D508–D516 *Nucleic Acids Research, 2013, Vol. 41, Database issue* *Published online 29 November 2012*
doi:10.1093/nar/gks1226

D^2P^2: database of disordered protein predictions

Matt E. Oates[1,2,*], Pedro Romero[3], Takashi Ishida[4], Mohamed Ghalwash[5,6], Marcin J. Mizianty[7], Bin Xue[8], Zsuzsanna Dosztányi[9], Vladimir N. Uversky[8,10], Zoran Obradovic[5], Lukasz Kurgan[7], A. Keith Dunker[3] and Julian Gough[1]

[1]Department of Computer Science, University of Bristol, Bristol BS8 1UB, UK, [2]Bristol Centre for Complexity Sciences, University of Bristol, Bristol BS8 1TR, UK, [3]Center for Computational Biology and Bioinformatics, Indiana University, Indianapolis, IN 46202-5122, USA, [4]Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology, Tokyo 152-8552, Japan, [5]Center for Data Analytics and Biomedical Informatics, College of Science and Technology, Temple University, Philadelphia, PA 19122, USA, [6]Mathematics Department, Faculty of Science, Ain Shams University, Cairo 11566, Egypt, [7]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta, Canada T6G 2V4, [8]Department of Molecular Medicine, University of South Florida, Tampa, FL MDC07, USA, [9]Institute of Enzymology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Budapest 1113, Hungary and [10]Institute for Biological Instrumentation, Russian Academy of Sciences, 142290 Pushchino, Moscow Region, Russia

Received August 15, 2012; Revised and Accepted October 31, 2012

ABSTRACT

We present the Database of Disordered Protein Prediction (D^2P^2), available at http://d2p2.pro (including website source code). A battery of disorder predictors and their variants, VL-XT, VSL2b, PrDOS, PV2, Espritz and IUPred, were run on all protein sequences from 1765 complete proteomes (to be updated as more genomes are completed). Integrated with these results are all of the predicted (mostly structured) SCOP domains using the SUPERFAMILY predictor. These disorder/structure annotations together enable comparison of the disorder predictors with each other and examination of the overlap between disordered predictions and SCOP domains on a large scale. D^2P^2 will increase our understanding of the interplay between disorder and structure, the genomic distribution of disorder, and its evolutionary history. The parsed data are made available in a unified format for download as flat files or SQL tables either by genome, by predictor, or for the complete set. An interactive website provides a graphical view of each protein annotated with the SCOP domains and disordered regions from all predictors overlaid (or shown as a consensus). There are statistics and tools for browsing and comparing genomes and their disorder within the context of their position on the tree of life.

INTRODUCTION

Intrinsically disordered or unstructured proteins (IDPs) exist as highly flexible polypeptide chains *in vivo* behaving as an ensemble of conformational states with no stable tertiary structure (1). Regions of IDP can exist as unfolded chains or molten globules with well-developed secondary structure and often function through transition between differently folded states (2). Mechanisms for functional conformational transition include binding with other proteins, nucleic acids, various small molecules and numerous posttranslational modifications, such as phosphorylation, which has been shown to be especially important (3,4). Biological functions of known IDPs are varied and their roles include the following: instigation of protein complex formation, molecular recognition as seen in nucleoporins of the nuclear pore complex (5), signal transduction, transcriptional regulation and many other functions (6,7) too numerous to list here.

Much work has been done on producing classification and annotation of known unstructured regions from 3D experimental data found in the PDB (8), DisProt (9) and IDEAL (10) resources. However, the past focus on structured protein domains have limited the total number of described IDP regions. For example: the current release of DisProt (v6.00; 1 July 2012) describes 667 proteins containing 1467 verified disordered regions; and the IDEAL database (as of 9 May 2012) describes 209 disordered proteins in detail, 97 of which have been experimentally verified to be structured and disordered over the same region under different conditions; and also MobiDB (11)

*To whom correspondence should be addressed. Tel: +44 7963 096805; Fax: +44 1179 545208; Email: Matt.Oates@bristol.ac.uk

has applied a method for identifying mobile regions from nuclear magnetic resonance (NMR) structures (12) to 26 933 proteins (v1.2.1 as of 1 November 2012). Because of the biases of structural resolution and the relative ease of task for disorder prediction methods compared with *de novo* fold prediction, many algorithms have been developed to discover novel regions of disorder from amino acid sequence alone (13,14).

Disorder prediction has been proven effective even in novel protein space, starting from the first predictor of protein disorder (15) in 1997 followed with over a decade of comparative study since the CASP5 assessment in 2002. The Database of Disordered Protein Prediction (D^2P^2) database aims to provide unified and exhaustive disorder predictions for all currently sequenced genomes with protein annotations. These data will both further development of current and future predictors; in addition to providing the biological community with ready access to results from predictors not chosen just for availability and ease of use, but for their predictive efficacy and coverage of protein disorder properties.

Two major classes of protein disorder can be described; short regions, typically $<\sim$15–20 residues often serving as flexible linkers between or within domains, and long regions of $>$30–50 residues. These two classes have different amino acid propensities and frequently two prediction methods or variants are required to get full coverage (16). This sort of behavior can be a barrier to the biological investigator who wishes to have quick access to results. Resources such as MetaDisorder (17) go some way to resolve this issue providing the naïve user with reliable results from a similar spectrum of predictors as D^2P^2. However, such meta-submission tools are of limited use if your study involves protein sequences at the scale of a whole genome or clade. The MobiDB resource has some similar goals to D^2P^2 including predictions from Espritz and IUPred pre-computed for 4 662 776 sequences (MobiDB v1.2.1 as of 1 November 2012) from UniProt. D^2P^2 provides a library of predictions (Figure 1 for example) for a set of 10 429 761 protein sequences from 1765 complete genomes (Table 1) and a growing collection of predictors. A focus of the D^2P^2 data is providing access to SCOP domain prediction alongside disorder to show the interplay of known structure and disorder. The Dichot system (18) provides complimentary disordered/structured data for UniProt Human sequences and the DISOPRED2 (19) disorder predictor: deemed too computationally expensive to be included with the full D^2P^2 dataset. Conclusions on the relation of structure and disorder made with Dichot on human sequence and those made with D^2P^2 are discussed later.

Users of D^2P^2 will be those asking basic science questions at the scale of whole genomes or the whole tree of life, or those seeking to develop methods for prediction and wishing to know the specific behaviors of each predictor over a large library of sequence. Additionally, D^2P^2 data highlight the inverse of well-folded structure and could be informative for developing better approaches to fold prediction, as well as screening novel domain families in conserved protein sequence awaiting crystallographic study.

MATERIALS AND METHODS

Sequences

The sequence library of complete genomes from SUPERFAMILY 1.75 as of 8 November 2011 was used as the basis for all prediction results. This includes 1765 complete genomes from 1256 distinct species from across the whole tree of cellular life (Table 1). Currently, viral genomes are not included in D^2P^2 but they will be included in future updates.

Predictions

D^2P^2 currently includes the following: PONDR VL-XT, PONDR VSL2b, PrDOS, PV2, Espritz (all variants) and IUPred (all variants) along with ANCHOR to predict disordered regions that undergo binding transitions during protein–protein interaction.

PONDR® VL-XT: PONDR (*P*redictor *O*f *N*atural *D*isordered *R*egions) is a set of neural network predictors of disordered regions on the basis of local amino acid composition, flexibility, hydropathy, coordination number and other factors. These predictors classify each residue within a sequence as either ordered or disordered. PONDR VL-XT integrates three feed forward neural networks: the Variously characterized Long, version 1 (VL1) predictor from Romero *et al*. (20), which predicts non-terminal residues, and the X-ray characterized N- and C- terminal predictors (XT) from Li *et al*. (21), which predicts terminal residues. Output for the VL1 predictor starts and ends 11 amino acids from the termini. The XT predictors output provides predictions up to 14 amino acids from their respective ends. A simple average is taken for the overlapping predictions; and a sliding window of 9 amino acids is used to smooth the prediction values along the length of the sequence. Unsmoothed prediction values from the XT predictors are used for the first and last four sequence positions.

PONDR® VSL2 is a combination of neural network predictors for both short and long disordered regions (16). A length limit of 30 residues divides short- and long-disordered regions. Each individual predictor is trained by the dataset containing sequences of that specific length. The final prediction is a weighted average determined by a second layer predictor (16). PONDR® VSL2 applies not only the sequence profile but also the result of sequence alignments from PSI-BLAST (22) and secondary structure prediction from PHD (23) and PSIPRED (24). This predictor is so far the most accurate predictor in the PONDR family (25).

PrDOS is composed of two predictors. The first predictor is implemented using a support vector machine with a position-specific profile of local amino acid sequence. A similar concept to how PSIPRED (24) predicts local secondary structure features. The second predictor assumes the conservation of intrinsic disorder in homologous protein domain families (19,26) and is implemented using PSI-BLAST (22) and a novel measure of disorder (27). The final prediction is taken as the combination of the results of the two predictors described.

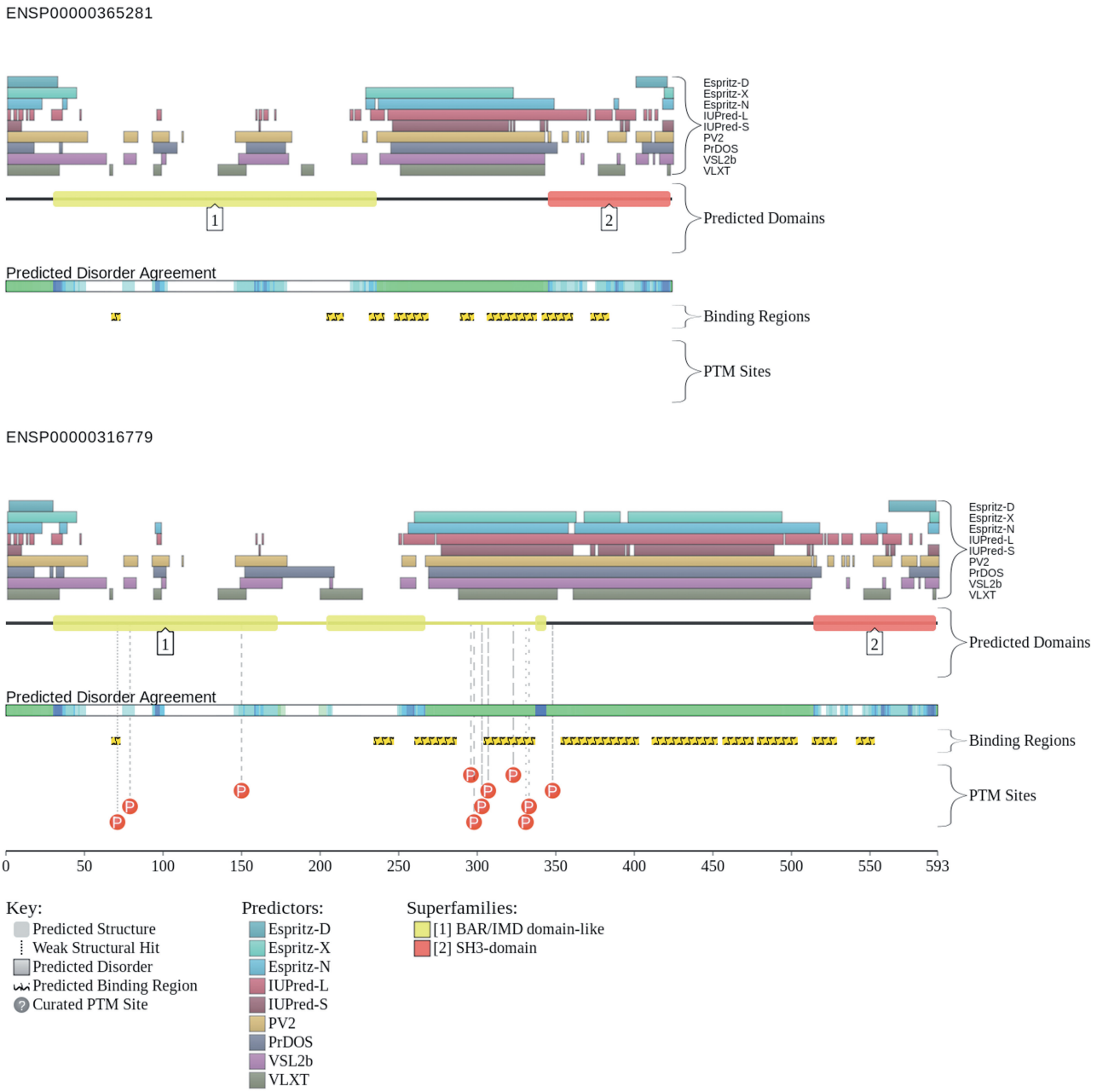


Figure 1. An example graphical report from the D^2P^2 website for two transcripts of the human gene BIN1. All disorder predictions (pastel-colored blocks) are stacked and aligned against the polypeptide chain in black. Their interplay with the predicted SCOP domains (bright-colored rounded blocks) is shown. The level of agreement between all of the disorder predictors is shown as color intensity in an aligned gradient bar below the stack of predictions. The green segments represent disorder that is not found within a predicted SCOP domain. The blue segments are where the disorder predictions intersect the SCOP domain prediction. Below the disorder agreement line, ANCHOR binding region predictions are displayed (yellow blocks with zigzag infill), along with PTM sites from PhosphoSitePlus when known (shown as lettered spheres hanging below other predictions).

Table 1. The number of genomes and sequences included in the database at the time of writing

Domain	Number of genomes	Reference species	Strains	Total sequences
Eukarya	352	298	54	5 746 620
Bacteria	1305	862	443	4 216 314
Archaea	108	96	12	238 232
Total	1765	1256	509	10 429 761

The intention is to expand this over time as new genomes are described.

PV2 is a meta-predictor that was built upon five prediction methodologies trained on different disordered protein datasets: logistic regression, a neural network, a support vector machine, a conditional random field and finally VSL2B to capture the correlation between the neighboring residues. The PV2 meta-prediction reports a residue as disordered if any two of the underlying methods agree on a disordered state (28). The meta-predictor PV2 achieved either higher or comparable accuracy with other methods in both CASP8 and CASP9 sequences

and it had a good balance between sensitivity and specificity. The PV2 meta-predictor was also reliable on the structured domains predictions and it was in the top eight disorder predictors in CASP9 (29) for balanced accuracy.

Espritz predicts three variants of disorder using bidirectional recursive neural networks trained on the following datasets: PDB X-ray crystallography of short disorder (Espritz-X), NMR mobility (Espritz-N) and DisProt data for long disorder (Espritz-D). Either method can be run with a fast or slower variant of the algorithm (30). Because of the wide genomic scale of this database the fast variant was used. Additionally, the following cut-offs were used for the scores (probabilities) generated by each Espritz flavor to yield 5% false positive rate: Esprit-X 0.1434, Espritz-N 0.3089 and Espritz-D 0.5072.

IUPred assumes that the core of a well-structured globular protein has amino acids that can make enough favorable contacts to form a stable 3D structure. A matrix of amino acid pairs holds estimates of their interaction energies which is then used with a position-specific scoring method to predict when stretches of amino acids are not contributing to a stable structure (31). Additionally, IUPred includes both a short (IUPred-S) and long (IUPred-L) variant of its scoring method.

ANCHOR is a predictor of binding sites within disordered regions. It uses the same energy estimation method that underlies IUPred to predict disordered regions in general. ANCHOR finds regions that cannot form enough favorable interactions on their own to form a stable structure, but could gain energy by interacting with a globular structure (32). These sites are often the basis of short linear motifs important in binding to the surface of partner proteins or structured domains in the same polypeptide chain. This property can be functional for both inhibition of an active site or for mediating dynamic protein–protein interactions and complex formation.

SUPERFAMILY is a library of hidden Markov models (HMMs) for SCOP structural domain classifications at the superfamily and family level (33). Assignments from the 1.75 version of the HMM library (34) were used to provide predictions of SCOP structural domains (35,36). *E* value cut offs used were identical to those in the SUPERFAMILY online resource with the assignments coming directly from a mirror of the source database. When new HMM models and SCOP classifications are added to SUPERFAMILY new annotations will automatically be shown in D^2P^2.

D^2P^2 Consensus was calculated at 25, 50, 75 and 100% agreement between all of the prediction methods and stored in the database. This allows a user to filter results based on conservation between prediction methodologies and for outputting likely regions of interest in query sequences online (taken at 75%). For a description of the consensus calculation see Figure 2.

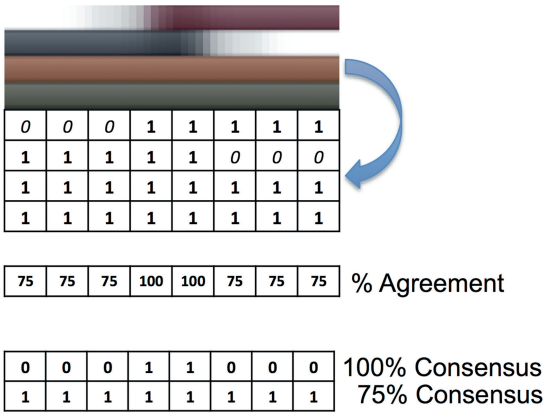


Figure 2. Toy example of the D^2P^2 predictor consensus calculation (see Figure 1 for a real example). The colored bars (top) represent real valued and binary disorder prediction output for four imagined predictors. Any real valued output is converted to a binary form by thresholding at a cut-off of 0.5 (as per CASP requirements) or at each prediction methods' advised cut-off minimizing false-positive rate. Next, a binary N × M matrix of per residue (N) and per predictor (M) results is created (blue arrow). The percentage from full agreement of a disordered state is calculated for each column of the binary matrix. This is then re-encoded as a binary matrix (bottom) for each threshold of agreement (or consensus) and further run-length encoded for storage in the database as a set of agreed upon regions of disorder. Taking a higher percentage cut-off of consensus yields a more conservative result with 100% likely under predicting. When searching online with D^2P^2 75% consensus is used to highlight regions of sequence that are likely disordered.

DATABASE

Data from the database are made available as tab-delimited files for maximum accessibility along with a MySQL schema file for anyone wishing to reconstruct the relational database tables.

Sequence

All protein sequences included in the database are provided along with their mapping to each genome with any comments from the source genome project made available.

Predictions

Disorder predictions for each predictor are available as well as per genome. SUPERFAMILY assignments are available direct from the SUPERFAMILY resource, but derived statistics from these assignments are included in the available data. All predictor outputs were consolidated into a single format in the database by thresholding any real valued result to a binary prediction, all predicted regions were then run-length encoded. Original real valued results are also included in the database for interested parties in the form of JSON arrays. A simple web service is available to obtain all binary predictions for a sequence as JSON by sequence ID query.

Search

Search for disorder in sequences of interest is provided through queries using lists of sequence ID either from the originating genome project or UniProt ID where applicable, free text search of the protein's comments and

sequence IDs from the genome project, as well as exactly matching whole sequences to all genomes. Included with finding exactly matched sequences CS-BLAST (37) is available to find the nearest matching protein in the database to identify likely disordered regions of novel sequence, though for this task some prediction methods included in D^2P^2 provide their own online prediction portals that are linked from the database online. For investigating disorder on a larger scale where a user does not have a protein of interest, a browse page is provided. The whole database can be inspected for proteins that come from a specific genome or taxon, as well as those that match specific content such as SCOP superfamily assignment, domain-centric Gene Ontology assignment, DisProt and IDEAL curated validation, the percentage of disorder content in the protein and the percentage agreement of all predictors agreeing on a given disordered region.

Statistics

Several pre-computed statistics are included in the database per predictor for each sequence, these include the following: the number of residues predicted disordered; the percentage of protein predicted disordered; the number of residues predicted disordered in a predicted SCOP domain; the percentage of the predicted disorder that lies in a SCOP domain and the percentage of the whole protein predicted disordered and inside a SCOP domain. Additionally, per sequence and per predictor pair-wise comparison statistics are included for the purpose of future predictor development: the number of residues both methods agree are disordered; the percentage of each methods total disorder that agrees with another method; the percentage of the whole protein the methods agree are disordered; the number of residues predicted with one method but not another; the percentage of all residues predicted in one method not found in another and the percentage of the whole protein one method predicts to be disordered that another method does not.

Reports

Graphical reports are available via the web of all disorder, SCOP structure, ANCHOR binding region and PhosphoSitePlus (38) post-translational modification (PTM) site assignments for a given set of sequences of interest. Additionally, where relevant experimental annotations and cross references are provided by the DisProt and IDEAL curated databases along with predicted disorder. Dependent on browser functionality a scalable vector graphics figure is made available with all prediction data embedded, mouse popups provide direct access to each region of interest. Additionally, publication ready figures are also one-click downloadable for any search result. In Figure 1, we see an example of such a report.

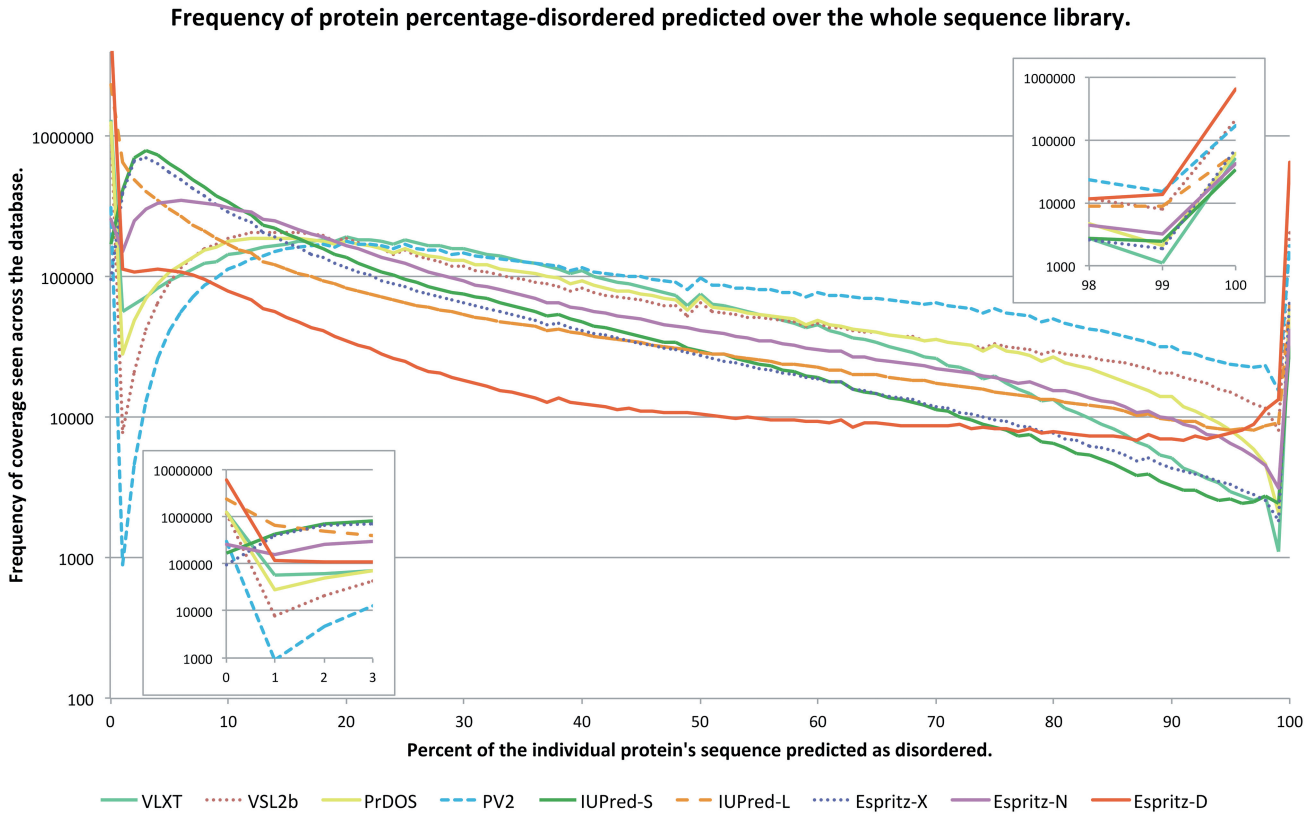


Figure 3. A graph showing the distribution of total disorder coverage per-protein over the whole database of protein sequences for each predictor. The *X* axis shows the percentage of a protein sequence that was covered with disorder prediction from a given predictor, binned at 1% intervals. The *Y* axis shows the frequency of observed sequences with a given percentage coverage of disorder, $\log_{10}$ scaled for ease of comparison. The inset (left) shows the first 3% zoomed for clarity of how each predictor treats more structured proteins, the inset (right) shows the final 3% where proteins are predicted to be profoundly disordered with little to no stable tertiary structure.

Source code

Perl source code for the website is available through Git at: https://github.com/MattOates/d2p2.pro.

RESULTS

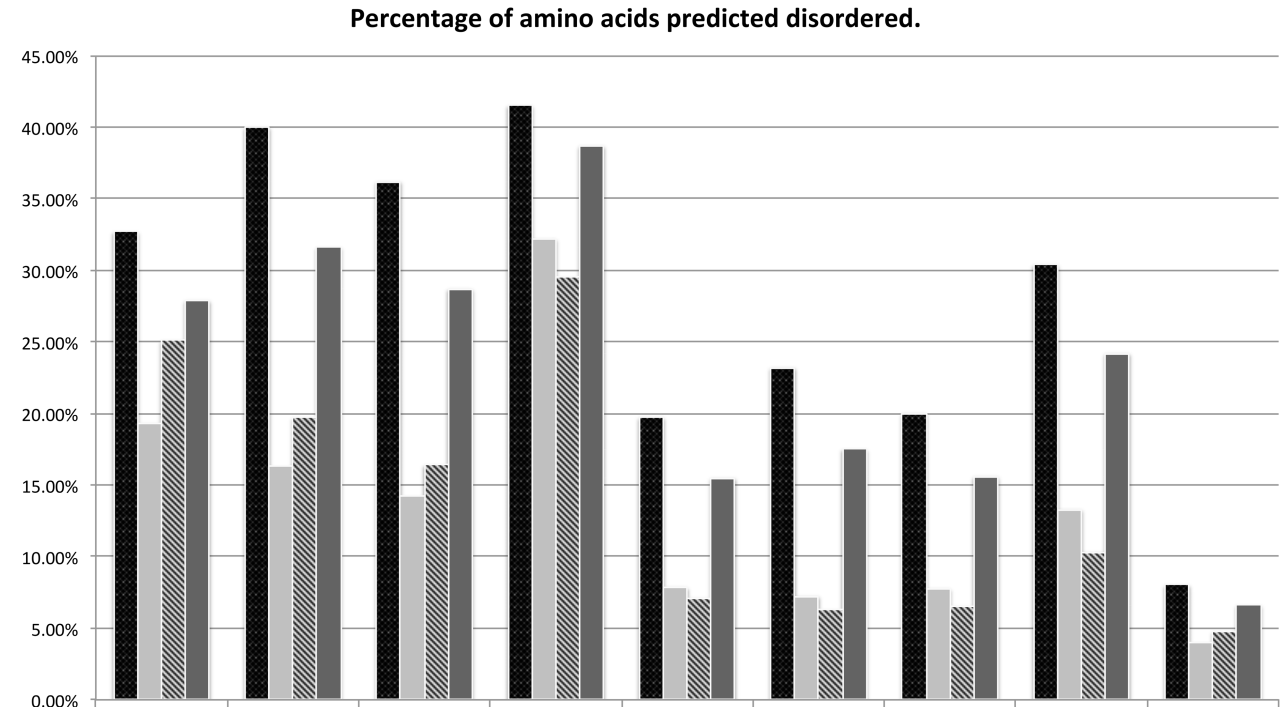
The real product of the work is the database itself, but we describe briefly below a first global look at the data.

Global comparison of disorder predictors

One aim of the D^2P^2 database is to provide statistics for improving disorder prediction. In Figure 3, we show each prediction method's coverage compared over all sequences in the database. Certain features are immediately apparent. At 0–1%, all predictors avoid stable globular structures, with a rapid change to a regime of unstructured regions covering 10–50% of a given protein being common. All prediction methods change trend toward higher frequencies at >98% coverage mark, representing families of profoundly unstructured proteins. IUPred-S (short variant) as expected has higher frequency of short sub-regions and lower frequency in longer regions, so too does VL-XT. PrDOS and VSL2b are relatively balanced toward long and short regions of disorder predicted, with PV2 predicting greater numbers of long disordered regions over short. An avenue of improvement might be to investigate the production of a meta-predictor that better handles short and long regions of disorder, perhaps including IUPred-S and PV2 with VL-XT to avoid over prediction; feasibility of such approaches was discussed recently by Peng and Kurgan (39). The aim of this work is not to develop a meta-predictor but to empower the prediction community to use D^2P^2 as a key information resource driving methods development.

Prediction by domain of cellular life

Figure 4 shows global statistics per predictor for each domain of cellular life. The general trend for all disorder predicted is that Eukarya have had a large expansion in the quantity of disordered sequence. The story for Archaea and Bacteria is less clear, where five methods out of nine show Archaea as having greater disordered content than Bacteria. The exception to this is that PV2, IUPred-L, Espritz-X and Espritz-N find Bacteria to have more disordered sequence than Archaea. This inversion in Bacterial and Archaeal disorder content between predictor variants such as seen with IUPred-S versus IUPred-L and Espritz-D versus other variants suggests that these two domains of life differ in the forms of disorder present if



	VLXT	VSL2b	PrDOS	PV2	IUPred-S	IUPred-L	Espritz-X	Espritz-N	Espritz-D
Eukarya	32.70%	40.00%	36.20%	41.50%	19.75%	23.20%	20.00%	30.40%	8.02%
Bacteria	19.30%	16.30%	14.20%	32.20%	7.86%	7.24%	7.76%	13.30%	3.96%
Archaea	25.10%	19.70%	16.40%	29.50%	7.09%	6.27%	6.51%	10.30%	4.77%
All	27.90%	31.60%	28.70%	38.70%	15.49%	17.50%	15.60%	24.20%	6.59%

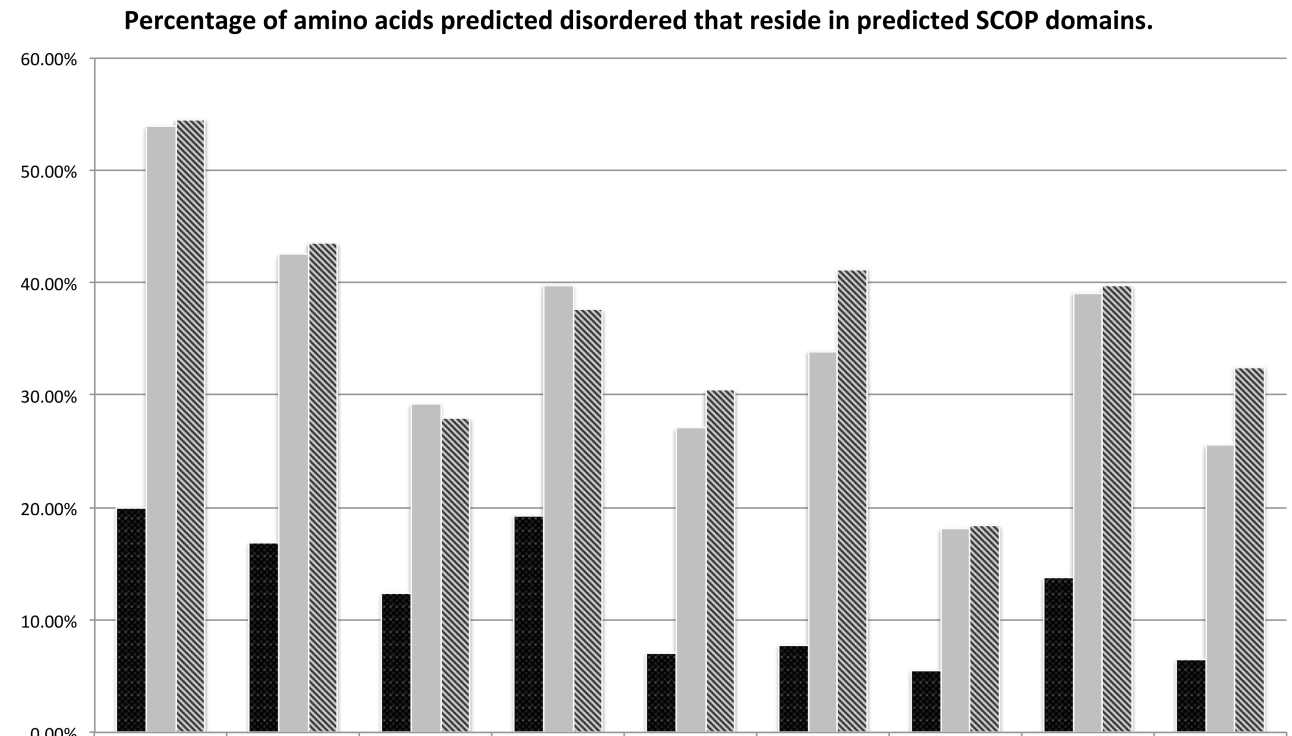
Figure 4. A bar chart grouped by prediction method of global percentage disorder predicted per domain of cellular life. The *X* axis shows results per domain grouped by predictor, the *Y* axis shows the percentage of all amino acid residues for a given domain of life predicted disordered by a given method.

not the quantity. Looking at the interplay between structure and disorder over evolution, we see in Figure 5 all predictors register a pronounced switch to disorder between SCOP domains rather than within domains for Eukarya. Archaea and Bacteria being similar to each other in reduced coverage outside of domains and proportionally twice as much coverage within domains. This overlap of SCOP prediction and disorder prediction does not imply incorrect prediction for either category, as short-type disorder has specific function within structured domains as regions that undergo dynamic structural transitions (2).

DISCUSSION

The main content of this database is fairly straightforward yet of great value: comprehensive disorder prediction on genomes shown alongside structural domains. Similar work was done previously for the human proteome (40). This prior work included sequence conservation as a third feature, leading to three types of proteins for the human proteome: structured (52%), disordered (35%) and cryptic domains (18%), where cryptic domains were defined as sequences with high evolutionary conservation that failed to match any known structured domains and were thus assumed to be structured domains for which the structures had not yet been determined. This conclusion was based on the assumption that all disordered regions show high sequence variability. However, there are reports of regions of disorder that show high sequence conservation (26,41). Thus, an important use of D^2P^2 will be to determine which cryptic domains are predicted to be structured by multiple predictors and which cryptic domains are predicted to be disordered, thus partitioning these regions into likely globular domains of currently unknown structure and into likely regions of disorder with high sequence conservation. This work is in progress and will be reported when completed. Current disorder findings from D^2P^2 data for human (ENSEMBL release 63) using all predictors shows ~37–50% of human amino acids predicted disordered with ~29–39% of the amino acids being intra-domain disorder i.e. not found within SCOP domains. Structured domains cover ~44% of amino acids leaving ~17–27% of the amino acids unassigned to either SCOP domains or intra-domain disorder.

The data in D^2P^2 have been made as accessible as possible, and is provided interactively via a website including a graphical display with a consensus plot. We are anxious to communicate with users with regard to future developments, so users should not hesitate to



	VLXT	VSL2b	PrDOS	PV2	IUPred-S	IUPred-L	Espritz-X	Espritz-N	Espritz-D
Eukarya	20.00%	16.90%	12.40%	19.30%	7.08%	7.79%	5.47%	13.80%	6.42%
Bacteria	53.90%	42.50%	29.20%	39.80%	27.09%	33.90%	18.10%	39.10%	25.50%
Archaea	54.50%	43.50%	27.90%	37.60%	30.43%	41.20%	18.40%	39.70%	32.50%

Figure 5. Amino acids which have been predicted to be disordered (Figure 4) were then sub-classified as either being inter- or intra-domain disorder. This figure shows a bar chart, with results grouped by predictor, of the percentage of disordered amino acids that reside within a predicted SCOP domain. The *X* axis shows results per domain grouped by predictor, the *Y* axis shows the percentage of all amino acid residues for a given domain of life predicted disordered by a given method.

suggest or provide additional tools or predictors to be added in the future.

Example use

D^2P^2 provides informative data for various types of biological investigation. A good example is the exploration of isoforms and their function. A recent study by Ellis *et al.* (42) showed that alternative splicing of proteins has rewired protein–protein interaction networks in neural tissue, and that these are important in tissue-specific function. Figure 1 shows the Bridging Integrator 1 (Bin1) gene from the study, and two of its most dimorphic isoforms (ENSP00000365281 and ENSP00000316779). It was change in disordered regions that were shown to alter Bin1 interaction with Dynamin 2 (Dnm2) facilitating endocytosis within neural-tissue. With D^2P^2 these forms of analyses can be automated with the addition of multiple sources of evidence for disordered regions. Additionally, the inclusion of PhosphoSitePlus curated PTM annotation lets us see that the disordered inserts between BIN1 isoforms also undergo posttranslational modifications as part of the regulatory process. The suggestion from the D^2P^2 data that Eukarya have a bias toward intra-domain disorder (Figure 5) suggests that this sort of study is likely to be increasingly important in characterizing the full complexity of protein interaction and regulation in Eukaryotes.

Further work

The principal future goal is to include more disorder predictors. Although the database has a substantial collection there are important predictors that need to be added, and furthermore important new predictors are likely to be developed over time. The other main future goal is to expand the sequences on which we have disorder predictions to include more genomes, e.g. thousands of viral genomes and other sequence sets that are already in SUPERFAMILY. We also intend to improve the interface and provide more tools for online analysis, e.g. tools to enable searches by Gene Ontology, tools for comparative genomics, analysis methods that take advantage of the domain-based sTOL (http://supfam.org/SUPERFAMILY/sTOL) and additional software that capitalizes on other tools attached to SUPERFAMILY.

ACKNOWLEDGEMENTS

The authors acknowledge the efforts of all original authors for each disorder prediction method used in D^2P^2 and their indirect contribution to the data they present. Additionally, this resource would not have been possible without contributions from many open source developers our software is predicated on. The Bolyai Janos fellowship for Z.D. is gratefully acknowledged.

FUNDING

Engineering and Physical Research Council (EPSRC) [EP/E501214/1 to M.E.O.]; Dissertation Fellowship awarded by the University of Alberta (to M.J.M.); Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery (to L.K.); Program of the Russian Academy of Sciences for ''Molecular and Cellular Biology'' (to V.N.U.); the US National Science Foundation [EF 0849803 to A.K.D. and V.N.U.]; Biotechnology and Biological Sciences Research Council (BBSRC) [BB/G022771/1 to J.G.]. Funding for open access charge: Engineering and Physical Research Council (EPSRC) [EP/E501214/1].

Conflict of interest statement. None declared.

REFERENCES

1. Uversky,V.N., Gillespie,J.R. and Fink,A.L. (2000) Why are ''natively unfolded'' proteins unstructured under physiologic conditions? *Proteins*, **41**, 415–427.
2. Uversky,V.N. (2002) Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.*, **11**, 739–756.
3. Iakoucheva,L.M., Radivojac,P., Brown,C.J., O'Connor,T.R., Sikes,J.G., Obradovic,Z. and Dunker,A.K. (2004) The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.*, **32**, 1037–1049.
4. Song,J., Lee,M.S., Carlberg,I., Vener,A.V. and Markley,J.L. (2006) Micelle-induced folding of spinach thylakoid soluble phosphoprotein of 9 kDa and its functional implications. *Biochemistry*, **45**, 15633–15643.
5. Yamada,J., Phillips,J.L., Patel,S., Goldfien,G., Calestagne-Morelli,A., Huang,H., Reza,R., Acheson,J., Krishnan,V.V., Newsam,S. *et al.* (2010) A bimodal distribution of two distinct categories of intrinsically-disordered structures with separate functions in FG nucleoporins. *Mol. Cell. Proteomics*, **9**, 2205–2224.
6. Dunker,A.K., Silman,I., Uversky,V.N. and Sussman,J.L. (2008) Function and structure of inherently disordered proteins. *Curr. Opin. Struct. Biol.*, **18**, 756–764.
7. Dyson,H.J. and Wright,P.E. (2005) Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.*, **6**, 197–208.
8. Rose,P.W., Beran,B., Bi,C., Bluhm,W.F., Dimitropoulos,D., Goodsell,D.S., Prlić,A., Quesada,M., Quinn,G.B., Westbrook,J.D. *et al.* (2011) The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res.*, **39**, D392–D401.
9. Sickmeier,M., Hamilton,J.A., LeGall,T., Vacic,V., Cortese,M.S., Tantos,A., Szabo,B., Tompa,P., Chen,J., Uversky,V.N. *et al.* (2007) DisProt: the database of disordered proteins. *Nucleic Acids Res.*, **35**, D786–D793.
10. Fukuchi,S., Sakamoto,S., Nobe,Y., Murakami,S.D., Amemiya,T., Hosoda,K., Koike,R., Hiroaki,H. and Ota,M. (2012) IDEAL: intrinsically disordered proteins with extensive annotations and literature. *Nucleic Acids Res.*, **40**, D507–D511.
11. Di Domenico,T., Walsh,I., Martin,A.J.M. and Tosatto,S.C.E. (2012) MobiDB: a comprehensive database of intrinsic protein disorder annotations. *Bioinformatics*, **28**, 2080–2081.
12. Martin,A.J.M., Walsh,I. and Tosatto,S.C.E. (2010) MOBI: a web server to define and visualize structural mobility in NMR protein ensembles. *Bioinformatics*, **26**, 2916–2917.
13. He,B., Wang,K., Liu,Y.-L., Xue,B., Uversky,V.N. and Dunker,A.K. (2009) Predicting intrinsic disorder in proteins: an overview. *Cell Res.*, **19**, 929–949.
14. Peng,Z.L. and Kurgan,L. (2012) Comprehensive comparative assessment of in-silico predictors of disordered regions. *Curr. Protein Pept. Sci.*, **13**, 6–18.
15. Romero,P., Obradovic,Z., Kissinger,C., Villifranca,J.E. and Dunker,A.K. (1997) Identifying disordered regions in proteins from amino acid sequence. *Proc. Int. Conf. Neural Networks*, **1**, 90–95.
16. Peng,K., Radivojac,P., Vucetic,S., Dunker,A.K. and Obradovic,Z. (2006) Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics*, **7**, 208.

17. Kozlowski,L.P. and Bujnicki,J.M. (2012) MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins. *BMC Bioinformatics*, **13**, 111.
18. Fukuchi,S., Homma,K., Minezaki,Y., Gojobori,T. and Nishikawa,K. (2009) Development of an accurate classification system of proteins into structured and unstructured regions that uncovers novel structural domains. *BMC Struct. Biol.*, **9**, 26.
19. Ward,J.J., Sodhi,J.S., McGuffin,L.J., Buxton,B.F. and Jones,D.T. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.*, **337**, 635–645.
20. Romero,P., Obradovic,Z., Li,X., Garner,E.C., Brown,C.J. and Dunker,A.K. (2001) Sequence complexity of disordered protein. *Proteins*, **42**, 38–48.
21. Li,X., Romero,P., Rani,M., Dunker,A.K. and Obradovic,Z. (1999) Predicting protein disorder for N-, C-, and internal regions. *Genome Inform Ser Workshop Genome Inform*, **10**, 30–40.
22. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
23. Rost,B., Sander,C. and Schneider,R. (1994) PHD—an automatic mail server for protein secondary structure prediction. *Comput. Appl. Biosci.*, **10**, 53–60.
24. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.
25. Xue,B., Dunbrack,R.L., Williams,R.W., Dunker,A.K. and Uversky,V.N. (2010) PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. *Biochem. Biophys. Acta*, **1804**, 996–1010.
26. Chen,J.W., Romero,P., Uversky,V.N. and Dunker,A.K. (2006) Conservation of intrinsic disorder in protein domains and families: I. A database of conserved predicted disordered regions. *J. Proteome Res.*, **5**, 879–887.
27. Ishida,T. and Kinoshita,K. (2007) PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Res.*, **35**, W460–W464.
28. Ghalwash,M.F., Dunker,A.K. and Obradovic,Z. (2012) Uncertainty analysis in protein disorder prediction. *Mol. Biosyst.*, **8**, 381–391.
29. Monastyrskyy,B., Fidelis,K., Moult,J., Tramontano,A. and Kryshtafovych,A. (2011) Evaluation of disorder predictions in CASP9. *Proteins: Struct., Funct., Bioinf.*, **79**, 107–118.
30. Walsh,I., Martin,A.J., Di Domenico,T. and Tosatto,S.C. (2012) ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics*, **28**, 503–509.
31. Dosztányi,Z., Csizmók,V., Tompa,P. and Simon,I. (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.*, **347**, 827–839.
32. Mészáros,B., Simon,I. and Dosztányi,Z. (2009) Prediction of protein binding regions in disordered proteins. *PLoS Comput. Biol.*, **5**, e1000376.
33. Gough,J., Karplus,K., Hughey,R. and Chothia,C. (2001) Assignment of homology to genome sequences using a library of hidden Markov Models that represent all proteins of known structure. *J. Mol. Biol.*, **313**, 903–919.
34. de Lima Morais,D., Fang,H., Rackham,O., Wilson,D., Pethica,R., Chothia,C. and Gough,J. (2011) SUPERFAMILY 1.75 including a domain-centric gene ontology method. *Nucleic Acids Res.*, **39**, D427–D434.
35. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
36. Andreeva,A., Howorth,D., Brenner,S.E., Hubbard,T.J.P., Chothia,C. and Murzin,A.G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.*, **32**, D226–D229.
37. Biegert,A. and Söding,J. (2009) Sequence context-specific profiles for homology searching. *Proc. Natl Acad. Sci. USA.*, **106**, 3770–3775.
38. Hornbeck,P.V., Kornhauser,J.M., Tkachev,S., Zhang,B., Skrzypek,E., Murray,B., Latham,V. and Sullivan,M. (2012) PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res.*, **40**, D261–D270.
39. Peng,Z. and Kurgan,L. (2012) On the complementarity of the consensus-based disorder prediction. *Pac. Symp. Biocomput.*, 176–187.
40. Fukuchi,S., Hosoda,K., Homma,K., Gojobori,T. and Nishikawa,K. (2011) Binary classification of protein molecules into intrinsically disordered and ordered segments. *BMC Struct. Biol.*, **11**:29.
41. Chen,J.W., Romero,P., Uversky,V.N. and Dunker,A.K. (2006) Conservation of intrinsic disorder in protein domains and families: II. Functions of conserved disorder. *J. Proteome Res.*, **5**, 888–898.
42. Ellis,J.D., Barrios-Rodiles,M., Çolak,R., Irimia,M., Kim,T., Calarco,J.A., Wang,X., Pan,Q., O'Hanlon,D., Kim,P.M. *et al.* (2012) Tissue-specific alternative splicing remodels protein-protein interaction networks. *Mol. Cell.*, **46**, 884–892.