

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

Gábor Erdős, Mátyás Pajkos and Zsuzsanna Dosztányi ^{*}

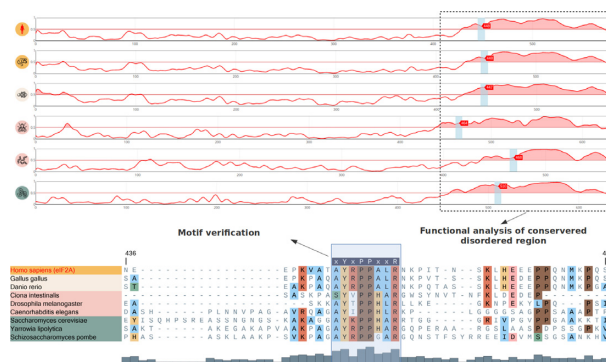
Department of Biochemistry, Eötvös Loránd University, Pázmány Péter stny 1/c, Budapest H-1117, Hungary

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.

*To whom correspondence should be addressed. Tel: +36 1372 2500/8537; Email: 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 analyses 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 20×20 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.

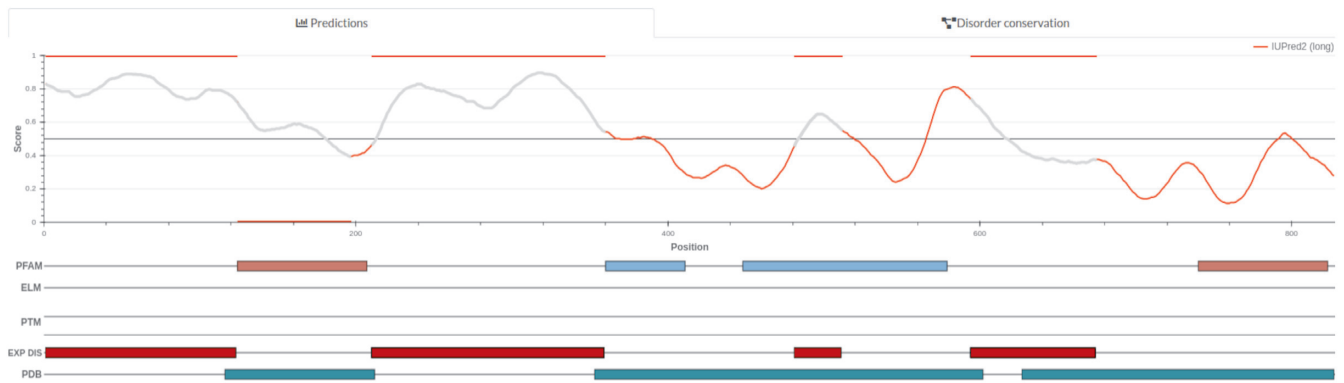


Figure 1. The output of IUPred3 for the repressor-activator protein 1 of *Saccharomyces cerevisiae*. The strong smoothing option was used to generate this plot. At the upper part of the figure the disordered and ordered unambiguous experimentally verified protein regions are marked by red lines at the 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) 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., Dosztanyi, 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., Kryshchak, A., Mout, 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., 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.
 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.