Predicting Protein Conformational Disorder and Disordered Binding Sites

Ketty Tamburrini, Giulia Pesce, Juliet Nilsson, Frank Gondelaud, Andrey Kajava, Jean-Guy Berrin, Sonia Longhi

▶ **To cite this version:**

Ketty Tamburrini, Giulia Pesce, Juliet Nilsson, Frank Gondelaud, Andrey Kajava, et al.. Predicting Protein Conformational Disorder and Disordered Binding Sites. Data Mining Techniques for the Life Sciences, 2449, Springer US, pp.95-147, 2022, Methods in Molecular Biology, 10.1007/978-1-0716-2095-3_4 . hal-03834097

HAL Id: hal-03834097
https://hal.science/hal-03834097v1

Submitted on 28 Oct 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Predicting protein conformational disorder and disordered binding sites

Ketty Tamburrini[1,2$], Giulia Pesce[1$], Juliet Nilsson[1], Frank Gondelaud[1], Andrey V. Kajava[3], Jean-Guy Berrin[2] and Sonia Longhi[1*]

[1] Aix Marseille Univ, CNRS, Architecture et Fonction des Macromolécules Biologiques, AFMB, UMR 7257, Marseille, France

[2] INRAE, Aix Marseille Univ, Biodiversité et Biotechnologie Fongiques (BBF), UMR 1163, 13288 Marseille, France

[3] Centre de Recherche en Biologie cellulaire de Montpellier, UMR 5237, CNRS, Université Montpellier ; Institut de Biologie Computationnelle, Université Montpellier, Montpellier, France

$ equal contribution to the work

To whom correspondence should be addressed

Sonia Longhi

AFMB, UMR 7257, CNRS and Aix-Marseille University

163, avenue de Luminy, Case 932, 13288 Marseille Cedex 09, France

Tel: (33) 4 91 82 55 80; Fax: (33) 4 91 26 67 20

E-mail sonia.longhi@univ-amu.fr

Summary

In the last two decades it has become increasingly evident that a large number of proteins adopt either a fully or a partially disordered conformation. Intrinsically disordered proteins are ubiquitous proteins that fulfill essential biological functions while lacking a stable 3D structure. Their conformational heterogeneity is encoded by the amino acid sequence, thereby allowing intrinsically disordered proteins or regions to be recognized based on their sequence properties. The identification of disordered regions facilitates the functional annotation of proteins and is instrumental for delineating boundaries of protein domains amenable to crystallization. This chapter focuses on the methods currently employed for predicting protein disorder and identifying intrinsically disordered binding sites.

Keywords: intrinsic disorder, intrinsically disordered proteins, intrinsically disordered regions, intrinsically disordered binding sites, MoREs, MoRFs, induced folding, prediction methods and tools, disorder databases and metaservers.

Running head: Protein conformational disorder

1. Introduction

During the last two decades there has been an increasing amount of experimental and computational evidence pointing out the abundance of protein disorder within the protein realm[1]. Indeed, the frequency and length of disordered regions increase with increasing organism complexity, with as much as one third of eukaryotic proteins containing long (i.e. ≥ 30 residues) intrinsically disordered regions (IDRs) [2] and 12% of them being fully disordered [3]. Intrinsically disordered proteins (IDPs) are functional proteins that fulfill essential biological functions while lacking constant or highly populated secondary and tertiary structure under physiological conditions [4]. Although there are IDPs that carry out their function while remaining disordered all the time (e.g. entropic chains), many of them undergo a disorder-to-order transition upon binding to their physiological partner(s), a process termed induced folding or folding coupled to binding [5].

The functional relevance of disorder resides in an increased plasticity that enables the binding of numerous, structurally distinct targets. Accordingly, intrinsic disorder is a distinctive and common feature of "hub" proteins, with disorder serving as a determinant of protein promiscuity [6]. As such, most IDPs/IDRs are involved in functions that imply multiple partner interactions, such as molecular recognition, molecular assembly, cell cycle regulation, signal transduction and transcription (for reviews on IDPs see [7-9]). Beyond this role in regulation and signaling, it has been recently recognized that intrinsic disorder also plays a critical role in "liquid–liquid phase separation" (LLPS) or condensation phenomena, which drive the formation of membrane-less organelles (MLOs) [10-14]. These biological condensates play a critical role in the spatio-temporal organization of the cell, where they exert a multitude of key biological functions, ranging from transcriptional regulation and silencing to control of signal transduction networks. Their dysfunction is tied to a number of pathological states, including age-related neurological disorders [10-14].

The identification of disordered regions has a practical interest as it facilitates the functional annotation of proteins [15] and is instrumental for delineating protein domains amenable to crystallization [16-18].

Statistical analyses showed that the amino acid sequences of IDRs are significantly different from those of ordered proteins, thus allowing IDRs to be predicted with a rather good accuracy. Specifically, IDRs *(i)* have a biased amino acid composition, being enriched in G, S, P and depleted in W, F, I, Y, V, L, *(ii)* have a low secondary structure content, *(iii)* tend to have a low sequence complexity, *(iv)* are on average much more variable in orthologous and paralogous proteins than ordered ones being more tolerant to substitutions due to the lack of structural constraints.

Based on these peculiar sequence features, a number of disorder predictors have been developed (for reviews see [16,19-22,18,23-26]. As a growing number of disorder predictors have started to become available, it has become increasingly clear that predictions benefit from the use of different predictors [27]. Moreover it was shown that since different disorder predictors are based on different definitions of disorder, the combination of several predictions reinforce the reliability of the overall predictions on a specific position or region [28,29]. This is the main reason for developing metapredictors that help users to deal with the growing number of available disorder predictors and allow combining the results provided by several predictors. Some of these metapredictors also include the prediction of structured regions as a way to improve disorder predictions (i.e., as a way to alleviate ambiguity for regions with dubious state). Recently, the first round of the Critical Assessment of protein Intrinsic Disorder prediction (CAID), which provides a systematic benchmarking for disorder predictors, was published. In this first round, the performance of the main predictors was tested, considering several aspects as the predictions for completely disordered proteins and binding regions [25,30]. The assessment confirmed the

usefulness of predictors, and provided a ranking of the various predictors with details of each one [25].

The pivotal importance of disordered regions in proteins (functional interactions, binding, protein conformation, molecular switch, phase separations...) led to a growing interest of the scientific community for IDRs. Consequently, the number of requests submitted to disorder prediction servers shoot up. This exponential increase in the number of requests and the demanding resources required for predicting disorder (variety of predictors to be used and compared) has forced various research groups to build their own databases dedicated to store annotations and predictions related to IDRs. These databases constitute valuable resources of information that have to be exploited when seeking data on disordered regions into a protein of interest. They gather experimentally assessed information and/or predictions from several disorder predictors thereby fastening the identification of disordered regions. These databases allow fast and easy retrieval of annotated proteins that exhibit sequence similarity *vis-à-vis* a query protein. Although in most cases additional analyses are necessary to achieve a detailed description of the modular organization of a query protein, these databases nevertheless provide useful hints on the possible presence of disordered regions in a protein of interest.

In this chapter, we present a general suggested procedure for disorder prediction based on the combination of various tools for protein disorder prediction.

2. Methods

2.1. Searching databases dedicated to IDPs

We recommend as a first step to check whether the protein of interest or a similar protein exists in publicly available databases dedicated to IDPs. The most efficient way to do this is to use the search engines by sequences that are provided by most of their interfaces.

Obviously, the higher is the level of similarity between the matching sequences from these databases and the query sequence, the more relevant is the information that can be obtained on the query protein.

- A search result with more than 90 % of sequence identity with a sequence from a database that contains experimentally assessed information is the ideal case but will rarely occurs since these databases have still few entries.
    - A similarly high sequence identity with an entry of a database for which annotations are based on predictions will have to be analyzed further: if all the disorder predictions stored are convergent with high confidence (i.e., with high probability) then the results obtained can be considered of sufficiently good quality.
- In all other cases, it will be necessary to gather from these databases all the information that make sense about structured and disordered regions (boundaries) of the matching proteins displaying a reasonable level of similarity, and then to proceed to the next step (3.2) to complement the analysis by further predictions.

In case the search returns distant homologs of the sequence query (note that an E-value below 1.e-11 can be of interest), it is likely that conserved regions and non-conserved regions can be identified, where the former will correspond to structured regions, and the latter have good chances to correspond to disordered regions because of the higher selection pressure exerted on structured regions [31]. Below we list the databases dedicated to IDPs/IDRs.

2.1.1. The Database of Disordered Protein Prediction (D2P2) (http://d2p2.pro) [32] contains (as of May 2021) disorder predictions of 10,429,761 protein sequences from 1765 complete proteomes corresponding to 1,265 species, and their variants generated by nine disorder prediction methods (see 2.2.1): VLXT, VSL2b, PrDOS, PV2, ESpritz-D, ESpritz-X, ESpritz-N, IUPred-L and IUPred-S. D2P2 is also connected to the DisProt and IDEAL databases, two databases that contain experimentally confirmed information about disordered regions (see 2.1.3 and 2.1.4). Also, it is associated with two other databases: SuperFamily2 [33] and PhosphoSitePlus [34]. As by May 2021, D2P2 does not cover all organisms (viral proteomes are not yet included for instance).

D2P2 uses a "Meta" approach by gathering in a single output the data from several predictors and databases dedicated to disordered regions in proteins. An example of D2P2 output is provided in **Figure 1**. Using D2P2 as a preliminary tool to search for disordered regions will speed up the analysis of the query protein.

1. Paste the sequence(s) (FASTAformat as default) of interest in the "Sequences" field of the "Match Amino Sequence" section of the search page and click on the "Find proteins" button.
2. On the result page are displayed the corresponding entries that match 100% of the query sequence(s). On the graphical part of the output, the matching entries from the SuperFamily2, IDEAL and DisProt databases, as well as the predictions of disordered regions from the panel of predictors are aligned. Moving the mouse pointer over the barswill display complementary information such as the positions of the boundaries. If IDEAL or DisProt entries are found, clicking on their representation shapes will lead the user to the corresponding entries in these databases. The bottom part of the graph displays the predicted disorder agreement (corresponding to regions predicted to be disordered by more than 75 % of the predictors) and show additional data such as phosphorylation sites or ANCHOR (see 2.3) binding sites.
3. Below the graphical output, click on the tab entitled "Disorder regions" to get a summary of the predicted disordered regions in the corresponding matching sequence. The left side of the page will display the predicted regions for which at least 75 % of the predictors agreed (that could be taken as a consensus), while on the right part of the page all predictions per predictor will be listed.

Alternatively, in the main search page, the user can also use the second form and enter a free text in the “Search for any phrase including genomes, superfamilies, Gene Ontology, gene names and gene descriptions” of the “Sequence IDs and Free Text” field and click on the “Search” button. Results will be given in the result page in the same format as described above.

2.1.2. MobiDB (http://mobidb.bio.unipd.it/) contains intrinsic disorder annotations for more than 189 millions of entries (covering the entire PDB and DisProt) and predictions from several disorder predictors (see 2.2.1), including ESpritz, IUPred, DisEMBL, GlobPlot, VSL2B, JRONN, among others [35,36].

Although MobiDB is devoid of a BLAST/sequence search engine, it is fully integrated into UniProt thus allowing for each UniProt entry running a MobiDB search. In addition, MobiDB has a search engine by keywords that can also use UniProt search syntax to retrieve an entry.

1. Enter the name of the protein of interest or a more specific UniProt search syntax (e.g., name:"Alpha-synuclein" AND organism:"human").
2. On the result page click on the protein that corresponds the most to the query (the column entitled "Disorder" shows the fraction of residues involved in long disordered

regions).

3. The page displaying the protein annotations shows a summary of the information available. Next, the protein sequence and the regions of disorder consensus are displayed. Move the mouse pointer over the colored shapes to get the positions of the boundaries. In addition, by clicking on "advanced" a list of all predictor results is available with the consensus region on the top of this list. Each tracker of the list has a characteristic square icon indicating the annotation quality. For each prediction, the "sequence viewer" icon enables retrieving the amino acid sequence in which the ordered and disordered regions are colored differently, thereby making it easy to copy/paste regions of interest. Also, selecting each region displays the visualization of the annotations in the PDB structure as well as the contact networks.

2.1.3. DisProt (https://www.disprot.org/) is historically the first database on disorder and the largest publicly available database of disordered proteins whose disorder has been experimentally assessed [37,38]. Although it contains only 1746 entries (as of May 2021), the information therein stored is highly valuable since experimentally assessed.

1. In the Browse section, select BLAST option and paste the sequence in the field (raw format).
2. Check the score of the best BLAST hit on the result page (note that an E-value above 1.e-11 may not hold promise).
3. If the score is consistent, analyze the alignment of the corresponding matching sequence and note the boundaries of matching/mismatching regions.
4. Click on the reference of the entry of interest on top of the result page to display the details of the corresponding entries.
5. Compare the annotations of the selected entry with the boundaries obtained in step 4.

2.1.4. IDEAL (http://www.ideal.force.cs.is.nagoya-u.ac.jp/IDEAL/blast.html) is the second database, in terms of size, dedicated to proteins whose disorder has been experimentally assessed [39]. The total number of proteins in IDEAL is 995 (as of July 2020). The IDEAL interface provides a BLAST engine enabling efficient retrieval of existing annotations related to potential disordered regions in the sequence of interest.

1. Paste the sequence (raw format) in the "Blast Search" field.
2. Check the score of the best BLAST hit on the result page (note that an E-value above 1.e-11 may not hold promise).
3. If the score is consistent, analyze the alignment of the corresponding matching sequence and note the boundaries of matching/mismatching regions.
4. Click on the reference of the entry of interest on top of the result page to display the details of the corresponding entries. The disordered regions of the current entry are displayed in red. Detailed information can be accessed by clicking on the colored shapes.
5. Compare the annotations of the selected entry with the boundaries determined in step 3.

2.1.5. DescribePROT (http://biomine.cs.vcu.edu/servers/DESCRIBEPROT/) provides 13 putative structural and functional properties at the amino acid level for 1,365,946 proteins from 83 complete proteomes (as by May 2021) from model organisms [40]. Among the features provided by the sever are sequence conservation, secondary structure, solvent accessibility, intrinsic disorder, disordered linkers, signal peptides, disordered binding sites (i.e. MoRFs, see 2.3) and

interactions with proteins, DNA and RNAs. The results are made available instantaneously. The predictions can be accessed via an interactive graphical interface that enables simultaneous analysis of multiple descriptors. Results can also be downloaded in png, csv and json format. An example of DescribeProt output is provided in **Figure 2**.

1. Enter either the sequence of the protein of interest in FASTA format or the UniProt accession number or UniProt entry name.
2. Click on "search".
3. The result page displays on the top a summary of the predicted features. Details of each prediction are provided below and can also be accessed by ticking the box adjacent to "click here for details" for the desired prediction. Move the mouse pointer over the bars and the graphs to get the boundaries and the scores.

2.1.6. The **PED (Protein Ensemble Database**) (https://proteinensemble.org/) is a database for the deposition of structural ensembles of IDPs and of denatured proteins based on nuclear magnetic resonance spectroscopy, small-angle X-ray scattering (SAXS) and other data measured in solution [41,42]. Each entry consists of (i) primary experimental data with descriptions of the acquisition methods and algorithms used for the ensemble calculations, and (ii) the structural ensembles consistent with these data, provided as a set of models in a Protein Data Bank format. The total number of entries is 169 as by May 2021. Although PED does not possess a BLAST/sequence search engine, one can search it by using various criteria, such as protein name, gene name, function, UniProt ID, GenBank ID, DisProt ID, ensemble ID and PDB code. If the PED contains data about the protein of interest, this constitutes of course a compelling evidence of disorder (unless the structural ensemble has been obtained under denaturing conditions). In case the PED stores data for a related protein, this should be taken as a strong indication of disorder.

1. Enter the name of the protein of interest or a more specific UniProt search syntax and then click on "search".
2. On the result page, experimental data and structural ensemble can be downloaded.

2.1.7. Although the **PDB (Protein Data Bank**) is a database dedicated to structured proteins and assemblies, it indirectly provides information on disordered regions. Indeed, it allows delineating disordered regions and discarding structured regions from the list of regions potentially considered as disordered. The PDB also provides some information on disorder under the mention "***REMARK465"***, where regions of missing electron density are listed. It should be noted however that these regions are generally short as long regions generally prevent crystallization.

Go to https://www.rcsb.org/search/advanced/sequence.

1. Paste the sequence (raw format) in the "Sequence" field and change the "Display results as" field from "structure" to "polymer entities" to obtain information about E-value, sequence identity and coverage. Finally, click on the search icon.
2. On the result page, check the score of the best blast hit (note that an E-value superior to 1.e-11 probably does not hold promise) and note the boundaries of the matching regions in the selected alignment.
3. Look at the PDB entry pages of interest.
4. Report the boundaries of matching regions in the alignments to the secondary structure annotation of the PDB entry page selected by the user. The regions for which a secondary structure element has been reported cannot be considered as disordered. Regions of missing electron density can be considered as disordered.

2.1.8. Other databases: The databases detailed above are examples of the most complete databases. There are, however, several databases with useful and specific information that worth to be consulted, such as:

- Disordered Binding Site (DIBS), a database that collects complexes between IDRs and ordered protein structures (http://dibs.enzim.ttk.mta.hu/) [43].
- Mutual Folding Induced by Binding (MFIB), a database of protein complexes involving exclusively IDPs with mutual folding (*i.e.* complexes made of IDPs that fold upon binding to each other) (http://mfib.enzim.ttk.mta.hu/) [44].
- FuzDB, a database of fuzzy protein complexes (*i.e.* complexes in which the IDP/IDR retains a considerable amount of residual disorder) (http://protdyn-database.org/index.php) [45].

2.2. Running disorder predictions

Several disorder predictors have been developed, which exploit the sequence bias of disordered proteins. Different types or "flavors" of protein disorder exist [46], differing in the extent (i.e. the amount of residual secondary and/or tertiary structure) and in the length of disorder. Since different predictors rely on different physico-chemical parameters, a given predictor can be more performant in detecting a given feature of a disordered protein. Hence, predictions good enough to decipher the modular organization of a protein can only be obtained by combining various predictors (for examples see [47,17,48,49,16,19,50]).

It is useful to distinguish three kinds of predictors: *(i)* those that have been trained on datasets of disordered proteins, *(ii)* those that have not been trained on any dataset, and *(iii)* metapredictors that blend the results of different predictors. Some predictors use multiple sequence alignments in the computation of their predictions and the most advanced ones include structural information from the PDB when available. As already mentioned, alignments with homologous proteins can provide additional information on potentially disordered regions by themselves since the pressure of selection in disordered regions is not as important as in structured regions. Accordingly, sequence alignments will tend to show lack of conservation within disordered regions.

While predictors trained on datasets of disordered regions identify disordered regions on the basis of the peculiar sequence properties, the others identify disorder as lack of ordered 3D structure. The second group of predictors avoid the shortcomings and biases associated to datasets of disordered regions. Therefore, they are expected to perform better than the former on disordered proteins presently under-represented in training datasets (i.e., fully or mostly disordered proteins).

As the performance of predictors depends on both the type of disorder they predict and the type of disorder against which they were trained, multiple prediction methods need to be combined to improve the accuracy and specificity of disorder predictions [16,19,18,22].

Metapredictors are particularly well suited to speed up the analysis of disorder since they combine the results of several predictors and provide a unified view on the different predictors used. However, since disorder-related databases already return (consensus) predictions from multiple predictors, the added value of running metapredictors mainly resides in the possibility of retrieving additional information from non-redundant predictors (i.e., predictors not already included in the above-described databases) so as to complement the information gathered during the previous step.

2.2.1. Individual disorder predictors

As metapredictors make use of previously developed individual disorder predictors, we have chosen to first provide a short description of the latter along with their philosophy and guidelines on how to run them. Whenever predictors can be downloaded and run in local mode, which is typically very useful for large-scale analyses, this is explicitly indicated. For those predictors that are not endowed with the option of a downloadable stand-alone version, it may be useful to directly contact the program developers to inquire about the possibility of being sent such a version or a script that can be compiled for any machine and operative system.

Predictors trained on datasets of disordered proteins

PreDisorder (http://sysbio.rnet.missouri.edu/predisorder.html) [51] (under group name: MULTICOM-CMFR) was ranked among the best predictors in disorder prediction during CASP8 [52]. The prediction is based on an *ab initio* neural network method (trained on datasets). A PSIPRED profile of the sequence along with the predicted secondary structure and solvent accessibility is fed into a 1D Recursive Neural Network (1D-RNN) that makes the disorder predictions.

1. Enter the e-mail address, the protein name and its sequence in the corresponding field and click on the "Predict" button.
2. Results take several hours to be computed and are sent by e-mail. Results are returned in the form of three lines: the first line displays the amino acid sequence, the second line (dis)order predictions (where residues predicted to be disordered and ordered are tagged with a “D” and “O” character, respectively), and the third line displays the probability of disorder. Residues are considered to be disordered if their disorder probability is above 0.5.

PONDR (Predictor of Natural Disordered Regions) (http://www.pondr.com/), a neural network based on local amino acid composition, flexibility, and other sequence features, was the first predictor to be developed [53]. While in the past, access to PONDR was limited, the predictor is now publicly available. PONDR is available in various versions, namely VLXT, XL1_XT, CAN_XT, VL3-BA and VSL2, plus Cumulative Distribution Function (CDF) (see 2.2.1) and charge-hydropathy plot (CHplot) (see 2.2.1). To overcome the poor accuracy of the first PONDR predictors for short disordered regions (<30 residues), the group of Dunker has developed the VSL2 predictor, which was aimed at providing accurate predictions irrespective of the length of the disordered region [54]. The VSL2 predictor is based on a support vector machine. VSL2 was ranked among the best predictors in CASP7 [55]. VSL2 turned out to behave equally well towards regions of >30 and of <30 residues and to be able to identify short disordered regions that were miss-predicted by the previous PONDR predictors. Notably, VLXT can highlight potential protein-binding regions, indicated by sharp drops in the middle of long disordered regions (see 2.3).

1. Enter the protein name and paste the sequence in raw (or FASTA) format and click on "submit".
2. The result is provided as a plot. The significance threshold above which residues are considered to be disordered is 0.5. Segments composed by more than 40 consecutive disordered residues are highlighted by a thick black line.

DisProt VL2, VL3, VL3H and VL3E. The DisProt server (https://www.dabi.temple.edu/external/disprot/predictor.php) provides access to several predictors, such as VL2, VL3, VL3H, VL3E. Although the server is presently not available due to

a hardware upgrade, we have chosen to describe it in this section with the hope that it will again be made available in a close future.

VL3 uses several features from a previously introduced PONDR VL2 predictor [46], but benefits from optimized predictor models and a slightly larger (152 versus 145) set of disordered proteins that was corrected for mislabeling errors found in the smaller set. The VL3 predictor is based on an ensemble of feed-forward neural networks whose training stage is done using a dataset, obtained from both DisProt and PDB. PONDR VL3H uses the same method as VL3 but it uses homologues of the disordered proteins in the training stage, while PONDR VL3P uses attributes derived from sequence profiles obtained by PSI-BLAST searches [56,54]. Requests are limited to 100 per IP address per day and the maximum length of a query sequence is limited to 5,000 residues. For the VL3E predictor, which results from the combination of VL3H and VL3P, up to 10 queries no longer than 500 residues can be processed per IP address per day. Predictions for VL3E are sent by e-mail upon completion.

1. Chose the predictor to be run among the possible choices.
2. Paste the sequence in raw format, enter the e-mail address and click on "submit".
3. Prediction results are returned online and the plot can be saved (png format) by clicking on it with the mouse right button. The output also provides a table with disorder probabilities *per* residue. The significance threshold above which residues are considered to be disordered is 0.5.

Globplot 2 (http://globplot.embl.de) uses the "Russell/Linding" scale that expresses the propensity for a given amino acid to be in "random coil" or in "regular secondary structure" [57]. It also provides an easy overview of modular organization of large proteins thanks to user-friendly, built-in SMART, PFAM and low complexity predictions. Note that in Globplot outputs, changes of slope often correspond to domain boundaries.

1. Paste the sequence in raw format or enter the SwissProt ID (or AC) in the foreseen field, enter Title (optional) and click on "GlobPlot now".
2. The result page provides a postscript (ps) file that can be downloaded. Below the graph, the amino acid sequence of the protein is given, with disordered residues colored in blue.

The Globplot package can be downloaded by selecting "download" and following the instructions provided. TISEAN (https://www.pks.mpg.de/tisean//) and Biopython (https://biopython.org/wiki/Download) are required for the package to function.

DisEMBL (http://dis.embl.de) is based on a neural network and consists of three separate predictors, trained on separate datasets, that comprise respectively residues within "loops/coils", "hot loops" (DisEMBL Hot loops) (loops with high B-factors – i.e. very mobile from X-ray crystal structure), or that are missing from the PDB X-ray structures (called "Remark 465") [58]. Among these, the only true disorder predictor is Remark 465 (DisEMBL 465), as the two others only predict regions devoid of regular secondary structure. DisEMBL also provides prediction of low sequence complexity (CAST predictor) and aggregation propensity (TANGO predictor).

1. Paste the sequence in raw format or enter the SwissProt ID (or AC) in the foreseen field, enter Title (optional), click on "DisEMBL protein".
2. The result page provides a postcript (ps) file that can be downloaded. Below the graph, the amino acid sequence of the protein is given, with residues in loops and hot loops being colored in blue and red, respectively. Disordered residues, as predicted by Remark 465, are shown in green.

The DisEMBL pipeline package is released under the GPL license. The latest DisEMBL version can be obtained by selecting "download" and following the instruction.

DISOPRED3 is based on support vector machine classifiers trained on PSI-BLAST profiles [59]. It therefore incorporates information from multiple sequence alignments since its inputs are derived from sequence profiles generated by PSI-BLAST. Hence, prediction accuracy is lower if there are few homologues. It is accessible from the PSIPRED4.0 web server page (http://bioinf.cs.ucl.ac.uk/psipred/).

1. Select the "DISOPRED3 (Disopred Prediction)" prediction method. Several predictions can be carried out at the same time (e.g., secondary structure prediction, domain and functions predictions). By default, PSIPRED4.0 method is selected, untick if needed.
2. Paste the raw protein sequence in the "submission details" part. Enter a job name (required) and optionally an e-mail address. Then click on "Submit".
3. Results are displayed as a sequence plot where amino acids are annotated as "Disordered" (blue empty squares) or "Disordered, protein binding" (green empty squares). A DISOPRED plot is also returned and displays the prediction confidence values per amino acids. Amino acids having low confidence values (< 0.5) are not annotated on the sequence plot. The figure can be downloaded in PNG or SVG format. Alternatively, results can be downloaded as text files. The COMB NN format output provides the predicted scores of ordered (marked ".") and disordered (marked "*") residues while the PBDAT file contains score of protein-binding disordered residues (marked "^") as well as disordered residues which not bind proteins (marked "-") and ordered residues (marked "."). Both of the COMB NN and PBDAT files can be opened with classical text-editing software.

Disopred3 can be downloaded from http://bioinf.cs.ucl.ac.uk/software_downloads/.

DisoMine (https://www.bio2byte.be/b2btools/disomine/) is a Neural Network (NN) based predictor of disordered regions [60]. The disorder is not directly predicted from the aminoacidic sequence but is the result of the predicted properties of amino acid residues (as their backbone dynamics). Predictors of these biophysical properties are based on independent experimental datasets with well-defined properties or quantitative information, whereas order/disorder categories are more difficult to capture. From these independently calculated properties, feature vectors of bounded continuous values are assembled, which will replace the amino acid sequence in machine learning methods. DisoMine uses as input the results of four different bioinformatic tools, i.e. PSIPRED (secondary structure predictor), DynaMine for backbone dynamics or side-chain dynamics predictions (see 2.2.1) and EFoldMine (predictor of early folding regions).

1. Paste the target protein sequences in FASTA format and click on "Submit".
2. The results will be displayed after a few seconds/minutes in an interactive graphical form. Click on "Click for more information" to display a table with explanations on how to interpret the graphical results. The obtained results can be downloaded in text format or viewed on a JavaScript applet.

MoreRONN [61] is the successor of RONN [62], which is currently offline. As its predecessor, MoreRONN is based on Bio-Basis Function Neural Network (BBFNN), but the training input library has been clustered based on weak sequence similarities and tested with a ten-fold cross-validation approach. Instead of using the entire protein, MoreRONN uses a sliding window approach of 15 residues, comparing overlapping windows. The information on which MoreRONN relies is a set of curated disordered sequences and it does not need any other

information (e.g., amino-acid characteristics or secondary structure predictions) which makes it a solid and fast predictor. The server can be found at https://moreronn.org/ website.

1. Paste the sequences in FASTA format in the box.
2. Click on “Submit Sequences”
3. Wait until the end of the analysis and scroll down to visualize the results. Although the obtained data are not stored, it is possible to download the text file of the results clicking on “download raw data”.

The MoreRONN source code is available on Github for download; this version can run locally on GNU/Linux without prediction limits.

IsUnstruct is based on a dynamic programming searching for both disordered regions and individual disordered residues within the protein chain. It is based on a physical model in which each residue can be in one state i.e., ordered (fixed) or disordered (free). The model is an approximation of the Ising model, where a penalty for the state change takes the place of the interaction terms between neighbours. Starting from a query sequence in FASTA format the program calculates a profile of probabilities of disorder for each residue [63]. Program and server of IsUnstruct is based on potentials obtained from the Disordered Residues Data Base (DRDB) and on a library of disordered patterns (http://bioinfo.protres.ru/IsUnstruct/pattern.html) [64].

1. Open the server at the web site http://bioinfo.protres.ru/IsUnstruct/ and paste the sequence in FASTA format. In the actual version of the program is possible to run three variants: using the entire library of disordered patterns, only with the HHHH pattern or without any pattern.
2. Start the prediction by clicking on “Pred”.
3. The server gives the results of short and long protocols, plus a graph.

The code source can be downloaded from the website by selecting "get source" in the bottom right-hand corner.

DFLpred (Disorder Flexible Linkers predictor) is the first method able to predict disorder flexible linkers (DFLs). This model combines values of four features empirically selected, using a linear function to generate the output propensities. The results are in the form of a numeric score (between 0 and 1) representing the propensity of each residue to be a DFL, with higher values indicating a higher propensity to be a DFL. For values lower than 0.18 residues are considered NDFL (non-disordered flexible linker) [65]. DFLpred is available as a webserver at http://biomine.cs.vcu.edu/servers/DFLpred/.

1. Paste the protein sequences in FASTA format or upload the file. Provide an e-mail address for the results.
2. Click on “Run DFLpred”
3. At the end of the run you will receive a notification by email and/or in the browser window, if still open. To visualize the results, download the text file. The server generates the propensities binary predictions (DFLs/NDFLs). Three lines will be present in the file: protein ID, aminoacidic sequence, and the propensities score for each residue.

DFLpred is also provided as a stand-alone package written in Java, with details being provided at please https://fanchi.github.io/DFLpred/ for details.

DISpro is available from the ***SCRATCH*** server (http://scratch.proteomics.ics.uci.edu/). It is based on a neural network [66]. It combines sequence profiles obtained by PSI-BLAST, secondary structure predictions and solvent accessibility. This predictor was trained on disordered sequences (i.e., regions of missing atomic coordinates) derived from the PDB.

1. Enter the e-mail address (required), the sequence name (optional), paste the sequence in raw format, select the disorder predictor (i.e., DISpro) and predictions to be run by ticking the appropriate box (e.g., SSpro for Secondary Structure or ABTMpro for Alpha Beta Transmembrane) and click on "Validate".
2. Prediction results are sent by e-mail. Residues predicted to be disordered or ordered are indicated by a "D" or an "O", respectively. Per residue disorder probabilities are also provided.

DisPro is also available as a stand-alone version for the analysis of large data sets, but only for Linux systems. It is also possible to download the SCRATCH suite of 1D predictors which includes protein secondary structure and relative solvent accessibility predictors.

CSpritz (http://protein.bio.unipd.it/cspritz/) takes into account sequence profiles obtained from PSI-BLAST and structure predictions. It is a disorder predictor for high-throughput applications, including NMR mobility.

CSpritz uses two separate predictors based on vector machines trained on different datasets [67]. The training dataset of short disordered regions (less than 45 residues) was derived from a subset of PDB sequences with short regions of missing density, while the training dataset of long regions was derived from both DisProt and from a subset of the PDB (i.e., PDBselect25). This server allows the submission of up to 10 sequences at one time and offers the possibility of choosing between predictions of short or long disordered regions.

1. Paste the sequence in FASTA format, enter the name of the query sequence (optional), and optionally the e-mail address.
2. Chose the data set for disorder prediction (i.e., X-ray, "short", or DisProt "long") and click on "Submit".
3. Prediction results are returned online. At the top of the page the user can find the links to download the results in different formats. Residues predicted to be disordered or ordered are indicated by a red "D" or a black "O", respectively. Disorder statistics are also shown (i.e., percentage of disorder, number of disordered regions of > 30 or of >50 residues in length, length distribution of segments). At the bottom of the page, secondary structure prediction and linear motifs from the ELM database are also given.

ESpritz (http://protein.bio.unipd.it/espritz/) is based on a machine learning method which does not require sliding windows or any complex sources of information (Bi-directional Recursive Neural Networks (BRNN) [68].

1. Enter the e-mail address (optional), the name of the query sequence (optional), and then paste the sequence in raw format.
2. Chose the type of disorder (i.e., X-ray, Disprot, or NMR) and the threshold (i.e., Best Sw or 5% False Positive Rate for conservative predictions), and then click on "Predict".
3. Prediction results are returned online and sent by e-mail. Residues predicted to be disordered are tagged with a D character. It is also possible to get disorder predictions (with disorder probability) in text format by using the corresponding link on the top of the result page.

A stand-alone version of the program that can work on local machines can be obtained after requesting it via the download form at http://old.protein.bio.unipd.it/download/.

PrDOS (http://prdos.hgc.jp/cgi-bin/top.cgi) is composed of two predictors: a predictor based on the local amino acid sequence, and one based on template proteins (or homologous

proteins for which structural information is available) [69]. The first part is implemented using support vector machine algorithm for the position specific score matrix (or profile) of the input sequence. More precisely, a sliding window is used to map individual residues into a feature space. A similar idea has already been used in secondary structure prediction, as in PSIPRED. The second part assumes the conservation of intrinsic disorder in protein families, and is simply implemented using PSI-BLAST and a specific measure of disorder. The final prediction is a combination of the results of the two predictors.

1. Paste the sequence in raw format, enter the sequence name and the e-mail address (optional) and click on "predict".
2. A new page appears where the estimated calculation time is indicated. The user is asked to confirm the submission by clicking on the OK button.
3. On the results page, the plot can be saved as an image (eps or svg format) or as a pdf by clicking on the chosen format in “Download plot in vector format” or as png format by clicking on it with the mouse right button. Residues with disorder probabilities higher than 0.5 are considered to be disordered. Above the graph, the amino acid sequence is shown and disordered residues are shown in red. Disorder probabilities per residue can be retrieved by clicking on the download button (below the graph), which yields an output in the casp or csv format.

The server can accept a Multiple FASTA formatted input limited to sequences. A stand-alone version for large-scale dataset is available upon request. The latter can be addressed by e-mail by clicking on “contact us”.

SPOT-Disorder2 (https://sparks-lab.org/server/spot-disorder2/) elaborates protein evolutionary information derived from the position-specific substitution matrix (PSSM) profile from PSI-BLAST [70], hidden Markov model (HMM) profile from HHblits and predicted structural properties from SPOT-1D [71] to predict protein intrinsic disorder [72]. These input features are computed with an ensemble of deep Squeeze-and-Excitation residual inception and long short-term memory (LSTM) networks. SPOT-Disorder2 and AUCpreD performed as the best available methods in the CAID experiment [30].

1. Copy and paste the protein sequence in FASTA format and click on submit
2. Prediction results are returned online and disordered residues and their disorder probability scores (between 0 and 9) are labeled in red along the sequence. The output can be downloaded in a compressed text file together with HMM/PSSM/SPOT-1D predictions.

The server allows a maximum of ten sequences to be analyzed at a time. For larger analyses, the program can be downloaded by clicking on “Downloads” on the left of the page.

SPOT-Disorder single (https://sparks-lab.org/server/spot-disorder-single/) is a single-sequence based technique developed to improve the prediction of protein sequences with limited evolutionary information, for which profile-based methods are not suitable. It is based on a combination of Convolutional Neural Networks and LSTM networks [73].

1. Copy and paste the protein sequence in FASTA format and click on submit.
2. Prediction results are returned online and disordered residues and the respective disorder probabilities are highlighted in red along the sequence. The output can be downloaded in a compressed text file.

The program is also available as a stand-alone version for the analysis of large data sets. Click on “Downloads” on the left of the page and then choose SPOT-Disorder Single under “Protein Local Structural Prediction” panel.

IDP-Seq2Seq (http://bliulab.net/IDP-Seq2Seq/) is based on sequence-to-sequence learning, a new class of RNN typically used to solve complex language problems. The authors consider the identification of disordered regions as a "syntactic" problem, in which the protein sequence represents the protein language and the predicted structure-based features and sequence-based features identify the "semantic space" where the algorithm maps the intrinsically disordered regions. In order to capture length-dependent characteristics, three predictors for long, short and both long and short prediction of disordered regions were fused in one single predictor [74].

1. Enter the protein sequence in FASTA format or upload your file and click on submit.
2. The results are returned online. Disordered residues are marked by a "1" red character. To download the result in text format, click on "Download" at the top of the page.

AUCpreD (http://raptorx.uchicago.edu/StructurePropertyPred/predict/) predicts disorder using Deep Convolutional Neural Fields (DeepCNF), an integration of Conditional Random Field and Deep Convolutional Neural Network. Instead of considering the disorder state of each residue independently, this method correlates the disorder states of adjacent amino acids. AUCpreD was trained on the UniProt90 CASP9, CASP10 and CAMEO test proteins by maximizing the area under the ROC curve (AUC) [75].

The predictor is available at the RaptorX Structure Protein Prediction server (http://raptorx.uchicago.edu/StructurePropertyPred/predict/).

1. Paste the protein sequence(s) or upload the sequence file in FASTA format. Chose the prediction with sequence profile prediction or not (the latter option is faster but less accurate). Before clicking on submit, it is strongly recommended to provide an email address to retrieve the results. Each user can analyze up to 500 sequences.
2. A result URL is returned online: click on it or paste the JobID in My Jobs section to retrieve the results. Your browser shall support HTML5 and allow JavaScript for interactive visualization.
3. The result page is divided into three parts. The first section shows a summary of the prediction result and the download button. The second one returns the prediction along five lines: the first one (SEQ) returns the input sequence, the second and third strings (SS3, SS8) give the secondary-structure prediction, the fourth (ACC) show the solvent accessibility and the last one (DISO) the disorder regions indicated with the symbol '*'. Finally, the user can click on the Disorder box in the third section which will show the predicted disorder/order propensity of each residue by red/blue bars, respectively. Hovering the mouse over a residue will display the disorder/order probability in percentage for the labeled amino acid.

The program is also available as a stand-alone version for the analysis of large data sets at https://github.com/realbigws/Predict_Property.

Predictors that have not been trained on disordered proteins

IUPred2A (https://iupred2a.elte.hu/) is a well establish disorder predictor which uses a novel algorithm that evaluates the energy resulting from inter-residues interactions [76]. Although it was derived from the analysis of the sequences of globular proteins only, it allows the recognition of disordered proteins based on their lower interaction energy. This provides a new way to look at the lack of a well-defined structure, which can be viewed as a consequence of a significantly lower capacity to form favorable contacts, correlating with studies by the group of Galzitskaya [77]. IUPred2A includes two tools, ANCHOR2, hence enabling the identification of disordered binding regions [78], and a new tool to predict the presence of conditionally redox sensitive regions. The

program is also available as a stand-alone version (in python) for the analysis of large data sets. It can be obtained by clicking on “Download” from the program main page.

1. Paste the amino acids query sequence in raw or FASTA format. Alternatively, enter the SWISS-PROT/TrEMBL ID or the UniprotKB accession number of the protein. Multiple sequences can be submitted in FASTA format from a local text file using the “Browse” button. In this case, results are returned by e-mail.
2. Chose the prediction type (long disorder, short disorder, structured regions), the context-dependent predictions to be run (i.e., ANCHOR2 for binding regions, redox state for conditionally redox-sensitive in disordered regions or none), and click on "Submit".
3. Prediction results are promptly returned online and the plot can be saved (png format) by clicking on it with the mouse right button. The output also provides a table with disorder probabilities per residue. If the ANCHOR2 prediction was selected, the output plot will show the probability profile of each residue to being part of a binding region (blue line). The significance threshold above which residues are considered to be disordered and/or being part of binding regions is 0.5. If the redox state-dependent prediction was selected, redox-sensitive disordered regions in the output are identified by a purple shading area on the graph and marked by a box below the graph. If a SWISS-PROT/TrEMBL ID or an UniprotKB accession number has been entered, several annotations are displayed such as reported eukaryotic linear motifs (ELM), post translational modifications (PTM) sites as well as available PDB structures. The figure can be saved (in png format) by clicking on the floppy disk icon present on the top right of the plots. Users can also retrieve the numerical scores by clicking on the “Download results” button in JSON or text formats.

FoldUnfold (http://bioinfo.protres.ru/ogu/) calculates the expected average number of contacts *per* residue from the amino acid sequence alone [77]. The average number of contacts per residue was computed from a dataset of globular proteins. A region is considered as natively unfolded when the expected number of close residues is less than 20.4 for its amino acids and the region is greater or equal in size to the averaging window. The user can define the size of the sliding window, but it is recommended to use averaging frame of 41 and 11 to find long and short disordered regions, respectively.

1. Paste the sequence in FASTA format, and click on the "Predict" button.
2. Prediction results are returned online. In the “short result” description, boundaries of disordered regions (unfolded) are given at the bottom of the page, while the “long result” allows to save the output plot. In the profile, disordered residues are shown in red.

DynaMine (http://dynamine.ibsquare.be/) provides prediction of protein backbone dynamics, using an input sequence in the form of backbone N-H S^2 order parameters. The values are between 0 (for highly dynamic, full random movement) and 1 (rigid conformation). The DynaMine S^2 estimates are based on NMR chemical shift values, covering a time scale from femtoseconds down to microseconds and low milliseconds [79,80]. Although it is a simple linear regression system it is able to accurately distinguish regions with different structural organizations present in the protein. DynaMine achieves the best performance without necessarily depending on prior knowledge of disorder or 3-D structural information, thus providing independent evidence between structural disorder and dynamics in protein regions.

1. Provide one or more protein sequences in FASTA format or enter a UNIPROT

identifier. Enter an e-mail address (facultative).

2. Click "Submit".
3. Open the link received by e-mail to get the access to the results page (your work will be stored for one week). The results will be present in three different forms: the annotated plot of backbone dynamics profile, a graphical representation of the protein sequence and a report for all the residues of the sequence. It is possible to download a zip file (link at the top of the page) with the results of all the sequences submitted.

The DynaMine predictor is also available as a stand-alone version. The program can be run in local mode through the command line (Beta – python 2) or a JSON API. An APY key is required to obtain the predictions from DyneMine server. The APY key can be obtained by providing an e-mail address and then by clicking on the "request APY key" button at http://dynamine.ibsquare.be/download/.

s2D (https://www-cohsoftware.ch.cam.ac.uk/index.php/s2D) is trained on solution-based NMR data, which allows to quantitatively distinguish disordered random coil regions from disordered regions with residual secondary structure elements. The predictor uses a combination of artificial neural network and extreme learning machines. The advantage of this tool is the characterization of the conformational properties of disordered states and the identification of regions involved in disorder-to-order transitions [81].

1. To use the software, the user shall register at https://www-cohsoftware.ch.cam.ac.uk/.
2. After logging in, click on s2D on the left menu and paste one raw protein sequence at a time. Tick "plot also random coil profile" and then submit.
3. The result page returns two links: a) a tab separated file displaying the probability of each residue of populating α-helix, β-strand or random coil; b) a plot of the secondary structure populations. The yellow line in the plot represents disordered residues in random coil states, while blue and green bars show the probability of each amino acid to be in α-helix and β-strand structure, respectively. Disordered regions have both α-helix and β-strand populations smaller than 0.5.

The source code and executable are freely available for download: click on "s2D_py3" to run both version 2 and 1on Python3; for Python 2.7 click either on "version2" or "published version".

Stand-alone disorder predictors

As there is an increasing need for investigation of disorder in proteomes, it is worth to mention stand-alone predictors that allow fast analysis and finer regulation of the parameters. We provide below a short description of these disorder predictors along with guidelines to download them.

MobiDB-lite 3.0 (http://old.protein.bio.unipd.it/mobidblite/) is an optimized method for highly specific predictions of long IDRs. The method uses eight different predictors to derive a consensus that is refined to remove short disordered regions and keep only those longer than 20 residues [82]. The metapredictor offers the possibility to characterize different flavors of disorder, by identifying either polyampholytic, positive or negative polyelectrolytic, low complexity regions (see 2.4) or regions enriched in cysteine, proline or glycine or polar residues [82]. MobiDB-lite works in Linux system and requires Java (version >=7.x) and Python (version 2 and 3). The program is easily called by the command line.

DisPredict2 employs Support Vector Machine with the kernel Radial Basis Function and calculates a position specific estimated energy (PSEE) for each amino acid based solely on the

primary sequence of a protein. The PSEE depends on the neighborhood region of each amino acid, the pairwise contact energies between different amino acid types and their predicted relative solvent accessibility. The predictor was trained on short and long disordered sequences from both PDB and DisProt [83]. The user can download the software from https://github.com/tamjidul/DisPredict2_PSEE.

To compile and execute the predictor, the following are required:

- PSI-BLAST (ftp://ftp.ncbi.nih.gov/blast/),
- SPINE X (http://sparks.informatics.iupui.edu/SPINE-X/),
- IUPred2A (https://iupred2a.elte.hu/)
- DAVAR [84]
- libSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm)
- GCC (http://gcc.gnu.org/)

rawMSA integrates different methods for the prediction of protein structure characteristics. The program works on the raw Multiple Sequence Alignment (hence the name) as input to the neural network to extract evolutionary information. The idea is that amino acids with context-dependent similarities will map together with an embedding layer, as it occurs in natural language processing [85]. The program is available at https://bitbucket.org/clami66/rawmsa/src/master/.

Rosetta ResidueDisorder is an application in the Rosetta suite developed to predict disordered regions from protein primary sequence or, for the first time, directly from the coordinates of a protein structure [86,87]. Starting from the primary sequence, the software generates an ensemble of conformation and average the energy scores of such conformations, highlighting disordered regions as the ones energetically less favorable. The last update of the software allows also to predict folding/unfolding events [87].

In order to use Rosetta ResidueDisorder, the user must download the Rosetta software from https://new.rosettacommons.org/docs/latest/getting_started/Getting-started. Practical instruction to use Rosetta ResidueDisorder can be found in the Supporting Material from [86].

Binary disorder predictors

The charge/hydropathy method and its derivative FoldIndex is a predictor that has not been trained on disordered proteins. It is based on the elegant reasoning that folding of a protein is governed by a balance between attractive forces (of hydrophobic nature) and repulsive forces (electrostatic, between similarly charged residues) [88]. Thus, globular proteins can be distinguished from unstructured ones based on the ratio of their net charge versus their hydropathy. The Mean Net Charge (R) of a protein is determined as the absolute value of the difference between the number of positively and negatively charged residues divided by the total number of amino acid residues. It can be calculated using the program ProtParam at the ExPASy server (https://www.expasy.org/resources/protparam). The Mean Hydrophobicity (H) is the sum of normalized hydrophobicities of individual residues divided by the total number of amino acid residues minus 4 residues (to take into account fringe effects in the calculation of hydrophobicity). Individual hydrophobicities can be determined using the ProtScale program at the ExPASy server, which provides 24 different predefined hydrophobicity scales derived from the literature. The most frequently used is the scale determined by Kyte-Doolittle [89]. To use it, it is sufficient to indicate the options "Hphob / Kyte & Doolittle", a window size of 5, and normalizing the scale from 0 to 1. The values computed for individual residues are then exported to a spreadsheet, summed and divided by the total number of residues minus four to yield (H). A protein is predicted as disordered

if H < [(R + 1.151) / 2.785]. Alternatively, charge/hydropathy analysis of a query sequence can be obtained by choosing this option on the main page of the PONDR server (see 2.2.1).

A drawback of this approach is that it is a binary predictor, i.e., it gives only a global (i.e., not positional) indication, which is not valid if the protein is composed of both ordered and disordered regions. It can be only applied to protein domains, implying that a prior knowledge of the modular organization of the protein is required.

A derivative of this method, FoldIndex (https://fold.weizmann.ac.il/fldbin/findex), solves this problem by computing the charge/hydropathy ratio using a sliding window along the protein [90]. However, since the default sliding window is set to 51 residues, FoldIndex does not provide reliable predictions for the N- and C-termini and is therefore not recommended for proteins with less than 100 residues.

1. Paste the sequence in raw format and click on "process".
2. The results page shows a plot that can be saved as an image (png format) by clicking on it with the mouse right button. Disordered regions are shown in red and have a negative "foldability" value, while ordered regions are shown in green and have a positive value. Disorder statistics (number of disordered regions, longest disordered region, number of disordered residues and scores) are given below the plot.

The cumulative distribution function (CDF) is another binary classification method [91,92]. The CDF analysis summarizes the per-residue predictions by plotting predicted disorder scores against their cumulative frequency, which allows ordered and disordered proteins to be distinguished based on the distribution of prediction scores [91,92]. A CDF curve gives the fraction of the outputs that are less than or equal to a given value. At any given point on the CDF curve, the ordinate gives the proportion of residues with a disorder score less than or equal to the abscissa. The outputs of predictors are unified to produce per-residue disorder scores ranging from 0 (ordered) to 1 (disordered). In this way, CDF curves for various disorder predictors always begin at the point (0, 0) and end at the point (1, 1) because disorder predictions are defined only in the range [0, 1] with values less than 0.5 indicating a propensity for order and values greater than or equal to 0.5 indicating a propensity for disorder. Fully disordered proteins have very low percentage of residues with low predicted disorder scores, as the majority of their residues possess high predicted disorder scores. On the contrary, the majority of residues in ordered proteins are predicted to have low disorder scores. Therefore, the CDF curve of a structured protein would increase very quickly in the domain of low disorder scores, and then goes flat in the domain of high disorder scores. For disordered proteins, the CDF curve would go upward slightly in the domain of low disorder scores, then increase quickly in the domain of high disorder scores. Fully ordered proteins thus yield convex CDF curves because a high proportion of the prediction outputs are below 0.5, while fully disordered proteins typically yield concave curves because a high proportion of the prediction outputs are above 0.5. Hence, theoretically, all fully disordered proteins should be located at the lower right half of the CDF plot, whereas all the fully ordered proteins should fall in the upper left half of this plot [91,92]. By comparing the locations of CDF curves for a group of fully disordered and fully ordered proteins, a boundary line between these two groups of proteins could be identified. This boundary line can therefore be used to separate ordered and disordered proteins with an acceptable accuracy, with proteins whose CDF curves are located above the boundary line being likely to be structured, and proteins with CDF curves below the boundary being likely to be disordered [91,92]. CDF-plots based on various disorder predictors have different accuracies [92]. PONDR® VSL2-based CDF was found to achieve the highest accuracy, which was up to 5-10% higher than the second best of the other five CDF functions for the separation of

fully disordered proteins from structured proteins also containing disordered loops or tails. As for the separation of fully structured from fully disordered proteins, the CDF curves derived from the various disorder predictors all were found to exhibit similar accuracies [92]. CDF analysis can be run from the PONDR server (see 2.2.1).

1. Enter the protein name and paste the sequence in raw (or FASTA) format, choose the disorder predictor to be run, tick CDF and click on "Submit Query".
2. The result is provided as a plot than can be saved (gif format) by clicking on it with the right mouse button. Disorder statistics (number of disordered regions, longest disordered region, percentage of disorder, number of disordered residues and scores) and CDF output scores are given below the plot.

The CH-CDF plot is an analytical tool combining the outputs of two binary predictors, the Charge-Hydropathy (CH) plot and the CDF plot, both predicting an entire protein as being ordered or disordered [93]. The CH-plot places each protein onto a 2D graph as a single point by taking the mean Kyte-Doolittle hydropathy of a protein as its X coordinate and the mean net charge of the same protein as its Y coordinate. In a CH-plot, structured, globular proteins and fully disordered, can be separated by a boundary line [88]. Proteins located above this boundary are likely to be disordered, while proteins located below this line are likely to be structured. The vertical distance on CH-plot from the location of the protein to the boundary line is then a scale of disorder (or order) tendency of the protein. This distance is referred to as the CH-distance. As explained above, in CDF-plots, ordered proteins curves tend to stay on the upper left half, whereas disordered proteins curves tend to locate at the lower right half of the plot. An approximately diagonal boundary line separating the two groups can be identified and the average distance of the CDF curves from this boundary is a measure of the disorder (order) status of a given protein and is referred to as CDF-distance. By putting together both the CH-distance and the CDF-distance, a new method called the CH-CDF plot was developed [93]. The CH-CDF plot provides very useful information on the general disorder status of a given protein. After setting up boundaries at CH=0 and CDF=0, the entire CH-CDF plot can be split into four quadrants. Starting from the upper right quadrant, by taking the clockwise sequence, the four quadrants are named Q1 (upper right), Q2 (lower right), Q3 (lower left), and Q4 (upper left). Proteins in Q1 are structured by CDF, but disordered by CH; proteins in Q2 are predicted to be structured by both CDF and CH; proteins in Q3 are disordered by CDF but structured by CH; and proteins in Q4 are predicted to be disordered by both methods. The location of a given protein in this CH-CDF plot gives information about its overall physical and structural characteristics. **Figure 3** shows how to build a CH-CDF plot.

Presently, there is no publicly available automated server for the generation of CH-CDF plots.

Non-conventional disorder predictors

The hydrophobic cluster analysis (HCA) is a non-conventional disorder predictor in that it provides a graphical representation of the sequence that helps in identifying disordered regions (see [94]). Although HCA was not originally intended to predict disorder, it is very useful for unveiling disordered regions [95]. HCA outputs can be obtained from http://mobyle.rpbs.univ-paris-diderot.fr/cgi-bin/portal.py?form=HCA#forms::HCA. HCA provides a two-dimensional helical representation of protein sequences in which hydrophobic clusters (i.e., represented by the most hydrophobic residues V, I, L, M, Y, W, F) are plotted along the sequence (**Figure 4**) [95]. As such, HCA is not *stricto sensu* a predictor. Disordered regions are recognizable as they are depleted in (or devoid of) hydrophobic clusters. HCA stands aside from other predictors, since it provides a

representation of the short-range environment of each amino acid, thus giving information not only on order/disorder but also on the folding potential (see 2.3). Although HCA does not provide a quantitative prediction of disorder and rather requires human interpretation, it provides additional, qualitative information as compared to automated predictors. In particular, HCA highlights coiled-coils, regions with a biased composition, regions with potential for induced folding and very short potential globular domains (for examples see [17,16,19]). Finally, it allows meaningful comparison with related protein sequences and enables a better definition of the boundaries of disordered regions. On the other hand, if HCA is a powerful tool to delineate regions devoid of regular secondary structure elements, it is poorly suited to recognize molten and premolten globules, i.e., proteins with a substantial amount of secondary structure but devoid of stable tertiary structure.

1. Paste the sequence (raw format) in the "Input Data" tab. Alternatively, upload a text file containing the query sequence.
2. Choose the output format (PDF or PostScript format) in the "Options" tab. The HCA plot can also be generated in black and white by clicking on "Yes" in the corresponding option.
3. Click on the "Run" button. The HCA plot is returned on line and can be saved by clicking on the floppy disk icon.

2.2.2. Metapredictors

GeneSilico MetaDisorder MD (http://genesilico.pl/metadisorder/) is a free open tool for prediction of protein disorder. The MetaDisorder web service is articulated into four different meta-predictors. The first one of them is *metadisorder*, built on numerous disorder predictors, which are DISOPRED2, DisEMBL (3 versions), Globplot, DISprot, RONN, iPDA, IUPred (2 versions), Poodle-s, Pdisorder, Poodle-l, PrDOS and Spritz (2 versions). This meta-server was among the best predictors in CASP8 (2008). The second one is *metadisorder3d*, based on fold recognition methods to find similar sequences, such as Phyre, Pcons, HHsearch, PSI-BLAST, FFAS, MGenThreader. A genetic algorithm is used to deduce protein disorder using gaps in alignments. Finally, the *metadisordermd* combines the previous two meta-methods in a new one.

The last and most famous is *metadisorderm2*, which corresponds to an improved version of the first *metadisorder* released in 2008 that instead of using the Sw score [Sw=(2ACC-1) where ACC=(Sensitivity+Specificity)/2], uses the so-called Sww score, which tries to capture the best features of the Sw score and AUC (Area Under a "Receiver Operating Characteristic", ROC, curve) that is indicative of the classifier accuracy. One interesting point to mention here is that among these predictors are also other metaservers. As such, MetaDisorderMD2 is an extreme application of the concept that "the combination of different disorder predictors helps in refining the predictions". This method is based on 13 different disorder predictors and 8-fold recognition methods. As a result, it provides the user with the raw CASP formatted output of each disorder predictor and corresponding alignments for the fold recognition methods, along with a computed consensus in the same format. It also displays a plot that allows one to compare the consensus to any other disorder predictor result [96].

Metadisorderm2 was among the best predictors of protein disorder evaluated during independent tests in CASP9 (2010).

1. Enter a title to the query, the e-mail address and paste sequence (raw format) in the corresponding field. Then click on the "Submit" button.
2. The results are displayed in an HTML page but can also be seen in raw text from a link available in the page results. An email is sent giving a link towards the result

page. On the graphical output, residues whose disorder probability is above 0.5 are considered as disordered. The results of all four metapredictors are shown.

MULTICOM is a protein tertiary structure prediction server. It provides information on secondary structure, solvent accessibility, disorder, domain boundaries, along with predicted models (in PDB format). It is a simple averaging approach that is different from other meta-methods based on consensus voting [51]. MULTICOM makes disorder predictions based on a five-layers architecture, divided into target identification and ranking, multiple-template combination, model generation and model refinement [97]. The MULTICOM protein structure prediction system was enhanced during 2018 CASP13 experiment with three components: distance-driven template-free modeling (*ab initio*), contact distance prediction based upon deep convolutional neural network, protein model ranking improved by deep learning and contact prediction. MULTICOM was ranked 3rd out of all 98 predictors in either template-based structure modeling and template-free in CASP13 [98]. The server can be reached from http://sysbio.rnet.missouri.edu/multicom_cluster/ and returns results by e-mail in a CASP/PDB format.

1. Enter a target name, the protein sequence in raw format and provide the email address in the corresponding field. Then click on the "Predict" button.
2. Open the result e-mail that contains model evaluation, model combination, and model refinement data in the CASP/PDB format.

The ***DEPICTER*** (DisorderEd PredictIon CenTER) server combines together ten disorder predictors to predict disorder, disorder functions, protein and nucleic acids bindings, linkers, and moonlighting regions. The prediction of disorder is based on a consensus calculated on three disorder predictors (SPOT-Disorder single, IUPred-long and IUPred-short). The consensus method works on 54 features extracted from each reside in the input protein sequence and the two flanking residues, using two sliding windows (e.g., prediction score for predicted residue, average of the prediction sores from the main window). These features are then analyzed with four machine learning algorithms to generate the final propensity score (values between 0 and 1). Residues with score greater than the threshold of 0.5 are considered as disordered. The predictor was trained on sequences from DisProt and PDB [99]. The DEPICTER server is available at http://biomine.cs.vcu.edu/servers/DEPICTER/.

1. Enter or upload a file with protein sequence, provide your e-mail address (optional).
2. Choose the methods to compute (e.g., disorder prediction, protein binding regions prediction, DNA-binding regions and so on - by default, all methods are selected) and click “RUN”.
3. Results are returned online and sent by email if an address is provided. The results page shows the binary prediction of disordered regions (in grey), disordered protein-binding regions (green), RNA/DNA-binding regions (light/dark blue) linkers (pink), and multifunctional regions (violet). For a complete prediction visualization, untick “Only binaries”. Hovering the mouse on the disordered region returns the amino acid position and its disorder propensity score. Results can be downloaded as a text file by clicking on “RESULTS.txt”.

MFDp (Multilayered Fusion-based Disorder predictor) is a metapredictor that is made of three support vector machines specialized for the prediction of disordered regions. It combines these results with multiple complementary disorder predictors, namely DISOclust, DISOPRED, IUPRED-L and IUPRED–S. In addition, MFDp also takes into account secondary structure predictions, solvent accessibility, backbone dihedral torsion angles and B-factors in order to

generate its consensus [100]. The web server can be found at http://biomine.cs.vcu.edu/servers/MFDp/. It accepts up to five sequences at a time.

1. Enter the protein sequence in FASTA format and provide the e-mail address in the corresponding field. Choose the predictor for which you want to see the result (DISOclust, DISOPRED, IUPRED L and/or IUPRED S) in addition to the MFDp prediction, and then click on the "Run" button.
2. Results can be accessed from a link displayed on the MFDp processing page. An e-mail is also sent giving a link towards the result page. Results are in the form of an alignment of the different predictor results and the consensus prediction built by MFDp. Disordered residues are marked by a red "D" character and the confidence values are reported below. In addition, results can also be downloaded in csv format.

The standalone version of MFDp is freely available to academic users upon request by sending an e-mail to mizianty@ualberta.ca.

MFDp2 (http://biomine.ece.ualberta.ca/MFDp2/) combines per-residue disorder probabilities predicted by MFDp with per-sequence disorder content predicted by DisCon. Then predictions are filtered with a post processing filter [101]. The server accepts up to 100 sequences.

1. Enter the protein sequence in FASTA format and provide the e-mail address in the corresponding field.
2. The output shows optimized per-residue disorder probability profiles, per-sequence disorder content, list (with analysis) of disordered segments, and several profiles that help in the interpretation of the results. The results are available online in a graphical format and can be also downloaded in a text-based (parsable) format.

DisCoP (http://biomine.cs.vcu.edu/servers/disCoP/) (Disorder Consensus-based Predictor) goes through a three stage process (i) sequence processing by SPINE-D, DISOclust, MD and DISOPRED2, (ii) elaboration of predictions to generate features and (iii) development of a regression model trained on the features to produce DisCoP's prediction [102,103]. The server accepts up to five protein sequences.

1. Enter or upload the sequence in FASTA format and provide an e-mail address. To start the prediction click on "Run disCoP".
2. The results page is returned by clicking on the given link in the status page, or on the link received by e-mail. It includes a visualization of the predictions and a text file with the prediction results (downloadable in csv format). The results are color-coded and provided as a real-valued confidence (propensity for disorder) and a binary prediction (disordered vs structured residue).

PONDR-FIT uses a consensus artificial neural network (ANN) prediction method that combines PONDR-VLXT, PONDR-VSL2, PONDR-VL3, FoldIndex, IUPred, and TopIDP [104]. The predictor can be run online for academic use only, from http://original.disprot.org/pondr-fit.php. Commercial users must obtain permission from http://www.pondr.com/.

1. Enter the sequence file in FASTA (or EMBL) format and then click on the "Submit" button.
2. The server returns a graphical plot of disorder probabilities for each amino acid position, along with a raw output file of the results.

Interestingly, the server enables specifying regions of interest using a specific syntax described in the main page.

PredictProtein **(**www.predictprotein.org) is a server based on a system of neural networks that combines the outputs from several original prediction methods, with the evolutionary profiles and sequence features that correlate with protein disorder such as predicted solvent accessibility and protein flexibility. Beyond providing predictions of secondary structure, trans-membrane regions and disulphide bridges among other features, the server also returns predictions of disorder. In particular, the NORSnet, UCON, PROFBval and MetaDisorder (MD) programs can be run from the PredictProtein server.

NORSnet is a neural network-based method for the identification of disordered loops [105]. NORSnet was trained to distinguish between very long contiguous segments with non-regular secondary structure (NORS regions) and well-folded proteins. The program is also provided as a Debian package that can be found at https://rostlab.org/owiki/index.php/Norsnet.

Ucon is a method that combines predictions for protein-specific contacts with a generic pairwise potential. This predictor was trained against the DisProt and the PDB. It performs well in predicting proteins with long disordered regions [106]. Ucon can also be downloaded as a Debian package from https://rostlab.org/owiki/index.php/Ucon.

MD (Meta Disorder) [107] runs a panel of four predictors carefully selected on the basis of their complementarity in predicting disorder, namely DISOPRED2, PROFbval [108], NORSnet and Ucon. Once it has gathered results from these predictors it calculates the arithmetic average over the four raw outputs. The results of MD that are included within the PredictProtein output, come in a raw format yielding the computed probability for the MD consensus associated to each distinct disorder predictor results. Like UCON and NORSp, MD can be also downloaded as a Debian package from http://rostlab.org/debian/pool/non-free/m/metadisorder/.

From the PredictProtein page:

1. Enter the amino acid sequence (raw data) and click on the "PredictProtein" button.
2. Either enter the e-mail address without creating an account (in which case you will run Open PredictProtein) or create an account that will allow you subsequently to login with a password. Note that Open PredictProtein does not store jobs. Alternatively, open the link shown and wait until the end of the analysis.
3. Upon completion of prediction, the user is sent an e-mail with a link to the result page. Boundaries of NORS regions are indicated above the annotated sequence in which solvent exposure, secondary structure elements, coils and trans-membrane regions are also indicated. On the left side of the result page, different layout options can be chosen. Clicking on “Protein Disorder and Flexibility” will give access to prediction results as provided by PROFBval, Ucon, NORSnet and MD in the form of colored boxes. Mouse over the different colored boxes to learn more about the annotations.

MeDor (MEtaserver of DisORder) (http://MeDor.afmb.univ-mrs.fr/) stands aside with respect to other metapredictors as *(i)* it provides an output in a specific format that can be annotated, saved and further modified, and *(ii)* was not originally intended to provide a consensus of disorder prediction and was rather conceived to speed up the disorder prediction step by itself and to provide a global overview of predictions [29]. It is presently being updated, and the updated version is conceived to also generate two consensus disorder predictions, one corresponding to regions predicted as disordered by *all* the predictors and one corresponding to regions predicted as disordered by the majority of predictors.

MeDor allows fast, simultaneous analysis of a query sequence by multiple predictors and easy comparison of the prediction results. It also enables a standardized access to disorder predictors and allows meaningful comparisons among various query sequences. It provides a

graphical interface with a unified view of the output of multiple disorder predictors. Beyond providing a graphical representation of the regions of predicted disorder, MeDor is also conceived to serve as a tool allowing to highlight specific regions of interest and to retrieve their sequence. In addition, MeDor outputs can be saved, modified and printed. Presently, the following programs are run by MeDor: a secondary structure prediction (SSP), based on the StrBioLib library of the Pred2ary program [109], HCA, DorA (an unpublished predictor developed in the AFMB lab that uses size and abundance of hydrophobic clusters in the HCA plot to predict disorder), MoreRONN, FoldUnfold, FoldIndex, MobiDB-lite, and Phobius. MobiDB-lite is a metapredictor that uses eight different disorder predictors (see 2.2.1). Phobius (http://phobius.sbc.su.se/index.html) predicts transmembrane regions. While SSP and HCA do not require a web connection, the other predictors are remotely launched through connection to the public web servers. Additional predictors could be nevertheless easily implemented in MeDor in the future. Predictors to be run can be selected from the MeDor input frame.

MeDor provides a graphical output, in which the sequence query and the results of the various predictors are featured horizontally, with a scroll bar allowing progression from the N-terminus to the C-terminus. All predictions are drawn along the sequence that is represented as a single, continuous horizontal line. MeDor also allows highlighting specific regions of interest and retrieving their sequence. Output files are in the specific (.med) format that is made of XML and thus can provide a graphical output for any program that return such a format. As XML is quite simple to access, it is also possible to edit the ".med" file manually to get a fully customized output that could even integrate additional predictions not initially provided. The (.med) file format can also be opened by any XML reader and the format is well described by the "xsd" file provided with the program. It is also possible to customize the output (highlight regions of interest, change colors, add and edit comments...) and to retrieve the predictor statistics values at each position, as well as the amino acid sequence of specific regions of interest.

1. Go to the MeDor home page (http://MeDor.afmb.univ-mrs.fr/)
2. Paste the sequence in either raw or FASTA format and optionally enter the sequence name
3. Click on "Start MeDor"
4. Alternatively, MeDor can be downloaded (chose the appropriate version according to your computer environment). Using the downloaded version of MeDor instead of the applet version enables the user to *(i)* print the results, *(ii)* save the output as an image, *(iii)* save (and load) files in the MeDor format, *(iv)* access the comment panel, (*v*) import a sequence by providing the SwissProt accession number.

QUARTERplus (QUality Assessment for pRotein inTrinsic disordEr pRedictions) (http://biomine.cs.vcu.edu/servers/QUARTERplus/) is a server that provides disorder predictions along with easy to interpret residue-level quality assessment scores (QA) that reliably quantify the residue-level predictive quality of the predictors [111]. QA scores complement the propensities produced by the implemented disorder predictors by identifying regions where disorder predictions are more likely to be correct. The deep neural network utilizes the QA scores to identify and fix the regions where the original disorder predictions are poor. From the server main page, the user can also choose the "Multiple Sequence Mode" that enables analyzing up to 50 sequences by choosing one among the following predictors: disEMBL-465, disEMBL-HL, GlobPlot, IUPred-long, IUPred-short and VSL2B. The server provides disorder prediction and QA scores for the selected predictor. The default option on the main page is the "Single Sequence Mode", which

provides disorder prediction and QA scores for QUARTERplus Meta Predictor, SPOT Disorder, Disopred3 and IUPRED-short.

1. Paste the protein sequence in FASTA format.
2. Enter your e-mail address.
3. Click on “Run”.
4. Upon completion of prediction, the user is sent an e-mail with a link to the result page and to a csv file. Results are in the form of a plot with per residue disorder scores for the four predictors (with residues with a score above 0.5 being considered as disordered). Below the graph, for each predictor is shown a bar with disordered regions shown in black and ordered regions shown in grey. Under each bar is another bar with pre-residue QA scores colored from red (poorly reliable) to green (highly reliable).

2.2.3. Combining predictors and experimental data

An extreme extension of the combined use of different predictors is the combined use of *in silico* and experimental approaches with the ultimate goal of inferring as many structural information as possible while limiting the experimental characterization to relatively low-demanding experiments. An illustration of such a combined approach can be found in [112], where far-UV circular dichroism and computational analyses were combined. In that study, the authors plotted the ratio between the Θ_{222} and Θ_{200} ($\Theta_{222}/\Theta_{200}$) of a set of IDPs under study, along with the $\Theta_{222}/\Theta_{200}$ ratio of a set of well-characterized random coil-like and premolten globule-like proteins [113]. The authors then set an arbitrary threshold of the $\Theta_{222}/\Theta_{200}$ ratio that allows discrimination between random coil-like IDPs and IDPs adopting a premolten-like conformation. Then, they generated a plot in which the distance of each IDP under study from this threshold was plotted as a function of its CH-distance in the CH plot (**Figure 5**). This analysis was intended to combine, and hence extend, the two methods previously introduced by Uversky and co-workers [88,113] to allow random coil-like forms to be readily and easily distinguished from premolten-like forms among proteins predicted to be intrinsically disordered by the hydropathy/charge method. In the resulting plot, increasingly negative CH distances designate proteins with increasing disorder, while increasingly positive $\Theta_{222}/\Theta_{200}$ distances designate IDPs becoming progressively more collapsed, as a consequence of an increased content in regular secondary structure. Thus, the left bottom quadrant is expected to correspond to IDPs adopting a random coil-like conformation, while the right bottom quadrant is supposed to designate IDPs adopting a premolten globule-like conformation.

2.3. Identifying disordered regions involved in binding to partners

IDPs bind to their target(s) through interaction-prone short segments that become ordered upon binding to partner(s). These regions are referred to as "Molecular Recognition Elements" (MoREs) or "Molecular Recognition Features" (MoRFs) [114-116] or "Intrinsically Disordered Binding" (IDB) sites [78].

Before specific predictors became publicly available, these regions could be successfully identified using tools that had not been specifically designed to this aim: indeed, PONDR VL-XT and HCA were found to be very helpful to identify disordered binding regions. Owing to its high sensitivity to local sequence peculiarities, PONDR VL-XT was noticed to be able to identify disorder-based interaction sites [114] (for examples see [117,118]).

HCA is similarly instrumental for the identification of regions undergoing induced folding, because burying of hydrophobic residues at the protein-partner interface is often the major driving force in protein folding [119,78]. In some cases, hydrophobic clusters are found within secondary structure elements that are unstable on their own in the free protein, but can stably fold upon binding to a partner. Therefore, HCA can be very informative in highlighting potential induced folding regions (for examples see [50,120,112]).

1. Perform HCA on the query sequence using either the Mobyle portal or the MeDor metaserver (see 2.2.2) and look for short hydrophobic clusters occurring within disordered regions.
2. Perform prediction using PONDR VL-XT (see 2.2.1) and look for sharp (and short) drops in the middle of disorder predictions.

In the last years, a few predictors aimed at identifying disorder-based regions have become publicly available. The majority of the MoRFs predictors are accessible via a web interface and they will be detailed below accompanied with a short description of their philosophy and details on how to run them. Additionally, some codes are freely available to be run in local mode, such as MoRFMPM (minimax probability machine) (https://github.com/HHJHgithub/MoRFs_MPM) [121], Predict-MoRF (https://github.com/roneshsharma/Predict-MoRFs) [122] and MoRFpred-plus (https://github.com/roneshsharma/MoRFpred-plus) [123]. Note that some methods, such as α-MoRFpred [114], α-MoRFpred-II [115], retro-MoRF [124] and the one developed by Fang and coworkers [125], are not available (as of May 2021) as web servers or as downloadable codes. Finally, some predictors, like MFSPSSMpred [126] and SPOT-MoRF [127], are currently not reachable.

ANCHOR2 (https://iupred2a.elte.hu/), accessible from the IUPred2A server page, seeks to identify segments that reside in disordered regions that cannot form enough favorable intrachain interactions to fold on their own and are likely to gain stabilizing energy by interacting with a structured protein partner. The underlying philosophy of ANCHOR relies on the pairwise energy estimation approach developed for IUPred [128]. Detailed description of how the predictor works is reported in section 3.2.2.2.1 and an example of the output is shown in **Figure 6A**.

The program is also available as a stand-alone version and can be run as a default sub-routine within IUPred2A. It can be obtained by clicking on “Download” from the program main page.

MoRFpred (http://biomine-ws.ece.ualberta.ca/MoRFpred/) identifies all types of MoRFs (α, β, coil and complex) [129]. MoRFpred uses a novel design in which annotations generated by sequence alignment are fused with predictions generated by a support vector machine, which uses a custom designed set of sequence-derived features. The features provide information about evolutionary profiles, selected physiochemical properties of amino acids, predicted disorder, solvent accessibility and B-factors.

1. Paste the sequence in FASTA format, provide the e-mail address (required). Up to five sequences can be entered.
2. Click on "Run MoRFpred".
3. Results are returned on line by clicking on a link to the results page (an e-mail is also sent as soon as results are available). The first line displays the query sequence, while the second and third lines show the predictions. The second row annotates Molecular Recognition Feature (MoRF) (marked as "M", in red) and non-MoRF (marked as "n", in green) residues, and the third row gives prediction scores (the higher the score the more likely it is that a given residue is a

MoRF). A horizontal scroll bar allows moving along the sequence. Results can also be downloaded in csv format.

fMoRFpred (http://biomine.cs.vcu.edu/servers/fMoRFpred/), for **f**ast **Mo**lecular **R**ecognition **F**eature **pred**ictor, can be used instead of its counterpart MorRFpred for faster prediction but with a slightly less accuracy [130].

1. Paste the sequence in FASTA format, provide an e-mail address (required). Up to 2000 sequences can be entered or alternatively browse for a local text file containing the sequences to analyze by clicking on "Choose a file".
2. Click on "Run fMoRFpred".
3. Results are returned as a downloadable text file in the result page. Prediction for each protein is given in 5 lines: 1) protein name, 2) protein sequence where uppercase and lowercase amino acids residues stand for predicted MoRF and non-MoRF, respectively, 3) MoRFs probability prediction, 4) MoRFs binary prediction where "1" and "0" stands for predicted MoRF and non-MoRF residue, respectively, 5) disorder binary prediction where "1" and "0" stands for predicted disordered and ordered residue, respectively.

MoRFchibi SYSTEM (https://morf.msl.ubc.ca/index.xhtml) is a set of three different MoRF predictors: MoRFCHiBi, MoRFCHiBi_Light and MoRFCHiBi_Web [131]. MoRFCHiBi, which is the fastest, uses only the physicochemical properties of amino acids. MoRFCHiBi_Light employs Bayes rule to integrate the MoRFCHiBi score as well as the disorder score generated by ESpritz. It accurately detects long MoRF sequences (> 30 residues). Finally, MoRFCHiBi_Web, which is the most accurate, integrates MoRFCHiBi scores as well as disorder propensity and conservation information using Bayes rule. Conservation information is based on the fact that MoRFs are more conserved than other regions in disordered proteins. Note that MoRFchibi SYSTEM is available as a web server, a RESTful web server and as a downloadable software.

1. Enter the sequence in FASTA format.
2. Click on "Submit Job". Provide an e-mail optionally.
3. Predictions usually are returned in less than a minute for a single sequence and appear on line as a green row under the sequence input field. To download the text results, click on "Ready" and then on "Download". The text file is composed of 8 columns: 1) the residue index, 2) the residue name, 3) the MoRFCHiBi_Web (MCW) scores, 4) the MoRFCHiBi_Light (MCL) scores, 5) the MoRFCHiBi (MC) scores, 6) another MoRF prediction, MoRFDC (MDC) based on the disorder prediction and the conservation score, 7) the disorder propensity score (IDP) and 8) the initial conservation propensity score (ICS).
4. Alternatively, MoRFchibi SYTEM provides a user-friendly graph to visualize predicted MoRFs (**Figure 6B**). The graphical results appear after clicking on "Graph". By default, the graph displays the MCW predictions (in red) but it's still possible to display the other predictions scores by clicking on the items on the right panel. To display the identified MoRFs sequences, click on "Toggle MoRF Bands" under the graph. The MoRFs appear as blue areas on the graph. Passing the mouse over the graph will display each amino acid residues for clear interpretation of the predictions. Graphical results can be downloaded by clicking on the "≡" symbol in PNG, JPEG, PDF or SVG formats.

DISOPRED3 is a disorder predictor (see 2.2.1) that has been implemented with a protein binding site predictor [132]. The predictor seeks to identify protein binding sites in disordered

regions that fold upon binding to a protein partner using a support vector machine (SVM) that takes into account sequence conservation, amino acids composition as well as the localization of the analyzed region along the sequence. Guidelines to run this predictor are found in section 3.2.1.1.6. An example of a DISOPRED3 output is shown in **Figure 6C**.

OPAL (http://www.alok-ai-lab.com/tools/opal/) combines the predictions of MoRFchibi and PROMIS (for Prediction of MoRFs Incorporating Structure) and computes a score to predict MoRFs with a size comprised between 5 and 25 residues [133]. MoRFchibi is based on the physicochemical properties of amino acids while the PROMIS model is used to discriminate between MoRF and non-MoRF residues according to structural attributes such as the half-sphere exposure (HSE), the solvent accessible surface area (ASA) and backbone angles of the disordered protein sequence.

1. Paste the raw protein sequence in the associated field. The sequence must be at least 26 residues long. E-mail address is optional.
2. Click on "Submit Job". A Job Id is given, enter it in the "Download Results" section and press "View Output". Refresh until results are returned or wait for an e-mail.
3. Predictions output are returned on line as a text format comprising 5 columns: 1) residue index, 2) residue name, 3) the OPAL scores, 4) the PROMIS score and 5) a binarized score where 0 corresponds to non-MoRF region and 1 to MoRFs. Results can be saved locally as a text file by copy-pasting the data.

Note that the ***OPAL+*** server (http://www.alok-ai-lab.com/tools/opal_plus) has been developed. It slightly outperforms its OPAL counterpart by 0.4 – 0.7 % in several MoRF test sets [134]. OPAL+ incorporates the hidden Markov model (HMM) profiles and physicochemical properties of MoRFs and their flanking regions. OPAL+ is available as a web server or downloadable from https://github.com/roneshsharma/OPAL-plus. The prediction takes longer than OPAL since OPAL+ required a HMM alignment file, which is obtained from the HHblits web server (https://toolkit.tuebingen.mpg.de/tools/hhblits), as well as hsa2, hsb2 and sp3 files. These last files are generated by the SPIDER3 server which is currently unavailable since the Spark lab, which is hosting SPIDER and several other predictors (e.g spot-MoRF), is currently moving to another server. SPIDER2 or SPIDER3 packages are nonetheless still downloadable from https://sparks-lab.org/downloads/.

DisoRDPbind (http://biomine.cs.vcu.edu/servers/DisoRDPbind/) predicts RNA, DNA and protein binding sites within disordered regions [135]. Four steps are performed by the predictor: 1) physicochemical properties of amino acids composing the input sequence are determined, the secondary structure and the disorder propensity are predicted and the sequence complexity is estimated, 2) numerical features proper to DNA, RNA and protein are generated from the previous step for the prediction of DNA, RNA and protein-binding residues, 3) the features are used as input for three logistic regression models (for each interaction type) to compute a propensity score for each amino acid, and 4) these scores are combined with functional annotations from sequence alignment generated by BLAST providing the final prediction.

1. Paste the input sequence in FASTA format and provide an e-mail address (required). Up to 5000 sequences can be entered or alternatively browse for a local text file containing the sequences to analyze by clicking on "Choose a file".
2. Click on "Run DisoRDPbind".
3. Results are returned in seconds as a downloadable text file in the result page. Prediction for each protein is given in 8 lines: 1) protein name, 2) protein sequence where uppercase amino acids are predicted to interact with DNA, RNA or protein

while the lowercase amino acids residues do not, 3) binary results for each amino acids for the RNA-binding prediction (1: predicted to bind, 0: doesn't bind), 4) RNA-binding propensity scores, 5) binary results for each amino acids for the DNA-binding prediction (1: predicted to bind, 0: doesn't bind), 6) DNA-binding propensity scores, 7) binary results for each amino acids for the protein-binding prediction (1: predicted to bind, 0: doesn't bind), 8) protein-binding propensity scores. Results can be saved locally as a text file by copy-pasting the data.

2.4. General procedure for disorder prediction

As already discussed, since the performance of predictors is dependent on both the type (*i.e.* short *versus* long, complete *versus* partial) of disorder they predict and on the type of disorder against which they were trained, multiple prediction methods need to be combined to improve the accuracy and specificity of disorder predictions. **Figure 7** illustrates a general sequence analysis procedure that integrates the peculiarities of each method to predict disordered regions.

1. Retrieve the amino acid sequence and the description file of the protein of interest by entering the protein name at the UniProt (http://www.uniprot.org) in the "Search" field.
2. Generate a multiple sequence alignment. A set of related sequences can be obtained by running HHblits (http://toolkit.tuebingen.mpg.de/hhblits). Click on the "get selected sequences" option and save them to a file in FASTA format. Use this file as input for building up a multiple sequence alignment using TCoffee (http://tcoffee.crg.cat/apps/tcoffee/do:regular). Mark variable regions as likely corresponding to flexible linkers or long disordered regions.
3. Search for long (>50 residues) regions devoid of predicted secondary structure using the PSIPRED (http://bioinf.cs.ucl.ac.uk/psipred/psiform.html) [136] and PredictProtein (http://www.predictprotein.org/) servers.
4. Using either the amino acid sequence or the UniProt ID, search the D2P2 and MobiDB databases. As D2P2 does not cover all organisms, and MobiDB does not include IDEAL entries, it is also recommended to search the IDEAL database. Search also the PED.
   In case no or incomplete information about disordered regions is obtained in this way, the analysis will have to be refined by performing the following steps.
5. Perform an analysis of sequence composition using the ProtParam ExPASy server (http://www.expasy.ch/tools/protparam.html) and compare the results with the average sequence composition of proteins within the UniProtKB/Swiss-Prot database (http://www.expasy.ch/sprot/relnotes/relstat.html).
6. Perform an analysis of sequence complexity using the SEG program [137]. Although the SEG program is implemented in many protein prediction servers (such as PredictProtein for instance), the program can also be downloaded from ftp://ftp.ncbi.nih.gov/pub/seg/seg, while simplified versions with default settings can be run at either http://mendel.imp.univie.ac.at/METHODS/seg.server.html or http://www.ncbi.nlm.nih.gov/BLAST or http://mendel.imp.ac.at/METHODS/seg.server.html. The stringency of the search for low-complexity segments is determined by 3 user-defined parameters: trigger window length [W], trigger complexity [K(1)] and extension complexity [K(2)].

Typical parameters for disorder prediction of long non-globular domains are [W]=45, [K(1)]=3.4 and [K(2)]=3.75, while for short non-globular domains are [W]=25, [K(1)]=3.0 and [K(2)]=3.3. Note however, that low complexity regions can also be found in ordered proteins, such as coiled-coils and other non-globular proteins like collagen.

7. Search for (i) signal peptides and transmembrane regions using the Phobius server (http://phobius.sbc.su.se/index.html) [138], (ii) leucine zippers using the 2ZIP server (http://2zip.molgen.mpg.de/) [139], (iii) coiled-coils using programs such as Coils (http://www.ch.embnet.org/software/COILS_form.html) [140] and (iv) regions forming collagen triple-helices. Note that the identification of coiled-coils is crucial since they can lead to miss-predictions of disorder (for examples see [16,19]). Likewise, it is important to identify regions forming collagen triple helices otherwise they will be predicted as IDRs due to their high content in Gly and Pro residues. It is also recommended to use DIpro (http://contact.ics.uci.edu/bridge.html) [141] to identify possible disulfide bridges and to search for possible metal-binding regions by looking for conserved Cys3-His or Cys2-His2 motifs in multiple sequence alignments. Indeed, the presence of conserved cysteines and/or of metal-binding motifs prevents meaningful local predictions of disorder within these regions, as they may display features typifying disorder while gaining structure upon disulfide formation or upon binding to metal ions [88].
8. Run Pfam and CATH HMMs to identify structured domains [142,143]. It is also recommended to run Pfam HMMs to identify regions forming collagen triple-helices, otherwise they will be predicted as IDRs due to their high content in Gly and Pro residues.
9. Run HCA to highlight regions devoid of hydrophobic clusters and with obvious sequence bias composition.
10. Run disorder predictions and identify a consensus of disorder. Since running multiple prediction methods is a time-consuming procedure, and since combining several predictors often allows achieving accuracies higher than those of each of the component predictors, it is recommended to perform predictions using metapredictors. As a first approach, we suggest to use the default parameters of each metapredictor, as they generally perform at best in terms of accuracy, specificity and sensitivity. Once a gross domain architecture for the protein of interest is established, the case of domains whose structural state is uncertain can be settled using the charge/hydropathy method, which has a quite low error rate. As a last step, boundaries between ordered and disordered regions can be refined using HCA, and regions with propensity to undergo folding coupled to binding can be identified using MoRFchibiSYSTEM or OPAL, which are the most accurate predictors as discussed in [26].

Figure legends

Figure 1. Output provided by the D2P2 database for human p53, a thoroughly investigated protein (UniProt ID P04637) containing intrinsically disordered regions. This output well illustrates the amount of information that can be obtained on both structural organization and post-translational modifications (PTM). The predicted Superfamily and Pfams domains (colored blocks) are shown. Regions predicted as disordered by the various predictors are shown along with a predicted disorder agreement. The level of agreement between all of the disorder predictors is shown as color intensity in an aligned gradient bar below the predictions (with a color code ranging from clear to deep blue with increasing agreement). The green segments represent disorder that is not found within a predicted SCOP domain. Below the disorder agreement line, ANCHOR binding sites are displayed (yellow blocks with zigzag infill), along with PTM sites from PhosphoSitePlus when known (shown as lettered spheres hanging below other predictions).

Figure 2. Output provided by DescribePROT for human p53 showing the various predicted features ranging from disorder, solvent accessibility secondary structure, protein and nucleic acid binding sites, signal peptides, conservation and linker regions.

Figure 3. The CH-CDF plot for an IDP (red) and a structured protein (green). (**A**) Disorder prediction curve by PONDR ® VSL2. The dashed line separates disordered from ordered residues. (**B**) CH plot for the two hypothetical proteins. The solid grey line represents the border between disordered and ordered proteins. The distance of each protein from the line is the Y-coordinate of that protein in the CH-CDF plot. (**C**) CDF plot of the two proteins. The average of the distances from the CDF curve to the boundary line (in gray) is the X-coordinate of that protein in the CH-CDF plot. D) CH-CDF plot of the two protein. The graph is divided in four quadrants (Q1-4) as explained in the text.

Figure 4. HCA plot of Hendra virus phosphoprotein (UniProt ID O55778). Hydrophobic amino acids (V, I, L, F, M, Y, W) are shown in green and are encircled and their contours are joined forming clusters. Clusters mainly correspond to regular secondary structures (α-helices and β-strands). The shape of the clusters is often typical of the associated secondary structures. Hence, horizontal and vertical clusters are mainly associated with α-helices and β-strands, respectively. A dictionary of hydrophobic clusters, gathering the main structural features of the most frequent hydrophobic clusters has been published helping the interpretation of HCA plots [144]. Sequence segments separating hydrophobic clusters (at least 4 non hydrophobic amino acids) mainly correspond to loops or linker (LNK) regions between globular domains. Long regions devoid of clusters correspond to disordered regions and small clusters within disordered regions correspond to putative MoREs. Coiled-coil regions have a peculiar and easily recognizable appearance in the form of long horizontal clusters. Symbols are used to represent amino acids with peculiar structural properties (stars for prolines, black diamonds for glycines, squares and dotted squares for threonines and serines, respectively). Basic and acidic residues are shown in blue and red, respectively.

Figure 5. CH-ellipticity plot of N_{TAIL} proteins from Hendra and Nipah virus (HeV, NiV). For each NT_{AIL} protein, the distance from the boundary in the CH plot, referred to as CH distance, has been plotted as a function of the distance from the boundary in the $\Theta_{222}/\Theta_{200}$ plot, where the latter is a threshold enabling separating random coil-like and premolten globule-like IDPs. Modified from [112].

Figure 6. **MoRF predictions output generated by four different on-line predictors.** (**A**) ANCHOR2 (blue line) and IUPred2A (red line) scores. (**B**) MoRFchibiSYSTEM results. Red curve: MoRFCHiBi_Web scores, blue areas: predicted MoRFs. (**C**) Annotated sequence plot generated by DISOPRED3. "Disorder" residues and "disorder, protein-binding" residues are framed in blue and green, respectively. Absence of annotations on residues reflects a low confidence level. Predictions have been carried out on a 150 residues artificial disordered protein having an average disorder propensity of 0.75 whose sequence has been generated with InSiDDe, a server for the generation of artificial protein sequences of desired length and disorder score (http://insidde.afmb.univ-mrs.fr/) [145].

Figure 7. Proposed general scheme for prediction of disordered regions in a protein.

References

1. Peng Z, Yan J, Fan X, Mizianty MJ, Xue B, Wang K, Hu G, Uversky VN, Kurgan L (2015) Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. Cellular and molecular life sciences : CMLS 72 (1):137-151. doi:10.1007/s00018-014-1661-9

2. 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. Journal of Molecular Biology 337 (3):635-645

3. Bogatyreva NS, Finkelstein AV, Galzitskaya OV (2006) Trend of amino acid composition of proteins of different taxa. J Bioinform Comput Biol 4 (2):597-608

4. Dunker AK, Babu MM, Barbar E, Blackledge M, Bondos SE, Dosztányi Z, Dyson HJ, Forman-Kay J, Fuxreiter M, Gsponer J, Han K-H, Jones DT, Longhi S, Metallo SJ, Nishikawa K, Nussinov R, Obradovic Z, Pappu RV, Rost B, Selenko P, Subramaniam V, Sussman JL, Tompa P, Uversky VN (2013) What's in a name? Why these proteins are intrinsically disordered. Intrinsically Disordered Proteins 1:e24157

5. Uversky VN (2015) The multifaceted roles of intrinsic disorder in protein complexes. FEBS letters. doi:10.1016/j.febslet.2015.06.004

6. Haynes C, Oldfield CJ, Ji F, Klitgord N, Cusick ME, Radivojac P, Uversky VN, Vidal M, Iakoucheva LM (2006) Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. PLoS Comput Biol 2 (8):e100

7. Habchi J, Tompa P, Longhi S, Uversky VN (2014) Introducing Protein Intrinsic Disorder. Chem Rev 114 (13):6561-6588. doi:10.1021/cr400514h

8. Babu MM (2016) The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. Biochemical Society transactions 44 (5):1185-1200. doi:10.1042/BST20160172

9. Uversky VN (2019) Intrinsically Disordered Proteins and Their "Mysterious" (Meta)Physics. Frontiers in Physics 7 (10). doi:10.3389/fphy.2019.00010

10. Uversky VN (2017) Intrinsically disordered proteins in overcrowded milieu: Membrane-less organelles, phase separation, and intrinsic disorder. Curr Opin Struct Biol 44:18-30. doi:10.1016/j.sbi.2016.10.015

11. Banani SF, Lee HO, Hyman AA, Rosen MK (2017) Biomolecular condensates: organizers of cellular biochemistry. Nature reviews Molecular cell biology 18 (5):285-298. doi:10.1038/nrm.2017.7

12. Shin Y, Brangwynne CP (2017) Liquid phase condensation in cell physiology and disease. Science 357 (6357). doi:10.1126/science.aaf4382

13. Boeynaems S, Alberti S, Fawzi NL, Mittag T, Polymenidou M, Rousseau F, Schymkowitz J, Shorter J, Wolozin B, Van Den Bosch L, Tompa P, Fuxreiter M (2018) Protein Phase Separation: A New Phase in Cell Biology. Trends in cell biology 28 (6):420-435. doi:10.1016/j.tcb.2018.02.004

14. Alberti S, Hyman AA (2021) Biomolecular condensates at the nexus of cellular stress, protein aggregation disease and ageing. Nature reviews Molecular cell biology 22 (3):196-213. doi:10.1038/s41580-020-00326-6

15. Lobley A, Swindells MB, Orengo CA, Jones DT (2007) Inferring function using patterns of native disorder in proteins. PLoS Comput Biol 3 (8):e162

16. Ferron F, Longhi S, Canard B, Karlin D (2006) A practical overview of protein disorder prediction methods. Proteins 65 (1):1-14

17. Ferron F, Rancurel C, Longhi S, Cambillau C, Henrissat B, Canard B (2005) VaZyMolO: a tool to define and classify modularity in viral proteins. Journal of General Virology 86 (Pt 3):743-749

18. Lieutaud P, Ferron F, Habchi J, Canard B, Longhi S (2013) Predicting protein disorder and induced folding : a practical approach. In: Dunn B (ed) Advances in Protein and Peptide Sciences, vol 1. Bentham Science Publishers, pp 441-492 (452)

19. Bourhis JM, Canard B, Longhi S (2007) Predicting protein disorder and induced folding: from theoretical principles to practical applications. Curr Protein Pept Sci 8 (2):135-149

20. Uversky VN, Radivojac P, Iakoucheva LM, Obradovic Z, Dunker AK (2007) Prediction of intrinsic disorder and its use in functional proteomics. Methods Mol Biol 408:69-92

21. He B, Wang K, Liu Y, Xue B, Uversky VN, Dunker AK (2009) Predicting intrinsic disorder in proteins: an overview. Cell Research. doi:cr200987 [pii]10.1038/cr.2009.87

22. Longhi S, Lieutaud P, Canard B (2010) Conformational disorder. Methods Molecular Biology 609:307-325

23. Meng F, Uversky VN, Kurgan L (2017) Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions. Cellular and molecular life sciences : CMLS 74 (17):3069-3090. doi:10.1007/s00018-017-2555-4

24. Liu Y, Wang X, Liu B (2019) A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. Briefings in bioinformatics 20 (1):330-346. doi:10.1093/bib/bbx126

25. Necci M, Piovesan D, Tosatto SCE (2021) Critical assessment of protein intrinsic disorder prediction. Nature methods 18 (5):472-481. doi:10.1038/s41592-021-01117-3

26. Katuwawala A, Peng Z, Yang J, Kurgan L (2019) Computational Prediction of MoRFs, Short Disorder-to-order Transitioning Protein Binding Regions. Computational and Structural Biotechnology Journal 17:454-462. doi:https://doi.org/10.1016/j.csbj.2019.03.013

27. Monastyrskyy B, Kryshtafovych A, Moult J, Tramontano A, Fidelis K (2014) Assessment of protein disorder region predictions in CASP10. Proteins 82 Suppl 2:127-137. doi:10.1002/prot.24391

28. Ishida T, Kinoshita K (2008) Prediction of disordered regions in proteins based on the meta approach. Bioinformatics 24 (11):1344-1348. doi:10.1093/bioinformatics/btn195

29. Lieutaud P, Canard B, Longhi S (2008) MeDor: a metaserver for predicting protein disorder. BMC Genomics 9 ((Suppl 2)):S25

30. Lang B, Babu MM (2021) A community effort to bring structure to disorder. Nature methods 18 (5):454-455. doi:10.1038/s41592-021-01123-5

31. Brown CJ, Johnson AK, Dunker AK, Daughdrill GW (2011) Evolution and disorder. Current Opinion in Structural Biology 21 (3):441-446. doi:S0959-440X(11)00036-4 [pii]10.1016/j.sbi.2011.02.005

32. Oates ME, Romero P, Ishida T, Ghalwash M, Mizianty MJ, Xue B, Dosztanyi Z, Uversky VN, Obradovic Z, Kurgan L, Dunker AK, Gough J (2013) D(2)P(2): database of disordered protein predictions. Nucleic Acids Research 41 (Database issue):D508-516. doi:10.1093/nar/gks1226.

33. Pandurangan AP, Stahlhacke J, Oates ME, Smithers B, Gough J (2019) The SUPERFAMILY 2.0 database: a significant proteome update and a new webserver. Nucleic Acids Res 47 (D1):D490-D494. doi:10.1093/nar/gky1130

34. Hornbeck PV, Zhang B, Murray B, Kornhauser JM, Latham V, Skrzypek E (2015) PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. Nucleic Acids Res 43 (Database issue):D512-520. doi:10.1093/nar/gku1267

35. Potenza E, Di Domenico T, Walsh I, Tosatto SC (2015) MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins. Nucleic Acids Research 43 (Database issue):D315-320. doi:10.1093/nar/gku982

36. Piovesan D, Necci M, Escobedo N, Monzon AM, Hatos A, Micetic I, Quaglia F, Paladin L, Ramasamy P, Dosztanyi Z, Vranken WF, Davey NE, Parisi G, Fuxreiter M, Tosatto SCE (2021) MobiDB: intrinsically disordered proteins in 2021. Nucleic Acids Res 49 (D1):D361-D367. doi:10.1093/nar/gkaa1058

37. Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A, Szabo B, Tompa P, Chen J, Uversky VN, Obradovic Z, Dunker AK (2007) DisProt: the Database of Disordered Proteins. Nucleic Acids Research 35 (Database issue):D786-793

38. Hatos A, Hajdu-Soltesz B, Monzon AM, Palopoli N, Alvarez L, Aykac-Fas B, Bassot C, Benitez GI, Bevilacqua M, Chasapi A, Chemes L, Davey NE, Davidovic R, Dunker AK, Elofsson A, Gobeill J, Foutel NSG, Sudha G, Guharoy M, Horvath T, Iglesias V, Kajava AV, Kovacs OP, Lamb J, Lambrughi M, Lazar T, Leclercq JY, Leonardi E, Macedo-Ribeiro S, Macossay-Castillo M, Maiani E, Manso JA, Marino-Buslje C, Martinez-Perez E, Meszaros B, Micetic I, Minervini G, Murvai N, Necci M, Ouzounis CA, Pajkos M, Paladin L, Pancsa R, Papaleo E, Parisi G, Pasche E, Barbosa Pereira PJ, Promponas VJ, Pujols J, Quaglia F, Ruch P, Salvatore M, Schad E, Szabo B, Szaniszlo T, Tamana S, Tantos A, Veljkovic N, Ventura S, Vranken W, Dosztanyi Z, Tompa P, Tosatto SCE, Piovesan D (2020) DisProt: intrinsic protein disorder annotation in 2020. Nucleic Acids Res 48 (D1):D269-D276. doi:10.1093/nar/gkz975

39. Fukuchi S, Amemiya T, Sakamoto S, Nobe Y, Hosoda K, Kado Y, Murakami SD, Koike R, Hiroaki H, Ota M (2014) IDEAL in 2014 illustrates interaction networks composed of intrinsically

disordered proteins and their binding partners. Nucleic Acids Research 42 (Database issue):D320-325. doi:10.1093/nar/gkt1010

40. Zhao B, Katuwawala A, Oldfield CJ, Dunker AK, Faraggi E, Gsponer J, Kloczkowski A, Malhis N, Mirdita M, Obradovic Z, Soding J, Steinegger M, Zhou Y, Kurgan L (2021) DescribePROT: database of amino acid-level protein structure and function predictions. Nucleic Acids Res 49 (D1):D298-D308. doi:10.1093/nar/gkaa931

41. Varadi M, Kosol S, Lebrun P, Valentini E, Blackledge M, Dunker AK, Felli IC, Forman-Kay JD, Kriwacki RW, Pierattelli R, Sussman J, Svergun DI, Uversky VN, Vendruscolo M, Wishart D, Wright PE, Tompa P (2014) pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. Nucleic Acids Research 42 (Database issue):D326-335. doi:10.1093/nar/gkt960

42. Lazar T, Martinez-Perez E, Quaglia F, Hatos A, Chemes LB, Iserte JA, Mendez NA, Garrone NA, Saldano TE, Marchetti J, Rueda AJV, Bernado P, Blackledge M, Cordeiro TN, Fagerberg E, Forman-Kay JD, Fornasari MS, Gibson TJ, Gomes GW, Gradinaru CC, Head-Gordon T, Jensen MR, Lemke EA, Longhi S, Marino-Buslje C, Minervini G, Mittag T, Monzon AM, Pappu RV, Parisi G, Ricard-Blum S, Ruff KM, Salladini E, Skepo M, Svergun D, Vallet SD, Varadi M, Tompa P, Tosatto SCE, Piovesan D (2021) PED in 2021: a major update of the protein ensemble database for intrinsically disordered proteins. Nucleic Acids Res 49 (D1):D404-D411. doi:10.1093/nar/gkaa1021

43. Schad E, Ficho E, Pancsa R, Simon I, Dosztanyi Z, Meszaros B (2018) DIBS: a repository of disordered binding sites mediating interactions with ordered proteins. Bioinformatics 34 (3):535-537. doi:10.1093/bioinformatics/btx640

44. Ficho E, Remenyi I, Simon I, Meszaros B (2017) MFIB: a repository of protein complexes with mutual folding induced by binding. Bioinformatics 33 (22):3682-3684. doi:10.1093/bioinformatics/btx486

45. Miskei M, Antal C, 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 (D1):D228-D235. doi:10.1093/nar/gkw1019

46. Vucetic S, Brown C, Dunker K, Obradovic Z (2003) Flavors of protein disorder. Proteins 52:573-584.

47. Karlin D, Ferron F, Canard B, Longhi S (2003) Structural disorder and modular organization in Paramyxovirinae N and P. Journal of General Virology 84 (Pt 12):3239-3252

48. Severson W, Xu X, Kuhn M, Senutovitch N, Thokala M, Ferron F, Longhi S, Canard B, Jonsson CB (2005) Essential amino acids of the hantaan virus N protein in its interaction with RNA. Journal of Virology 79 (15):10032-10039

49. Llorente MT, Barreno-Garcia B, Calero M, Camafeita E, Lopez JA, Longhi S, Ferron F, Varela PF, Melero JA (2006) Structural analysis of the human respiratory syncitial virus phosphoprotein: characterization of an a-helical domain involved in oligomerization. Journal of General Virology 87:159-169

50. Habchi J, Mamelli L, Darbon H, Longhi S (2010) Structural Disorder within Henipavirus Nucleoprotein and Phosphoprotein: From Predictions to Experimental Assessment. PLoS ONE 5 (7):e11684. doi:10.1371/journal.pone.0011684

51. Deng X, Eickholt J, Cheng J (2009) PreDisorder: ab initio sequence-based prediction of protein disordered regions. BMC Bioinformatics 10:436. doi:1471-2105-10-436 [pii]10.1186/1471-2105-10-436

52. Noivirt-Brik O, Prilusky J, Sussman JL (2009) Assessment of disorder predictions in CASP8. Proteins 77 Suppl 9:210-216. doi:10.1002/prot.22586

53. Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK (2001) Sequence complexity of disordered proteins. Proteins 42 (1):38-48

54. Obradovic Z, Peng K, Vucetic S, Radivojac P, Dunker AK (2005) Exploiting heterogeneous sequence properties improves prediction of protein disorder. Proteins 61 (Suppl. 7):176-182

55. Bordoli L, Kiefer F, Schwede T (2007) Assessment of disorder predictions in CASP7. Proteins 69 Suppl 8:129-136. doi:10.1002/prot.21671

56. Obradovic Z, Peng K, Vucetic S, Radivojac P, Brown CJ, Dunker AK (2003) Predicting intrinsic disorder from amino acid sequence. Proteins 53 Suppl 6:566-572

57. Linding R, Russell RB, Neduva V, Gibson TJ (2003) GlobPlot: Exploring protein sequences for globularity and disorder. Nucleic Acids Research 31 (13):3701-3708

58. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB (2003) Protein disorder prediction: implications for structural proteomics. Structure (Camb) 11 (11):1453-1459

59. Ward JJ, McGuffin LJ, Bryson K, Buxton BF, Jones DT (2004) The DISOPRED server for the prediction of protein disorder. Bioinformatics 20 (13):2138-2139

60. Orlando G, Raimondi D, Codice F, Tabaro F, Vranken W (2020) Prediction of disordered regions in proteins with recurrent Neural Networks and protein dynamics. bioRxiv:2020.2005.2025.115253. doi:10.1101/2020.05.25.115253

61. Ramraj V (2014) Exploiting Whole-PDB Analysis in Novel Bioinformatics Applications. University of Oxford,

62. Yang ZR, Thomson R, McNeil P, Esnouf RM (2005) RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. Bioinformatics 21 (16):3369-3376. doi:10.1093/bioinformatics/bti534

63. Lobanov MY, Galzitskaya OV (2011) The Ising model for prediction of disordered residues from protein sequence alone. Physical biology 8 (3):035004. doi:10.1088/1478-3975/8/3/035004

64. Lobanov MY, Sokolovskiy IV, Galzitskaya OV (2013) IsUnstruct: prediction of the residue status to be ordered or disordered in the protein chain by a method based on the Ising model. Journal of biomolecular structure & dynamics 31 (10):1034-1043. doi:10.1080/07391102.2012.718529

65. Meng F, Kurgan L (2016) DFLpred: High-throughput prediction of disordered flexible linker regions in protein sequences. Bioinformatics 32 (12):i341-i350. doi:10.1093/bioinformatics/btw280

66. Cheng J, Sweredoski M, Baldi P (2005) Accurate Prediction of Protein Disordered Regions by Mining Protein Structure Data, Data Mining and Knowledge Discovery. 11:213-222

67. Pollastri G, McLysaght A (2005) Porter: a new, accurate server for protein secondary structure prediction. Bioinformatics 21 (8):1719-1720

68. Walsh I, Martin AJ, Di Domenico T, Tosatto SC (2012) ESpritz: accurate and fast prediction of protein disorder. Bioinformatics 28 (4):503-509. doi:10.1093/bioinformatics/btr682

69. Ishida T, Kinoshita K (2007) PrDOS: prediction of disordered protein regions from amino acid sequence. Nucleic Acids Research 35 (Web Server issue):W460-464. doi:gkm363 [pii]10.1093/nar/gkm363

70. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25 (17):3389-3402. doi:10.1093/nar/25.17.3389

71. Hanson J, Paliwal K, Litfin T, Yang Y, Zhou Y (2019) Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks. Bioinformatics 35 (14):2403-2410. doi:10.1093/bioinformatics/bty1006

72. Hanson J, Paliwal KK, Litfin T, Zhou Y (2019) SPOT-Disorder2: Improved Protein Intrinsic Disorder Prediction by Ensembled Deep Learning. Genomics, proteomics & bioinformatics 17 (6):645-656. doi:10.1016/j.gpb.2019.01.004

73. Hanson J, Paliwal K, Zhou Y (2018) Accurate Single-Sequence Prediction of Protein Intrinsic Disorder by an Ensemble of Deep Recurrent and Convolutional Architectures. Journal of chemical information and modeling 58 (11):2369-2376. doi:10.1021/acs.jcim.8b00636

74. Tang YJ, Pang YH, Liu B (2020) IDP-Seq2Seq: Identification of Intrinsically Disordered Regions based on Sequence to Sequence Learning. Bioinformatics. doi:10.1093/bioinformatics/btaa667

75. Wang S, Ma J, Xu J (2016) AUCpreD: proteome-level protein disorder prediction by AUC-maximized deep convolutional neural fields. Bioinformatics 32 (17):i672-i679. doi:10.1093/bioinformatics/btw446

76. Meszaros B, Erdos G, Dosztanyi Z (2018) IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. Nucleic Acids Res 46 (W1):W329-W337. doi:10.1093/nar/gky384

77. Galzitskaya OV, Garbuzynskiy SO, Lobanov MY (2006) FoldUnfold: web server for the prediction of disordered regions in protein chain. Bioinformatics 22 (23):2948-2949

78. Meszaros B, Simon I, Dosztanyi Z (2009) Prediction of protein binding regions in disordered proteins. PLoS Comput Biol 5 (5):e1000376. doi:10.1371/journal.pcbi.1000376

79. Cilia E, Pancsa R, Tompa P, Lenaerts T, Vranken WF (2014) The DynaMine webserver: predicting protein dynamics from sequence. Nucleic Acids Res 42 (Web Server issue):W264-270. doi:10.1093/nar/gku270

80. Cilia E, Pancsa R, Tompa P, Lenaerts T, Vranken WF (2013) From protein sequence to dynamics and disorder with DynaMine. Nature communications 4:2741. doi:10.1038/ncomms3741

81. Sormanni P, Camilloni C, Fariselli P, Vendruscolo M (2015) The s2D method: simultaneous sequence-based prediction of the statistical populations of ordered and disordered regions in proteins. J Mol Biol 427 (4):982-996. doi:10.1016/j.jmb.2014.12.007

82. Necci M, Piovesan D, Clementel D, Dosztanyi Z, Tosatto SCE (2020) MobiDB-lite 3.0: fast consensus annotation of intrinsic disorder flavours in proteins. Bioinformatics. doi:10.1093/bioinformatics/btaa1045

83. Iqbal S, Hoque MT (2016) Estimation of Position Specific Energy as a Feature of Protein Residues from Sequence Alone for Structural Classification. PLoS ONE 11 (9):e0161452. doi:10.1371/journal.pone.0161452

84. Faraggi E, Xue B, Zhou Y (2009) Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network. Proteins 74 (4):847-856. doi:10.1002/prot.22193

85. Asgari E, Mofrad MR (2015) Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. PLoS ONE 10 (11):e0141287. doi:10.1371/journal.pone.0141287

86. Kim SS, Seffernick JT, Lindert S (2018) Accurately Predicting Disordered Regions of Proteins Using Rosetta ResidueDisorder Application. The journal of physical chemistry B 122 (14):3920-3930. doi:10.1021/acs.jpcb.8b01763

87. Seffernick JT, Ren H, Kim SS, Lindert S (2019) Measuring Intrinsic Disorder and Tracking Conformational Transitions Using Rosetta ResidueDisorder. The journal of physical chemistry B 123 (33):7103-7112. doi:10.1021/acs.jpcb.9b04333

88. Uversky VN, Gillespie JR, Fink AL (2000) Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins 41 (3):415-427

89. Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157 (1):105-132. doi:10.1016/0022-2836(82)90515-0

90. Zeev-Ben-Mordehai T, Rydberg EH, Solomon A, Toker L, Auld VJ, Silman I, Botti S, Sussman JL (2003) The intracellular domain of the Drosophila cholinesterase-like neural adhesion protein, gliotactin, is natively unfolded. Proteins 53 (3):758-767

91. Oldfield CJ, Cheng Y, Cortese MS, Brown CJ, Uversky VN, Dunker AK (2005) Comparing and Combining Predictors of Mostly Disordered Proteins. Biochemistry 44 (6):1989-2000

92. Xue B, Oldfield CJ, Dunker AK, Uversky VN (2009) CDF it all: consensus prediction of intrinsically disordered proteins based on various cumulative distribution functions. FEBS letters 583 (9):1469-1474. doi:S0014-5793(09)00260-9 [pii]10.1016/j.febslet.2009.03.070

93. Mohan A, Sullivan WJ, Jr., Radivojac P, Dunker AK, Uversky VN (2008) Intrinsic disorder in pathogenic and non-pathogenic microbes: discovering and analyzing the unfoldomes of early-branching eukaryotes. Mol Biosyst 4 (4):328-340

94. Bitard-Feildel T, Lamiable A, Mornon JP, Callebaut I (2018) Order in Disorder as Observed by the "Hydrophobic Cluster Analysis" of Protein Sequences. Proteomics 18 (21-22):e1800054. doi:10.1002/pmic.201800054

95. Callebaut I, Labesse G, Durand P, Poupon A, Canard L, Chomilier J, Henrissat B, Mornon JP (1997) Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. Cellular and Molecular Life Sciences 53 (8):621-645.

96. Kozlowski LP, Bujnicki JM (2012) MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins. BMC Bioinformatics 13 (1):111. doi:1471-2105-13-111 [pii]10.1186/1471-2105-13-111

97. Li J, Deng X, Eickholt J, Cheng J (2013) Designing and benchmarking the MULTICOM protein structure prediction system. BMC structural biology 13:2. doi:10.1186/1472-6807-13-2

98. Hou J, Wu T, Cao R, Cheng J (2019) Protein tertiary structure modeling driven by deep learning and contact distance prediction in CASP13. Proteins 87 (12):1165-1178. doi:10.1002/prot.25697

99. Barik A, Katuwawala A, Hanson J, Paliwal K, Zhou Y, Kurgan L (2020) DEPICTER: Intrinsic Disorder and Disorder Function Prediction Server. J Mol Biol 432 (11):3379-3387. doi:10.1016/j.jmb.2019.12.030

100. Mizianty MJ, Stach W, Chen K, Kedarisetti KD, Disfani FM, Kurgan L (2010) Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. Bioinformatics 26 (18):i489-496. doi:btq373 [pii]10.1093/bioinformatics/btq373

101. Mizianty MJ, Uversky V, Kurgan L (2014) Prediction of intrinsic disorder in proteins using MFDp2. Methods Mol Biol 1137:147-162. doi:10.1007/978-1-4939-0366-5_11

102. Fan X, Kurgan L (2014) Accurate prediction of disorder in protein chains with a comprehensive and empirically designed consensus. Journal of biomolecular structure & dynamics 32 (3):448-464. doi:10.1080/07391102.2013.775969

103. Oldfield CJ, Fan X, Wang C, Dunker AK, Kurgan L (2020) Computational Prediction of Intrinsic Disorder in Protein Sequences with the disCoP Meta-predictor. Methods Mol Biol 2141:21-35. doi:10.1007/978-1-0716-0524-0_2

104. Xue B, Dunbrack RL, Williams RW, Dunker AK, Uversky VN (2010) PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. Biochimica et Biophysica Acta (BBA) - Bioenergetics 1804 (4):996-1010. doi:S1570-9639(10)00013-0 [pii]10.1016/j.bbapap.2010.01.011

105. Schlessinger A, Liu J, Rost B (2007) Natively unstructured loops differ from other loops. PLoS Comput Biol 3 (7):e140. doi:10.1371/journal.pcbi.0030140

106. Schlessinger A, Punta M, Rost B (2007) Natively unstructured regions in proteins identified from contact predictions. Bioinformatics 23 (18):2376-2384

107. Schlessinger A, Punta M, Yachdav G, Kajan L, Rost B (2009) Improved disorder prediction by combination of orthogonal approaches. PLoS ONE 4 (2):e4433. doi:10.1371/journal.pone.0004433

108. Schlessinger A, Yachdav G, Rost B (2006) PROFbval: predict flexible and rigid residues in proteins. Bioinformatics 22 (7):891-893. doi:10.1093/bioinformatics/btl032

109. Chandonia JM (2007) StrBioLib: a Java library for development of custom computational structural biology applications. Bioinformatics 23 (15):2018-2020

110. Necci M, Piovesan D, Dosztanyi Z, Tosatto SCE (2017) MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins. Bioinformatics 33 (9):1402-1404. doi:10.1093/bioinformatics/btx015

111. Katuwawala A, Ghadermarzi S, Hu G, Wu Z, Kurgan L (2021) QUARTERplus: Accurate disorder predictions integrated with interpretable residue-level quality assessment scores. Computational and Structural Biotechnology Journal 19:2597-2606. doi:https://doi.org/10.1016/j.csbj.2021.04.066

112. Blocquel D, Habchi J, Gruet A, Blangy S, Longhi S (2012) Compaction and binding properties of the intrinsically disordered C-terminal domain of Henipavirus nucleoprotein as unveiled by deletion studies. Mol Biosyst 8 (1):392-410. doi:10.1039/c1mb05401e

113. Uversky VN (2002) Natively unfolded proteins: a point where biology waits for physics. Protein Science 11 (4):739-756.

114. Oldfield CJ, Cheng Y, Cortese MS, Romero P, Uversky VN, Dunker AK (2005) Coupled folding and binding with alpha-helix-forming molecular recognition elements. Biochemistry 44 (37):12454-12470. doi:10.1021/bi050736e

115. Cheng Y, Oldfield CJ, Meng J, Romero P, Uversky VN, Dunker AK (2007) Mining alpha-helix-forming molecular recognition features with cross species sequence alignments. Biochemistry 46 (47):13468-13477. doi:10.1021/bi7012273

116. Vacic V, Oldfield CJ, Mohan A, Radivojac P, Cortese MS, Uversky VN, Dunker AK (2007) Characterization of molecular recognition features, MoRFs, and their binding partners. J Proteome Res 6 (6):2351-2366

117. Bourhis J, Johansson K, Receveur-Bréchot V, Oldfield CJ, Dunker AK, Canard B, Longhi S (2004) The C-terminal domain of measles virus nucleoprotein belongs to the class of intrinsically disordered proteins that fold upon binding to their physiological partner. Virus Research 99:157-167

118. John SP, Wang T, Steffen S, Longhi S, Schmaljohn CS, Jonsson CB (2007) Ebola virus VP30 is an RNA binding protein. Journal of Virology 81 (17):8967-8976

119. Meszaros B, Tompa P, Simon I, Dosztanyi Z (2007) Molecular principles of the interactions of disordered proteins. Journal of Molecular Biology 372 (2):549-561

120. Habchi J, Blangy S, Mamelli L, Ringkjobing Jensen M, Blackledge M, Darbon H, Oglesbee M, Shu Y, Longhi S (2011) Characterization of the interactions between the nucleoprotein and the phosphoprotein of Henipaviruses. Journal of Biological Chemistry 286 (15):13583-13602

121. He H, Zhao J, Sun G (2019) Computational prediction of MoRFs based on protein sequences and minimax probability machine. BMC Bioinformatics 20 (1):529. doi:10.1186/s12859-019-3111-z

122. Sharma R, Kumar S, Tsunoda T, Patil A, Sharma A (2016) Predicting MoRFs in protein sequences using HMM profiles. BMC Bioinformatics 17 (Suppl 19):504. doi:10.1186/s12859-016-1375-0

123. Sharma R, Bayarjargal M, Tsunoda T, Patil A, Sharma A (2018) MoRFPred-plus: Computational Identification of MoRFs in Protein Sequences using Physicochemical Properties and HMM profiles. Journal of theoretical biology 437:9-16. doi:10.1016/j.jtbi.2017.10.015

124. Xue B, Dunker AK, Uversky VN (2010) Retro-MoRFs: identifying protein binding sites by normal and reverse alignment and intrinsic disorder prediction. International journal of molecular sciences 11 (10):3725-3747. doi:10.3390/ijms11103725

125. Fang C, Moriwaki Y, Zhu D, Shimizu K (2018) Identifying MoRFs in Disordered Proteins Using Enlarged Conserved Features. Paper presented at the Proceedings of the 2018 6th International Conference on Bioinformatics and Computational Biology, Chengdu, China,

126. Fang C, Noguchi T, Tominaga D, Yamana H (2013) MFSPSSMpred: identifying short disorder-to-order binding regions in disordered proteins based on contextual local evolutionary conservation. BMC Bioinformatics 14:300. doi:10.1186/1471-2105-14-300

127. Hanson J, Litfin T, Paliwal K, Zhou Y (2020) Identifying molecular recognition features in intrinsically disordered regions of proteins by transfer learning. Bioinformatics 36 (4):1107-1113. doi:10.1093/bioinformatics/btz691

128. Dosztanyi Z, Meszaros B, Simon I (2009) ANCHOR: web server for predicting protein binding regions in disordered proteins. Bioinformatics 25 (20):2745-2746. doi:btp518 [pii]10.1093/bioinformatics/btp518

129. Disfani FM, Hsu WL, Mizianty MJ, Oldfield CJ, Xue B, Dunker AK, Uversky VN, 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 (12):i75-83. doi:bts209 [pii]10.1093/bioinformatics/bts209

130. Yan J, Dunker AK, Uversky VN, Kurgan L (2016) Molecular recognition features (MoRFs) in three domains of life. Mol Biosyst 12 (3):697-710. doi:10.1039/c5mb00640f

131. Malhis N, Jacobson M, Gsponer J (2016) MoRFchibi SYSTEM: software tools for the identification of MoRFs in protein sequences. Nucleic Acids Res 44 (W1):W488-493. doi:10.1093/nar/gkw409

132. Jones DT, Cozzetto D (2015) DISOPRED3: precise disordered region predictions with annotated protein-binding activity. Bioinformatics 31 (6):857-863. doi:10.1093/bioinformatics/btu744

133. Sharma R, Raicar G, Tsunoda T, Patil A, Sharma A (2018) OPAL: prediction of MoRF regions in intrinsically disordered protein sequences. Bioinformatics 34 (11):1850-1858. doi:10.1093/bioinformatics/bty032

134. Sharma R, Sharma A, Raicar G, Tsunoda T, Patil A (2019) OPAL+: Length-Specific MoRF Prediction in Intrinsically Disordered Protein Sequences. Proteomics 19 (6):e1800058. doi:10.1002/pmic.201800058

135. Peng Z, Kurgan L (2015) High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. Nucleic Acids Res 43 (18):e121. doi:10.1093/nar/gkv585

136. McGuffin LJ, Bryson K, Jones DT (2000) The PSIPRED protein structure prediction server. Bioinformatics 16 (4):404-405.

137. Wootton JC (1994) Non-globular domains in protein sequences: automated segmentation using complexity measures. Comput Chem 18 (3):269-285.

138. Kall L, Krogh A, Sonnhammer EL (2007) Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. Nucleic Acids Research 35 (Web Server issue):W429-432

139. Bornberg-Bauer E, Rivals E, Vingron M (1998) Computational approaches to identify leucine zippers. Nucleic Acids Research 26 (11):2740-2746

140. Lupas A, Van Dyke M, Stock J (1991) Predicting coiled coils from protein sequences. Science 252 (5009):1162-1164

141. Baldi P, Cheng J, Vullo A (2004) Large-scale prediction of disulphide bond connectivity. Adv Neural Inf Process Syst 17:97-104

142. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar Gustavo A, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, Finn RD, Bateman A (2020) Pfam: The protein families database in 2021. Nucleic Acids Research 49 (D1):D412-D419. doi:10.1093/nar/gkaa913

143. Sillitoe I, Bordin N, Dawson N, Waman VP, Ashford P, Scholes HM, Pang CSM, Woodridge L, Rauer C, Sen N, Abbasian M, Le Cornu S, Lam SD, Berka K, Varekova Ivana H, Svobodova

R, Lees J, Orengo CA (2020) CATH: increased structural coverage of functional space. Nucleic Acids Research 49 (D1):D266-D273. doi:10.1093/nar/gkaa1079

144. Eudes R, Le Tuan K, Delettre J, Mornon JP, Callebaut I (2007) A generalized analysis of hydrophobic and loop clusters within globular protein sequences. BMC structural biology 7:2. doi:10.1186/1472-6807-7-2

145. Schramm A, Lieutaud P, Gianni S, Longhi S, Bignon C (2017) InSiDDe: A Server for Designing Artificial Disordered Proteins. International journal of molecular sciences 19 (1). doi:10.3390/ijms19010091