 Z. Dosztányi, V. Csizmók, P. Tompa and I. Simon, *J. Mol. Biol.*, 2005, **347**, 827–839.
- 41 K. Peng, P. Radivojac, S. Vucetic, A. K. Dunker and Z. Obradovic, *BMC Bioinf.*, 2006, **7**, 208.
- 42 S. J. Furney, D. G. Higgins, C. A. Ouzounis and N. Lopez-Bigas, *BMC Genomics*, 2006, **7**, 3.
- 43 Y. Xia, E. A. Franzosa and M. B. Gerstein, *PLoS Comput. Biol.*, 2009, **5**, e1000413.
- 44 C. J. Brown, S. Takayama, A. M. Campen, P. Vise, T. W. Marshall, C. J. Oldfield, C. J. Williams and A. K. Dunker, *J. Mol. Evol.*, 2002, **55**, 104–110.
- 45 C. J. Brown, A. K. Johnson, A. K. Dunker and G. W. Daughdrill, *Curr. Opin. Struct. Biol.*, 2011, **21**, 441–446.
- 46 B. Mészáros, P. Tompa, I. Simon and Z. Dosztányi, *J. Mol. Biol.*, 2007, **372**, 549–561.
- 47 B. Mészáros, I. Simon and Z. Dosztányi, *Phys. Biol.*, 2011, **8**, 035003.
- 48 B. Mészáros, I. Simon and Z. Dosztányi, *PLoS Comput. Biol.*, 2009, **5**, e1000376.
- 49 Z. Dosztányi, B. Mészáros and I. Simon, *Bioinformatics*, 2009, **25**, 2745–2746.
- 50 Y. Cheng, C. J. Oldfield, J. Meng, P. Romero, V. N. Uversky and A. K. Dunker, *Biochemistry*, 2007, **46**, 13468–13477.
- 51 R. D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J. E. Pollington, O. L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E. L. Sonnhammer, S. R. Eddy and A. Bateman, *Nucleic Acids Res.*, 2010, **38**, D211–D222.
- 52 C. A. Brady and L. D. Attardi, *J. Cell Sci.*, 2010, **123**, 2527–2532.
- 53 T. Soussi and K. G. Wiman, *Cancer Cell*, 2007, **12**, 303–312.
- 54 A. C. Joerger and A. R. Fersht, *Adv. Cancer Res.*, 2007, **97**, 1–23.
- 55 S. Bell, C. Klein, L. Muller, S. Hansen and J. Buchner, *J. Mol. Biol.*, 2002, **322**, 917–927.
- 56 C. Addison, J. R. Jenkins and H. W. Sturzbecher, *Oncogene*, 1990, **5**, 423–426.
- 57 M. C. Hollander, G. M. Blumenthal and P. A. Dennis, *Nat. Rev. Cancer*, 2011, **11**, 289–301.
- 58 T. Maehama and J. E. Dixon, *Trends Cell Biol.*, 1999, **9**, 125–128.
- 59 J. O. Lee, H. Yang, M. M. Georgescu, A. Di Cristofano, T. Maehama, Y. Shi, J. E. Dixon, P. Pandolfi and N. P. Pavletich, *Cell (Cambridge, Mass.)*, 1999, **99**, 323–334.
- 60 W. Feng, H. Wu, L. N. Chan and M. Zhang, *J. Biol. Chem.*, 2008, **283**, 23440–23449.
- 61 L. C. Trotman and P. P. Pandolfi, *Cancer Cell*, 2003, **3**, 97–99.
- 62 D. J. Freeman, A. G. Li, G. Wei, H. H. Li, N. Kertesz, R. Lesche, A. D. Whale, H. Martinez-Diaz, N. Rozengurt, R. D. Cardiff, X. Liu and H. Wu, *Cancer Cell*, 2003, **3**, 117–130.
- 63 K. Kurose, K. Gilley, S. Matsumoto, P. H. Watson, X. P. Zhou and C. Eng, *Nat. Genet.*, 2002, **32**, 355–357.
- 64 L. Shapiro, *Structure (London)*, 1997, **5**, 1265–1268.
- 65 B. M. Gumbiner, *Curr. Opin. Cell Biol.*, 1995, **7**, 634–640.
- 66 A. H. Huber, W. J. Nelson and W. I. Weis, *Cell (Cambridge, Mass.)*, 1997, **90**, 871–882.
- 67 G. Wu, G. Xu, B. A. Schulman, P. D. Jeffrey, J. W. Harper and N. P. Pavletich, *Mol. Cell*, 2003, **11**, 1445–1456.
- 68 M. Al-Fageeh, Q. Li, W. M. Dashwood, M. C. Myzak and R. H. Dashwood, *Oncogene*, 2004, **23**, 4839–4846.
- 69 P. Polakis, *Biochim. Biophys. Acta*, 1997, **1332**, F127–F147.
- 70 K. W. Kinzler and B. Vogelstein, *Cell (Cambridge, Mass.)*, 1996, **87**, 159–170.
- 71 B. Rubinfeld, B. Souza, I. Albert, O. Muller, S. H. Chamberlain, F. R. Masiarz, S. Munemitsu and P. Polakis, *Science*, 1993, **262**, 1731–1734.
- 72 L. K. Su, B. Vogelstein and K. W. Kinzler, *Science*, 1993, **262**, 1734–1737.
- 73 Y. Xing, W. K. Clements, I. Le Trong, T. R. Hinds, R. Stenkamp, D. Kimelman and W. Xu, *Mol. Cell*, 2004, **15**, 523–533.
- 74 K. Eklof Spink, S. G. Fridman and W. I. Weis, *EMBO J.*, 2001, **20**, 6203–6212.
- 75 M. Guharoy and P. Chakrabarti, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 15447–15452.
- 76 M. Landau, I. Mayrose, Y. Rosenberg, F. Glaser, E. Martz, T. Pupko and N. Ben-Tal, *Nucleic Acids Res.*, 2005, **33**, W299–W302.
- 77 M. Fuxreiter, P. Tompa and I. Simon, *Bioinformatics*, 2007, **23**, 950–956.
- 78 J. Gsponer, M. E. Futschik, S. A. Teichmann and M. M. Babu, *Science*, 2008, **322**, 1365–1368.
- 79 M. M. Babu, R. van der Lee, N. S. de Groot and J. Gsponer, *Curr. Opin. Struct. Biol.*, 2011, **21**, 432–440.
- 80 J. Z. Sanborn, S. C. Benz, B. Craft, C. Szeto, K. M. Kober, L. Meyer, C. J. Vaske, M. Goldman, K. E. Smith, R. M. Kuhn, D. Karolchik, W. J. Kent, J. M. Stuart, D. Haussler and J. Zhu, *Nucleic Acids Res.*, 2011, **39**, D951–D959.
- 81 M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock, *Nat. Genet.*, 2000, **25**, 25–29.
- 82 B. Aranda, P. Achuthan, Y. Alam-Faruque, I. Armean, A. Bridge, C. Derow, M. Feuermann, A. T. Ghanbarian, S. Kerrien, J. Khadake, J. Kerssemakers, C. Leroy, M. Menden, M. Michaut, L. Montecchi-Palazzi, S. N. Neuhauser, S. Orchard, V. Perreau, B. Roechert, K. van Eijk and H. Hermjakob, *Nucleic Acids Res.*, 2010, **38**, D525–D531.



## Article

# Mutations of Intrinsically Disordered Protein Regions Can Drive Cancer but Lack Therapeutic Strategies

Bálint Mészáros <sup>1,2</sup>, Borbála Hajdu-Soltész <sup>1</sup>, András Zeke <sup>3</sup> and Zsuzsanna Dosztányi <sup>1,\*</sup>

<sup>1</sup> Department of Biochemistry, ELTE Eötvös Loránd University, H-1117 Budapest, Hungary; balint.meszáros@embl.de (B.M.); soltesz.borbala@gmail.com (B.H.-S.)

<sup>2</sup> EMBL Heidelberg, Meyerhofstraße 1, 69117 Heidelberg, Germany

<sup>3</sup> Institute of Enzymology, RCNS, P.O. Box 7, H-1518 Budapest, Hungary; zeke.andras@ttk.mta.hu

\* Correspondence: zsuzsanna.dosztanyi@ttk.elte.hu; Tel.: +36-1-372 2500/8537

**Abstract:** Many proteins contain intrinsically disordered regions (IDRs) which carry out important functions without relying on a single well-defined conformation. IDRs are increasingly recognized as critical elements of regulatory networks and have been also associated with cancer. However, it is unknown whether mutations targeting IDRs represent a distinct class of driver events associated with specific molecular and system-level properties, cancer types and treatment options. Here, we used an integrative computational approach to explore the direct role of intrinsically disordered protein regions driving cancer. We showed that around 20% of cancer drivers are primarily targeted through a disordered region. These IDRs can function in multiple ways which are distinct from the functional mechanisms of ordered drivers. Disordered drivers play a central role in context-dependent interaction networks and are enriched in specific biological processes such as transcription, gene expression regulation and protein degradation. Furthermore, their modulation represents an alternative mechanism for the emergence of all known cancer hallmarks. Importantly, in certain cancer patients, mutations of disordered drivers represent key driving events. However, treatment options for such patients are currently severely limited. The presented study highlights a largely overlooked class of cancer drivers associated with specific cancer types that need novel therapeutic options.

**Keywords:** intrinsically disordered regions; protein modules; short linear motifs; molecular switches; cancer genomics; driver gene identification; cancer hallmarks; drug targets



**Citation:** Mészáros, B.; Hajdu-Soltész, B.; Zeke, A.; Dosztányi, Z. Mutations of Intrinsically Disordered Protein Regions Can Drive Cancer but Lack Therapeutic Strategies. *Biomolecules* **2021**, *11*, 381. <https://doi.org/10.3390/biom11030381>

Academic Editor: Prakash Kulkarni

Received: 25 January 2021

Accepted: 24 February 2021

Published: 4 March 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

The identification of cancer driver genes and understanding their mechanisms of action is necessary for developing efficient therapeutics [1]. Many cancer-associated genes encode proteins that are modular, containing not only globular domains but also intrinsically disordered proteins/regions (IDPs/IDRs) [2–4]. IDRs can be characterized by conformational ensembles; however, the detailed properties of these ensembles can vary greatly from largely random-like behavior to exhibiting strong structural preferences, with the length of these segments ranging from a few residues to domain-sized segments [5–7]. The function of IDRs relies on their inherent conformational heterogeneity and plasticity, enabling them to act as flexible linkers or entropic chains, mediate transient interactions through linear motifs, direct the assembly of macromolecular assemblies or even drive the formation of membraneless organelles through liquid–liquid phase separation [5–8]. In general, disordered regions are core components of interaction networks and fulfill critical roles in regulation and signaling [4]. In accordance with their crucial functions, IDPs are often associated with various diseases [9], in particular with cancer. The prevalence of protein disorder among cancer-associated proteins was generally observed [10]. However, cancer-associated missense mutations showed a strong preference for ordered regions, which indicates that the association between protein disorder and cancer might be indirect [11]. Nevertheless, a direct link between protein disorder and cancer was suggested in

the case of two common forms of generic alterations: chromosomal rearrangements [12] and copy number variations [13]. Cancer mutations were shown to occur within linear motif sites located in IDRs [14]. In a specific case, the creation of IDR-mediated interactions was suggested to lead to tumorigenesis [15]. However, it has not been systematically analyzed whether mutations of IDRs can have a direct role driving cancer development or what the main molecular functions and biological processes altered by such events are.

In recent years, thousands of human cancer genomes have become available through large-scale sequencing efforts. The collected genetic variations revealed that cancer samples are heterogeneous and contain a large number of randomly occurring, so-called passenger mutations. Therefore, one of the main challenges for the interpretation of cancer genomics data is the identification of genes whose mutations actively contribute to cancer development, the so-called driver genes. When samples are analyzed in combination, various patterns start to emerge that enable the identification of cancer driving genes [16]. These signals can highlight genes which are frequently mutated in specific types of cancer [17,18], biological processes/pathways that are commonly altered in tumor development [19,20] or traits that govern tumorigenic transformation of cells [21]. The positional accumulation of mutations within specific ordered structures, domains or interaction surfaces was also shown to be a strong indicator of cancer driver roles [22–26]. The number of driver genes is currently estimated to be in the low to mid hundreds [27], but this number could increase with the growing number of sequenced cancer genomes [18]. However, most of the known, well-characterized driver genes are associated with ordered domains of proteins. Overall, the structural and functional properties of the affected proteins determine their oncogenic or tumor suppressor roles, which, in the case of context-dependent genes, can also depend on tissue type or the stage of tumor progression.

The complex relationship between protein disorder and cancer can be demonstrated through two well-characterized examples, p53 (corresponding to gene TP53) and  $\beta$ -catenin (CTNNB1). As a tumor suppressor, p53 is most commonly altered by truncating mutations, but it also contains a large number of missense variations. Mutations collected from multiple patients across different cancer types tend to cluster within the central region of p53 which corresponds to the ordered DNA-binding domain [28]. In contrast, significantly fewer mutations correspond to the disordered N- and C-terminal regions which are involved in numerous, sometimes overlapping protein–protein interactions [29]. In particular, almost no mutations are located within the N-terminal region corresponding to a so-called degron motif, a linear motif site recognized by the E3 ligase MDM2 that plays a critical role in regulating the degradation of p53 [30]. Furthermore, the tetramerization domain in the C-terminal part is also less affected by cancer mutations. This region represents a so-called disordered domain, a conserved region that forms a well-defined structure in its oligomeric form. The tetrameric ordered structure masks a nuclear export signal, which needs to become exposed for the proper function of p53, highlighting the intrinsic dynamic properties of this region [31]. The oncogenic  $\beta$ -catenin presents a completely different scenario. In terms of domain organization,  $\beta$ -catenin also contains a disordered N- and C-terminal and an ordered domain in between [32]. However, in this case the cancer mutations are largely localized to a short segment within the N-terminal disordered region which corresponds to the key degron motif regulating the cellular level of  $\beta$ -catenin in the absence of Wnt signalling [11,14].

The aim of this work was to explore if other IDRs, similarly to  $\beta$ -catenin, play a potential driver role in cancer. Based on cancer mutations collected from genome-wide screens and targeted studies [33], we identified significantly mutated protein regions [34] and classified them into ordered and disordered regions by integrating experimental structural knowledge and predictions. Automated and high-quality manually curated information was gathered for the collected examples to gain better insights into their functional and system-level properties, and to confirm their roles in tumorigenic processes. We aimed to answer the following questions: What are the characteristic molecular mechanisms, biological processes and protein–protein interaction network roles associated with proteins

mutated at IDRs? At a more generic level, how fundamental is the contribution of IDPs to tumorigenesis? Are IDP mutations just accessory events, or can they be the dominant molecular background to the emergence of cancer? Is there a characteristic difference in terms of treatment options between patient samples targeted mostly within ordered and disordered regions?

## 2. Material and Methods

### 2.1. Identification of Driver Regions in Cancer-Associated Proteins

To collect mutation data, cancer mutations were retrieved from the v83 version of COSMIC (Catalogue Of Somatic Mutations In Cancer) [33] and the v6.0 version of TCGA (The Cancer Genome Atlas). Mutations used from both databases included only missense mutations and in-frame insertions and deletions. Mutations were filtered similarly to the procedure described in [34]. Mutations from samples with over 100 mutations were discarded to avoid the inclusion of hypermutated samples. Samples including a large number of mutations in pseudogenes or mutations indicated as possible sequencing/assembly errors in [35] were also discarded. Redundant samples were filtered out. Mutations falling into positions of known common polymorphisms [36] or genomically unconserved regions based on the PhastCons method [37] were filtered out. The final set of COSMIC mutations used as an input to region identification consisted of 599,137 missense mutations, 4189 insertions and 12,670 deletions from 253,568 samples. The final set of TCGA mutations used as an input to region identification consisted of 274,109 missense mutations, 2775 insertions and 2900 deletions from 7058 samples.

Driver regions were identified using iSiMPRe [34] with the filtered mutations from COSMIC and TCGA, separately. Then, regions obtained from COSMIC and TCGA mutations were merged, and *p*-values for significance were kept from the dataset with the higher significance. Only regions with high significance (*p*-values lower than  $10^{-6}$ ) were kept.

### 2.2. Structural Categorization of Driver Regions

Regions were assigned ordered or disordered status based on the structural annotation of the corresponding functional unit, incorporating experimental data as well as predictions. For this, we collected experimentally verified annotations for disorders from the DisProt [38] and IDEAL [39] databases, and for disordered binding regions from the DIBS [40] and MFIB [31] databases. We also mapped known PDB structures [41]. Structure of a monomeric single domain protein chain was taken as a direct evidence for order. In contrast, missing residues in case of X-ray structures and mobile regions calculated for NMR ensembles using the CYRANGE method [42] were taken as indication of disorder. Pfam families annotated as the domain type were considered as ordered, while families annotated as disordered were assigned as disordered. All these types of evidence were extended by homology transfer.

Pfam entities with no instances overlapping with any protein regions with a clear structural designation were annotated using predictions, together with protein residues not covered by known structural modules. Such protein regions were defined as ordered or disordered using predictions from IUPred [43,44] and ANCHOR [45,46]. Residues predicted to be disordered or to be part of a disordered binding region, together with their 10 residue flanking regions, were considered to form disordered modules. Regions shorter than 10 residues were discarded. Regions annotated as disordered were also checked using additional prediction methods using the MobiDB database [47] and structure prediction using HHPred [48]. The final ordered/disordered status of the identified regions was based on manual assertion, taking into account information from the literature if available (Supplementary Table S1). For the disordered regions, the level of supporting information for the disordered region is also included (Supplementary Table S2). Please note that we use gene symbols to refer to their protein products throughout the manuscript, with corresponding names of protein products also specified in the Supplementary Table S2.

### 2.3. System-Level Analyses

Gene Ontology terms (GO) [49,50] were used to quantify interaction capabilities, involvement in various biological processes, molecular toolkits and hallmarks of cancer. In each case, a separate collection of GO terms (termed GO Slim) was compiled. Each GO Slim features a manual selection of GO terms that are independent from each other, meaning that they are neither child or parent terms of each other. Terms were assigned a level showing the fewest number of successive parent terms that include the root term of the ontology namespace (considered to be level 0).

GO term enrichments in a set of proteins were calculated by first obtaining expected values. Expected mean occurrence values for GO terms together with standard deviations were calculated by assessing randomly selected protein sets from the background (the full human proteome) 1000 times. The enrichment in the studied set is expressed as the difference from the expected mean in standard deviation units.

GO for molecular toolkits: biological\_process terms attached to proteins with identified regions were filtered for ancestry. The resulting set was manually filtered, yielding 93 terms which were manually grouped into 16 toolkits. Enrichments for toolkits were calculated as the ratio of the sum of expected and observed values for individual terms. Individual terms and enrichments for each toolkit are shown in Supplementary Table S3.

GO Slim for assessing interaction capacity: Terms from levels 1–4 from the molecular\_function namespace were filtered for ancestry and only the more specific terms were kept, i.e., terms from levels 1–3 were only included if they had no child terms. Only terms describing interactions containing the keyword “binding” were kept. Individual terms are shown in Supplementary Table S4.

GO for the assessment of process overlaps: Terms from levels 1–4 from the biological\_process namespace were filtered for ancestry and only the most specific terms were kept. Only those terms were considered that were attached to at least one protein from the set studied (full human proteome, ordered drivers or disordered drivers). Individual terms are shown in Supplementary Table S5.

GO for hallmarks of cancer: Terms were chosen from the biological\_process namespace via manual curation using the GO annotations of known cancer genes as a starting point. Terms were only kept if they showed a significant ( $p < 0.01$ ) enrichment on proteins in the full census cancer driver set compared to randomly selected human proteins. Individual terms and enrichments for each hallmark are shown in Supplementary Table S6.

To characterize the network properties of the selected examples, binary protein–protein interactions for the human proteome were downloaded from the IntAct database [51] on 06 May 2018. Data were filtered for human–human interactions, where interaction partners were identified by UniProt accessions. Interactions from spoke expansions were excluded. Interactions were kept in an undirected way. (Values for disordered drivers are quoted in Supplementary Table S2).

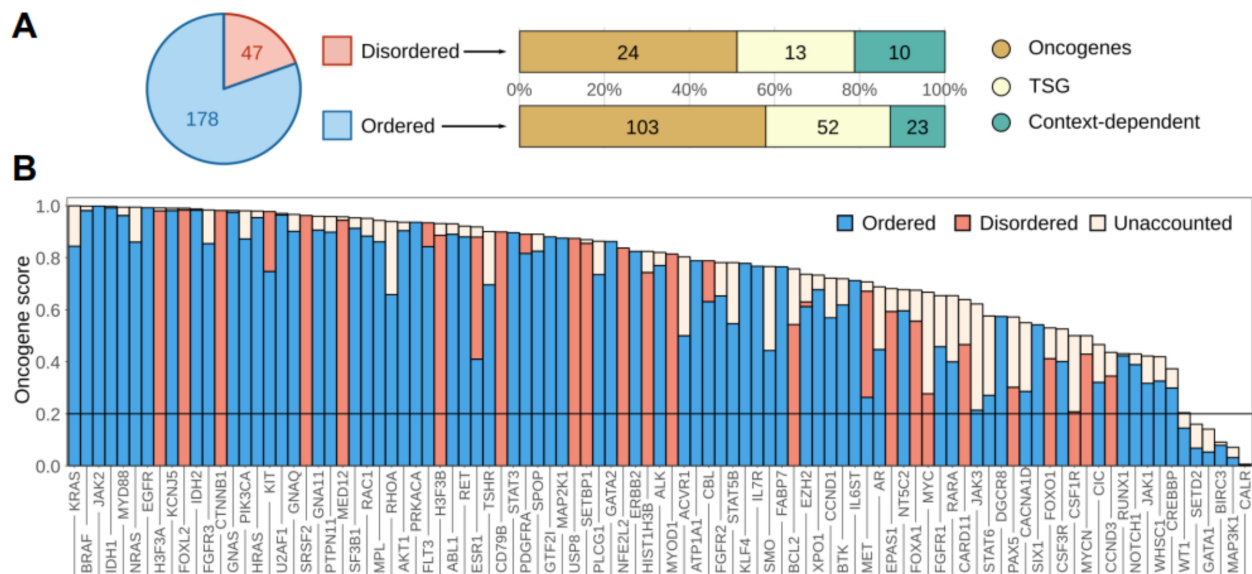
## 3. Results

### 3.1. Disordered Protein Modules Are Targets for Tumorigenic Mutations

For the purpose of our analysis, it was necessary to use an approach that could identify not only cancer drivers, but also the specific regions directly targeted by cancer mutations. We used the iSiMPRe [34] method, which can highlight significantly mutated regions without prior assumptions about the type or the size of such regions and was shown to perform similarly to other methods in identifying cancer drivers [52]. Cancer mutations were collected from the COSMIC and TCGA databases and were pre-filtered (see Section 2.1). The filtering steps were necessary to eliminate cases with a random accumulation of mutations with no biological significance, especially in the case of IDRs [34]. We restricted our analysis to high-confidence cases to minimize the chance of false positives. The order/disorder status of the identified significantly mutated regions was determined based on experimental data or homology transfer, when available, or by using a combination of prediction

approaches (See Data and Methods). Cancer drivers were manually characterized as tumor suppressors (TSGs), oncogenes and context-dependent cancer genes based on the literature.

Altogether, we identified 178 ordered and 47 disordered driver regions in 145 proteins from the human proteome (Supplementary Table S1, Figure 1A). The ratio of disordered driver regions was lower than expected on the ratio of disordered residues (21% vs. 30%). This was the case for both oncogenes and tumor suppressor genes, but not for context-dependent genes. Further underlining the relevance of IDRs, context-dependent cancer drivers also had more residues and mutations within disordered regions in general, together with a slightly higher proportion of disordered drivers (see Supplementary Figure S1).



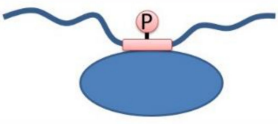
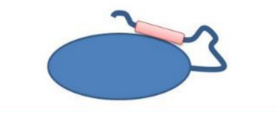
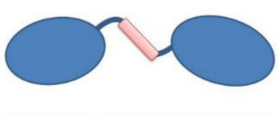
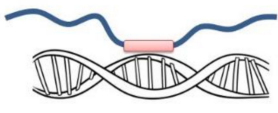


**Figure 1.** The distribution of ordered and disordered driver protein regions. **(A)** The distribution of ordered and disordered driver protein regions and their distribution among oncogenes, tumor suppressor genes (TSG) and context-dependent genes. **(B)** Oncogene scores of full genes and oncogene scores explained by the identified regions in oncogenes and context-dependent driver genes. “Unaccounted” corresponds to the fraction of mutations not in the identified, high significance regions.

The identified driver regions typically represent compact modules, usually not covering more than 10% or 20% of the sequences in the case of oncogenes and tumor suppressors, respectively (Supplementary Figure S2). It was suggested that true oncogenes are recognizable from mutation patterns according to the 20/20 rule, having a higher than 20% fraction of missense point mutations in recurring positions (termed the oncogene score [53]). In contrast, tumor suppressors have lower oncogene scores, and predominantly contain truncating mutations. Figure 1B shows that the 20/20 rule holds true for the vast majority of the identified region-harboring oncogenes and context-dependent genes, even based on the oncogene scores calculated from the identified regions alone. This underlines that the identified driver regions are the main source of the oncogenic effect in almost all cases. While most drivers contain both ordered and disordered modules, oncogenesis is typically mediated through either ordered or disordered mutated regions. This effectively partitions cancer drivers into “ordered drivers” and “disordered drivers,” regardless of the exact structural composition of the full protein.

While many of the disordered drivers have already been identified previously as cancer drivers, our analysis identified 13 additional examples that were not included in the list of identified cancer drivers collected recently [27]. However, even in these cases there is literature data supporting their importance in driving cancer (Supplementary Table S2).

### 3.2. Disordered Drivers Function via Distinct Molecular Mechanisms

We collected available information about the possible mechanisms of action of the disordered regions that are altered in cancer (Figure 2, Supplementary Table S2). Although this information was partially incomplete in several cases, it still allowed us to highlight the distinct properties of the identified disordered drivers.

	Disordered functional unit	Tumor suppressors	Context dependent genes	Oncogenes
Linear motif / PTM		p14 <sup>ARF</sup> RSP15	EPAS1 NRF2 ESR1 FOXO1 FOXL2	CTNNB1 CCND3 MYC MYCN SETBP1 CD79B MET USP8 CSF1R histone H3s
Auto-regulatory			EZH2	
Flexible linker			CBL PAX5	KIT FLT3 PDGFRA
DNA/RNA binding		EIF1AX CEBPA		FOXA1 SRSF2
Disordered domain		APC ID3 VHL TP53 SMARCB1	MED12	MYOD1 CARD11 CALR
Unknown		ASXL1 MLH1 EP300		BCL2

**Figure 2.** IDP regions mutated in cancer. The classification of identified disordered cancer drivers. Protein names in gray indicate known switching mechanisms either via post-translational modifications (PTMs) or overlapping functions. In protein architecture schematics, blue ovals represent folded domains, blue lines represent disordered regions and red rectangles represent disordered driver modules. Boxes placed between two categories indicate dual functions. For detailed mutation profiles for each gene, see the online visualization links in Supplementary Table S2.

Several of the identified highly mutated disordered regions correspond to linear motifs, including sites for protein–protein interactions (e.g., USP8, FOXO1 and ESR1) or degron motifs that regulate the degradation of the protein (e.g. CTNNB1, CCND3 and CSF1R). However, other types of disordered functional modules can also be targeted by cancer mutations. IDRs with autoinhibitory roles (e.g., modulating the function of adjacent folded domains) are represented by EZH2, a component of the polycomb repressive complex 2. While the primary mutation site in this case is located in the folded SET domain, cancer mutations are also enriched within the disordered C-terminus that normally regulates the substrate binding site on the catalytic domain. Another category corresponds to regions involved in DNA and RNA binding. The highly flexible C-terminal segment of the winged helix domain is altered in the case of FOXA1, interfering with the high affinity DNA binding. For the splicing factor SRSF2, mutations affect the RNA binding region (Figure 2).

Larger functional disordered modules, often referred to as intrinsically disordered domains (IDDs), can also be the primary sites of cancer mutations. Mutated IDDs exhibit varied structure and sequence features. In VHL, the commonly mutated central region adopts a molten globule state in isolation [54]. The mutated region of APC incorporates several repeats containing multiple linear motif sites, which are likely to function collectively as part of the  $\beta$ -catenin destruction complex [55]. In CALR, cancer mutations alter the C-terminal domain-sized low complexity region, altering  $\text{Ca}^{2+}$  binding and protein localization [56].

Linker IDRs, not directly involved in molecular interactions, are also frequent targets of cancer mutations. The juxtamembrane regions located between the transmembrane segment and the kinase domain of KIT and related kinases are the main representatives of this category. Similarly, the regulatory linker region connecting the substrate- and the E2-binding domains is one of the dominant sites of mutations in the case of the E3 ubiquitin ligase CBL.

One of the recurring themes among cancer-related IDP regions is the formation of molecular switches (Supplementary Table S2). The most commonly occurring switching mechanism involves various post-translational modifications (PTMs), including serine or threonine phosphorylation (e.g., CCND3, MYC and APC), tyrosine phosphorylation (e.g., CBL, CD79B, and CSF1R), methylation (e.g., histone H3s [H3F3A/H3F3B/HIST1H3B]) or acetylation (e.g., ESR1). An additional way of forming molecular switches involves overlapping functional modules (Figure 2). In the case of PAX5, the mutated flexible linker region is also involved in the high-affinity binding of the specific DNA binding site [57]. Cancer mutations of the bZip domain of CEBPA disrupt not only the DNA binding function, but the dimerization domain as well [58]. In addition to their linker function, the juxtamembrane regions of kinases are also involved in autoinhibition and trans-phosphorylation, regulating degradation and downstream signaling events [59,60].

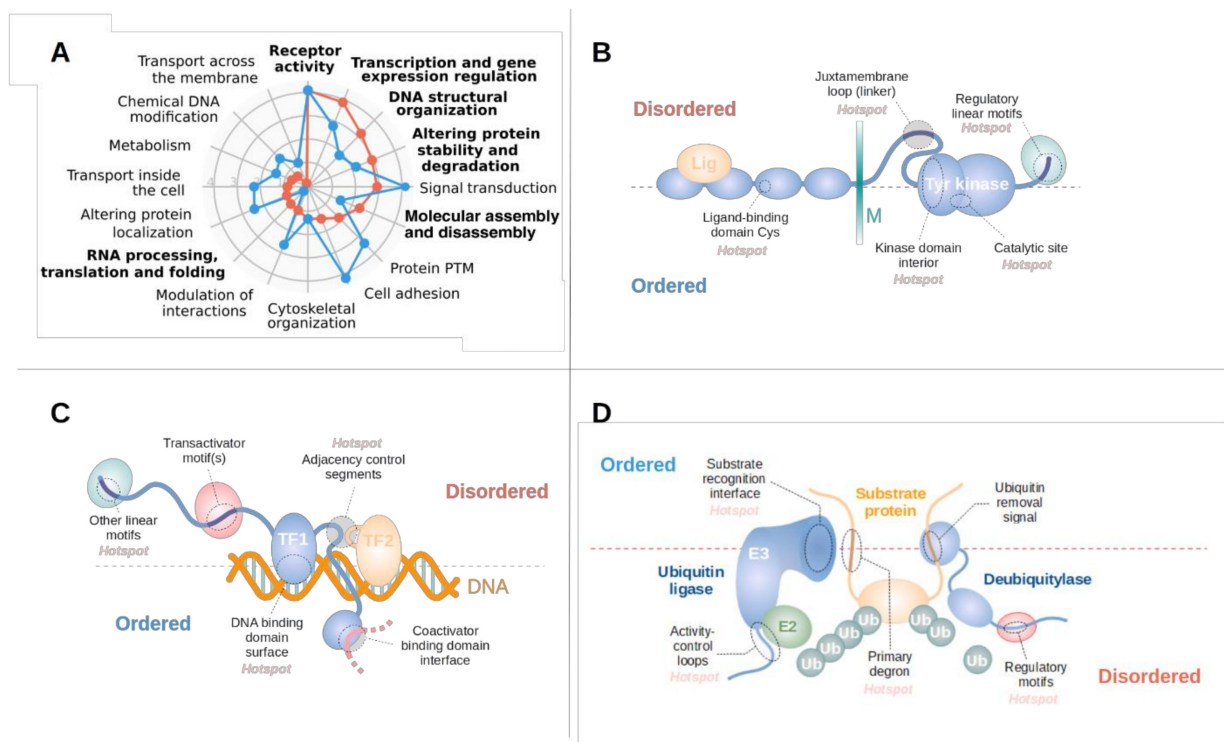
The collected examples of disordered regions mutated in cancer cover both oncogenes and tumor suppressors, as well as context-dependent genes (Figure 2). There is a slight tendency for tumor suppressors to be altered via longer functional modules, such as IDDs. Nevertheless, with the exception of linkers in tumor suppressors and IDDs in context-dependent genes, every other combination occurs even within our limited set.

### *3.3. Disordered Driver Mutations Preferentially Modulate Receptor Tyrosine Kinases, DNA-Processing and The Degradation Machinery*

Disordered and ordered drivers can employ different molecular mechanisms in order to fulfill their associated biological processes. To quantify these differences, we assembled a set of molecular toolkits integrating Gene Ontology terms (see Data and Methods and Supplementary Table S3). Based on this, we calculated the enrichment of each molecular toolkit in both disordered and ordered drivers in comparison with the full human proteome, highlighting enriched and possibly driver class-specific toolkits (Figure 3A). Receptor activity is the most enriched function for both types of drivers, owing at least partially to the fact that receptor tyrosine kinases can often be modulated via both ordered domains and IDRs (Figure 1B). In contrast, the next three toolkits enriched for disordered drivers are highly characteristic of them. These are gene expression regulation and the modulation of DNA structural organization—together representing the structural and the information content-related aspects of DNA processing—and the degradation of proteins, mainly through the ubiquitin-proteasome system. In addition, RNA processing, translation and folding is also characteristic of disordered drivers; and while this toolkit is not highly enriched compared to the human proteome in general, ordered drivers are almost completely devoid of this toolkit.

Among the highlighted functional groups, receptor tyrosine kinases (RTKs) are well-known to be major players in tumorigenesis [61]. While for several RTKs the major mutational events are oncogenic kinase domain mutations, there are also RTKs that contain a secondary disordered mutation site with lower incidence rates, or an alternative primary site which usually shows context dependent behavior. Several RTKs are clear examples of

this context dependence: gastrointestinal stromal tumor mutations prefer IDR mutations in both KIT and PDGFRA [62], while leukaemia prefers catalytic site mutations in KIT. Group III receptor tyrosine kinases in general (including KIT, FLT3 and PDGFRA) are especially prone to be mutated at their disordered juxtamembrane regions (Figure 3B). In some cases, such as FLT3, these IDRs are the main sites for tumorigenic mutations [63]. However, RTK IDR mutations are not restricted to group III receptor tyrosine kinases, as MET also often harbors missense mutations at its juxta-membrane loop region. These mutations include missense mutations affecting the Tyr1010 phosphorylation site and exon 19 skipping, removing a degron located within this region [64]. In contrast, CSF1R mutations accumulate in the negative regulatory motifs (a c-Cbl ubiquitin ligase binding motif) in the receptor tail, leading to the overactivation of the receptor [65] in various haematopoietic cancers.



**Figure 3.** Pathways and processes modulated by disordered driver mutations. (A) Overrepresentation of molecular toolkits defined based on gene ontology (GO) terms for ordered (blue) and disordered (red) drivers, compared to the average of the whole human proteome. Categories enriched in disordered drivers represented in bold. B–D: schematic examples of receptor tyrosine kinases (RTKs) (B), transcription factors (C) and components of the ubiquitin ligase machinery (D) that are modulated through disordered driver regions. Typically, these proteins have a modular architecture. Functional modules that are mutated preferentially in disordered or ordered regions are placed above or below the middle line.

Cancer mutations often target various elements of the transcriptional machinery, including transcription factors, repressors, transcriptional regulators and coactivators/corepressors [66] (Figure 3C). In most cases, transcription factors are targeted through linear motifs that regulate the degradation (EPAS1, CTNNB1, MYC and NMYC) or localization of the protein (FOXO1). Mutated IDRs can also directly affect the activity of the protein. These regions often work in conjunction with a separate DNA-binding domain and can shift affinities for various DNA-binding events (such as FOXA1 mutations preferentially affecting low-affinity DNA binding [67]), or can disrupt interaction with cofactors (such as the SMAD3 interaction of the FOXL2 [68]). In the case of bZip-type dimeric transcription factors, mutations can affect the interaction through the modulation of the disordered dimerization domain. Depending on the activating/repressive function of individual



transcription factors, IDR-mutated proteins can be both oncogenes (with MYOD1 mutations promoting the dimerization with MYC [69]), or tumor suppressors (with ID3 mutations impairing its repressor activity [70]). Disordered mutational hotspots also target other elements of the transcription machinery, affecting either covalent or non-covalent histone modifications and altering histone PTMs or histone exchange/movement along the DNA. However, the exact role of several other proteins involved in chromatin remodelling is still somewhat unclear (SETBP1 or ASXL1).

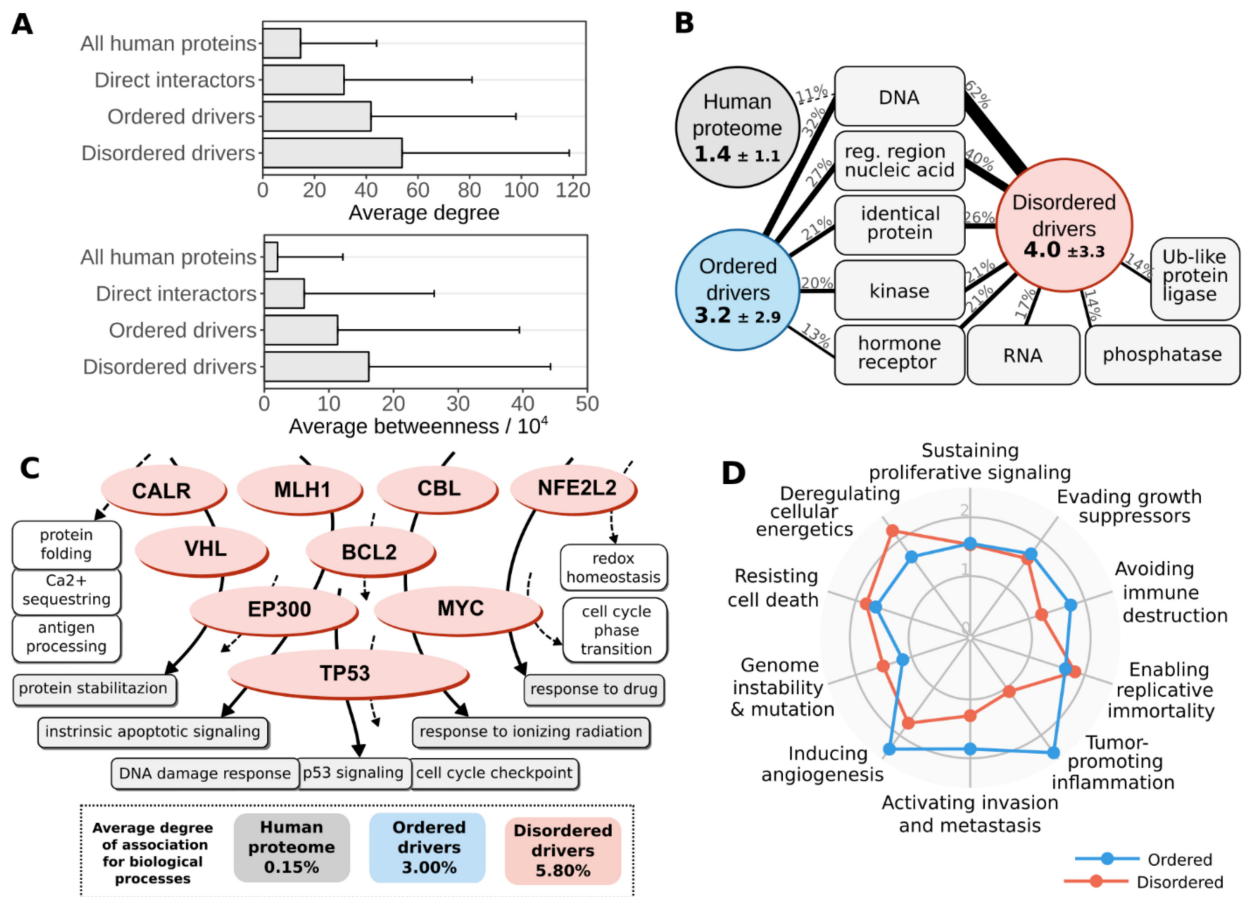
The alteration of protein abundance through the ubiquitin-proteasomal system (UPS) is a central theme in tumorigenesis [30]. Interestingly, ubiquitination sites are seldom mutated directly. More commonly cancer mutations directly alter degron motifs which typically reside in disordered protein regions (Figure 3D). Such mutations lead to increased abundance of the protein by disrupting the recognition by the corresponding E3 ligase. Complementing degron mutations, ubiquitin ligases are also implicated in tumorigenesis (Figure 3D). These enzymes are typically highly modular and can harbor driver mutations in both ordered and disordered regions (Supplementary Table S1). FBXW7 is mutated at its ordered substrate-binding domain, paralleled with target degron mutations in its substrates, MYC and MYCN. In contrast, VHL, which is the substrate recognition component of the cullin-2 E3 ligase complex, is targeted through a large disordered driver region, with its target EPAS1 bearing a mutant degron. The activity of CBL, the main E3 ligase responsible for the regulation of turnover for RTKs, is targeted through a disordered linker/autoregulatory region in acute myeloid leukemia (AML) and other hematopoietic disorders. In addition to the disruption of ubiquitination, the enhancement of deubiquitination can also provide a tumorigenic effect. USP8, the deubiquitinase required for entry into the S phase, is mutated at its disordered 14-3-3-binding motif, enhancing deubiquitinase activity in lung cancer [71].

#### *3.4. Disordered Mutations Give Rise to Cancer Hallmarks by Targeting Central Elements of Biological Networks*

Almost all of the analyzed IDRs are involved in binding to a molecular partner, even some of the linkers owing to their multifunctionality. Therefore, we analyzed known protein–protein interactions of ordered and disordered cancer drivers in more detail (see Data and Methods). Our results indicate that both sets of drivers are involved in a large number of interactions and show increased betweenness values compared to average values of the human proteome, even compared to the direct interaction partners of cancer drivers (Figure 4A). However, this trend is even more pronounced for disordered drivers. The elevated interaction capacity could also be detected at the level of molecular function annotations using Gene Ontology (see Supplementary Table S4 and Data and Methods). Figure 4B shows the average number of types of molecular interaction partners for both disordered and ordered drivers contrasted with the average for the human proteome. The main interaction partners are similar for both types of drivers, often binding to nucleic acids, homodimerizing or binding to receptors. However, disordered drivers are able to physically interact with a wider range of molecular partners, and are also able to more efficiently interact with RNA and the effector enzymes of the post-translational modification machinery. This, in particular, can offer a way to more easily integrate and propagate signals through the cell, relying on the spatio-temporal regulation of interactions via previously demonstrated switching mechanisms (Supplementary Table S2).

The high interaction capacity and central position of disordered drivers allows them to participate in several biological processes. The association between any two processes can be assessed by quantifying the overlap between their respective protein sets (see Data and Methods). We analyzed the average overlap between various processes using a set of non-redundant human-related terms of the Gene Ontology (Supplementary Table S5). The average overlap of proteins for two randomly chosen processes is 0.15%, showing that as expected, in general biological processes utilize characteristically different gene/protein sets. Restricting proteins to the identified drivers and only considering processes connected to at least one of them, the average overlap between processes increased to 3.00% for or-

dered drivers and 5.80% for disordered drivers (Figure 4C). This shows that the integration of various biological processes is a distinguishing feature of cancer genes in general and for disordered drivers in particular, and that IDPs targeted in cancer are efficient integrators of a wide range of processes.



**Figure 4.** Characteristics of cancer drivers at the network/pathway and cellular levels. (A) Average degree (top) and betweenness (bottom) of all human proteins, showing the direct interaction partners of drivers, ordered drivers and disordered drivers. (B) The average occurrence of various types of interaction partners for the whole human proteome (grey circle), ordered drivers (blue circle) and disordered drivers (red circle). Values in circles show the average number of types of interactions together with standard deviations. The most common interaction types are shown in grey boxes, with connecting lines showing the fraction of proteins involved in that binding mode. Only interaction types present for at least 1/8th of the proteins are shown. (C) Top: An example subset of disordered drivers with associated biological processes marked with arrows (dashed and solid arrows marking processes involving only one or several disordered drivers). Bottom: Average values of overlap between protein sets of various biological processes, considering the full human proteome (grey), ordered drivers (blue) and disordered drivers (red). Process names in grey represent processes that involve at least two disordered drivers, names in white boxes mark other processes attached to disordered drivers. (D) Overrepresentation of hallmarks of cancer for ordered (blue) and disordered (red) drivers compared to all census drivers.

Cancer hallmarks describe ubiquitously displayed traits of cancer cells [21]. In order to quantify the contribution of drivers to each of the ten hallmarks, we manually curated sets of biological process terms from the Gene Ontology that represent separate hallmarks (see Data and Methods and Supplementary Table S6). Enrichment analysis of these terms shows that all hallmarks are significantly overrepresented in census cancer drivers compared to the human proteome (Supplementary Figure S3A), serving as a proof-of-concept for the used hallmark quantification scheme. Furthermore, comparing drivers with identified regions to all census cancer drivers shows a further enrichment (Supplementary Figure S3B), indicating that the applied region identification protocol of iSiMPRe is able

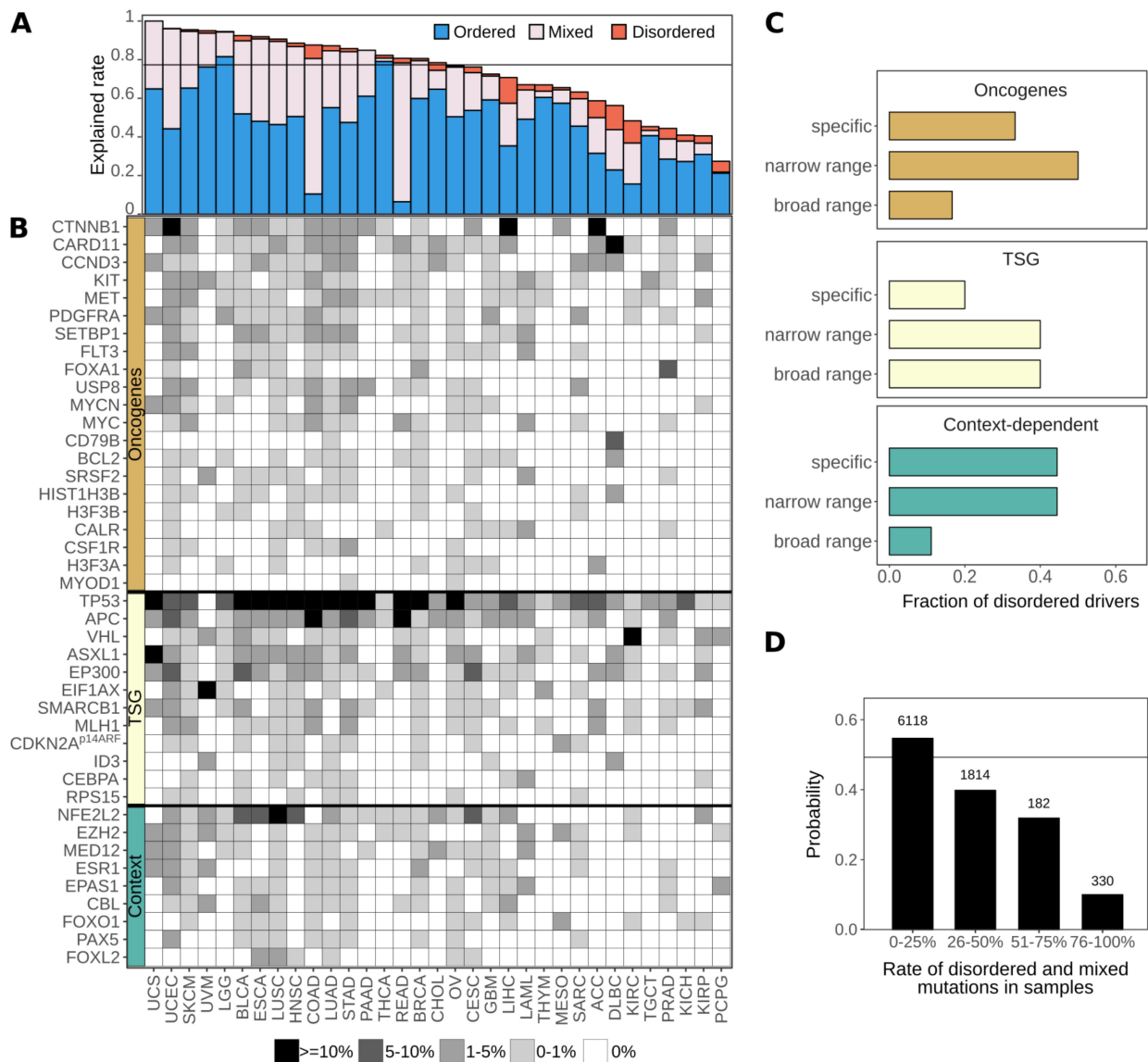
to pick up on the main tumorigenic signal by pinpointing strong driver genes. Separate enrichment calculations for ordered and disordered drivers show that despite subtle differences in enrichments, in general all ten hallmarks are overrepresented in both driver groups (Figure 4D). This indicates that while the exact molecular mechanisms through which ordered domain and IDR mutations contribute to cancer are highly variable, both types of genetic modulation can give rise to all necessary cellular features of tumorigenic transformation. Hence, IDR mutations provide a mechanism that is sufficient on its own for cancer formation.

### 3.5. Disordered Drivers Can Be the Dominant Players at The Patient Sample Level

We assessed the role of the identified drivers at the patient level using whole-genome sequencing data from TCGA; 10,197 tumor samples containing over three and a half million genetic variations were considered to delineate the importance of disordered drivers at the sample level across the 33 cancer types covered in TCGA. In driver region identification, we only considered mutations with a local effect (missense mutations and in frame indels), which naturally yielded only a restricted subset of all true drivers. However, in patient-level analyses, we also considered other types of genetic alterations of the same gene in order to get a more complete assessment of the alteration of identified driver regions per cancer type (see Data and Methods).

In spite of the incompleteness of the identified set of driver genes, we still found that on average about 80% of samples contain genetic alterations that affect at least one identified ordered or disordered driver region. Thus, the identified regions are able to describe the main players of tumorigenesis at the molecular level (Figure 5A). While at the protein level typically either ordered or disordered regions are modulated (Figure 1B), at the patient level most samples show a mixed structural background, most notably in colorectal cancers (COAD and READ). Some cancer types, however, show distinct preferences for the modulation of a single type of structural element. For thyroid carcinoma (THCA) or thymoma (THYM), the molecular basis is almost always the exclusive mutation of ordered protein regions. At the other extreme, the modulation of disordered regions is enough for tumor formation in a considerable fraction of cases of liver hepatocellular, adrenocortical, and renal cell carcinomas, together with diffuse large B-cell lymphoma (LIHC, ACC, KIRC and DLBC). These results, in line with the previous hallmark analyses, show that IDR mutations can constitute a complete set of tumorigenic alterations. Hence, there are specific subsets of patients that carry predominantly or exclusively disordered driver mutations in their exome.

Whole genome sequencing data was also used to assess the cancer type specificity of disordered drivers (Figure 5B). Basically, all studied cancer types have at least one disordered driver that is mutated in at least 1% of cases, with the exception of thyroid carcinoma (THCA). There are only four disordered drivers that can be considered as generic drivers, being mutated in a high number of cancer types. p53 presents a special case in this regard, as it is the main tumor suppressor gene in humans and thus is most often affected by gene loss or truncations which are likely to eliminate the corresponding protein product. These alterations abolish the function of both the ordered and disordered driver regions at the same time (the DNA-binding domain and the tetramerization region). In contrast, the other three generic disordered drivers are predominantly altered via localized mutations in their disordered regions: the degrons of  $\beta$ -catenin and NRF2 and the central region of APC, and hence these are true disordered drivers which are commonly mutated in several cancer types. However, the majority of disordered drivers show a high degree of selectivity for tumor types, being mutated only in very specific cancer types. This specificity is strongly connected to the tumorigenic roles of disordered drivers (Figure 5C). Considering 1% of patient samples as the cutoff, tumor suppressors are typically implicated in a broad range of cancer types, while oncogenes on average show a high cancer type specificity. Context-dependent disordered drivers are often mutated in only a very restricted set of cancers.



**Figure 5.** Therapeutic options for targeting disordered drivers. **(A)** Fraction of samples that contain altered driver genes per cancer type. Samples can contain mutations affecting only ordered drivers (blue), only disordered drivers (red) or both (mixed, gray). **(B)** Percentage of cancer samples, grouped by cancer types, containing genetic alterations that target the identified disordered driver regions. **(C)** The distribution of disordered drivers from the three classes of cancer genes (oncogenes, tumor suppressor genes (TSG) and context dependent genes) categorized into specific, narrow and broad range based on the frequency of samples they are mutated in (see Data and Methods). **(D)** The probability of having an available FDA-approved drug for at least one mutation-affected gene for patients, as a function of the ratio of affected disordered genes compared to all mutated genes in the sample. The horizontal black line represents the total fraction of targetable samples (0.49) from 8444 samples.

Strikingly, the identified disordered drivers can have an even more dominant role. In several rarer cancers or more specific cancer subtypes which are not included in the broad classes described in TCGA (including both malignant and benign cases), mutations in a specific disordered driver are the main, or one of the main, driver events (Table 1). Altogether, this list includes 18 of our disordered cancer drivers. In the collected cancer types, targeting disordered regions can have a potentially huge treatment advantage, and in many cases, the counteraction of these IDR mutations may be the only viable therapeutic strategy.

**Table 1.** Cancer types with mutation incidence rates around or above 10% in the disordered driver gene of interest per total patients studied.

Tumor Type (Name)	Implicated Gene Product	Malignancy	Incidence	Reference
Diffuse large B-cell lymphoma (ABC subtype)	CARD11	malignant	9.6–10.8% (7/73, 4/37)	[72,73]
Burkitt lymphoma	CCND3	malignant	14.6% (6/41)	[74]
Diffuse large B-cell lymphoma (ABC subtype)	CCND3	malignant	10.7% (3/28)	[74]
Diffuse large B-cell lymphoma (PCNS subtype)	CD79B	malignant	31.6% (6/19)	[75]
Acute myeloid leukaemia	CEBPA	malignant	15% (16/104)	[76]
Myelodysplasia and acute myeloblastic leukemia	CSF1R	malignant	12.7% (14/110)	[77]
Endometrioid endometrial carcinoma (low-grade)	CTNNB1	malignant	87.0% (47/54)	[78]
Ovarian endometrioid carcinomas (low-grade)	CTNNB1	malignant	53.3% (16/30)	[79]
Hepatocellular carcinoma (HBV/HCV related)	CTNNB1	malignant	26% (32/122)	[80]
Desmoid tumor	CTNNB1	benign	73% (106/145)	[81]
Juvenile nasopharyngeal angiofibroma	CTNNB1	benign	75% (12/16)	[82]
Paraganglioma	EPAS1	possibly malignant	17% (7/41)	[83]
Adult granulosa cell tumors of the ovary	FOXL2	malignant	93–97% (52/56, 86/89)	[84,85]
Pediatric anaplastic astrocytoma/glioblastoma	H3F3A	malignant	17.9–27.1% (5/28, 35/129)	[86]
Giant cell tumor of bone (stromal cell)	H3F3A	benign	92% (49/53)	[87]
Chondroblastoma (stromal cell)	H3F3B	benign	95% (73/77)	[87]
GIST	KIT	malignant	47% (57/121)	[88]
Extrauterine leiomyoma and leiomyosarcoma	MED12	(possibly) malignant	19% (6/32)	[89]
Phyllodes tumor of breast	MED12	possibly malignant	49% (41/83)	[90]
Uterine leiomyoma	MED12	benign	70% (159/225)	[91]
Rhabdomyosarcoma	MYOD1	malignant	20% (10/49)	[92]
Esophageal squamous cell carcinoma	NFE2L2	malignant	9.6% (47/490)	[93]
B-cell progenitor acute lymphoblastic leukemia	PAX5	malignant	34–39% (40/117, 94/242)	[94,95]
Chronic myelomonocytic leukemia	SETBP1	malignant	25% (14/56)	[96]
Atypical Chronic Myeloid Leukemia	SETBP1	malignant	24.3% (17/70)	[97]
Chronic myelomonocytic leukaemia	SRSF2	malignant	47% (129/275)	[98]
Pituitary adenoma	USP8	possibly malignant	14% (6/42)	[99]

### 3.6. Cancer Incidences Arising through Disordered Drivers Lack Effective Drugs

Next, we addressed how well disordered drivers are targetable by current FDA approved drugs, as collected by the OncoKB database [100]. This database currently contains 83 FDA-approved anticancer drugs, either as part of standard care or efficient off-label use (see Data and Methods). These drugs have defined exome mutations that serve as indications for their use. The majority of these drugs target ordered domains, mostly inhibiting kinases. Currently only seven drugs are connected to disordered region mutations, which correspond to only four sites in FGFR and c-Met. These drugs act indirectly, targeting ordered kinase domains, to counteract the effect of the listed activating disordered mutations.

This represents a clear negative treatment option bias against patients whose tumor genomes contain disordered drivers. Considering all mutations in patient samples gathered in TCGA, the fraction of disordered driver mutations actually serves as an indicator of whether there are suitable drugs available. Patients with mostly ordered driver mutations have a roughly 50% chance that an FDA-approved drug can be administered with the expected therapeutic effect. This chance drops to 10% for patients with predominantly disordered mutations (Figure 5D). Thus, incidences of cancer arising through disordered driver mutations are currently heavily under-targeted, highlighting the need for efficient targeting strategies for IDP-driven cancers.

## 4. Discussion

In recent years, cancer genome projects have revealed the genomic landscapes of many common forms of human cancer. As a result, several hundred cancer driver genes have been identified whose genetic alterations can be directly linked to tumorigenesis [27]. Only a few of these genes correspond to “mutation mountains,” i.e., genes that are commonly altered in different tumor types, while most of the cancer drivers are altered infrequently [53]. Cancer driver genes are associated with a set of core cellular processes, also termed hallmarks [21]. At a more detailed level, however, drivers are surprisingly heterogeneous in terms of molecular functions and cellular roles. In this work we showed that cancer drivers are also diverse in terms of their structural properties. Using an integrated computational approach, we identified a set of cancer drivers that are specifically targeted by mutation in a disordered region. IDRs represent around 30% of residues in the human proteome and are also an integral part of many cancer-associated proteins. Despite the critical roles of these regions, they are often not the main sites of driver mutations [11]. Our results confirmed that driver mutations that alter the proper functioning of ordered domains of the encoded protein are slightly overrepresented compared to those that modulate the function of disordered regions. Nevertheless, in a significant number of cases, corresponding to around 20% of the mutated drivers, cancer mutations specifically target disordered regions (Figure 1A).

The critical role of these disordered drivers in tumorigenesis is supported not only by the enrichment of single nucleotide variations and in-frame insertions and deletions, but also by literature data (Supplementary Table S2). Disordered drivers are associated with known cancer hallmarks through specific biological processes (Figure 3A) and show strong evolutionary conservation [101]. Driver mutations within IDRs are present in samples across a wide range of cancer types, and can also be the main, or one of the main, driver events for several tumor subclasses, including both malignant and benign cases (Table 1). Our work highlighted several novel drivers that are not yet included in the previous collections of cancer driver genes previously assembled based on a combination of computational methods [27], indicating a hidden bias in the identification of driver genes.

The collection of disordered cancer drivers highlighted many interesting examples that carry out important functions without relying on a well-defined structure, extending the list of IDR with disease relevance. Many of the collected cases correspond to linear motif sites which mediate interactions with globular domains, regulating interactions and localization or cellular fate of proteins. However, the collected examples represent a

broader set of functional mechanisms, encompassing DNA- and RNA-binding regions, linkers, autoinhibitory segments and disordered domains. These functional modules can also regulate the assembly of large macromolecular complexes and regulate the activity of neighboring domains. The key to the proper functioning of the targeted IDRs is their structural disorder, which enables them to undergo drastic conformational changes depending on context-dependent regulation. While in most cases it has been clarified how mutations of the critical IDR disrupt the balance between the different functional states, our understanding of this mechanism is still incomplete for several examples (Figure 2, Supplementary Table S2). For instance, the mutation and conservation pattern of MLH1 highlights a novel linear motif site within the disordered linker region of MLH1 with unknown function. In the case of p14Arf, the functional role of the mutated region needs to be revisited in the light of recent evidence on the relevance of phase separation organizing the nucleolus [102]. ASXL1 and EP300 are both involved in chromatin remodelling, but little is known about the functional roles of the disordered regions targeted by cancer mutations.

At the patient level, samples in general contain a combination of genetic alterations that involve both ordered and disordered drivers. However, patients with mostly IDR mutations typically have significantly limited treatment options. Most current anticancer drugs target ordered protein domains, and are inhibitors designed against enzyme activity (using either competitive or noncompetitive inhibition) [103–105]. In general, current successful drug development efforts mainly focus on ordered protein domains derived within the framework of structure-based rational drug design [106]. However, IDPs can potentially offer new directions for cancer therapeutics [107]. Currently tested approaches include the direct targeting of IDPs by specific small compounds, or blocking the globular interaction partner of IDPs [108,109]. The successful identification of disordered drivers and corresponding tumor types provides the first step in providing the means for new therapeutic interventions in cancer types that currently lack treatment options.

## 5. Conclusions

In this work, we went beyond a simple association between IDRs and cancer by taking advantage of the avalanche of data produced by systematic analyses and large-scale sequencing projects of cancer genomes. Our work underlines the direct driver role of IDRs in cancer. It provides fundamental insights into the specific molecular mechanisms and regulatory processes altered by cancer mutations targeting IDRs, highlighting important regions that need further structural and functional characterizations. Furthermore, we showed that many already known cancer drivers rely on intrinsic flexibility for their function and identified novel cancer drivers that had been overlooked by current driver identification approaches, revealing a structure-centric bias that still exists in these methods. Importantly, our work also demonstrates the relevance of disordered drivers at the patient level and highlights a strong need to expand treatment options for IDRs. By looking at the timeline of the COSMIC database, we can observe a steady growth of disordered drivers with every new release (Supplementary Figure S4). Nevertheless, our study was restricted to cases that were targeted by point mutations or in-frame insertions or deletions, therefore the location of alterations can be directly linked to the perturbed functional module. However, there are additional disordered drivers that are altered via more complex genetic mechanisms in cancer, such as specific frameshift mutations (e.g., NOTCH1 [110]), chromosomal translocations (e.g., BCR [111], ERG [112]) or copy number variations (e.g., p14ARF [113]). Altogether these observations suggest that we can expect the emergence of further examples of genetic alterations of driver genes that interfere with structurally disordered regions as the number of cancer studies increase/ Furthermore, this paper also highlights cancer types where novel drug design strategies targeting disordered regions are needed.

**Supplementary Materials:** The following are available online at <https://www.mdpi.com/2218-273X/11/3/381/s1>, Figure S1: The distribution of residues and cancer mutations, Figure S2: Identified regions are compact functional units, Figure S3: Overrepresentation of cancer hallmarks, Figure S4: Growth of disordered cancer drivers, Table S1: List of regions identified using iSiMPRe,

based on both COSMIC and TCGA mutations, Table S2: Identified disordered driver genes with all annotations, Table S3: Gene Ontology terms used in the quantification of molecular toolkits used by cancer driver genes, Table S4: Gene Ontology terms used in the quantification of interaction capabilities, Table S5: Gene Ontology terms used in the quantification of biological process overlaps, Table S6: Gene Ontology terms used in the quantification of hallmarks of cancer.

**Author Contributions:** B.M. contributed to conceptualization, development of methodology and software, formal analysis, investigation of the findings, developing resources, data curation, writing the manuscript, visualization of data. B.H.-S. contributed to developing software, formal analysis, investigation of the findings, writing the manuscript, and visualization of data. A.Z. contributed to conceptualization, investigation of the findings, writing the manuscript, visualization of data, and acquisition of funding. Z.D. contributed to conceptualization, development of methodology and software, investigation of the findings, developing resources, data curation, writing the manuscript, supervision, project administration, and acquisition of funding. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the “Lendület” grant from the Hungarian Academy of Sciences (LP2014-18) (Z.D.), OTKA grants (K108798 and K129164) (Z.D.) and the grant PD-120973 (A.Z) of the National Research, Development and Innovation office of Hungary and the EMBO|EuropaBio fellowship 7544 (B.M.).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The authors declare that the data supporting the findings of this study are available within the paper and its Supplementary Information Files.

**Acknowledgments:** The authors thank Mark Adamsbaum and Toby J. Gibson, Péter Tompa and László Buday for the critical reading of and their constructive comments on the manuscript.

**Conflicts of Interest:** The authors declare no competing interests.

## References

1. Nussinov, R.; Jang, H.; Tsai, C.-J.; Cheng, F. Review: Precision medicine and driver mutations: Computational methods, functional assays and conformational principles for interpreting cancer drivers. *PLoS Comput. Biol.* **2019**, *15*, e1006658.
2. Babu, M.M. The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. *Biochem. Soc. Trans.* **2016**, *44*, 1185–1200. [[CrossRef](#)]
3. Van Roey, K.; Uyar, B.; Weatheritt, R.J.; Dinkel, H.; Seiler, M.; Budd, A.; Gibson, T.J.; Davey, N.E. Short Linear Motifs: Ubiquitous and Functionally Diverse Protein Interaction Modules Directing Cell Regulation. *Chem. Rev.* **2014**, *114*, 6733–6778. [[CrossRef](#)]
4. Wright, P.E.; Dyson, H.J. Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* **2015**, *16*, 18–29. [[CrossRef](#)] [[PubMed](#)]
5. Tompa, P. Intrinsically unstructured proteins. *Trends Biochem. Sci.* **2002**, *27*, 527–533. [[CrossRef](#)]
6. Van Der Lee, R.; Buljan, M.; Lang, B.; Weatheritt, R.J.; Daughdrill, G.W.; Dunker, A.K.; Fuxreiter, M.; Gough, J.; Gsponer, J.; Jones, D.T.W.; et al. Classification of Intrinsically Disordered Regions and Proteins. *Chem. Rev.* **2014**, *114*, 6589–6631. [[CrossRef](#)] [[PubMed](#)]
7. Dyson, H.J.; Wright, P.E. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* **2005**, *6*, 197–208. [[CrossRef](#)]
8. Uversky, V.N. Intrinsically disordered proteins in overcrowded milieu: Membrane-less organelles, phase separation, and intrinsic disorder. *Curr. Opin. Struct. Biol.* **2017**, *44*, 18–30. [[CrossRef](#)] [[PubMed](#)]
9. Babu, M.M.; van der Lee, R.; de Groot, N.S.; Gsponer, J. Intrinsically disordered proteins: Regulation and disease. *Curr. Opin. Struct. Biol.* **2011**, *21*, 432–440. [[CrossRef](#)] [[PubMed](#)]
10. Iakoucheva, L.M.; Brown, C.J.; Lawson, J.D.; Obradovic, Z.; Dunker, A.K. Intrinsic Disorder in Cell-signaling and Cancer-associated Proteins. *J. Mol. Biol.* **2002**, *323*, 573–584. [[CrossRef](#)]
11. Pajkos, M.; Mészáros, B.; Simon, I.; Dosztányi, Z. Is there a biological cost of protein disorder? Analysis of cancer-associated mutations. *Mol. Biosyst.* **2011**, *8*, 296–307. [[CrossRef](#)]
12. Hegyi, H.; Buday, L.; Tompa, P. Intrinsic Structural Disorder Confers Cellular Viability on Oncogenic Fusion Proteins. *PLoS Comput. Biol.* **2009**, *5*, e1000552. [[CrossRef](#)] [[PubMed](#)]
13. Vavouri, T.; Semple, J.I.; Garcia-Verdugo, R.; Lehner, B. Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity. *Cell* **2009**, *138*, 198–208. [[CrossRef](#)]
14. Uyar, B.; Weatheritt, R.J.; Dinkel, H.; Davey, N.E.; Gibson, T.J. Proteome-wide analysis of human disease mutations in short linear motifs: Neglected players in cancer? *Mol. Biosyst.* **2014**, *10*, 2626–2642. [[CrossRef](#)]



15. Meyer, K.; Kirchner, M.; Uyar, B.; Cheng, J.-Y.; Russo, G.; Hernandez-Miranda, L.R.; Szymborska, A.; Zauber, H.; Rudolph, I.-M.; Willnow, T.E.; et al. Mutations in Disordered Regions Can Cause Disease by Creating Dileucine Motifs. *Cell* **2018**, *175*, 239–253. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Weinstein, J.N.; The Cancer Genome Atlas Research Network; Collisson, E.A.; Mills, G.B.; Shaw, K.R.M.; Ozenberger, B.A.; Ellrott, K.; Shmulevich, I.; Sander, C.; Stuart, J.M. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **2013**, *45*, 1113–1120. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Lawrence, M.S.; Stojanov, P.; Polak, P.; Kryukov, G.V.; Cibulskis, K.; Sivachenko, A.; Carter, S.L.; Stewart, C.; Mermel, C.H.; Roberts, S.A.; et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nat. Cell Biol.* **2013**, *499*, 214–218. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Lawrence, M.S.; Stojanov, P.; Mermel, C.H.; Robinson, J.T.; Garraway, L.A.; Golub, T.R.; Meyerson, M.L.; Gabriel, S.B.; Lander, E.S.; Getz, G. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nat. Cell Biol.* **2014**, *505*, 495–501. [\[CrossRef\]](#)
19. Copeland, N.G.; Jenkins, N.A. Deciphering the genetic landscape of cancer—from genes to pathways. *Trends Genet.* **2009**, *25*, 455–462. [\[CrossRef\]](#)
20. Ali, M.A.; Sjöblom, T. Molecular pathways in tumor progression: From discovery to functional understanding. *Mol. BioSyst.* **2009**, *5*, 902–908. [\[CrossRef\]](#)
21. Hanahan, D.; Weinberg, R.A. Hallmarks of Cancer: The Next Generation. *Cell* **2011**, *144*, 646–674. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Yang, F.; Petsalaki, E.; Rolland, T.; Hill, D.E.; Vidal, M.; Roth, F.P. Protein Domain-Level Landscape of Cancer-Type-Specific Somatic Mutations. *PLoS Comput. Biol.* **2015**, *11*, e1004147. [\[CrossRef\]](#)
23. Tokheim, C.; Bhattacharya, R.; Niknafs, N.; Gygax, D.M.; Kim, R.; Ryan, M.; Masica, D.L.; Karchin, R. Exome-Scale Discovery of Hotspot Mutation Regions in Human Cancer Using 3D Protein Structure. *Cancer Res.* **2016**, *76*, 3719–3731. [\[CrossRef\]](#)
24. Porta-Pardo, E.; Garcia-Alonso, L.; Hrabe, T.; Dopazo, J.; Godzik, A. A Pan-Cancer Catalogue of Cancer Driver Protein Interaction Interfaces. *PLoS Comput. Biol.* **2015**, *11*, e1004518. [\[CrossRef\]](#)
25. Engin, H.B.; Kreisberg, J.F.; Carter, H. Structure-Based Analysis Reveals Cancer Missense Mutations Target Protein Interaction Interfaces. *PLoS ONE* **2016**, *11*, e0152929. [\[CrossRef\]](#)
26. Kamburov, A.; Lawrence, M.S.; Polak, P.; Leshchiner, I.; Lage, K.; Golub, T.R.; Lander, E.S.; Getz, G. Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, E5486–E5495. [\[CrossRef\]](#)
27. Bailey, M.H.; Tokheim, C.; Porta-Pardo, E.; Sengupta, S.; Bertrand, D.; Weerasinghe, A.; Colaprico, A.; Wendl, M.C.; Kim, J.; Reardon, B.; et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **2018**, *173*, 371–385. [\[CrossRef\]](#)
28. Giacomelli, A.O.; Yang, X.; Lintner, R.E.; McFarland, J.M.; Duby, M.; Kim, J.; Howard, T.P.; Takeda, D.Y.; Ly, S.H.; Kim, E.; et al. Mutational processes shape the landscape of TP53 mutations in human cancer. *Nat. Genet.* **2018**, *50*, 1381–1387. [\[CrossRef\]](#) [\[PubMed\]](#)
29. Gibson, T.J. Cell regulation: Determined to signal discrete cooperation. *Trends Biochem. Sci.* **2009**, *34*, 471–482. [\[CrossRef\]](#) [\[PubMed\]](#)
30. Mészáros, B.; Kumar, M.; Gibson, T.J.; Uyar, B.; Dosztányi, Z. Degrons in cancer. *Sci. Signal.* **2017**, *10*, eaak9982. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Fichó, E.; Reményi, I.; Simon, I.; Mészáros, B. MFIB: A repository of protein complexes with mutual folding induced by binding. *Bioinformatics* **2017**, *33*, 3682–3684. [\[CrossRef\]](#) [\[PubMed\]](#)
32. Xu, W.; Kimelman, D. Mechanistic insights from structural studies of beta-catenin and its binding partners. *J. Cell Sci.* **2007**, *120*, 3337–3344. [\[CrossRef\]](#)
33. Forbes, S.; Beare, D.; Bindal, N.; Bamford, S.; Ward, S.; Cole, C.; Jia, M.; Kok, C.; Boutselakis, H.; De, T.; et al. COSMIC: High-Resolution Cancer Genetics Using the Catalogue of Somatic Mutations in Cancer. *Curr. Protoc. Hum. Genet.* **2016**, *91*, 10–11. [\[CrossRef\]](#)
34. Mészáros, B.; Zeke, A.; Reményi, A.; Simon, I.; Dosztányi, Z. Systematic analysis of somatic mutations driving cancer: Uncovering functional protein regions in disease development. *Biol. Direct* **2016**, *11*, 1–23. [\[CrossRef\]](#)
35. Buljan, M.; Blattmann, P.; Aebersold, R.; Boutros, M. Systematic characterization of pan-cancer mutation clusters. *Mol. Syst. Biol.* **2018**, *14*, e7974. [\[CrossRef\]](#)
36. Smigielski, E.M.; Sirotkin, K.; Ward, M.; Sherry, S.T. dbSNP: A database of single nucleotide polymorphisms. *Nucleic Acids Res.* **2000**, *28*, 352–355. [\[CrossRef\]](#)
37. Siepel, A.; Bejerano, G.; Pedersen, J.S.; Hinrichs, A.S.; Hou, M.; Rosenbloom, K.; Clawson, H.; Spieth, J.; Hillier, L.W.; Richards, S.; et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **2005**, *15*, 1034–1050. [\[CrossRef\]](#) [\[PubMed\]](#)
38. Piovesan, D.; Tabaro, F.; Mičetić, I.; Necci, M.; Quaglia, F.; Oldfield, C.J.; Aspromonte, M.C.; Davey, N.E.; Davidović, R.; Dosztányi, Z.; et al. DisProt 7.0: A major update of the database of disordered proteins. *Nucleic Acids Res.* **2017**, *45*, D1123–D1124. [\[CrossRef\]](#) [\[PubMed\]](#)
39. Fukuchi, S.; Amemiya, T.; Sakamoto, S.; Nobe, Y.; Hosoda, K.; Kado, Y.; Murakami, S.D.; Koike, R.; Hiroaki, H.; Ota, M. IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners. *Nucleic Acids Res.* **2014**, *42*, D320–D325. [\[CrossRef\]](#)
40. Schad, E.; Fichó, E.; Pancsa, R.; Simon, I.; Dosztányi, Z.; Mészáros, B. DIBS: A repository of disordered binding sites mediating interactions with ordered proteins. *Bioinformatics* **2017**, *34*, 535–537. [\[CrossRef\]](#) [\[PubMed\]](#)

41. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242. [[CrossRef](#)]
42. Kirchner, D.K.; Güntert, P. Objective identification of residue ranges for the superposition of protein structures. *BMC Bioinform.* **2011**, *12*, 1–11. [[CrossRef](#)]
43. Mészáros, B.; Erdős, G.; Dosztányi, Z. IUPred2A: Context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* **2018**, *46*, W329–W337. [[CrossRef](#)]
44. Dosztányi, Z.; Csizmók, V.; Tompa, P.; Simon, I. The Pairwise Energy Content Estimated from Amino Acid Composition Discriminates between Folded and Intrinsically Unstructured Proteins. *J. Mol. Biol.* **2005**, *347*, 827–839. [[CrossRef](#)]
45. Mészáros, B.; Simon, I.; Dosztányi, Z. Prediction of Protein Binding Regions in Disordered Proteins. *PLoS Comput. Biol.* **2009**, *5*, e1000376. [[CrossRef](#)]
46. Dosztányi, Z.; Mészáros, B.; Simon, I. ANCHOR: Web server for predicting protein binding regions in disordered proteins. *Bioinformatics* **2009**, *25*, 2745–2746. [[CrossRef](#)] [[PubMed](#)]
47. Piovesan, D.; Tabaro, F.; Paladin, L.; Necci, M.; Mičetić, I.; Camilloni, C.; Davey, N.; Dosztányi, Z.; Mészáros, B.; Monzon, A.M.; et al. MobiDB 3.0: More annotations for intrinsic disorder, conformational diversity and interactions in proteins. *Nucleic Acids Res.* **2017**, *46*, D471–D476. [[CrossRef](#)]
48. Zimmermann, L.; Stephens, A.; Nam, S.-Z.; Rau, D.; Kübler, J.; Lozajic, M.; Gabler, F.; Söding, J.; Lupas, A.N.; Alva, V. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *J. Mol. Biol.* **2018**, *430*, 2237–2243. [[CrossRef](#)]
49. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene Ontology: Tool for the unification of biology. *Nat. Genet.* **2000**, *25*, 25–29. [[CrossRef](#)] [[PubMed](#)]
50. Antonazzo, G.; Attrill, H.; Brown, N.; Marygold, S.J.; McQuilton, P.; Ponting, L.; Millburn, G.H. The Gene Ontology Consortium Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.* **2017**, *45*, D331–D338.
51. Orchard, S.; Ammari, M.; Aranda, B.; Breuza, L.; Briganti, L.; Broackes-Carter, F.; Campbell, N.H.; Chavali, G.; Chen, C.; del-Toro, N.; et al. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **2014**, *42*, D358–D363. [[CrossRef](#)] [[PubMed](#)]
52. Tokheim, C.J.; Papadopoulos, N.; Kinzler, K.W.; Vogelstein, B.; Karchin, R. Evaluating the evaluation of cancer driver genes. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 14330–14335. [[CrossRef](#)] [[PubMed](#)]
53. Vogelstein, B.; Papadopoulos, N.; Velculescu, V.E.; Zhou, S.; Diaz, L.A.; Kinzler, K.W. Cancer Genome Landscapes. *Science* **2013**, *339*, 1546–1558. [[CrossRef](#)] [[PubMed](#)]
54. Sutovsky, H.; Gazit, E. The von Hippel-Lindau tumor suppressor protein is a molten globule under native conditions: Implications for its physiological activities. *J. Biol. Chem.* **2004**, *279*, 17190–17196. [[CrossRef](#)]
55. Aoki, K.; Taketo, M.M. Adenomatous polyposis coli (APC): A multi-functional tumor suppressor gene. *J. Cell Sci.* **2007**, *120*, 3327–3335. [[CrossRef](#)]
56. Elf, S.; Abdelfattah, N.S.; Chen, E.; Perales-Patón, J.; Rosen, E.A.; Ko, A.; Peisker, F.; Florescu, N.; Giannini, S.; Wolach, O.; et al. Mutant Calreticulin Requires Both Its Mutant C-terminus and the Thrombopoietin Receptor for Oncogenic Transformation. *Cancer Discov.* **2016**, *6*, 368–381. [[CrossRef](#)]
57. Garvie, C.W.; Hagman, J.; Wolberger, C. Structural Studies of Ets-1/Pax5 Complex Formation on DNA. *Mol. Cell* **2001**, *8*, 1267–1276. [[CrossRef](#)]
58. Paz-Priel, I.; Friedman, A. C/EBP $\alpha$  dysregulation in AML and ALL. *Crit. Rev. Oncog.* **2011**, *16*, 93–102. [[CrossRef](#)]
59. Hubbard, S.R. Juxtamembrane autoinhibition in receptor tyrosine kinases. *Nat. Rev. Mol. Cell Biol.* **2004**, *5*, 464–471. [[CrossRef](#)] [[PubMed](#)]
60. Li, E.; Hristova, K. Receptor tyrosine kinase transmembrane domains: Function, dimer structure and dimerization energetics. *Cell Adh. Migr.* **2010**, *4*, 249–254. [[CrossRef](#)]
61. Sangwan, V.; Park, M. Receptor Tyrosine Kinases: Role in Cancer Progression. *Curr. Oncol.* **2006**, *13*, 191–193. [[CrossRef](#)]
62. Oppelt, P.J.; Hirbe, A.C.; Van Tine, B.A. Gastrointestinal stromal tumors (GISTs): Point mutations matter in management, a review. *J. Gastrointest. Oncol.* **2017**, *8*, 466–473. [[CrossRef](#)]
63. Deeb, K.K.; Smonskey, M.T.; DeFedericis, H.; Deeb, G.; Sait, S.N.; Wetzler, M.; Wang, E.S.; Starostik, P. Deletion and deletion/insertion mutations in the juxtamembrane domain of the FLT3 gene in adult acute myeloid leukemia. *Leuk. Res. Rep.* **2014**, *3*, 86–89. [[CrossRef](#)]
64. Pilotto, S.; Gkoutakos, A.; Carbognin, L.; Scarpa, A.; Tortora, G.; Bria, E. MET exon 14 juxtamembrane splicing mutations: Clinical and therapeutical perspectives for cancer therapy. *Ann. Transl. Med.* **2017**, *5*, 2. [[CrossRef](#)]
65. Chase, A.; Schultheis, B.; Kreil, S.; Baxter, J.; Hidalgo-Curtis, C.; Jones, A.; Zhang, L.; Grand, F.H.; Melo, J.V.; Cross, N.C.P. Imatinib sensitivity as a consequence of a CSF1R-Y571D mutation and CSF1/CSF1R signaling abnormalities in the cell line GDM. *Leukemia* **2008**, *23*, 358–364. [[CrossRef](#)] [[PubMed](#)]
66. Lee, T.I.; Young, R.A. Transcriptional Regulation and Its Misregulation in Disease. *Cell* **2013**, *152*, 1237–1251. [[CrossRef](#)]
67. Robinson, J.L.L.; Holmes, K.A.; Carroll, J.S. FOXA1 mutations in hormone-dependent cancers. *Front. Oncol.* **2013**, *3*, 20. [[CrossRef](#)] [[PubMed](#)]

68. Roybal, L.L.; Hambarchyan, A.; Meadows, J.D.; Barakat, N.H.; Pepa, P.A.; Breen, K.M.; Mellon, P.L.; Coss, D. Roles of Binding Elements, FOXL2 Domains, and Interactions With cJUN and SMADs in Regulation of FSH $\beta$ . *Mol. Endocrinol.* **2014**, *28*, 1640–1655. [[CrossRef](#)] [[PubMed](#)]
69. Van Antwerp, M.E.; Chen, D.G.; Chang, C.; Prochownik, E.V. A point mutation in the MyoD basic domain imparts c-Myc-like properties. *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 9010–9014. [[CrossRef](#)] [[PubMed](#)]
70. Project, T.I.M.-S.; Richter, J.; Schlesner, M.; Hoffmann, S.; Kreuz, M.; Leich, E.; Burkhardt, B.; Rosolowski, M.; Ammerpohl, O.; Wagener, R.; et al. Recurrent mutation of the ID3 gene in Burkitt lymphoma identified by integrated genome, exome and transcriptome sequencing. *Nat. Genet.* **2012**, *44*, 1316–1320. [[CrossRef](#)] [[PubMed](#)]
71. Byun, S.; Lee, S.-Y.; Lee, J.; Jeong, C.-H.; Farrand, L.; Lim, S.; Reddy, K.; Kim, J.Y.; Lee, M.-H.; Lee, H.J.; et al. USP8 Is a Novel Target for Overcoming Gefitinib Resistance in Lung Cancer. *Clin. Cancer Res.* **2013**, *19*, 3894–3904. [[CrossRef](#)]
72. Lenz, G.; Davis, R.E.; Ngo, V.N.; Lam, L.; George, T.C.; Wright, G.W.; Dave, S.S.; Zhao, H.; Xu, W.; Rosenwald, A.; et al. Oncogenic CARD11 Mutations in Human Diffuse Large B Cell Lymphoma. *Science* **2008**, *319*, 1676–1679. [[CrossRef](#)]
73. Compagno, M.; Lim, W.K.; Grunn, A.; Nandula, S.V.; Brahmachary, M.; Shen, Q.; Bertoni, F.; Ponzoni, M.; Scandurra, M.; Califano, A.; et al. Mutations of multiple genes cause deregulation of NF-kappaB in diffuse large B-cell lymphoma. *Nature* **2009**, *459*, 717–721. [[CrossRef](#)]
74. Schmitz, R.; Young, R.M.; Ceribelli, M.; Jhavar, S.; Xiao, W.; Zhang, M.; Wright, G.L.; Shaffer, A.L.; Hodson, D.J.; Buras, E.; et al. Burkitt lymphoma pathogenesis and therapeutic targets from structural and functional genomics. *Nat. Cell Biol.* **2012**, *490*, 116–120. [[CrossRef](#)] [[PubMed](#)]
75. Zheng, M.; Perry, A.M.; Bierman, P.; Loberiza, F.; Nasr, M.R.; Szwajcer, D.; Del Bigio, M.R.; Smith, L.M.; Zhang, W.; Greiner, T.C. Frequency of MYD88 and CD79B mutations, and MGMT methylation in primary central nervous system diffuse large B-cell lymphoma. *Neuropathology* **2017**, *37*, 509–516. [[CrossRef](#)] [[PubMed](#)]
76. Lin, L.-I.; Chen, C.-Y.; Lin, D.-T.; Tsay, W.; Tang, J.-L.; Yeh, Y.-C.; Shen, H.-L.; Su, F.-H.; Yao, M.; Huang, S.-Y.; et al. Characterization of CEBPA Mutations in Acute Myeloid Leukemia: Most Patients with CEBPA Mutations Have Biallelic Mutations and Show a Distinct Immunophenotype of the Leukemic Cells. *Clin. Cancer Res.* **2005**, *11*, 1372–1379. [[CrossRef](#)] [[PubMed](#)]
77. Ridge, S.A.; Worwood, M.; Oscier, D.; Jacobs, A.; Padua, R.A. FMS mutations in myelodysplastic, leukemic, and normal subjects. *Proc. Natl. Acad. Sci. USA* **1990**, *87*, 1377–1380. [[CrossRef](#)]
78. Liu, Y.; Patel, L.; Mills, G.B.; Lu, K.H.; Sood, A.K.; Ding, L.; Kucherlapati, R.; Mardis, E.R.; Levine, D.A.; Shmulevich, I.; et al. Clinical Significance of CTNNB1 Mutation and Wnt Pathway Activation in Endometrioid Endometrial Carcinoma. *J. Natl. Cancer Inst.* **2014**, *106*, dju245. [[CrossRef](#)]
79. McConechy, M.K.; Ding, J.; Senz, J.; Yang, W.; Melnyk, N.; Tone, A.A.; Prentice, L.M.; Wiegand, K.C.; McAlpine, J.N.; Shah, S.P.; et al. Ovarian and endometrial endometrioid carcinomas have distinct CTNNB1 and PTEN mutation profiles. *Mod. Pathol.* **2014**, *27*, 128–134. [[CrossRef](#)]
80. Pezzuto, F.; Izzo, F.; Buonaguro, L.; Annunziata, C.; Tatangelo, F.; Botti, G.; Buonaguro, F.M.; Tornesello, M.L. Tumor specific mutations in TERT promoter and CTNNB1 gene in hepatitis B and hepatitis C related hepatocellular carcinoma. *Oncotarget* **2016**, *7*, 54253–54262. [[CrossRef](#)] [[PubMed](#)]
81. Mullen, J.T.; DeLaney, T.F.; Rosenberg, A.E.; Le, L.; Iafrate, A.J.; Kobayashi, W.; Szymonifka, J.; Yeap, B.Y.; Chen, Y.-L.; Harmon, D.C.; et al.  $\beta$ -Catenin mutation status and outcomes in sporadic desmoid tumors. *Oncologist* **2013**, *18*, 1043–1049. [[CrossRef](#)] [[PubMed](#)]
82. Mishra, A.; Singh, V.; Verma, V.; Pandey, S.; Trivedi, R.; Singh, H.P.; Kumar, S.; Dwivedi, R.C.; Mishra, S.C. Current status and clinical association of beta-catenin with juvenile nasopharyngeal angiofibroma. *J. Laryngol. Otol.* **2016**, *130*, 907–913. [[CrossRef](#)] [[PubMed](#)]
83. Comino-Méndez, I.; De Cubas, A.A.; Bernal, C.; Álvarez-Escolá, C.; Sánchez-Malo, C.; Ramírez-Tortosa, C.L.; Pedrinaci, S.; Rapizzi, E.; Ercolino, T.; Bernini, G.; et al. Tumoral EPAS1 (HIF2A) mutations explain sporadic pheochromocytoma and paraganglioma in the absence of erythrocytosis. *Hum. Mol. Genet.* **2013**, *22*, 2169–2176. [[CrossRef](#)] [[PubMed](#)]
84. Jamieson, S.; Butzow, R.; Andersson, N.; Alexiadis, M.; Unkila-Kallio, L.; Heikinheimo, M.; Fuller, P.J.; Anttonen, M. The FOXL2 C134W mutation is characteristic of adult granulosa cell tumors of the ovary. *Mod. Pathol.* **2010**, *23*, 1477–1485. [[CrossRef](#)] [[PubMed](#)]
85. Shah, S.P.; Köbel, M.; Senz, J.; Morin, R.D.; Clarke, B.A.; Wiegand, K.C.; Leung, G.; Zayed, A.; Mehl, E.; Kalloger, S.E.; et al. Mutation of FOXL2 in Granulosa-Cell Tumors of the Ovary. *N. Engl. J. Med.* **2009**, *360*, 2719–2729. [[CrossRef](#)]
86. Gielen, G.H.; Gessi, M.; Hammes, J.; Kramm, C.M.; Waha, A.; Pietsch, T. H3F3A K27M mutation in pediatric CNS tumors: A marker for diffuse high-grade astrocytomas. *Am. J. Clin. Pathol.* **2013**, *139*, 345–349. [[CrossRef](#)] [[PubMed](#)]
87. Behjati, S.; Tarpey, P.S.; Presneau, N.; Scheipl, S.; Pillay, N.; Van Loo, P.; Wedge, D.C.; Cooke, S.L.; Gundem, G.; Davies, H.; et al. Distinct H3F3A and H3F3B driver mutations define chondroblastoma and giant cell tumor of bone. *Nat. Genet.* **2013**, *45*, 1479–1482. [[CrossRef](#)]
88. Xu, Z.; Huo, X.; Tang, C.; Ye, H.; Nandakumar, V.; Lou, F.; Zhang, D.; Jiang, S.; Sun, H.; Dong, H.; et al. Frequent KIT Mutations in Human Gastrointestinal Stromal Tumors. *Sci. Rep.* **2015**, *4*, 5907. [[CrossRef](#)]
89. Ravegnini, G.; Mariño-Enriquez, A.; Slater, J.; Eilers, G.; Wang, Y.; Zhu, M.; Nucci, M.R.; George, S.; Angelini, S.; Raut, C.P.; et al. MED12 mutations in leiomyosarcoma and extrauterine leiomyoma. *Mod. Pathol.* **2013**, *26*, 743–749. [[CrossRef](#)] [[PubMed](#)]

90. Laé, M.; Gardrat, S.; Rondeau, S.; Richardot, C.; Caly, M.; Chemlali, W.; Vacher, S.; Couturier, J.; Mariani, O.; Terrier, P.; et al. MED12 mutations in breast phyllodes tumors: Evidence of temporal tumoral heterogeneity and identification of associated critical signaling pathways. *Oncotarget* **2016**, *7*, 84428–84438. [[CrossRef](#)] [[PubMed](#)]
91. Mäkinen, N.; Mehine, M.; Tolvanen, J.; Kaasinen, E.; Li, Y.; Lehtonen, H.J.; Gentile, M.; Yan, J.; Enge, M.; Taipale, M.; et al. MED12, the Mediator Complex Subunit 12 Gene, Is Mutated at High Frequency in Uterine Leiomyomas. *Science* **2011**, *334*, 252–255. [[CrossRef](#)] [[PubMed](#)]
92. Rekhi, B.; Upadhyay, P.; Ramteke, M.P.; Dutt, A. MYOD1 (L122R) mutations are associated with spindle cell and sclerosing rhabdomyosarcomas with aggressive clinical outcomes. *Mod. Pathol.* **2016**, *29*, 1532–1540. [[CrossRef](#)]
93. Du, P.; Huang, P.; Huang, X.; Li, X.; Feng, Z.; Li, F.; Liang, S.; Song, Y.; Stenvang, J.; Brünner, N.; et al. Comprehensive genomic analysis of Oesophageal Squamous Cell Carcinoma reveals clinical relevance. *Sci. Rep.* **2017**, *7*, 1–9. [[CrossRef](#)]
94. Familiades, J.; Bousquet, M.; Lafage-Pochitaloff, M.; Béné, M.-C.; Beldjord, K.; De Vos, J.; Dastugue, N.; Coyaoud, E.; Struski, S.; Quelen, C.; et al. PAX5 mutations occur frequently in adult B-cell progenitor acute lymphoblastic leukemia and PAX5 haploinsufficiency is associated with BCR-ABL1 and TCF3-PBX1 fusion genes: A GRAALL study. *Leukemia* **2009**, *23*, 1989–1998. [[CrossRef](#)]
95. Mullighan, C.G.; Goorha, S.; Radtke, I.; Miller, C.B.; Coustan-Smith, E.; Dalton, J.D.; Girtman, K.; Mathew, S.; Ma, J.; Pounds, S.B.; et al. Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nat. Cell Biol.* **2007**, *446*, 758–764. [[CrossRef](#)]
96. Ouyang, Y.; Qiao, C.; Chen, Y.; Zhang, S.-J. Clinical significance of CSF3R, SRSF2 and SETBP1 mutations in chronic neutrophilic leukemia and chronic myelomonocytic leukemia. *Oncotarget* **2017**, *8*, 20834–20841. [[CrossRef](#)]
97. Piazza, R.; Valletta, S.; Winkelmann, N.; Redaelli, S.; Spinelli, R.; Pirola, A.; Antolini, L.; Mologni, L.; Donadoni, C.; Papaemmanuil, E.; et al. Recurrent SETBP1 mutations in atypical chronic myeloid leukemia. *Nat. Genet.* **2012**, *45*, 18–24. [[CrossRef](#)]
98. Meggendorfer, M.; Roller, A.; Haferlach, T.; Eder, C.; Dicker, F.; Grossmann, V.; Kohlmann, A.; Alpermann, T.; Yoshida, K.; Ogawa, S.; et al. SRSF2 mutations in 275 cases with chronic myelomonocytic leukemia (CMML). *Blood* **2012**, *120*, 3080–3088. [[CrossRef](#)] [[PubMed](#)]
99. Ballmann, C.; Thiel, A.; Korah, H.E.; Reis, A.-C.; Saeger, W.; Stepanow, S.; Köhrer, K.; Reifenberger, G.; Knobbe-Thomsen, C.B.; Knappe, U.J.; et al. USP8 Mutations in Pituitary Cushing Adenomas—Targeted Analysis by Next-Generation Sequencing. *J. Endocr. Soc.* **2018**, *2*, 266–278. [[CrossRef](#)] [[PubMed](#)]
100. Chakravarty, D.; Gao, J.; Phillips, S.M.; Kundra, R.; Zhang, H.; Wang, J.; Rudolph, J.E.; Yaeger, R.; Soumerai, T.; Nissan, M.H.; et al. OncoKB: A Precision Oncology Knowledge Base. *JCO Precis. Oncol.* **2017**, *2017*, 1–16. [[CrossRef](#)] [[PubMed](#)]
101. Pajkos, M.; Zeke, A.; Dosztányi, Z. Ancient Evolutionary Origin of Intrinsically Disordered Cancer Risk Regions. *Biomolecules* **2020**, *10*, 1115. [[CrossRef](#)]
102. Mitrea, D.M.; Kriwacki, R.W. On the relationship status for Arf and NPM1—It’s complicated. *FEBS J.* **2018**, *285*, 828–831. [[CrossRef](#)] [[PubMed](#)]
103. Scatena, R.; Bottoni, P.; Pontoglio, A.; Mastrototaro, L.; Giardina, B. Glycolytic enzyme inhibitors in cancer treatment. *Expert Opin. Investig. Drugs* **2008**, *17*, 1533–1545. [[CrossRef](#)] [[PubMed](#)]
104. Griffith, D.; Parker, J.P.; Marmion, C.J. Enzyme inhibition as a key target for the development of novel metal-based anti-cancer therapeutics. *Anti-Cancer Agents Med. Chem.* **2010**, *10*, 354–370. [[CrossRef](#)] [[PubMed](#)]
105. Pathania, S.; Bhatia, R.; Baldi, A.; Singh, R.; Rawal, R.K. Drug metabolizing enzymes and their inhibitors’ role in cancer resistance. *Biomed. Pharmacother.* **2018**, *105*, 53–65. [[CrossRef](#)]
106. Lounnas, V.; Ritschel, T.; Kelder, J.; McGuire, R.; Bywater, R.P.; Foloppe, N. Current progress in structure-based rational drug design marks a new mindset in drug discovery. *Comput. Struct. Biotechnol. J.* **2013**, *5*, e201302011. [[CrossRef](#)]
107. Kulkarni, P. Intrinsically disordered proteins and prostate cancer: Pouring new wine in an old bottle. *Asian J. Androl.* **2016**, *18*, 659–661. [[CrossRef](#)] [[PubMed](#)]
108. Neira, J.L.; Bintz, J.; Arruebo, M.; Rizzuti, B.; Bonacci, T.; Vega, S.; Lanas, A.; Velázquez-Campoy, A.; Iovanna, J.L.; Abián, O. Identification of a Drug Targeting an Intrinsically Disordered Protein Involved in Pancreatic Adenocarcinoma. *Sci. Rep.* **2017**, *7*, 39732. [[CrossRef](#)]
109. Metallo, S.J. Intrinsically disordered proteins are potential drug targets. *Curr. Opin. Chem. Biol.* **2010**, *14*, 481–488. [[CrossRef](#)]
110. Wang, N.J.; Sanborn, Z.; Arnett, K.L.; Bayston, L.J.; Liao, W.; Proby, C.M.; Leigh, I.M.; Collisson, E.A.; Gordon, P.B.; Jakkula, L.; et al. Loss-of-function mutations in Notch receptors in cutaneous and lung squamous cell carcinoma. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 17761–17766. [[CrossRef](#)]
111. Ballerini, P.; Struski, S.; Cresson, C.; Prade, N.; Toujani, S.; Deswarte, C.; Dobbelsstein, S.; Petit, A.; Lapillonne, H.; Gautier, E.-F.; et al. RET fusion genes are associated with chronic myelomonocytic leukemia and enhance monocytic differentiation. *Leukemia* **2012**, *26*, 2384–2389. [[CrossRef](#)] [[PubMed](#)]
112. An, J.; Ren, S.; Murphy, S.J.; Dalangood, S.; Chang, C.; Pang, X.; Cui, Y.; Wang, L.; Pan, Y.; Zhang, X.; et al. Truncated ERG Oncoproteins from TMPRSS2-ERG Fusions Are Resistant to SPOP-Mediated Proteasome Degradation. *Mol. Cell* **2015**, *59*, 904–916. [[CrossRef](#)] [[PubMed](#)]
113. Lesueur, F.; French Familial Melanoma Study Group; De Lichy, M.; Barrois, M.; Durand, G.; Bombled, J.; Avril, M.-F.; Chompret, A.; Boitier, F.; Lenoir, G.M.; et al. The contribution of large genomic deletions at the CDKN2A locus to the burden of familial melanoma. *Br. J. Cancer* **2008**, *99*, 364–370. [[CrossRef](#)] [[PubMed](#)]



Article

# Ancient Evolutionary Origin of Intrinsically Disordered Cancer Risk Regions

Mátyás Pajkos<sup>1</sup>, András Zeke<sup>2</sup> and Zsuzsanna Dosztányi<sup>1,\*</sup>

<sup>1</sup> Department of Biochemistry, ELTE Eötvös Loránd University, Pázmány Péter stny 1/c, H-1117 Budapest, Hungary; matyaspajkos@caesar.elte.hu

<sup>2</sup> Research Centre for Natural Sciences, Magyar tudósok körútja 2, H-1117 Budapest, Hungary; zeke.andras@ttk.mta.hu

\* Correspondence: dosztanyi@caesar.elte.hu

Received: 21 June 2020; Accepted: 20 July 2020; Published: 28 July 2020



**Abstract:** Cancer is a heterogeneous genetic disease that alters the proper functioning of proteins involved in key regulatory processes such as cell cycle, DNA repair, survival, or apoptosis. Mutations often accumulate in hot-spots regions, highlighting critical functional modules within these proteins that need to be altered, amplified, or abolished for tumor formation. Recent evidence suggests that these mutational hotspots can correspond not only to globular domains, but also to intrinsically disordered regions (IDRs), which play a significant role in a subset of cancer types. IDRs have distinct functional properties that originate from their inherent flexibility. Generally, they correspond to more recent evolutionary inventions and show larger sequence variations across species. In this work, we analyzed the evolutionary origin of disordered regions that are specifically targeted in cancer. Surprisingly, the majority of these disordered cancer risk regions showed remarkable conservation with ancient evolutionary origin, stemming from the earliest multicellular animals or even beyond. Nevertheless, we encountered several examples where the mutated region emerged at a later stage compared with the origin of the gene family. We also showed the cancer risk regions become quickly fixated after their emergence, but evolution continues to tinker with their genes with novel regulatory elements introduced even at the level of humans. Our concise analysis provides a much clearer picture of the emergence of key regulatory elements in proteins and highlights the importance of taking into account the modular organisation of proteins for the analyses of evolutionary origin.

**Keywords:** intrinsically disordered regions; linear motifs; gene duplications; de novo; evolutionary origin

## 1. Introduction

Most human genes are thought to have an extensive and very deep evolutionary history. In line with the thought “Nature is a tinkerer, not an inventor” [1], major human gene families date back to the earliest Eukaryotic evolutionary events, or even beyond. The very oldest layers of human genes encode metabolically, structurally, or otherwise essential proteins that typically go back to unicellular evolutionary stages. Mutations to this core biochemical apparatus can prove disruptive to all aspects of cellular life, and indeed, there are known mutational targets associated with genome stability and cancer. In contrast to these “caretaker” genes, a more novel set of genes have emerged at the transition to a multicellular stage. These “gatekeeper” proteins are involved in cell-to-cell communication, especially in early embryonic development and tissue regeneration. Gatekeeper genes that control cell division are among the best known cancer-associated oncogenes and tumor suppressors [2].

In order to establish the evolutionary origins of cancer genes, Domazet-Loaso and Tautz carried out a systematic analysis based on phylostratigraphic tracking [3]. By correlating the evolutionary

origin of genes with particular macroevolutionary transitions, they found that a major peak connected to the emergence of cancer genes corresponds to the level where multicellular animals have emerged. However, many cancer genes have a more ancient origin and can be traced back to unicellular organisms. These trends seem to apply to the appearance of disease genes [4] and novel genes in general as well [5]. These studies were based on the evolutionary history of the founder domains. However, new genes can also be generated by duplication either in whole or from part of existing genes, when the duplicate copy of a gene becomes associated with a different phenotype to its paralogous partner. This mechanism can also influence the emergence of disease genes [5].

By taking advantage of the flux of cancer genome data, several new proteins have been identified to play a direct role in driving tumorigenesis during recent years [6]. One of the key signatures of cancer drivers is the presence of mutation hotspot regions, where many different patients might show a similarly recurrent pattern of mutations [7]. These hotspots are typically located within well-folded, structured domains. However, many cancer associated proteins have a complex modular architecture, incorporating not only globular domains, but also intrinsically disordered segments, which can also be sites of cancer mutations. In our recent work, we systematically collected disordered regions that are directly targeted by cancer mutations and analyzed their basic functional and system level properties. [8]. While only a relatively small subset of such disordered cancer drivers was identified, their mutations can be the main driver event in certain cancer types. These disordered regions can function in a variety of ways including post-transcriptional modification sites (PTMs), linear motifs, linkers, and larger sized functional modules typically involved in binding to macromolecular complexes. These disordered cancer drivers have a characteristic functional repertoire and increased interaction potential, and their perturbation can give rise to all ten hallmarks of cancer independently of ordered drivers [8].

In general, owing to the lack of structural constraints, disordered segments show more evolutionary variability [9]. In particular, linear motifs can easily emerge to a previously non-functional region of protein sequence by only a few mutations, or disappear as easily, leaving little trace after millions or billions of years [10]. However, elements fulfilling a critical regulatory function might linger on for a longer time. So far, the evolutionary origin of intrinsically disordered regions that have a critical function proven by a human disease association has not been analyzed.

In the current study, we studied the evolutionary origin of disordered cancer risk regions. For this, we used a dataset of cancer driving proteins in which cancer mutations specifically targeted intrinsically disordered regions [8]. We retrieved phylogeny data from the ENSEMBL Compara database. Using a novel conservation and phylogenetic-based strategy, we determined the evolutionary origin not only at the gene level, but also at the region level. In addition, we also investigated the emergence mechanism of disordered cancer risk regions and how evolutionary constraints, selection, and gene duplications events influenced the fate of these examples. Finally, we presented interesting case studies that demonstrate the ancient evolutionary origin of these examples and the continuing evolution of their genes built around the critical conserved functional module.

## 2. Materials and Methods

### 2.1. Dataset

We used a subset of the previously identified disordered cancer risk regions [8]. These regions were identified based on genetic variations collected from the COSMIC database [11] using the method that located specific regions that are enriched in cancer mutations [7]. Disorder status of these regions was verified based on experimental data collected from dedicated databases and from the literature when available, or based on consensus disorder prediction methods [8]. Mapping was not feasible for CDKN2A isoform (Tumor suppressor ARF), because it was not present in the ENSEMBL database we used in our study, hence this protein was excluded from the further analyses. Proteins in which both disordered and ordered cancer regions were identified were filtered out in order to be able to

focus clearly on the disordered regions. Regions that were primarily mutated by in-frame insertion and deletion and contained less than 15 missense mutations were also excluded because of our conservation calculation method (see below). Finally, histone proteins were merged, keeping the single entry of HIST1H3B. Ultimately, we obtained a list of 36 disordered cancer risk regions of 32 proteins APC (Adenomatous polyposis coli protein): 1284–1537, ASXL1 (Polycomb group protein ASXL1): 1102–1107, BCL2 (Apoptosis regulator Bcl-2): 2–80, CALR(Calreticulin): 358–384, CARD11 (Caspase recruitment domain-containing protein 11): 111–134; 207–266; 337–436, CBL (E3 ubiquitin-protein ligase CBL): 365–374, CCND3 (G1/S-specific cyclin-D3): 278–290, CD79B (B-cell antigen receptor complex-associated protein beta chain): 191–199, CEBPA (CCAAT/enhancer-binding protein alpha): 293–327, CSF1R (Macrophage colony-stimulating factor 1 receptor): 969–969, CTNNB1 (Catenin beta-1): 32–45, EIF1AX (Eukaryotic translation initiation factor 1A, X-chromosomal): 4–15, EPAS1 (Endothelial PAS domain-containing protein 1): 529–539, ESR1 (Estrogen receptor): 303–303, FOXA1 (Hepatocyte nuclear factor 3-alpha): 248–268, FOXL2 (Forkhead box protein L2): 134–134, FOXO1 (Forkhead box protein O1): 19–26, HIST1H3B (Histone H3.1): 28–28, ID3 (DNA-binding protein inhibitor ID-3): 48–70, MED12 (Mediator of RNA polymerase II transcription subunit 12): 44–44, MLH1 (DNA mismatch repair protein Mlh1): 379–385, MYC (Myc proto-oncogene protein): 57–60, MYCN(N-myc proto-oncogene protein): 44–44, MYOD1(Myoblast determination protein 1): 122–122, NFE2L2 (Nuclear factor erythroid 2-related factor 2): 20–38; 75–82, PAX5 (Paired box protein Pax-5): 75–80, RPS15 (40S ribosomal protein S15): 129–145, SETBP 1 (SET-binding protein): 858–880, SMARCB1 (SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily B member 1): 368–381, SRSF2 (Serine/arginine-rich splicing factor 2): 95–95, USP8 (Ubiquitin carboxyl-terminal hydrolase 8): 713–736, VHL (von Hippel-Lindau disease tumor suppressor): 54–136; 144–193.

## 2.2. Evolutionary Framework

In this work, we calculated the evolutionary origin of cancer risk regions within our dataset of disordered proteins. Our approach focused on the age of orthologous gene families, instead of focusing on the evolutionary origin of founder domains. Assignment of age of human gene families (origin) was carried out using the ENSEMBL genome browser database. To identify the origin of individual human gene families, we fetched the phylogenies and analysed the evolutionary supertrees built by the pipeline of the ENSEMBL Compara multi-species comparisons project [12,13]. The used release (99) of the project contained 282 reference species including 277 vertebrata, 4 eumetazoa, and 1 opisthokonta (*S. cerevisiae*) species. Note that, in these phylogenies, the most ancient node can be the ancestor of yeast. The origin of the gene family was identified by taking the taxonomy level of the most ancient node of the phylogenetic supertrees. Taxonomy levels were broken into major nested age categories (mammals, vertebrates, eumetazoa, opisthokonta), similarly to previous studies [14].

To define the evolutionary origin of regions, we built a customized pipeline that included collecting and mapping mutations from COSMIC database to ENSEMBL entries, constructing multiple sequence alignments of protein families, and mapping the cancer regions among orthologs and paralogs. According to the ENSEMBL supertrees, protein sequences of human paralogs (including the cancer gene) and their orthologs were queried from the database using the Rest API function. Then, multiple sequence alignments of the corresponding sequences were created with MAFFT (default settings) [15]. On the basis of the sequence alignments, cancer regions were mapped onto the sequences. In the mapping step, cancer regions were considered as functional units (linear motifs, linkers, disordered domains) and borders of the regions were defined according to this. When the highly mutated regions covered only a single residue, it was extended to cover the known functional linear motif or using its sequence neighbourhood. On this basis, the subset of paralogs, in which the mapped cancer region was found to be conserved, was identified.

Next, the set of sequences containing regions that showed evolutionary similarity to the mutated regions were identified among the collected orthologs and paralogs. Conservation of the regions among paralogs was evaluated relying on two strategies, by calculating the similarity of mutated positions in

the cancer risk regions (see below) and based on HMM profiles. This consideration was taken into account in order to reduce the chance of false conservation interpretation arising from the difficulty of aligning disordered proteins. The HMM profiles were built from conserved cancer regions of vertebrate model organisms using the HMMER (version 3.3) method [16]. The identified region hits were manually checked to minimize the chance of false positives or negatives. Next, we identified the evolutionarily most distant relative in which the cancer region was declared to be conserved. As a result, the origin of the region could differ from the origin of the orthologous gene family, when paralogue sequences that contained the conserved motif had a more ancient origin. Basically, we treated the cancer risk regions as the founder of the family. The taxonomy level of this ortholog was defined as the level in which the cancer region emerged in the common ancestor of this ortholog and *H. sapiens*.

### 2.3. Region Conservation

Within the identified cancer risk region, some of the positions could be more heavily mutated and are likely to be more critical for the function of this region. We took this into account when calculating the region conservation. Mutations for each position collected from the COSMIC database were mapped to the corresponding ENSEMBL human entry. On the basis of the sequence alignment corresponding to the cancer risk regions, we identified the positions that were similar to the reference sequence. Two positions were considered similar when the substitution score was non-negative according to the BLOSUM62 substitution matrix. A given cancer region was considered to be conserved between homologs, when the conserved residues carried more than 50% of missense mutations.

### 2.4. Positive Selection: Selectome and McDonald and Kreitman (MK) Test Results

For each entry in our dataset, we collected information about positive selection using the Selectome database (current version 6) [17]. This database contains collected sites of positive selection detected on a single branch of the phylogeny using the systematic branch-site test of the CODEML algorithm from the PAML [18] phylogenetic package version 4b. The ratio of non-synonymous and synonymous substitutions ( $\omega$ ) can be interpreted as a measurement of selective pressure indicating purifying ( $\omega$  values  $< 1$ ), neutral ( $\omega$  values = 1), or positive ( $\omega$  values  $> 1$ ) selection. In our work, positions under positive selection that have a posterior probability higher than 0.9 were extracted from the database and mapped onto our gene set.

However, the branch-site model generally cannot detect species-specific positive selection. Potential cases of human-specific positive selection may be detected effectively by comparing divergence to polymorphism data, as in the McDonald and Kreitman (MK) test. Human-specific positive selection detected by MK test previously calculated [19] was mapped onto our dataset of disordered cancer genes.

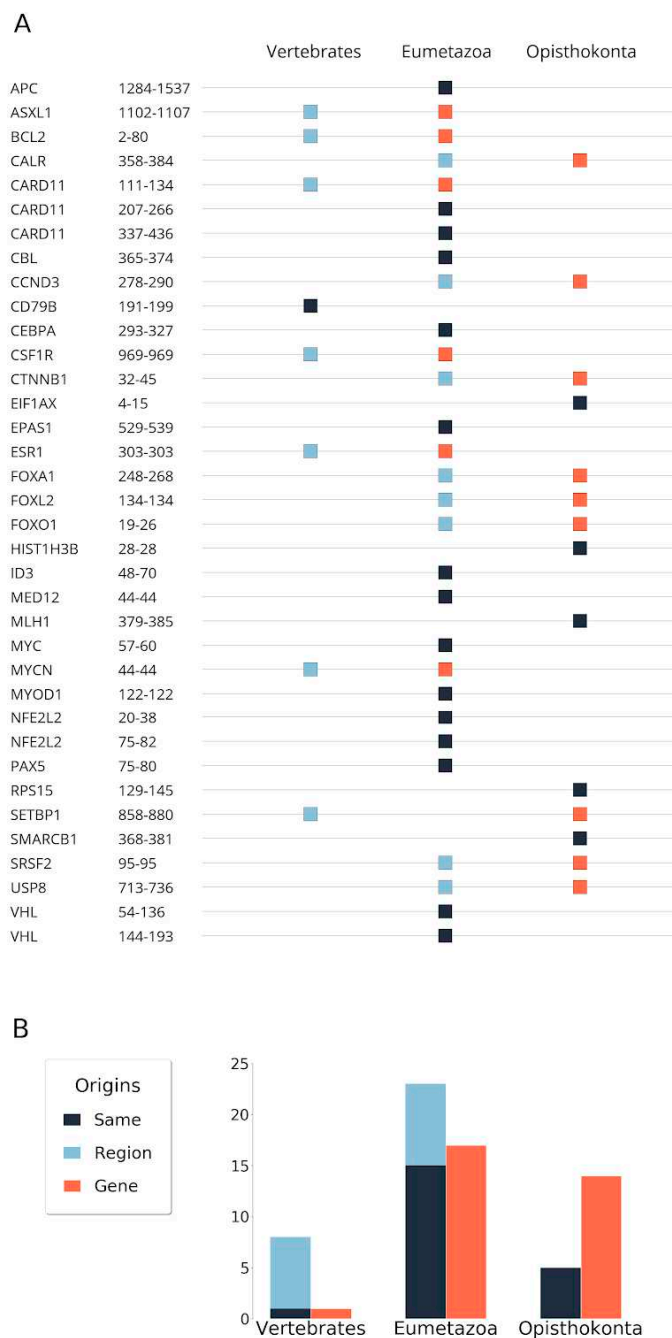
## 3. Results

### 3.1. Evolutionary Origin of Genes and Regions

Altogether, we collected 36 cancer risk regions of 32 disordered proteins and investigated the evolutionary origin at the level of genes and regions. The age estimation of disordered cancer genes was obtained using the last common ancestor of descendants using the ENSEMBL supertrees, which includes phylogeny of gene families returning not only individual gene history, but also relationships of ancient paralogs and their history (see Material and Methods). Using this strategy instead of analysing the evolution of individual genes or simply the emergence of the founder domain, we could define the origin of regions more precisely, even the ancient ones, without introducing any bias of overprediction of origins. However, some ambiguity still remained and was manually checked (Supplementary Materials 1). The genes were traced back to opisthokonta (in accordance with the ENSEMBL database) and divided into four major phylostratigraphic groups, which are associated with the emergence of unicellular, multicellular organisms, vertebrates, and mammals.



Previous results identified the level of eumetazoa as the main age for the emergence of cancer associated proteins [3]. We observed a similar trend in the case of disordered cancer proteins. Specifically, we found that 21 disordered cancer proteins, the majority of cases, have emerged at the level of eumetazoa (Figure 1). Fourteen cases were found to be even more ancient and could be traced back to single cell organisms, at least to opisthokonta. The only protein that emerged more recently, at the level of vertebrates, was CD79B, the B-cell antigen receptor complex-associated protein  $\beta$  chain. Its appearance is in agreement with the birth of many immune receptors [20] and is assumed to be driven by the insertion of transposable elements.

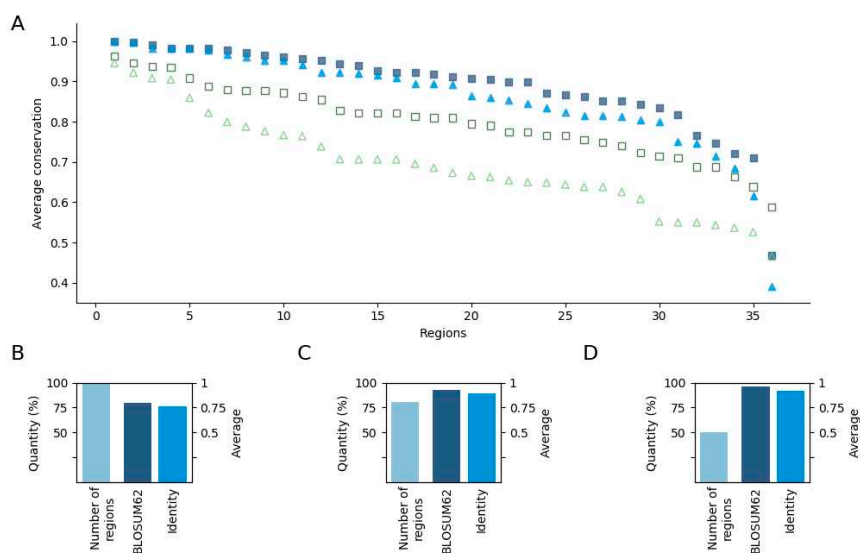


**Figure 1.** Conservation-based evolutionary origin of disordered cancer regions and genes. (A) The orange and sky blue squares represent the origin of genes and regions, respectively. Gunmetal squares indicate the same evolutionary origin at both region and gene levels. (B) Summary barchart of origins in the three gene-age categories.

In around half of the cases (21), the emergence of the mutated region was the same as the emergence of the protein (Figure 1). Strikingly, these included five cases (EIF1AX, HIST1H3B, MLH1, RPS15, SMARCB1) where not only the gene/protein, but also the region primarily mutated in human cancers were very ancient and could be traced back to unicellular organisms. Fifteen regions with Eumetazoa and one with Vertebrata origin could be traced back to the same level as their corresponding gene. However, in several cases, the emergence of the region was a more recent event compared with the emergence of the gene. Of these, eight regions emerged at the Eumetazoa and seven at the Vertebrate level. In general, there was only one level difference between the emergence of the gene and the region at this resolution. The only exception was SETBP1. In this case, the region itself emerged at the vertebrate level. However, the gene could be traced back to opisthokonta level, although the eumetazoa origin cannot be completely ruled out (see Supplementary Materials 1). Overall, many of the disordered regions were more recent evolutionary inventions compared with the origin of their genes, and date back to the common ancestors of eumetazoans or vertebrates. Nevertheless, the ancestors of all of the regions were already present from the vertebrate level.

### 3.2. Position Conservation

Overall, these results point to the ancient evolutionary origin of disordered regions involved in cancer, not only at the gene level, but also at the region level. To take a closer look, we also calculated the conservation of individual positions within the regions based both in terms of homologous substitutions and identity. The results show that these residues are highly conserved even compared with the conservation of the whole sequence (Figure 2). Here, 86% of the regions have more than 0.8 average conservation value even based on identities (Figure 2A). Among the cases with the four lowest values, the conservation of VHL, CALR, and APC, which all correspond to relatively longer segments, was still relatively high. The only outlier was BCL2. In this case, the mutations are distributed along the N-terminal, encompassing the highly conserved BH4 motif, as well as the linker region between the BH4 and C-terminal part, which is conserved only in mammals (Figure S1).



**Figure 2.** Representation of average conservation values. (A) Sorted conservation values for each region having positions with at least one mutation and for the whole protein. Squares (dark blue—region, green—whole sequence) and triangles (light blue—regions, green—full sequence) represent BLOSUM62 and identity based conservation values, respectively. The outlier at the very end of the sequence corresponds to the region of BCL2. (B–D) The number of regions and average conservation value of regions having positions with at least 1, 15, and 25 mutations, respectively. The conservation values are based on BLOSUM62 and identity, and the number of regions are colored by dark, medium, and sky blue, respectively.

Next, we investigated how this average value is altered when only the highly mutated positions are considered. We repeated that analysis for positions that had at least 15 and 25 missense mutations, which slightly decreased the number of regions considered. The remaining 28 and 17 regions with positions having at least 15 and 25 mutations had 0.93, 0.89, 0.96, and 0.92 average conservation values based on substitutions and identity, respectively (Figure 2C,D). This reflects a very clear trend with positions with a higher number of cancer mutations showing higher evolutionary conservation.

We also collected sites of potential positive selection mapped onto our genes based on the Selectome database [17], which provides information on likely molecular selection both at the level of the evolutionary branch and the sequence position based on the ratio of non-synonymous and synonymous substitutions ( $\omega$ ). According to these results, positive selection affected only three genes on the human lineage in our dataset, CALR, CTNNB1, and VHL. All of these selections could be mapped onto the vertebrates division with multiple positions (see Material and Methods) (Table 1).

**Table 1.** Positive selection within disordered cancer genes. Positions within cancer risk regions are colored blue. The numbers in brackets are the posterior probability of positive selection for each position.

Gene	Positions under Positive Selection Referring to the Human Protein Sequence
CALR	83(0.971), 155(0.971), 177(0.990), 267(0.995), 307(0.994), 336(0.991), 360(0.999)
CTNNB1	121(0.999), 206(0.993), 250(0.998), 287(0.991), 411(0.998), 433(0.993), 525(0.997), 552(0.998), 556(0.916)
VHL	127(0.957), 132(0.942), 141(0.923), 171(0.947), 183(0.963), 185(0.920)

However, these positions showed limited overlap with the mutated regions. In the case of CTNNB1, none of the positions under selection overlapped with the cancer mutated region. In the case of CALR, there was only a single position under selection within the cancer risk region, but it was not directly targeted by cancer mutations. In the case of VHL, six positions were detected with selective pressure and five of them were situated within the significantly mutated region. However, none of them corresponded to a highly mutated residue.

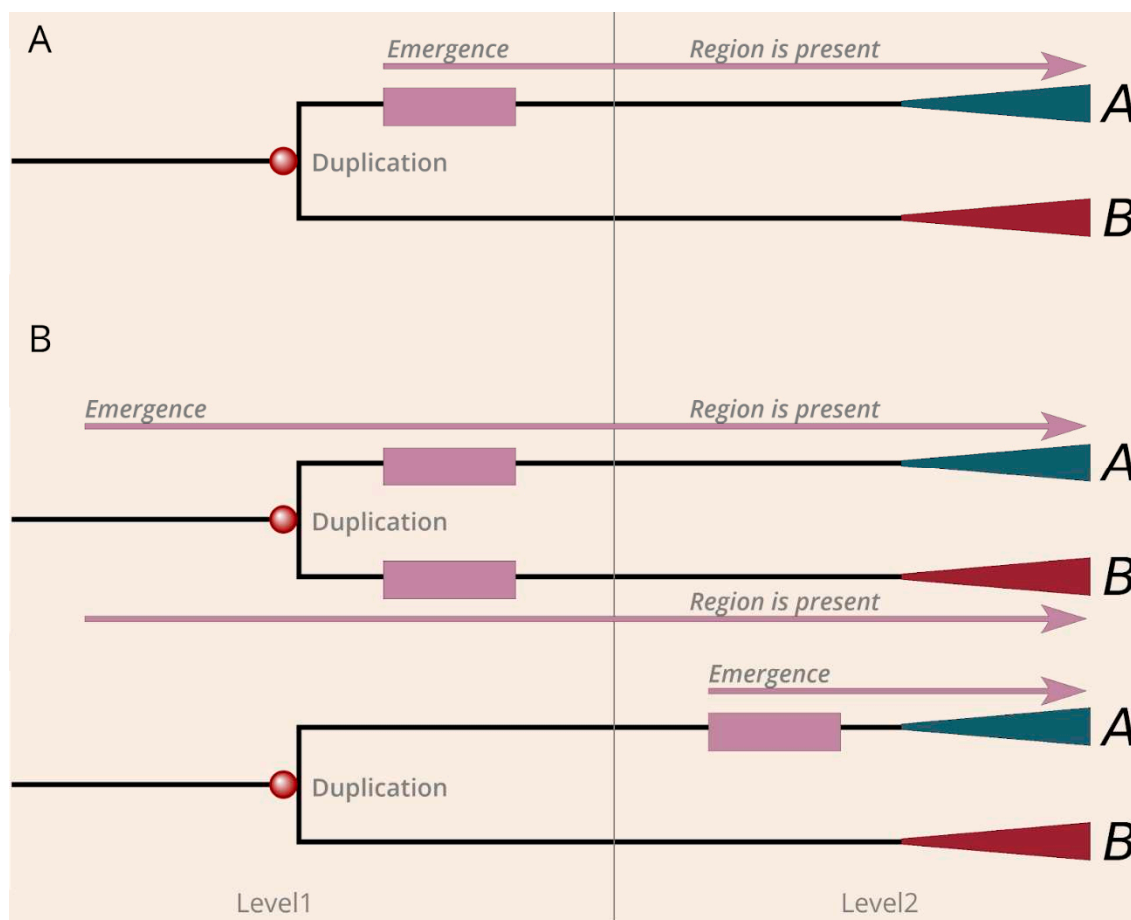
Taking advantage of an earlier analysis [19], we also analyzed if there was any human specific positive selection. As the  $\omega$  based approach can not be used without uncertainty to identify human-specific positive selection, this work relied on the McDonald and Kreitman (MK) test, which compares the divergence to polymorphism data using closely related species, such as human and chimp. There was only a single entry in our database, ESR1, that showed human specific evolutionary changes (see case studies).

### 3.3. Contribution of Duplications to the Emergence of Disease Risk Regions

Gene duplications often drive the appearance of a novel function through the process called neofunctionalization. In these cases, after a duplication event, one copy may acquire a novel, beneficial function that becomes preserved by natural selection. Here, we have evaluated whether the emergence of disordered cancer regions corresponds to such neofunctionalization events. For this analysis, we collected paralog sequences and evaluated if there were regions present in these sequences that showed clear evolutionary similarity to the cancer mutated region.

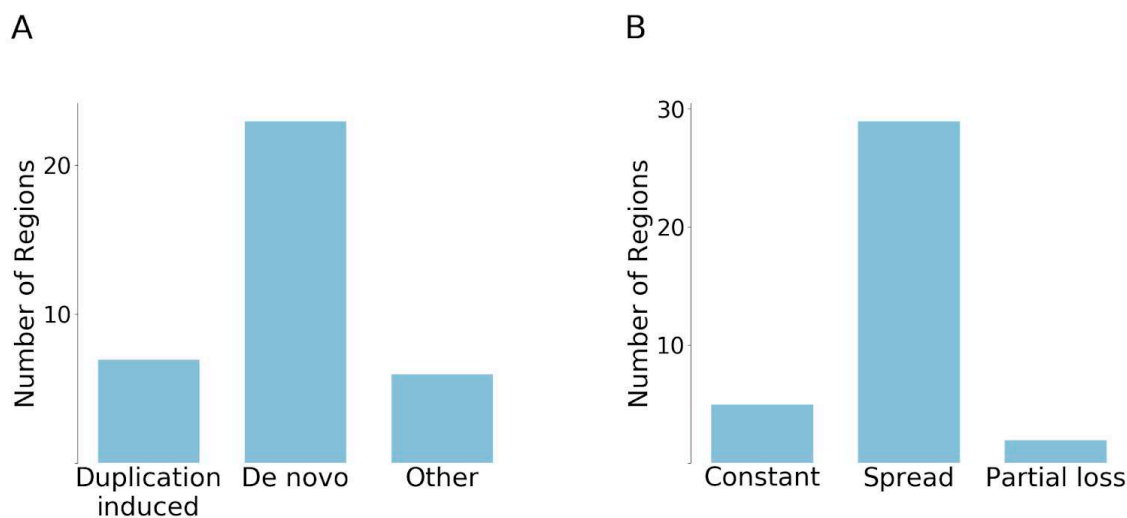
The evolutionary history of many genes is quite complex and can involve multiple duplication events. We focused on the level where the cancer regions emerged and distinguished the following scenarios based on the relationship between the duplication and the presence of the region among the paralogs. The first scenario corresponds to duplication induced neofunctionalization. In this case, an ancient cancer region emerged directly after a given gene duplication and became preserved in only one of the branches that appeared after the duplication (Figure 3A). There are two basic scenarios in which the duplication cannot be directly linked with the emergence of the regions. One possible scenario is when both branches contain the region, which indicates that the region must have emerged before the duplication (Figure 3B). The other possible scenario is when the region emerged at a later

evolutionary stage after a duplication, and duplication cannot be directly linked to neofunctionalization (Figure 3B).



**Figure 3.** The mechanisms of emergence of regions by neofunctionalization and de novo. (A) Demonstration of the model of duplication induced (neofunctionalization) cancer region emergence. (B) Depiction of the two sub-scenarios of the de novo region emergence. Mallow boxes and arrows explain the evolution of the region. Red and green triangles symbolize the further evolution of paralogs after gene duplications.

Surprisingly, the duplication induced neofunctionalization was much less common than we expected, with only seven cases showing this behaviour. One example for this scenario is presented by the  $\beta$ -catenin family, where the degron motif [21] based cancer risk region that emerged after duplication is present only on the branch of  $\beta$ -catenin and junctional plakoglobin (JUP). In contrast, we found that 23 regions have evolved by de novo emergence, which seemed to be the dominant mechanisms for the emergence of the analyzed cancer mutated disordered regions (Figure 4A). For example, ID3 underwent multiple duplications, but all paralogs contain the cancer risk region, which indicates that the region emerged prior to the duplication. Another example is ESR1, in which case the paralogs were born at the level of eumetazoa; however, this event is not directly linked to the emergence of the cancer region, which appeared only at the level of the ancient vertebrates. In addition, there were two singletons in our dataset, RPS15 and SMARCB1, which did not have any detectable paralogs. In the cases of ASXL1, CCND3, SETBP1, and the first region of CARD11, the evolutionary scenarios could not be unambiguously established. These six examples formed the “Other” group.



**Figure 4.** Categorization of emergence scenarios and evolutionary fates of cancer regions. (A) The number of regions that have emerged by duplication or de novo. Six regions were not categorized (Other). (B) Classification of cancer regions in terms of their evolutionary fate after emergence.

We also analyzed if additional duplication events occurred after the emergence of regions and whether the novel paralogues retained the regions. There are basically three scenarios that can occur: (i) the region is preserved without any further duplications; (ii) the region spreads and becomes preserved in all of the novel duplicates; (iii) partial loss scenario, that is, the region is preserved in some duplicates, but is lost in others. Our results show that the most common evolutionary fate is the second one (Figure 4B). In 29 cases, at least one duplication that inherited the region can be observed after the emergence of the cancer region. In contrast, only five regions were not duplicated. Some ancient cases, such as MLH1 and USP8, are also included among the non-duplicated ones, which means that the reason for the lack of duplications is not the short evolutionary time. The partial loss scenario was observed in only two cases, in the case of VHL and NFE2L2. For instance, in the case of VHL, there was a relatively recent gene duplication at the level of mammals. While the N-terminal segment is present on both paralogs (VHL and VHLL), the C-terminal segment is only present in VHL, but was lost from VHLL. In a similar fashion, NFE2L2 underwent a more recent gene duplication at the level of vertebrates, but the newly emerged paralog did not retain the two linear motifs that are primarily targeted by cancer mutations.

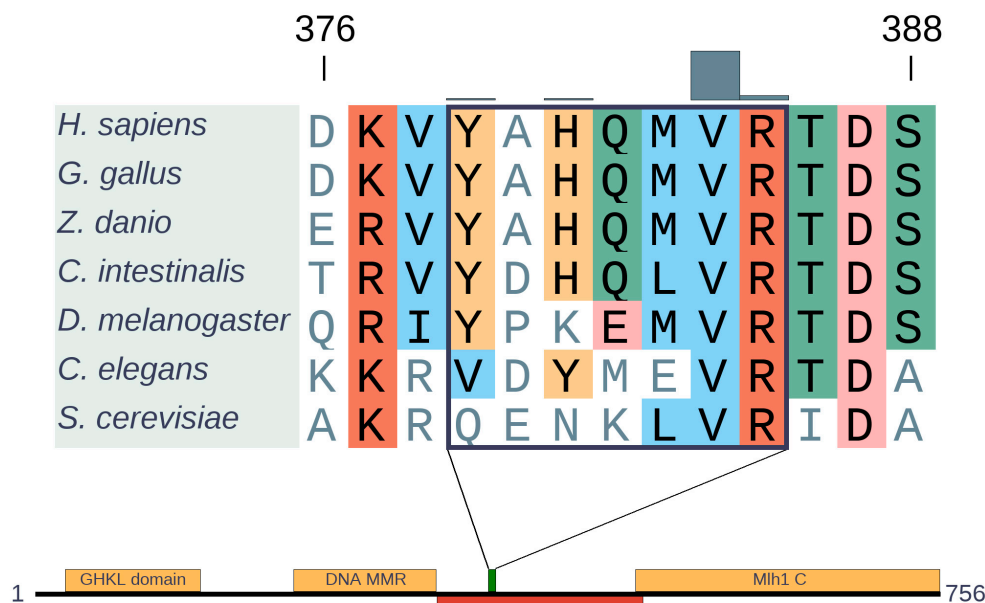
### 3.4. Case Studies

#### 3.4.1. MLH1

One of the most ancient examples in our dataset corresponds to MLH1 (MutL Homolog 1), an essential protein in DNA mismatch repair (MMR). As one of the classic examples of a caretaker function, mutations of MLH1 can lead to cancer by increasing the rate of single-base substitutions and frameshift mutations [22]. Several positions of MLH1 are mutated in people with Lynch syndrome, also known as hereditary nonpolyposis colorectal cancer (HNPCC). However, according to the COSMIC database of somatic cancer mutations, the most common mutation of MLH1 is V384D. Mutational studies of V384D using yeast assays and in vitro MMR assay did not indicate a strong phenotype, but still showed a limited decrease of MMR activity [23]. However, it was shown that the (mostly germline) V384D variant is clearly associated with increased colorectal cancer susceptibility [24], and it is highly prevalent in HER2-positive luminal B breast cancer [25].

MLH1 is an ancient protein that is present from bacteria to humans. It has a highly conserved domain organization that involves ordered N- and C-terminal domains connected by a disordered linker [26] (Figure 5). This underlines the functional importance not only of the structured domains,

but also of the connecting disordered region. In our previous work, we identified the region from 379 to 385 to be significantly mutated [7], which is located within the disordered segment. Recently, it was shown that the linker can regulate both DNA interactions and enzymatic activities of neighboring structured domains [27]. In agreement with the linker function, both the composition and length of this intrinsically disordered region (IDR) are critical for efficient MMR. Overall, most of the linker shows relatively low sequence conservation, however, the identified cancer risk region is highly conserved from across all eukaryotic sequences (Figure 5), in an island-like manner. Although the exact function of this region is not known, the strong evolutionary conservation indicates a highly important function, not yet explored in detail.



**Figure 5.** Alignment of MLH1 orthologs generated with MAFFT [15] and domain structure of human MLH1. The segment of the alignment represents the cancer region (highlighted by a rectangle) with the missense mutation distribution depicted by gray bars. Domains are depicted by yellow, disordered regions by red boxes, while the green box indicates the cancer risk region.

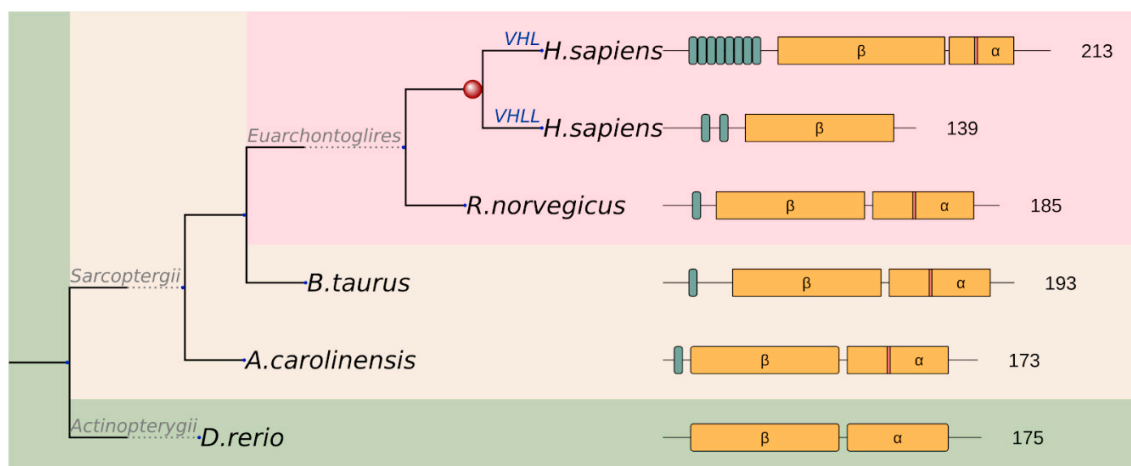
### 3.4.2. VHL

VHL, the Von Hippel-Lindau disease tumor suppressor protein possesses an E3 ligase activity. It plays a key role in cellular oxygen sensing by targeting hypoxia-inducible factors for ubiquitylation and proteasomal degradation. To carry out its function, VHL forms a complex with elongin B, elongin C, and cullin-2 and the RING finger protein RBX1 [28,29]. VHL has an  $\alpha$ -domain (also known as the VHL-box, residues 155 to 192) that forms the principal contacts with elongin C, and a larger  $\beta$ -domain (residues 63 to 154) that directly binds the proline hydroxylated substrate, HIF1 $\alpha$ . The positions mutated across various types of cancers cover a large part of the protein, including both the  $\alpha$  and  $\beta$  domains. While these regions form a well-defined structure in complex with elongin B, elongin C, and cullin-2, they are disordered in isolation and rapidly degraded [30].

The VHL gene emerged de novo at the level of Eumetazoa together with HIF $\alpha$  and PHD, the other key components of the hypoxia regulatory pathway. However, more recently, the gene underwent various evolutionary events. The VHL gene showed slightly higher evolutionary variations compared with other cancer risk regions (Figure 2). Some positions, including K171, showed signs of positive selection at the level of Sarcopterygii, which might implicate the occurrence of an important evolutionary event. It was shown that the SUMO E3 ligase PIASy interacts with VHL and induces VHL SUMOylation on lysine residue 171 [31]. VHL also undergoes ubiquitination on K171 (and K196), which is blocked by PIASy. In the proposed model of the dynamic regulation of VHL, the interaction

of VHL with PIASy results in VHL nuclear localization, SUMOylation, and stability for blocking ubiquitylation of VHL. Meanwhile, PIASy dissociation with VHL or attenuation of VHL SUMOylation facilitates VHL nuclear export, ubiquitylation, and instability. This dynamic process of VHL with reversible modification acts in concert to inhibit HIF1 $\alpha$  [32].

A novel acidic repeat region appeared at the N-terminal region of the protein at the level of Sarcopterygii, and this region underwent further repeat expansion in the lineage leading up to humans (Figure 6). These GxEEEx repeats are generally thought to confer additional regulation to the long isoform of VHL (translated from the first methionine), with a number of putative (USP7) or experimentally detected (p14ARF) interactors [33]. Although poorly studied, this repetitive region also seems to harbour casein kinase 2 (CK2) phosphorylation as well as proteolytic cleavage sites, regulating VHL half-life (consistent with a deubiquitinase, such as USP7 binding role) [34]. As a result of a recent gene duplication, the human genome even encodes a VHL-like protein (VHLL), which has lost the C-terminal segment including the  $\alpha$  domain. Consequently, VHLL cannot nucleate the multiprotein E3 ubiquitin ligase complex. Instead, it was suggested that VHLL functions as a dominant-negative VHL to serve as a protector of HIF1 $\alpha$  [35]. This example demonstrates that, while the basic cancer risk region remains largely unchanged during evolution, additional regulatory mechanisms can emerge to further fine-tune the function of the protein.



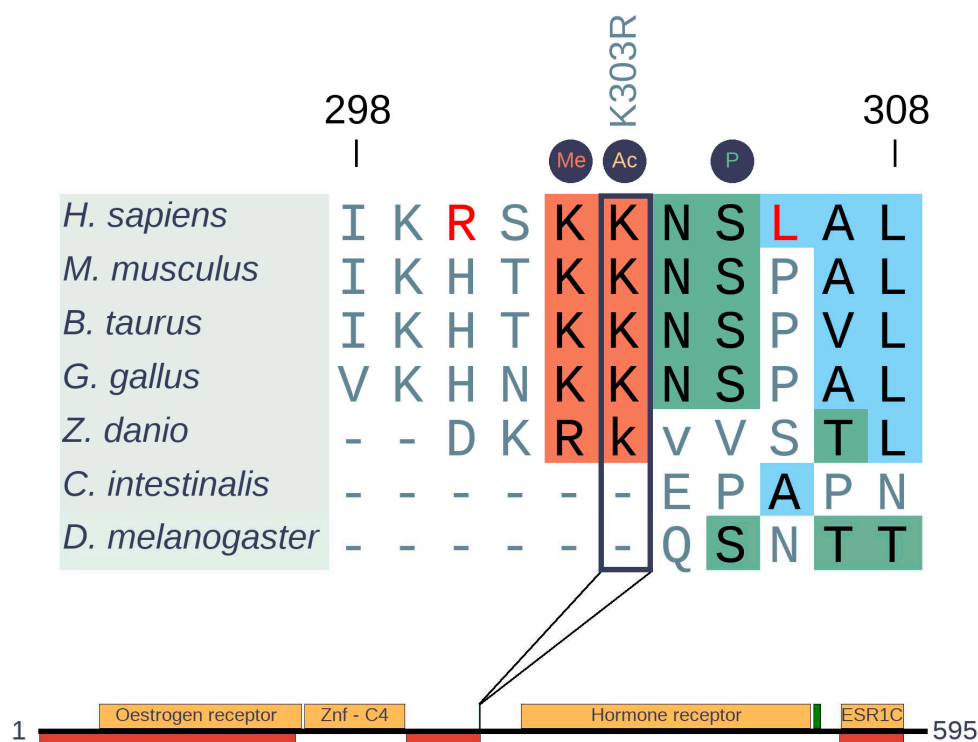
**Figure 6.** Schematic representation of the evolutionary scenario of the VHL family and the functional units of the members. Repeat units in varying numbers and the  $\alpha$  and  $\beta$  core domains are depicted by green and yellow boxes, respectively. Red stripe in the  $\alpha$  domain of human VHL indicates K171 identified to emerge by positive selection on the Sarcopterygii branch (mapped K171 to other Sarcopterygii are also indicated by red stripes).

### 3.4.3. ESR1

Estrogen receptor 1 (ESR1) is a member of the nuclear hormone receptor family with eumetazoan origin. The most common mutation in both primary and tamoxifen therapy associated samples corresponds to a single mutation (K303R). This single site emerged more recently (Figure 7) and is located in a rather complex switch region adjacent to the ligand-binding domain (Figure S2). The highly mutated K303 of ESR1 (more than 200 K303R missense mutations are seen in COSMIC) is a part of a motif-based molecular switch region involving several mutually exclusive PTMs. At positions 302, 303, and 305, methylation by SET7/9, acetylation by p300, and phosphorylation by PKA or PAK1 were observed in previous studies, respectively [36–40]. Our results show that this region is conserved only in Sarcopterygii, which indicates a relatively young evolutionary origin of the switching mechanism. However, while the methylation and acetylation sites are well conserved, the phosphorylation motif appears to be specific only to *H. sapiens*. We came to this conclusion because R300 and K302 as well as L306 are required for the protein kinase A (PKA) phosphorylation consensus and the oncogenic

mutation K303R is expected to turn this region into an even better PKA substrate [41,42]. Curiously, these residues are not found in any other mammal, supposing species specific adaptive changes.

Comparison of substitutions and polymorphic sites is a powerful approach to identify specific changes in a pair of closely related species, like *H. sapiens* and chimpanzee. Relying on this approach, 198 of 9785 analyzed genes were identified to show human-specific changes including ESR1 [19]. In ESR1, there are three more changes besides R300 and K306 (L44, Q502, S559) between *H. sapiens* and chimp that are also thought to be adaptive substitutions according to the MK test. Phosphorylation of S559 was experimentally identified, suggesting this residue is also a *H. sapiens* specific PTM [43,44], but there is no specific data in the literature about the biological function of L44 and Q502. Yet, we know that phosphorylation of S305 allows the increase of estrogen sensitivity by external stimuli other than steroids, and permits ESR1 activity even when the canonical estrogen effect is completely blocked by tamoxifen [40,42]. In mice, ESR1 activity is essential for the estrogen effect and normal estrous episodes [45,46]. Although we lack information, we theorize that this human-specific signaling crosstalk might somehow be connected to the continuous menstrual cycle of *H. sapiens* (quite unusual among mammals), or some other human-specific reproductive adaptation.



**Figure 7.** Insertion-free sequence alignment of estrogen receptor 1 (ESR1) orthologs and domain structure of human ESR1. The alignment generated with MAFFT [15] represents the cancer region with sites of post-translational modifications. Borders of non-depicted insertion of zebrafish are indicated by lower case letters. The highly mutated position (K303R) is highlighted by a rectangle. PTM sites are indicated by circles above the alignment. *H. sapiens* specific changes are colored in red. Domains are depicted in yellow, disordered regions are depicted by red boxes, while the green boxes indicate the cancer risk regions.

#### 4. Discussion

In our study, we aimed to estimate the evolutionary origin of disordered regions that are specifically targeted in cancer. Intrinsically disordered protein regions play essential roles in a wide-range of biological processes and can function as linear motifs, linkers, or other intrinsically disordered domain-sized segments [47]. They are integral parts of many cancer associated proteins and, in a smaller number of cases, they can also be the direct targets of cancer driving mutations. In general,



IDRs are believed to be of more recent evolutionary origin, and exhibit higher rates of evolutionary variations compared with that of folded globular domains [9]. However, this is not what we see in the case of disordered cancer genes. Instead, we observed that cancer-targeted disordered regions are extremely conserved with deep evolutionary origins, which underlines their critical function. The two main ages for emergence of disordered cancer genes can be linked to unicellular organisms and the emergence of multicellularity, in agreement with the result of phylostratigraphic tracking of cancer genes in general [3].

One of the most unexpected findings of our study is the examples of disordered cancer genes that can be traced back to unicellular organisms. Mechanistically, the group of cancer genes that emerged in unicellular organisms were suggested to play a caretaker role and contribute to tumorigenesis by increasing mutation rates and genome instability. In contrast, cancer genes that emerged at the level of multicellularity were suggested to typically have a gatekeeper function and promote tumour progression directly by changing cell differentiation, growth, and death rates [48]. MLH1 is one of the best characterized examples of a gene with a caretaker function [49]. It is involved in mismatch repair (MMR) of DNA bases that have been misincorporated during DNA replication. Thus, disruptive mutations of MLH1 greatly increase the rate of point mutations in genes and underline various inherited forms of cancer. However, the most commonly seen alterations in patients are located in the flexible internal linker. Mutational studies indicate that this highly conserved segment might not be directly involved in MMR, but likely has an important, currently uncharacterized function. The other ancient examples are also involved in basic cellular processes, however, they are associated with a broader set of functions. HIST1H3B, SMARCB1, and SETBP1 are involved in epigenetic regulation and their mutations can alter gene expression patterns [50,51]. Mutations of EIF1AX and RPS15 are likely to perturb translation events [52,53]. However, SRSF2, which is responsible for orchestrating splicing events, can also have a global influence on cellular states [54]. Therefore, the caretaker function is also a subject of evolution and some of its components emerged as a result of more recent evolutionary events.

A clear novelty of our approach is to focus at the origin of sub-gene elements; that is, regulatory regions, modules, and domains, instead of full genes. The genes can be built around founder genes that have an extremely ancient origin, but their biological function and regulation can change fundamentally during subsequent evolution. In several cases, the origin of the cancer mutated region was substantially more recent than the origin of the gene. Nevertheless, after their emergence, disordered cancer regions were fixated rapidly and showed little variations afterwards. However, their evolution at the gene level was not set in stone and there are several indications that this process continues indefinitely. In several cases, the cancer genes underwent gene duplications, further regulatory regions were added, or fine-tuned by changing some of the less critical positions. We highlighted a fascinating case when such an event occurred when our species, *H. sapiens*, separated from its primate relatives.

In general, the rate of gene duplications is very high (0.01 per gene per million years) over evolution, which provides the source of emergence of evolutionary novelties [55]. According to the general view, paralogs go through a brief period of relaxed selection directly after duplications—this time ensures the acquisition of novelties—and subsequently experience strong purifying selection, preserving the newly developed function. However, our results showed that only a few disordered cancer regions have emerged in a duplication induced manner and the vast majority of disordered cancer regions emerged de novo, independent of duplications. The evolution of disordered regions is better described by the ex-nihilo motif theory, which is based on the rapid disappearance and emergence of linear motifs by the change of only a few residues within a given disordered protein segment [10]. This evolutionary phenomenon is commonly observed in the case of linear motifs, for example, in the case of NFE2L2. This protein carries a pair of crucial linear motifs that have emerged in the ancient eumetazoa, but are not preserved in the most recent duplicates. In an evolutionary biology aspect, our results suggest that the evolution of functional novelties in the case of disordered region mediated functions requires a more complex model.

Exploring the evolutionary origin of cancer genes is an important step to understand how this disease can emerge. This knowledge can also have important implications of how their regulatory networks are disrupted during tumorigenesis and can be incorporated into developing improved treatment options [56]. In this work, we focused on a subset of cancer genes that belong to the class of intrinsic disordered proteins, which rely on their inherent flexibility to carry out their important functions. While the selected examples represent only a small subset of cancer genes, they are highly relevant for several specific cancer types [8]. In general, disordered proteins are evolutionarily more variable compared with globular proteins, however, the disordered cancer risk regions showed remarkable conservation with ancient evolutionary origin, highlighting their importance in core biological processes. Nevertheless, we found several examples where the region specifically targeted by cancer mutations emerged at a later stage compared with the origin of the gene family. Our results highlight the importance of taking into account the complex modular architecture of cancer genes in order to get a more complete understanding of their evolutionary origin.

**Supplementary Materials:** The following are available online at <http://www.mdpi.com/2218-273X/10/8/1115/s1>, Figure S1: Sequence alignment of BCL2 cancer region. Figure S2: Schematic representation of the evolutionary scenario and functional units of the ESR1 and ESR2 proteins. Supplementary Materials 1: Evolutionary origins of selected cases.

**Author Contributions:** Conceptualization, M.P., A.Z. and Z.D.; Data curation, M.P.; Formal analysis, M.P.; Funding acquisition, Z.D.; Investigation, M.P., A.Z. and Z.D.; Methodology, M.P.; Supervision, Z.D.; Visualization, M.P.; Writing—original draft, M.P., A.Z. and Z.D. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the “FIEK” grant from the National Research, Development, and Innovation Office (FIEK16-1-2016-0005) and the ELTE Thematic Excellence Programme (ED-18-1-2019-003) supported by the Hungarian Ministry for Innovation and Technology.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Jacob, F. Evolution and tinkering. *Science* **1977**, *196*, 1161–1166. [[CrossRef](#)] [[PubMed](#)]
2. Kinzler, K.W.; Vogelstein, B. Cancer-susceptibility genes. Gatekeepers and caretakers. *Nature* **1997**, *386*, 761–763. [[CrossRef](#)] [[PubMed](#)]
3. Domazet-Lošo, T.; Tautz, D. Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. *BMC Biol.* **2010**, *8*, 66. [[CrossRef](#)]
4. Domazet-Lošo, T.; Tautz, D. An ancient evolutionary origin of genes associated with human genetic diseases. *Mol. Biol. Evol.* **2008**, *25*, 2699–2707. [[CrossRef](#)] [[PubMed](#)]
5. Dickerson, J.E.; Robertson, D.L. On the origins of mendelian disease genes in man: The impact of gene duplication. *Mol. Biol. Evol.* **2012**, *29*, 2284. [[CrossRef](#)]
6. Bailey, M.H.; Tokheim, C.; Porta-Pardo, E.; Sengupta, S.; Bertrand, D.; Weerasinghe, A.; Colaprico, A.; Wendl, M.C.; Kim, J.; Reardon, B.; et al. Comprehensive characterization of cancer driver genes and mutations. *Cell* **2018**, *173*, 371–385.e18. [[CrossRef](#)]
7. Mészáros, B.; Zeke, A.; Reményi, A.; Simon, I.; Dosztányi, Z. Systematic analysis of somatic mutations driving cancer: Uncovering functional protein regions in disease development. *Biol. Direct* **2016**, *11*, 23. [[CrossRef](#)]
8. Mészáros, B.; Hajdu-Soltész, B.; Zeke, A.; Dosztányi, Z. Intrinsically disordered protein mutations can drive cancer and their targeted interference extends therapeutic options. *Bioinform. bioRxiv* **2020**, 2443. [[CrossRef](#)]
9. Brown, C.J.; Johnson, A.K.; Dunker, A.K.; Daughdrill, G.W. Evolution and disorder. *Curr. Opin. Struct. Biol.* **2011**, *21*, 441–446. [[CrossRef](#)]
10. Davey, N.E.; Cyert, M.S.; Moses, A.M. Short linear motifs—Ex nihilo evolution of protein regulation. *Cell Commun. Signal.* **2015**, *13*, 43. [[CrossRef](#)]
11. Sondka, Z.; Bamford, S.; Cole, C.G.; Ward, S.A.; Dunham, I.; Forbes, S.A. The COSMIC cancer gene census: Describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **2018**, *18*, 696–705. [[CrossRef](#)] [[PubMed](#)]

12. Flicek, P.; Amode, M.R.; Barrell, D.; Beal, K.; Brent, S.; Carvalho-Silva, D.; Clapham, P.; Coates, G.; Fairley, S.; Fitzgerald, S.; et al. Ensembl 2012. *Nucleic Acids Res.* **2011**, *40*, D84–D90. [[CrossRef](#)]
13. Herrero, J.; Muffato, M.; Beal, K.; Fitzgerald, S.; Gordon, L.; Pignatelli, M.; Vilella, A.J.; Searle, S.M.J.; Amode, R.; Brent, S.; et al. Ensembl comparative genomics resources. *Database* **2016**, *2016*, bav096. [[CrossRef](#)] [[PubMed](#)]
14. Liebeskind, B.J.; McWhite, C.D.; Marcotte, E.M. Towards consensus gene ages. *Genome Biol. Evol.* **2016**, *8*, 1812–1823. [[CrossRef](#)] [[PubMed](#)]
15. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780. [[CrossRef](#)]
16. Eddy, S.R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **2011**, *7*, e1002195. [[CrossRef](#)]
17. Moretti, S.; Laurenczy, B.; Gharib, W.; Castella, B.; Kuzniar, A.; Schabauer, H.; Studer, R.A.; Valle, M.; Salamin, N.; Stockinger, H.; et al. Selectome update: Quality control and computational improvements to a database of positive selection. *Nucleic Acids Res.* **2013**, *42*, D917–D921. [[CrossRef](#)]
18. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* **2007**, *24*, 1586–1591. [[CrossRef](#)]
19. Gayà-Vidal, M.; Alba, M.M. Uncovering adaptive evolution in the human lineage. *BMC Genom.* **2014**, *15*, 1–12. [[CrossRef](#)]
20. Berg, T.K.V.D.; Yoder, J.A.; Litman, G. On the origins of adaptive immunity: Innate immune receptors join the tale. *Trends Immunol.* **2004**, *25*, 11–16. [[CrossRef](#)]
21. Wu, G.; Xu, G.; Schulman, B.A.; Jeffrey, P.D.; Harper, J.W.; Pavletich, N.P. Structure of a beta-TrCP1-Skp1-beta-catenin complex: Destruction motif binding and lysine specificity of the SCF(beta-TrCP1) ubiquitin ligase. *Mol. Cell* **2003**, *11*, 1445–1456. [[CrossRef](#)]
22. Shcherbakova, P.V.; Kunkel, T.A. Mutator Phenotypes conferred by MLH1 Overexpression and by Heterozygosity for mlh1 Mutations. *Mol. Cell. Biol.* **1999**, *19*, 3177–3183. [[CrossRef](#)] [[PubMed](#)]
23. Takahashi, M.; Shimodaira, H.; Andreutti-Zaugg, C.; Iggo, R.; Kolodner, R.D.; Ishioka, C. Functional analysis of human MLH1 variants using yeast and in vitro mismatch repair Assays. *Cancer Res.* **2007**, *67*, 4595–4604. [[CrossRef](#)] [[PubMed](#)]
24. Akagi, Ohsawa, T.; Sahara, T.; Muramatsu, S.; Nishimura, Y.; Yathuoka, T.; Tanaka, Y.; Yamaguchi, K.; Ishida, H. Colorectal cancer susceptibility associated with the hMLH1 V384D variant. *Mol. Med. Rep.* **2009**, *2*, 887–891. [[CrossRef](#)]
25. Lee, S.E.; Lee, H.S.; Kim, K.-Y.; Park, J.-H.; Roh, H.; Park, H.Y.; Kim, W.S. High prevalence of the MLH1 V384D germline mutation in patients with HER2-positive luminal B breast cancer. *Sci. Rep.* **2019**, *9*, 10966. [[CrossRef](#)] [[PubMed](#)]
26. Gueneau, E.; Dhérin, C.; Legrand, P.; Tellier-Lebègue, C.; Gilquin, B.; Bonnesoeur, P.; Londino, F.; Quemener, C.; Le Du, M.-H.; A Márquez, J.; et al. Structure of the MutL $\alpha$  C-terminal domain reveals how Mlh1 contributes to Pms1 endonuclease site. *Nat. Struct. Mol. Biol.* **2013**, *20*, 461–468. [[CrossRef](#)]
27. Kim, Y.; Furman, C.M.; Manhart, C.M.; Alani, E.; Finkelstein, I.J. Intrinsically disordered regions regulate both catalytic and non-catalytic activities of the MutL $\alpha$  mismatch repair complex. *Nucleic Acids Res.* **2018**, *47*, 1823–1835. [[CrossRef](#)]
28. Kamura, T.; Maenaka, K.; Kotoshiba, S.; Matsumoto, M.; Kohda, D.; Conaway, R.C.; Conaway, J.W.; Nakayama, K.I. VHL-box and SOCS-box domains determine binding specificity for Cul2-Rbx1 and Cul5-Rbx2 modules of ubiquitin ligases. *Genes Dev.* **2004**, *18*, 3055–3065. [[CrossRef](#)]
29. Cardote, T.A.; Gadd, M.S.; Ciulli, A. Crystal structure of the Cul2-Rbx1-EloBC-VHL Ubiquitin Ligase complex. *Structure* **2017**, *25*, 901–911.e3. [[CrossRef](#)]
30. Sutovsky, H.; Gazit, E. The von Hippel-Lindau tumor suppressor protein is a molten globule under native conditions. *J. Biol. Chem.* **2004**, *279*, 17190–17196. [[CrossRef](#)]
31. Cai, Q.; Verma, S.C.; Kumar, P.; Ma, M.; Robertson, E.S. Hypoxia Inactivates the VHL tumor suppressor through PIASy-Mediated SUMO modification. *PLoS ONE* **2010**, *5*, e9720. [[CrossRef](#)] [[PubMed](#)]
32. Cai, Q.; Robertson, E.S. Ubiquitin/SUMO modification regulates VHL protein stability and nucleocytoplasmic localization. *PLoS ONE* **2010**, *5*, e12636. [[CrossRef](#)] [[PubMed](#)]
33. Minervini, G.; Mazzotta, G.; Masiero, A.; Sartori, E.; Corrà, S.; Potenza, E.; Costa, R.; Tosatto, S.C.E. Isoform-specific interactions of the von Hippel-Lindau tumor suppressor protein. *Sci. Rep.* **2015**, *5*, 12605. [[CrossRef](#)] [[PubMed](#)]

34. German, P.; Bai, S.; Liu, X.-D.; Sun, M.; Zhou, L.; Kalra, S.; Zhang, X.; Minelli, R.; Scott, K.L.; Mills, G.B.; et al. Phosphorylation-dependent cleavage regulates von Hippel Lindau proteostasis and function. *Oncogene* **2016**, *35*, 4973–4980. [[CrossRef](#)] [[PubMed](#)]
35. Qi, H.; Gervais, M.L.; Li, W.; DeCaprio, J.A.; Challis, J.R.G.; Ohh, M. Molecular cloning and characterization of the von Hippel-Lindau-like protein. *Mol. Cancer Res.* **2004**, *2*, 43–52.
36. Dhayalan, A.; Kudithipudi, S.; Rathert, P.; Jeltsch, A. Specificity analysis-based identification of new methylation targets of the SET7/9 protein lysine methyltransferase. *Chem. Biol.* **2011**, *18*, 111–120. [[CrossRef](#)]
37. Wang, C.; Fu, M.; Angeletti, R.H.; Siconolfi-Baez, L.; Reutens, A.T.; Albanese, C.; Lisanti, M.P.; Katzenellenbogen, B.S.; Kato, S.; Hopp, T.; et al. Direct Acetylation of the Estrogen receptor  $\alpha$  hinge region by p300 regulates transactivation and hormone sensitivity. *J. Biol. Chem.* **2001**, *276*, 18375–18383. [[CrossRef](#)]
38. Wang, R.-A.; Mazumdar, A.; Vadlamudi, R.K.; Kumar, R. P21-activated kinase-1 phosphorylates and transactivates estrogen receptor- $\alpha$  and promotes hyperplasia in mammary epithelium. *EMBO J.* **2002**, *21*, 5437–5447. [[CrossRef](#)]
39. Michalides, R.; Griekspoor, A.; Balkenende, A.; Verwoerd, D.; Janssen, L.; Jalink, K.; Floore, A.; Velds, A.; vant Veer, L.; Neeffjes, J. Tamoxifen resistance by a conformational arrest of the estrogen receptor  $\alpha$  after PKA activation in breast cancer. *Cancer Cell* **2004**, *5*, 597–605. [[CrossRef](#)]
40. Cui, Y.; Zhang, M.; Pestell, R.; Curran, E.M.; Welshons, W.V.; Fuqua, S.A.W. Phosphorylation of estrogen receptor  $\alpha$  blocks its Acetylation and regulates estrogen sensitivity. *Cancer Res.* **2004**, *64*, 9199–9208. [[CrossRef](#)]
41. Rust, H.L.; Thompson, P.R. Kinase consensus sequences: A breeding ground for crosstalk. *ACS Chem. Biol.* **2011**, *6*, 881–892. [[CrossRef](#)] [[PubMed](#)]
42. De Leeuw, R.; Flach, K.; Toaldo, C.B.; Alexi, X.; Canisius, S.; Neeffjes, J.; Michalides, R.; Zwart, W. PKA phosphorylation redirects ER $\alpha$  to promoters of a unique gene set to induce tamoxifen resistance. *Oncogene* **2012**, *32*, 3543–3551. [[CrossRef](#)] [[PubMed](#)]
43. Atsriku, C.; Britton, D.J.; Held, J.M.; Schilling, B.; Scott, G.K.; Gibson, B.W.; Benz, C.C.; Baldwin, M.A. Systematic mapping of posttranslational modifications in human estrogen receptor- $\alpha$  with emphasis on novel phosphorylation sites. *Mol. Cell. Proteom.* **2008**, *8*, 467–480. [[CrossRef](#)] [[PubMed](#)]
44. Williams, C.C.; Basu, A.; El-Gharbawy, A.; Carrier, L.; Smith, C.L.; Rowan, B.G. Identification of four novel phosphorylation sites in estrogen receptor  $\alpha$ : Impact on receptor-dependent gene expression and phosphorylation by protein kinase CK2. *BMC Biochem.* **2009**, *10*, 36. [[CrossRef](#)]
45. Walker, V.R.; Korach, K. Estrogen receptor knockout mice as a model for endocrine research. *ILAR J.* **2004**, *45*, 455–461. [[CrossRef](#)] [[PubMed](#)]
46. Porteous, R.; Herbison, A.E. Genetic deletion of Esr1 in the mouse preoptic area disrupts the LH surge and estrous cyclicity. *Endocrinology* **2019**, *160*, 1821–1829. [[CrossRef](#)]
47. Van Der Lee, R.; Buljan, M.; Lang, B.; Weatheritt, R.J.; Daughdrill, G.W.; Dunker, A.K.; Fuxreiter, M.; Gough, J.; Gsponer, J.; Jones, D.T.; et al. Classification of intrinsically disordered regions and proteins. *Chem. Rev.* **2014**, *114*, 6589–6631. [[CrossRef](#)] [[PubMed](#)]
48. Lengauer, C.; Kinzler, K.W.; Vogelstein, B. Genetic instabilities in human cancers. *Nature* **1998**, *396*, 643–649. [[CrossRef](#)]
49. Ellison, A.R.; Lofing, J.; Bitter, G.A. Human MutL homolog (MLH1) function in DNA mismatch repair: A prospective screen for missense mutations in the ATPase domain. *Nucleic Acids Res.* **2004**, *32*, 5321–5338. [[CrossRef](#)]
50. Duchatel, R.J.; Jackson, E.R.; Alvaro, F.; Nixon, B.; Hondermarck, H.; Dun, M.D. Signal transduction in diffuse intrinsic Pontine Glioma. *Proteomics* **2019**, *19*, e1800479. [[CrossRef](#)]
51. Piazza, R.; Magistroni, V.; Redaelli, S.; Mauri, M.; Massimino, L.; Sessa, A.; Peronaci, M.; Lalowski, M.M.; Soliymani, R.; Mezzatesta, C.; et al. SETBP1 induces transcription of a network of development genes by acting as an epigenetic hub. *Nat. Commun.* **2018**, *9*, 2192. [[CrossRef](#)] [[PubMed](#)]
52. Martin-Marcos, P.; Zhou, F.; Karunasiri, C.; Zhang, F.; Dong, J.; Nanda, J.; Kulkarni, S.D.; Sen, N.D.; Tamame, M.; Zeschnigk, M.; et al. eIF1A residues implicated in cancer stabilize translation preinitiation complexes and favor suboptimal initiation sites in yeast. *eLife* **2017**, *6*, e31250. [[CrossRef](#)] [[PubMed](#)]

53. Bretones, G.; Álvarez, M.G.; Arango, J.R.; Rodríguez, D.; Nadeu, F.; Prado, M.A.; Valdés-Mas, R.; Puente, D.A.; Paulo, J.A.; Delgado, J.; et al. Altered patterns of global protein synthesis and translational fidelity in RPS15-mutated chronic lymphocytic leukemia. *Blood* **2018**, *132*, 2375–2388. [[CrossRef](#)]
54. Masaki, S.; Ikeda, S.; Hata, A.; Shiozawa, Y.; Kon, A.; Ogawa, S.; Suzuki, K.; Hakuno, F.; Takahashi, S.-I.; Kataoka, N. Myelodysplastic syndrome-associated SRSF2 mutations cause splicing changes by altering binding motif sequences. *Front. Genet.* **2019**, *10*, 338. [[CrossRef](#)] [[PubMed](#)]
55. Assis, R.; Bachtrog, D. Rapid divergence and diversification of mammalian duplicate gene functions. *BMC Evol. Biol.* **2015**, *15*, 138. [[CrossRef](#)]
56. Trigos, A.S.; Pearson, R.B.; Papenfuss, A.T.; Goode, D. How the evolution of multicellularity set the stage for cancer. *Br. J. Cancer* **2018**, *118*, 145–152. [[CrossRef](#)] [[PubMed](#)]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).