BIOINFORMATICS APPLICATIONS NOTE

Vol. 21 no. 1 2005, pages 137–140
doi:10.1093/bioinformatics/bth476

DisProt: a database of protein disorder

Slobodan Vucetic[1], Zoran Obradovic[1], Vladimir Vacic[1], Predrag Radivojac[2], Kang Peng[1], Lilia M. Iakoucheva[3], Marc S. Cortese[2], J. David Lawson[4], Celeste J. Brown[5], Jason G. Sikes[6], Crystal D. Newton[6] and A. Keith Dunker[2,∗]

[1]Center for Information Science and Technology, Temple University, Philadelphia, PA 19122, USA, [2]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN 46202, USA, [3]Laboratory of Statistical Genetics, The Rockefeller University, New York, NY 10021, USA, [4]Concurrent Pharmaceuticals, 502 W. Office Center Dr, Fort Washington, PA 19034, USA, [5]Department of Biological Sciences, University of Idaho, ID 83844, USA and [6]School of Molecular Biosciences, Washington State University, Pullman, WA 99164, USA

Received on February 5, 2004; revised on July 2, 2004; accepted on August 3, 2004
Advance Access publication August 13, 2004

ABSTRACT

Summary: The Database of Protein Disorder (DisProt) is a curated database that provides structure and function information about proteins that lack a fixed three-dimensional (3D) structure under putatively native conditions, either in their entirety or in part. Starting from the central premise that intrinsic disorder is an important structural class of protein and in order to meet the increasing interest thereof, DisProt is aimed at becoming a central repository of disorder-related information. For each disordered protein, the database includes the name of the protein, various aliases, accession codes, amino acid sequence, location of the disordered region(s), and methods used for structural (disorder) characterization. If applicable, most entries also list the biological function(s) of each disordered region, how each region of disorder is used for function, as well as provide links to PubMed abstracts and major protein databases.

Availability: www.disprot.org

Contact: kedunker@iupui.edu

INTRODUCTION

Intrinsic order and disorder are defined herein at the amino acid residue level. The backbone atoms, hence also the residues, of ordered regions or proteins undergo small-amplitude, thermally-driven motions about their equilibrium positions determined as time-averaged values. In some cases, ordered regions cooperatively switch between two or more specific conformations. In contrast, intrinsically disordered proteins or regions exist as dynamic ensembles in which the atom positions and the backbone Ramachandran angles vary significantly over time with no specific equilibrium values and typically involve non-cooperative conformational changes. Thus, the existence of disorder is determined by a protein's dynamical properties, and not necessarily by the presence or absence of local secondary structure. We define an intrinsically disordered protein as one that contains at least one disordered region.

Despite the fact that intrinsically disordered proteins fail to form fixed three dimensional (3D) structure under physiological conditions, existing instead as ensembles of conformations, they carry out critically important biological functions (Dunker *et al*., 2002a; Iakoucheva *et al*., 2002). Several recent reviews (Wright and Dyson, 1999; Uversky, 2002; Dyson and Wright, 2002; Tompa, 2002) attest to the growing interest in these proteins. In addition, whole-cell NMR experiments demonstrate that intrinsic disorder can exist *in vivo* (Dedmon *et al*., 2002) and thus does not result merely from the failure to find the correct conditions or ligand for folding to occur. A collection of about 100 intrinsically disordered examples, many of which were characterized by two or more experimental methods, were assigned one of 28 specific functions and grouped into four functional classes: (1) molecular recognition; (2) molecular assembly; (3) protein modification; and (4) entropic chain activities (Dunker *et al*., 2002a). Intrinsically disordered regions are typically involved in regulation, signaling and control pathways in which interactions with multiple partners and high-specificity/low-affinity interactions are often requisite. In this way, the functional diversity provided by disordered regions complements those of ordered protein regions.

At the structural level, we proposed that intrinsically disordered regions may exist in molten globule-like (collapsed) and random coil-like (extended) forms (Dunker *et al*., 2001). Another form of disorder, the pre-molten globule, has

[∗]To whom correspondence should be addressed.

been proposed (Uversky, 2002), but it is unclear whether it is truly distinct from the random coil-like class. We also suggested that functions of disordered proteins may arise from the specific disorder form, from interconversion of disordered forms, or from transitions between disordered and ordered conformations (Dunker *et al*., 2001). These function-associated conformational changes may be brought about by alterations in environmental or cellular conditions (e.g. disorder-to-order transition upon binding during signal transduction).

A major hindrance in the study of intrinsically disordered proteins is the absence of an organized database for these proteins. The lack of a curated, systematized database impedes the dissemination of knowledge regarding disorder examples. An additional complication is the diversity of language used to describe similar phenomena. For example, while the absence of a fixed 3D structure is used herein as the definition of disorder, and thus molten globule-like proteins and molten globule-like regions would be included, others prefer to view disordered proteins and regions as operationally identified as missing coordinates in the Protein Data Bank (PDB) (Linding *et al*., 2003b), as essentially devoid of helix or sheet (Liu *et al*., 2002), as dynamically flexible ensembles (Ward *et al*., 2004), or as high B-factor coils (Linding *et al*., 2003a,b). Thus, a coordinated, widely used database could help to unify the terminology in this field. One notable attempt to provide such a resource of protein disorder is the ProDDO database (Sim *et al*., 2001); however, it is not curated, its contents are limited to PDB entries, and it does not provide structural or functional annotation of disorder. For these reasons, we embarked on the construction of the Database of Protein Disorder, known as 'DisProt', in which the available structural/functional information was obtained through exhaustive searches of the relevant literature and biological databases (Dunker *et al*., 2002a).

PROTEIN SELECTION CRITERIA

DisProt contains both proteins that lack a fixed 3D structure along their entire lengths as well as proteins that have local regions lacking a fixed 3D structure. The criteria for the inclusion of a protein in DisProt are discussed in the following paragraphs.

Several methods have been used to identify proteins wholly lacking fixed 3D structures, including several variants of NMR spectroscopy, circular dichroism (CD) spectroscopy, small-angle X-ray scattering (SAXS) and estimation of hydrodynamic radius. The details of each method are described elsewhere (Rose, 2002; Daughdrill *et al*., 2004). NMR spectroscopy provides the most direct estimation of the internal motions that characterize disordered ensembles and can be used to identify each residue lacking a fixed structure if peak assignments are made. However, for many proteins a global characterization is made without making individual NMR peak assignments. CD, SAXS and hydrodynamic radius measurements do not provide residue-by-residue information. For proteins characterized by these methods, part of a given protein might fold into a structured domain that is missed because it represents a small fraction of the total sequence. Thus, inclusion of such proteins in the database characterized by these methods introduces the possibility that regions of ordered structure are misclassified as regions that lack fixed structures. While proteins characterized by a single method are included, an effort is being made to identify proteins characterized by multiple methods. Such well-characterized proteins can serve as useful prototypes for proteins in particular classes.

One of the main methods for characterizing disorder is by missing electron density in X-ray structures, which leads to missing coordinates in the corresponding PDB files. For example, about two-thirds of the structured chains in the PDB contain localized regions of missing electron density (Obradovic *et al*., 2003); these regions vary in length from single residues to long segments containing more than 100 consecutive residues. For inclusion in this database, a given residue needs to lack electron density for all of its backbone atoms, not just the side chain atoms. For long regions of missing coordinates in X-ray structures there is the possibility that they are actually ordered domains that move as rigid bodies and thus fail to scatter X-rays coherently. Such regions have been called 'wobbly domains' (Dunker *et al*., 2001; Daughdrill *et al*., 2004). Although such unobserved regions would be misclassified as disordered in the present database, similarities in amino acid compositions between disordered regions characterized by different methods indicate that wobbly domains are the exception rather than the rule (Dunker *et al*., 2002b). Over time, an effort will be made to identify such misclassified, wobbly-domain regions and remove them from the database. For example, protease digestion typically releases wobbly domains as large, protease resistant fragments, while disordered regions tend to be cut at multiple loci and fail to yield such large fragments. Thus, a simple procedure based on protease digestion could be used to test whether a long unobserved region is a wobbly domain or not. In addition, NMR could be used if the entire protein is not too large.

Several NMR techniques can indicate regions of disorder within otherwise ordered protein structures. When modeling the 3D structure to fit the NMR data, some regions give multiple solutions to the NMR restraints and so appear as ensembles of possible structures. The standard view is that such regions are ordered, but the data are simply insufficient to identify the particular 3D structure. The alternative possibility is that such regions actually lack particular 3D structures because they are disordered. Such ambiguities can be sorted out by additional NMR relaxation experiments that directly measure the motions of the residues in such regions (Bracken, 2001).

DATABASE CONTENT

DisProt was designed as a companion to major online protein repositories; hence, whenever possible, DisProt provides links to PDB, SWISS-PROT and TrEMBL, GenBank, and PIR databases. The release 1.2 of the DisProt knowledge base consists of 154 proteins with 190 disordered regions, 164 of which are longer than 30 consecutive residues. Currently, DisProt is essentially non-redundant: no two proteins have sequence identity $>50\%$, and only four pairs and one triple of the proteins have sequence identity between 30 and 50%. The non-redundant nature of the database, however, was unintentional. Further releases of DisProt will include similar sequences, but an effort will be made to provide a representative, non-redundant subset to the community. Although relatively small, the set of DisProt proteins provides a considerable coverage of protein sequence space. For example, a BLAST search (with BLOSUM62, gap penalties of 11/1, and E-value threshold of 10^{-2}) revealed that 6720 or 5.1% of 132 648 SWISS-PROT sequences were covered by our database. Since low-complexity regions were not filtered by this procedure due to the positive correlation between disorder and low-complexity sequences, we observe that the 5.1% coverage can serve as an upper limit estimate.

The basic unit in DisProt is a protein chain that has been characterized as having one or more disordered regions. For each listed protein, the database provides the name and various aliases, accession codes and links to other databases, amino acid sequence, and the location(s) of the disordered region(s). Additional fields were allocated for the following information: (1) detection method(s), e.g. X-ray diffraction, NMR, circular dichroism; (2) function of the disordered region, e.g. containing a site for phosphorylation, binding to a specific partner; (3) structural transitions if known, i.e. whether a disorder-to-order or an order-to-disorder transition has been associated with the function of the disordered region; and (4) references, specifically describing the structural characterization or functions of the disordered region including direct links to PubMed citations.

DATABASE ACCESS

DisProt provides a browsable list of current proteins, as well as lists of proteins classified by their functions, possible structural transitions and the method(s) by which they were characterized. Every protein in our database can be viewed in HTML format directly at http://www.disprot.org/protein.php?id=disprot_id. In addition to the HTML representation, we provide the data in XML format for each disordered protein. Finally, a reduced view of the data is made available in FASTA format.

DisProt allows standard and advanced search. The standard search mode provides full-text search of the entire data base, while the advanced mode provides an interface for a fast, sequence-based search of DisProt, an option rarely found in similar databases. Through this option, a user enters a query sequence and is provided with a list of DisProt proteins with significant similarity to the query sequence.

Communication and feedback

A major function of DisProt is to serve as a central repository for submission information related to intrinsically disordered protein. Through a ‘Comment’ feature we encourage feedback from the bioscience community in terms of verifying the information we provide, as well as obtaining relevant data either from the ongoing research or from the previous research, both published and unpublished. It will be the responsibility of biological experts from our group to verify the submitted information and to update the database accordingly.

Implementation

DisProt was built using PostgreSQL relational database management system running on a Suse Linux platform. Its web interface is based on an Apache web server scripting with PHP language.

Supplementary data

Very often, gaps in the ATOM list of PDB records are indicative of protein disorder. We have parsed the August 5, 2003 release of PDB and extracted all protein chains with gaps of three or more residues in the ATOM list of X-ray characterized proteins. We provided this data for download through DisProt (missingXray.080503.txt file lists chains with missing residues in FASTA format). This data is expected to facilitate future growth of DisProt. Please note that the residue numbering herein sometimes does not match that of the corresponding PDB entry, so comparisons with the PDB entry should be carried out using sequence matching, not using the residue numbers.

Future developments

We will continue to expand the database in terms of the number of protein disorder records. We will also work towards improving the content of DisProt records. We are already in the process of collecting information for a new functional narrative field in the protein records. This field will contain not only a summarizing description of the function of the disordered region, but also an indication of the biological context associated with the particular function. Various functional classification schemes (Dunker *et al.*, 2002a; Tompa, 2002) will be included in future releases of the database. In addition, to minimize the problems associated with different views on disordered proteins, we will make an effort to clearly indicate whether a given protein or region has molten globule-like, pre-molten globule-like or random coil-like properties.

ACKNOWLEDGEMENTS

Support from NIH R01 LM007688-01 to A.K.D. and Z.O. is gratefully acknowledged. A.K.D. thanks the Indiana Genomics Initiative (INGEN) for its generous support; INGEN is funded in part by the Lilly Endowment, Inc.

REFERENCES

Bracken,C. (2001) NMR spin relaxation methods for characterization of disorder and folding in proteins. *J. Mol. Graph. Model.*, **19**, 3–12.

Daughdrill,G.W., Pielak,G.J., Uversky,V.N., Cortese,M.S. and Dunker,A.K. (2004) In Buchner,J. and Kiefhaber,T.1 (eds), *Protein Folding Handbook.* Wiley-VCH, Weinheim.

Dedmon,M.M., Patel,C.N., Young,G.B. and Pielak,G.J. (2002) FlgM gains structure in living cells. *Proc. Natl Acad. Sci., USA*, **99**, 12681–12684.

Dunker,A.K. Lawson,J.D., Brown,C.J., Williams,R.M., Romero,P., Oh,J.S., Oldfield,C.J., Campen,A.M., Ratliff,C.M., Hipps,K.W. *et al.* (2001) Intrinsically disordered protein. *J. Mol. Graph. Model.*, **19**, 26–59.

Dunker,A.K., Brown,C.J., Lawson,J.D., Iakoucheva,L.M. and Obradovic,Z. (2002a) Intrinsic disorder and protein function. *Biochemistry*, **41**, 6573–6582.

Dunker,A.K., Brown,C.J. and Obradovic,Z. (2002b) Identification and functions of usefully disordered proteins. *Adv. Protein Chem.*,**62**, 25–49.

Dyson,H.J. and Wright,P.E. (2002) Coupling of folding and binding for unstructured proteins. *Curr. Opin. Struct. Biol.*, **12**, 54–60.

Iakoucheva,L.M., Brown,C.J., Lawson,J.D., Obradovic,Z. and Dunker,A.K. (2002) Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.*, **323**, 573–584.

Linding,R., Jensen,L.J., Diella,F., Bork,P., Gibson,T.J. and Russell,R.B. (2003a) Protein disorder prediction: implications for structural proteomics. *Structure*, **11**, 1453–1459.

Linding,R., Russell,R.B., Neduva,V. and Gibson,T.J. (2003b) GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res.*, **31**, 3701–3708.

Liu,J., Tan,H. and Rost,B. (2002) Loopy proteins appear conserved in evolution. *J. Mol. Biol.*, **322**, 53–64.

Obradovic,Z., Peng,K., Vucetic,S., Radivojac,P., Brown,C.J. and Dunker,A.K. (2003) Predicting intrinsic disorder from amino acid sequence. *Proteins*, **53** (Suppl. 6), 566–572.

Rose,G.D. (ed.) (2002) *Advances in Protein Chemistry*, Vol. 62. Academic Press, NY.

Sim,K.L., Uchida,T. and Miyano,S. (2001) ProDDO: a database of disordered proteins from the Protein Data Bank (PDB). *Bioinformatics*, **17**, 379–380.

Tompa, P. (2002) Intrinsically unstructured proteins. *Trends Biochem. Sci.*, **27**, 527–33.

Uversky, V.N. (2002) Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.*, **11**, 739–756.

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.

Wright,P.E. and Dyson,H.J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.*, **293**, 321–331.