PROTEINS: Structure, Function, and Bioinformatics 65:1–14 (2006)

REVIEW

A Practical Overview of Protein Disorder Prediction Methods

François Ferron,[1,2] Sonia Longhi,[1*] Bruno Canard,[1] and David Karlin[3]
[1] *Architecture et Fonction des Macromolécules Biologiques, UMR 6098 CNRS et Universités Aix-Marseille I et II, Marseille, France*
[2] *Boston Biomedical Research Institute, Watertown, Massachusetts 02472*
[3] *Ecole de l'ADN, INMED, Marseille, France*

ABSTRACT **In the past few years there has been a growing awareness that a large number of proteins contain long disordered (unstructured) regions that often play a functional role. However, these disordered regions are still poorly detected. Recognition of disordered regions in a protein is important for two main reasons: reducing bias in sequence similarity analysis by avoiding alignment of disordered regions against ordered ones, and helping to delineate boundaries of protein domains to guide structural and functional studies. As none of the available method for disorder prediction can be taken as fully reliable on its own, we present an overview of the methods currently employed highlighting their advantages and drawbacks. We show a few practical examples of how they can be combined to avoid pitfalls and to achieve more reliable predictions. Proteins 2006;65:1–14.**

Key words: protein disorder; prediction methods

INTRODUCTION

Intrinsically unstructured/disordered or natively unfolded proteins have a broad occurrence in living organisms. They are characterized by the lack of stable secondary and tertiary structure under physiological conditions and in the absence of a binding partner/ligand. Intrinsically disordered proteins fulfil essential functions, which are often linked with their disordered structural state.

A protein region is defined as disordered if it is devoid of stable secondary structure and if it has a large number of conformations as seen using methods such as X-ray crystallography (lack of electron density), nuclear magnetic resonance (NMR), circular dichroism (CD), small-angle X-ray scattering, and various hydrodynamic measurements.[1,2] However, this definition embraces several categories of disorder: molten globules, partially unstructured proteins (premolten globules), and random coils (by increasing mobility and decreasing residual secondary structure content (see Uversky[3]).

What is the practical interest of identifying disordered regions? Disorder prediction is an essential prerequisite to protein sequence analysis. Disordered regions often have a biased amino acid composition that can lead to spurious sequence similarity with unrelated proteins. The recognition of regions of disorder is thus crucial to avoid spurious sequence alignments with sequences of globular proteins (for examples, see Iyer et al.[4]). Moreover, the recognition of disordered regions facilitates the identification of Eukaryotic Linear Motifs (ELMs), which are short functional motifs occurring mainly (>70%) within disordered regions (e.g., SH3, PDZ, phosphorylation sites[5–7]).

Second, disordered regions often prevent crystallization of proteins, or the generation of interpretable NMR data. Therefore, structural biologists use disorder predictions to delineate compact domains to solve their 3D structure, or to dissect target sequences into a set of independently folded domains in order to facilitate tertiary structure and threading predictions.[8]

Although the identification of disordered regions of less than 20 residues in length is generally thought to be less accurate,[9] recent results suggest that progress has been made in predicting short disordered regions.[10,11] Accordingly, we herein consider also short (i.e., less than 20 residues) regions of disorder.

As in other areas of bioinformatics, the reliability of disorder prediction benefits from the use of several methods based on different concepts, different physicochemical parameters, or different implementations. Using a single disorder predictor is not sufficient to achieve predictions good enough to decipher the modular organization of a protein (for examples, see refs. 12–16). Herein, we briefly review the sequence features of disordered proteins. Disor-

*Correspondence to: Sonia Longhi, Architecture et Fonction des Macromolécules Biologiques, UMR 6098 CNRS et Universités Aix-Marseille I et II, Case 932, 163 Avenue de Luminy, 13288 Marseille Cedex 09, France. E-mail: longhi@afmb.cnrs-mrs.fr

Received 8 April 2005; Revised 22 December 2005; Accepted 4 April 2006

Published online 19 July 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21075

TABLE I. Main Features of Software Tools for Disorder Prediction

Predictor	What is predicted	Based on	Generates and uses multiple sequence alignment?
PONDR[23,68,69] http://www.pondr.com	All regions that are not rigid including random coils, partially unstructured regions, and molten globules	Local aa composition, flexibility, hydropathy, etc	No
SEG[24] http://mendel.imp.univie.ac.at/METHODS/seg.server.html http://www.ncbi.nlm.nih.gov/BLAST (simplified version with default settings) ftp://ftp.ncbi.nih.gov/pub/seg/seg (to download program)	Low-complexity segments that is, “simple sequences” or “compositionally biased regions”	Locally optimized low-complexity segments are produced at defined levels of stringency and then refined according to the equations of Wootton and Federhen[24]	No
Disopred2[11] http://bioinf.cs.ucl.ac.uk/disopred	Regions devoid of ordered regular secondary structure	Cascaded support vector machine classifiers trained on PSI-BLAST profiles	Yes
Globplot[18] http://globplot.embl.de	Regions with high propensity for globularity on the Russell/Linding scale (see next)	Russell/Linding scale of disorder (propensities for secondary structures and random coils)	No
Disembl[70] http://dis.embl.de	LOOPS (regions devoid of regular secondary structure); HOT LOOPS (highly mobile loops); REMARK465 (regions lacking electron density in crystal structure)	Neural networks trained on X-ray structure data	No
NORSp[71] http://cubic.bioc.columbia.edu/services/NORSp	Regions with No Ordered Regular Secondary Structure (NORS). Most, but not all, are highly flexible	Secondary structure and solvent accessibility	Yes
FoldIndex[72] http://bip.weizmann.ac.il/fldbin/findex	Regions that have a low hydrophobicity and high net charge (either loops or unstructured regions)	Charge/hydropathy analyzed locally using a sliding window	No
Charge/hydropathy method[58] http://www.pondr.com	Fully unstructured domains (random coils)	Global sequence composition (hydrophobicity versus net charge)	No
HCA (Hydrophobic Cluster Analysis)[73] http://smi.snv.jussieu.fr/hca/hca-seq.html	Hydrophobic clusters, which tend to form secondary structure elements	Helical visualization of amino acid sequence	No
PreLink[74] http://genomics.eu.org	Regions that are expected to be unstructured in all conditions, regardless of the presence of a binding partner.	Compositional bias and low hydrophobic cluster content.	No
IUPred[43] http://iupred.enzim.hu	Regions that lack a well-defined 3D structure under native conditions	Energy resulting from interresidue interactions, estimated from local amino acid composition	No
RONN[44] http://www.strubi.ox.ac.uk/RONN	Regions that lack a well-defined 3D structure under native conditions.	Bio-basis function neural network trained on disordered proteins	No

der prediction methods are described in Table I, with tips and caveats concerning each predictor listed in Table II. We present a general scheme for disorder prediction in Figure 1, while Figure 2 illustrates a possible pitfall in disorder prediction. We applied the general scheme described in Figure 1 to an in-depth analysis of two well-characterized proteins, the nucleoprotein of measles virus (Figs. 3 and 4) and the Ubiquitin-like protein domain of hPLIC-2 (Fig. 5), and show how prediction methods need to be combined to achieve accurate disorder predictions.

SEQUENCE FEATURES OF DISORDERED PROTEINS

Sequence Composition

Intrinsically disordered proteins generally have a biased amino acid composition. A consensus of two independent studies, focusing respectively on the amino acids preferred at the surface of globular proteins or on those found less frequently in secondary structures[17,18] established the following empirical rule: G, S, and P are disorder-promoting amino acids, W, F, I, Y, V, and L are order-

TABLE II. Tips and Caveats of Disorder Prediction Methods

	Tips	Caveats
PONDR	PONDR comes in several versions: VL-XT may be used to look for binding sites, indicated by sharp drops in the middle of long disordered regions. VSL1 performs better to identify short regions of disorder. VL3 should be preferred to delineate domains as it gives smoother predictions.	
SEG	The stringency of the search for low-complexity segments is determined by three user-defined parameters: trigger window length [W], trigger complexity [K(1)] and extension complexity [K(2)] Parameters for disorder prediction: for long nonglobular domains, use long window lengths, typically: seg sequence 45 3.4 3.75 for shorter nonglobular domains, typically use: seg sequence 25 3.0 3.3	
Disopred2		Prediction accuracy is lower if there are few homologues
Globplot	Gives easy overview of modular organization of large proteins thanks to user-friendly, built-in SMART, PFAM, and low-complexity predictions. Changes of slope often correspond to domain boundaries	
Disembl	Gives predictions of low complexity regions and of aggregation propensity	Use the Loop predictor only as a filter to remove false disorder predictions of Hot Loops and Remark465, i.e., if Loop predicts that a region is ordered whereas Hot Loops or Remark 465 predict the opposite, Loop should be trusted
NORSp		Beware: some NORS are rigid, whereas some highly mobile regions have predicted secondary structure Prediction accuracy is lower if there are few homologs available
FoldIndex	Highlights some regions that are probably short loops better than a simple hydrophobicity plot	Foldindex does not provide results on the N- and C-termini. Therefore, it is not convenient to use it on small proteins (<100 aa).
Charge/hydropathy method	This method has not been trained on disordered proteins. Therefore, it is expected to recognize types of disordered proteins that are underrepresented in the disordered protein databases.	Requires prior knowledge of modular organization of protein. Applicable only to domains without disulfide bonds and without metal-binding regions.
HCA	Highlights coiled coils and regions with a biased composition Highlights regions with potential for induced folding Highlights very short potential globular domains Allows meaningful comparison with related proteins Allows a better definition of the boundaries of disordered regions	User's interpretation required
Prelink		Prelink generally predicts as ordered unstructured regions that have the potential to be ordered in the presence of a partner (i.e., to undergo induced folding)
IUPred	IUPRED uses a novel algorithm that was not trained on disordered proteins. Therefore, it is expected to recognize types of disordered proteins that are underrepresented in the disordered protein databases.	Applicable only to proteins without disulfide bonds and without metal-binding regions.
RONN		If RONN is being used to search for short regions of disorder it is advisable to inspect the plot for regions close to, but below, the threshold.

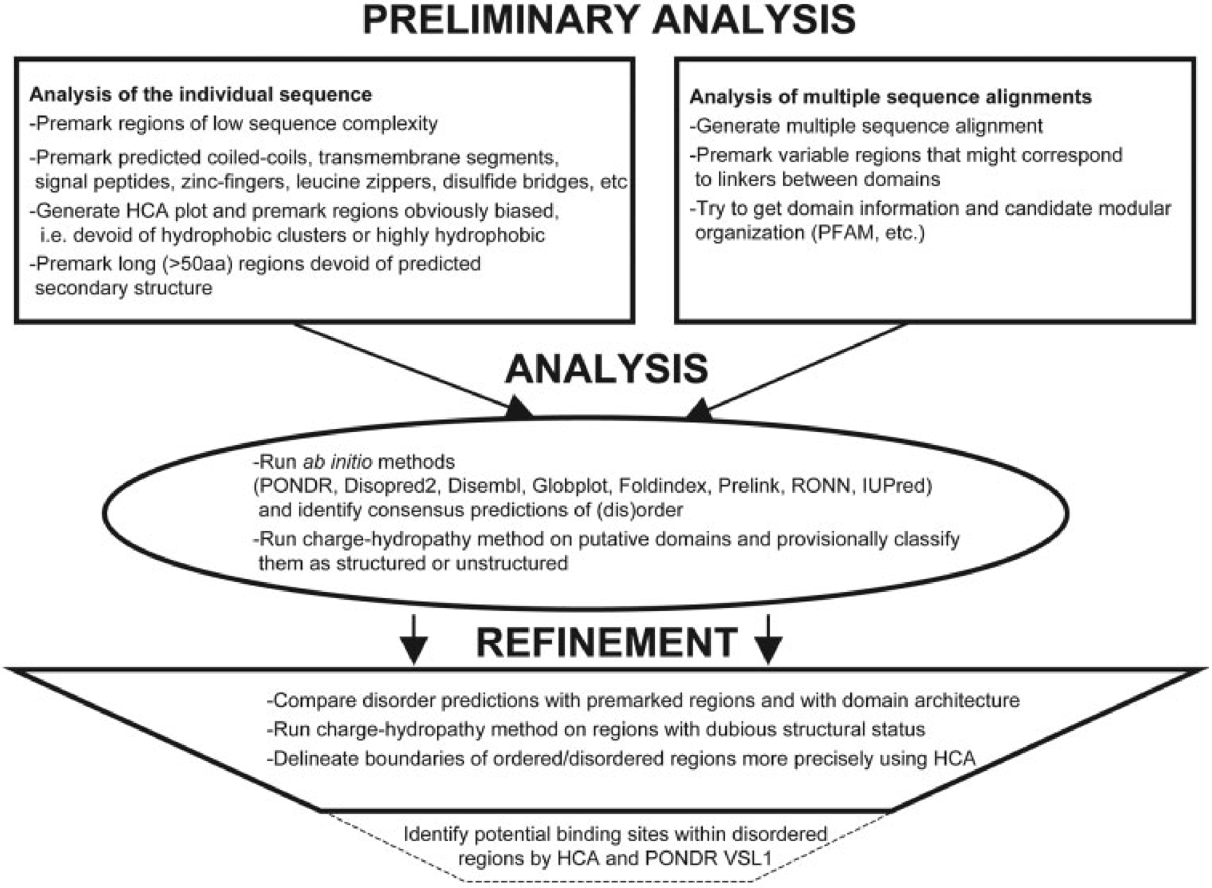

PRELIMINARY ANALYSIS
Analysis of the individual sequence
-Premark regions of low sequence complexity
-Premark predicted coiled-coils, transmembrane segments, signal peptides, zinc-fingers, leucine zippers, disulfide bridges, etc
-Generate HCA plot and premark regions obviously biased, i.e. devoid of hydrophobic clusters or highly hydrophobic
-Premark long (>50aa) regions devoid of predicted secondary structure
Analysis of multiple sequence alignments
-Generate multiple sequence alignment
-Premark variable regions that might correspond to linkers between domains
-Try to get domain information and candidate modular organization (PFAM, etc.)
ANALYSIS
-Run ab initio methods (PONDR, Disopred2, Disembl, Globplot, Foldindex, Prelink, RONN, IUPred) and identify consensus predictions of (dis)order
-Run charge-hydropathy method on putative domains and provisionally classify them as structured or unstructured
REFINEMENT
-Compare disorder predictions with premarked regions and with domain architecture
-Run charge-hydropathy method on regions with dubious structural status
-Delineate boundaries of ordered/disordered regions more precisely using HCA
Identify potential binding sites within disordered regions by HCA and PONDR VSL1

Fig. 1. General scheme for prediction of disordered and ordered regions in a protein.

promoting amino acids, while H and T are considered neutral with respect to disorder. Using sequence composition as the sole predictive parameter of disorder is not reliable. For instance, the RNA cap 2′-O-methyltransferase domain of dengue virus polymerase is structured[19] and yet is heavily depleted in some order-promoting residues and markedly enriched in some disorder-promoting residues (data not shown). However, Weathres et al.[20] recently reported that amino acid composition alone could allow recognition of intrinsically disordered proteins with a good accuracy. In any case, it is recommended to always analyze the sequence composition of proteins prior to further sequence analysis.[21]

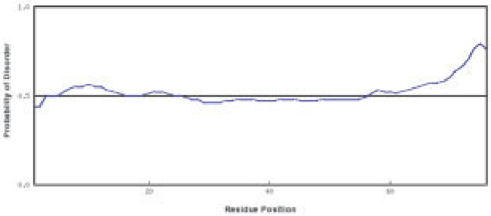
Low Predicted Secondary Structure Content

Secondary structure prediction is based on the propensity of each amino acid to belong to each type of secondary structure element, computed along sliding windows. Long (>70 aa) regions devoid of predicted secondary structure elements (as judged by using a combination of methods) are generally disordered. There are a few exceptions, called “loopy proteins,” which have no regular secondary structure and yet are ordered, like the Kringle domain, a triple-looped, disulphide-linked domain, found in some serine proteases and in some plasma proteins.[22]

Low-Sequence Complexity

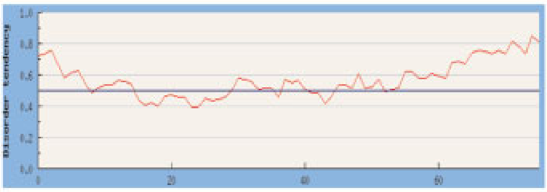
Low Complexity Regions (LCRs) are regions with a biased composition (homopolymeric runs, short-period repeats, and more subtle overrepresentation of a few residues), making use of fewer types of amino acids. Intrinsically disordered proteins tend to have a low sequence complexity, although it is not a general rule.[23,24] It has been shown recently that the more low-complexity regions an eukaryotic protein has, the less it is likely to be solubly expressed in bacteria. This might be related to the fact that low-complexity regions, which are more frequent in eukaryotic proteins than in bacterial proteins, are more sensitive to proteolytic degradation.[25] Given the tendency of IDPs to have low complexity, one could expect that they are less soluble than globular proteins. However, this is not a general rule. For instance, the intrinsically disordered N-terminal domain of the measles virus phosphoprotein is even more soluble than the structured C-terminal domain (see Karlin et al.[26] and Longhi et al.[27]). Likewise, the intrinsically disordered C-terminal domain of the measles virus nucleoprotein has a solubility comparable to that of its structured domain (see Longhi et al.[27]). Furthermore, intrinsically disordered proteins are less prone to aggregation compared to globular proteins,[28,29] which facilitates their purification and conservation.

RONN



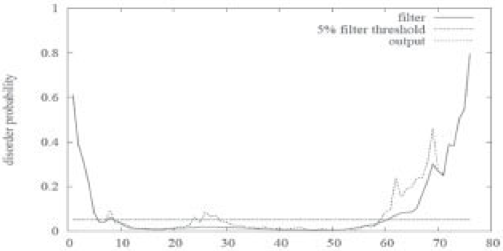
aa 1-56 have borderline disorder (probability very close to 0.5). Region 57-76 is predicted as disordered

IUPred

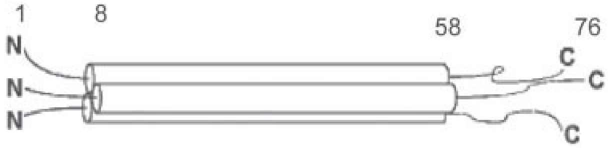


aa 10-53 have borderline disorder (probability very close to 0.5). The remaining regions (N and C-termini) are predicted as disordered

Disopred2

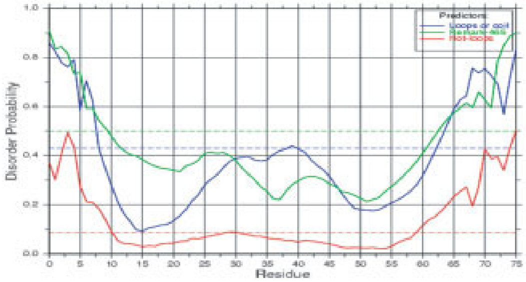


aa 7-60 are predicted as ordered. The remaining regions (N and C-termini) are predicted as disordered



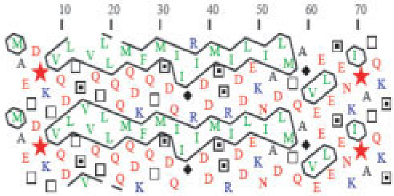
Structural model of the Heat Shock Factor-binding protein

Disembl



The 3 predictors (coil, Remark465, Hot loops) predict the region around aa 10-61 to be ordered. The remaining regions (N and C-termini) are predicted as disordered

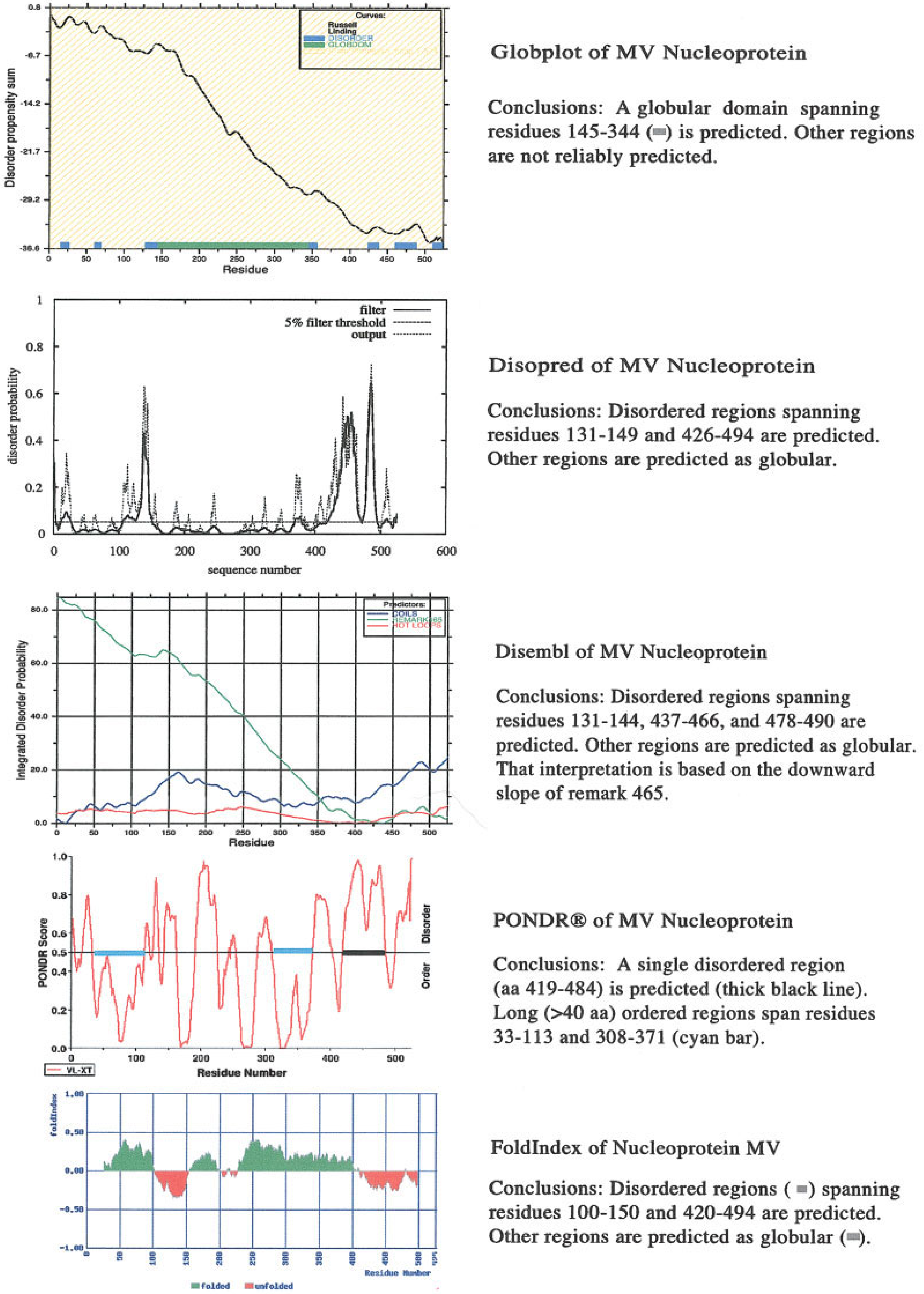
HCA



Typical coiled-coil pattern for aa 8-56 (long, horizontal cluster). This region **reg** is strikingly rich in Q, D (in red) .

Predictor	Disordered region
Prelink	66 - 76
Globplot	3 - 7 , 72 - 76
PONDR VSL1	1 - 76.
PONDR VL-XT	1 - 11, 50 - 76
Foldindex	Not applicable (protein is too short)
SEG 25 3.0 3.3 (to detect short non-globular regions)	7 - 33
SEG 45 3.4 3.75 (to detect long non-globular regions)	

Fig. 2. Analysis of the human heat shock factor-binding protein 1 (Genbank accession number: AF068754) using different predictors. The graphical output of each method and the corresponding interpretation is shown. The precise boundaries of ordered and disordered regions were derived from the corresponding text output (not shown). Bottom, structural model of the protein reproduced from ref. 59 with permission of Richard Morimoto and the American Society for Biochemistry and Molecular Biology. The N- and C-termini (thin lines) are disordered, whereas the central region forms a triple coiled-coil (cylinders). Numbers correspond to the amino acid boundaries of these regions. SEG parameters are explained in Table II.



Globplot of MV Nucleoprotein

Conclusions: A globular domain spanning residues 145-344 (▬) is predicted. Other regions are not reliably predicted.

Disopred of MV Nucleoprotein

Conclusions: Disordered regions spanning residues 131-149 and 426-494 are predicted. Other regions are predicted as globular.

Disembl of MV Nucleoprotein

Conclusions: Disordered regions spanning residues 131-144, 437-466, and 478-490 are predicted. Other regions are predicted as globular. That interpretation is based on the downward slope of remark 465.

PONDR® of MV Nucleoprotein

Conclusions: A single disordered region (aa 419-484) is predicted (thick black line). Long (>40 aa) ordered regions span residues 33-113 and 308-371 (cyan bar).

FoldIndex of Nucleoprotein MV

Conclusions: Disordered regions (▬) spanning residues 100-150 and 420-494 are predicted. Other regions are predicted as globular (▬).

Fig. 3. Measles virus (MV) nucleoprotein (N) (accession number: P35972) analyzed with different predictors. The graphical output of each method and the corresponding interpretation is shown. The precise boundaries of ordered and disordered regions were derived from the corresponding text output (not shown). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

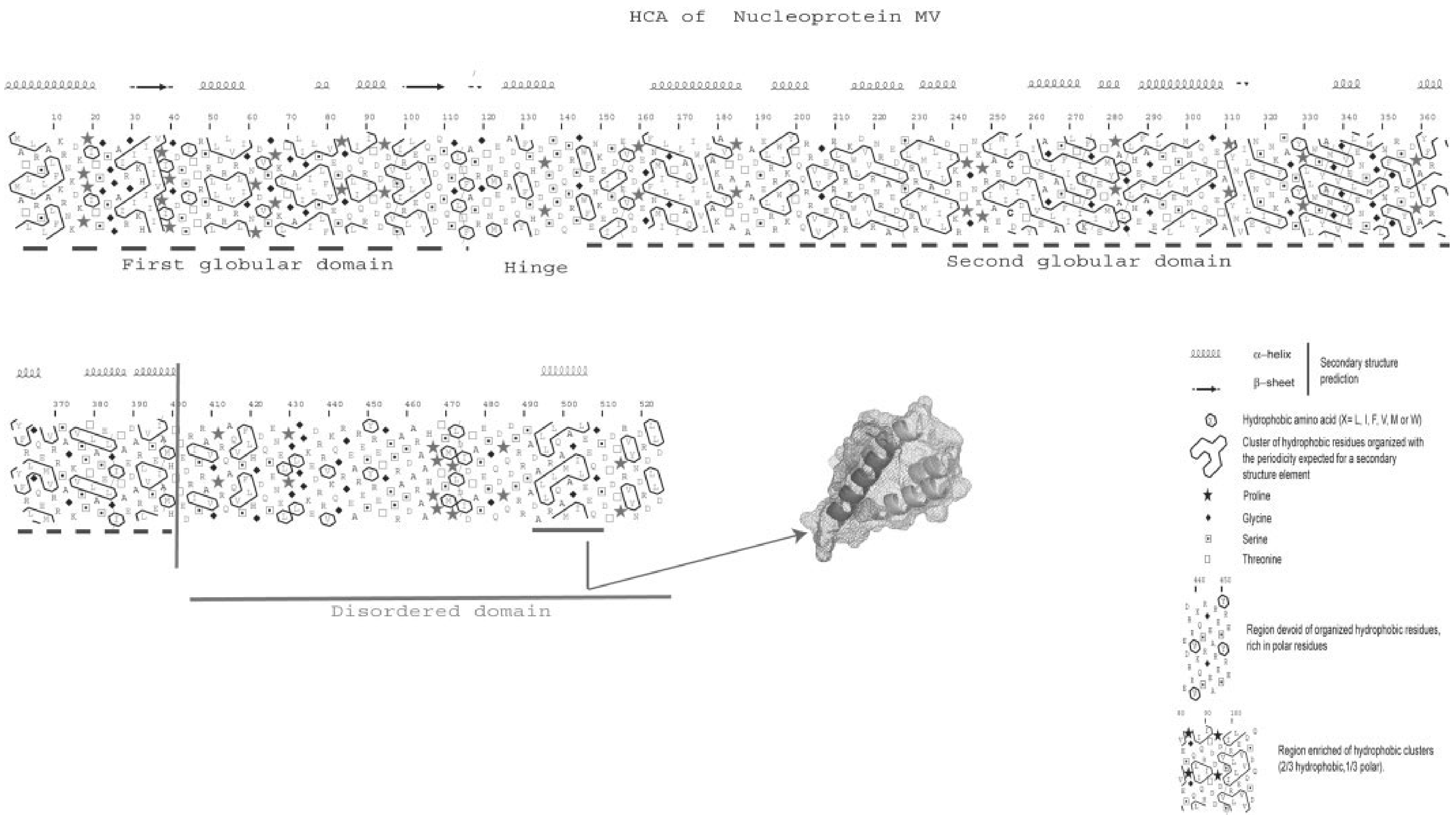
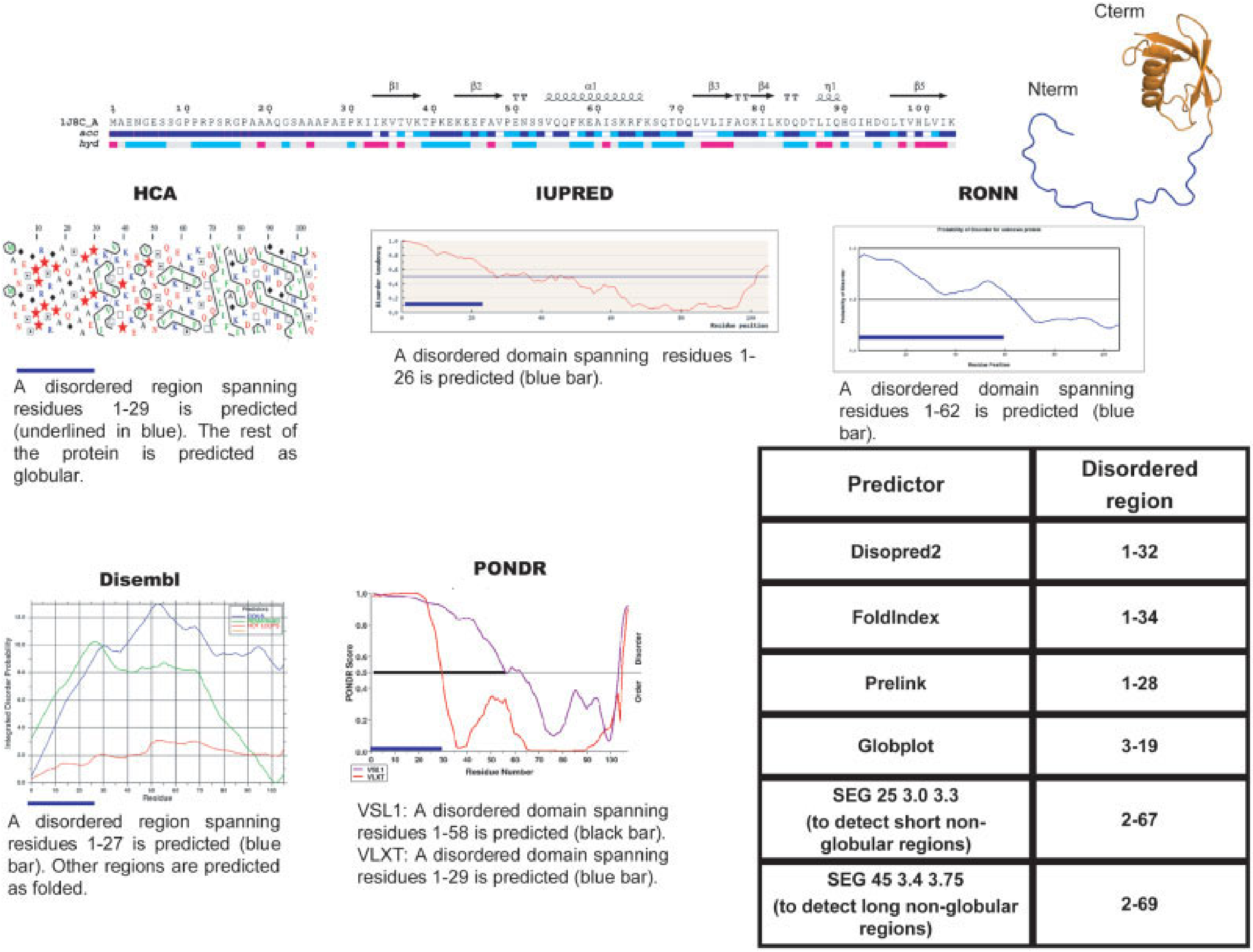

HCA of Nucleoprotein MV
First globular domain
Hinge
Second globular domain
Disordered domain
α-helix
β-sheet
Secondary structure prediction
Hydrophobic amino acid (X= L, I, F, V, M or W)
Cluster of hydrophobic residues organized with the periodicity expected for a secondary structure element
Proline
Glycine
Serine
Threonine
Region devoid of organized hydrophobic residues, rich in polar residues
Region enriched of hydrophobic clusters (2/3 hydrophobic, 1/3 polar).

Fig. 4. HCA plot of measles virus nucleoprotein. Conventions are explicited in the caption. Globular regions (framed) are characterized by a thick distribution of hydrophobic clusters, while unstructured regions are poor or devoid of hydrophobic clusters. Long disordered regions and predicted secondary structure elements are shown. There is no low-complexity region in measles virus N. The induced folding region is underlined, and the corresponding structure (dark gray α-helix) is presented in complex with the C-terminal domain of the measles virus P[75] (PDB code: 1T6O). The picture of the measles virus P was obtained using Pymol.



A disordered region spanning residues 1-29 is predicted (underlined in blue). The rest of the protein is predicted as globular.

A disordered domain spanning residues 1-26 is predicted (blue bar).

A disordered domain spanning residues 1-62 is predicted (blue bar).

A disordered region spanning residues 1-27 is predicted (blue bar). Other regions are predicted as folded.

VSL1: A disordered domain spanning residues 1-58 is predicted (black bar).
VLXT: A disordered domain spanning residues 1-29 is predicted (blue bar).

Predictor	Disordered region
Disopred2	1-32
FoldIndex	1-34
Prelink	1-28
Globplot	3-19
SEG 25 3.0 3.3 (to detect short non-globular regions)	2-67
SEG 45 3.4 3.75 (to detect long non-globular regions)	2-69

Fig. 5. Analysis of the Ubiquitin-like protein domain of hPLIC-2 (accession number: Q9UHD9) using different predictors. The protein sequence, with the secondary structure elements derived from the structure (PDB code 1J8C), is shown. The solvent accessibility of each residue is plotted below the sequence. The ribbon representation of the structure is shown; the globular domain is colored in gold while the disordered region is in blue. The picture was obtained using Pymol. The graphical output of various prediction methods and the corresponding interpretation are shown. The precise boundaries of ordered and disordered regions were derived from the corresponding text output (not shown). A blue bar highlights the disordered region for each graphical output, except for VSL1 for which the disordered region is highlighted with a black line. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Some special cases of low-complexity sequences are found in proteins with a certain amino acid periodicity (such as coiled-coils) and other nonglobular, yet ordered proteins (collagen, for example). It is recommended to always look for LCRs, coiled-coils, and repeats in a protein prior to further sequence analysis (using programs such as Paircoil[30] and Multicoil[31]). More subtle parameters to discriminate between globular and nonglobular proteins using the program SEG are discussed in refs. 21 and 24 (see also Table I).

High-Sequence Variability

Disordered regions are on average much more variable than ordered ones.[32] The reason why they evolve faster is not clear at present. The relationship between sequence variability and flexibility is well known by crystallographers: when a protein does not crystallize despite repeated attempts, crystallographers are used to removing hypervariable regions, presumed to be flexible linkers. High-sequence variability is not by itself an evidence of disorder, but only an indicator. A simple method to appreciate sequence variability is visual inspection of a multiple sequence alignment. However, it can sometimes be misleading. Programs that rely on nucleotide substitution rate (as described by Brown et al.[32]) can be very informative and should be used for a more rigorous analysis.[33]

PREDICTORS OF DISORDER

Several programs have been developed to predict disordered regions using the sequence features reviewed above. They are presented with their philosophy in Table I, and their salient points are briefly discussed below. A detailed description of each predictor is outside the scope of this review, and the reader interested in more details on a specific predictor is invited to refer to the relevant article, indicated in Table I.

Predictors of Disorder

Different predictors rely on different physicochemical parameters. Therefore, a given predictor can be more performant in detecting a given feature of a disordered protein. Thus, predictors are complementary, a point illustrated in the section focused on practical examples (see below). As there is no consensus on what disorder means, it is necessary to know precisely what is predicted by each method. For instance, "long disordered regions" predicted by PONDR correspond to regions that are not rigid including random coils, partially unstructured regions, and molten globules (Table I). On the contrary, if a protein is predicted to be unstructured by the charge/hydropathy method, it means that it is probably fully unstructured (random coil) (Table I). This issue has been stressed recently by a systematic comparison between these two prediction methods.[34]

Most predictors rely on training against a dataset of disordered protein regions. These datasets are either entirely built up by the authors or represent improvements of existing datasets. Despite continuous efforts, these datasets retain some inconsistencies and are necessarily biased, because large regions of disorder can prevent crystallization. Furthermore, these datasets contain relatively few disordered proteins. Indeed, Disprot (http://www.disprot.org/),[35] which is the largest publicly available database of disordered proteins whose disorder has been experimentally assessed, contains only about 400 entries. For these reasons, it is useful to distinguish two kinds of predictors: those that have been trained on datasets of disordered proteins (PONDR, Globplot, Disembl, Disopred2, RONN, PreLink), and those that have not, namely the charge/hydropathy method (and its derivative Foldindex), NORSp, and IUPred. The latter avoid the shortcomings and biases associated to the disordered datasets. Therefore, they are expected to perform better than the former methods on disordered proteins presently under-represented in training datasets.

PONDR, a neural network based on local amino acid composition, flexibility, and other sequence features, was the first predictor to be developed (Table I). It is now available in various versions, each having its own specificities (e.g., VL-XT allows highlighting of potential protein-binding regions), and Table II suggests which version should be chosen according to the user's goal. Noteworthy, PONDR has found an application in the study of structured proteins. Indeed, there is a strong inverse correlation between the VL-XT score within *ordered* regions and the presence of dehydrons, which are underwrapped backbone hydrogen bonds, recently identified as a major determinant of protein–protein interactions.[36,37]

Globplot uses a new scale called "Russell/Linding," specially developed to express the propensity for a given amino acid to be in "random coil" or in "regular secondary structure" (Table I). In Globplot, changes of slope often correspond to domain boundaries (Table II).

Disembl consists of three separate predictors, trained on separate datasets, that respectively comprise residues within "loops/coils" (as defined by DSSP[38]), "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") (Table I). The partitioning of residues into different flexibility groups is very useful depending upon the user's goals (i.e., "hot loop" may be used to correlate certain functional aspects of proteins with mobile loops, while "Remark 465" may be used to detect linkers likely to affect crystallization).

Disopred2 (Table I) is also based on a neural network, but incorporates information from multiple sequence alignments because its inputs are derived from sequence profiles generated by PSI-BLAST.

RONN (Table I) uses a novel approach, a bio-basis function neural network. It relies on the calculation of "distances," as determined by sequence alignment, from well-characterized prototype sequences (ordered, disordered, or a mixture of both). Its key feature is that amino acid side chain properties are not considered at any stage.

Prelink (Table I) relies on amino acid composition and on low hydrophobic cluster content. In this respect, it is a derivative of HCA, a powerful approach that is discussed below. PreLink is the first predictor that statistically proved the ability of HCA to detect linkers, an ability that had long been noticed before but never previously demonstrated.

The charge/hydropathy analysis 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), Thus, globular proteins can be distinguished from unstructured ones based on the ratio of their net charge versus their hydropathy (Table I). A drawback of this approach is that it gives only a global indication, not valid if the protein is composed of both ordered and disordered regions (Table II). A derivative of that method, Foldindex, solves this problem by computing the charge/hydropathy ratio along the protein (Table I).

NORSp (Table I) relies on the principle that long regions predicted to be devoid of secondary structure and accessible to the solvent are generally unstructured. However, this is not always true, as in the case of the Kringle domain mentioned above.

IUPred uses a novel algorithm that evaluates the energy resulting from interresidues interactions. 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 results of another study.[39]

The program SEG (Table I), which computes sequence complexity, has not been developed to detect disordered regions but has been used successfully in that aim by the group of Koonin.[21] Typical SEG parameters for disorder prediction are found in Table II.

Table I also includes a nonautomated method that is very useful for unveiling unstructured regions: Hydrophobic Cluster Analysis (HCA).[40] HCA makes use of a twodimensional helical representation of protein sequences in which hydrophobic clusters are plotted along the sequence (the reader is invited to refer to the excellent review by Callebaut et al.[40]). HCA stands aside from other predictors, because they only give insights on the extent of disorder/order, but do not correlate this information with the sequence by itself. Furthermore, there is little one can actually *learn* from comparing the output of these predictors for homologous proteins. In contrast, HCA 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 paragraph on induced folding below). Although HCA does not provide a quantitative prediction of disorder and rather presently requires human interpretation, it provides additional, qualitative information compared to automated predictors, a point illustrated in the section focused on practical examples (see below).

Error Rate of Predictors

A general error rate is difficult to evaluate, because it depends of the definition of disorder used, on the evaluation set, and on the criteria of evaluation. These points are well illustrated by the evaluation of disorder predictors within the recent Critical Assessment of Protein Structure Prediction, CASP6, where very different rankings were obtained as a function of the criteria used.[41] Moreover, the accuracy of a given predictor can be limited when predicting a type of disorder different from that against which it was trained.

Another reason that prevents the meaningful calculation of a precise error rate is the fact that a protein can be disordered by itself, and yet adopt a structure either in a cellular context (when binding to a partner, a phenomenon called "induced folding") or because of artefacts [crystal contacts during crystallization, for instance, or structure solved in a nonaqueous medium such as trifluoroethanol (TFE)]. Many prediction "errors" fall, in fact, in these categories, as discussed in two recent articles by the groups of Poupon and of Dunker (see references therein cited, and Figs. 4–5 in reference[42] and Figure 6 in ref. 34).

Because of all the reasons stated above, error rates are not given in Table I. In general, predictors are more reliable in predicting order than in predicting disorder, as (1) ordered sequences comprise only a very narrow portion of sequence space, that is, their sequence properties are much more recognizable; and (2) because of the limited number of disordered protein sequences available for predictor training. From the authors' personal communications, a conservative accuracy for ab initio methods, such as Disopred2, Disembl, and PONDR, is around 60–70% for predicting disorder, and about 80% for predicting order. Reportedly, the charge/hydropathy method has the best overall accuracy (83%[39]). However, this method requires prior knowledge of the domain boundaries (see Table I). Despite the inherent difficulty of estimating meaningful error rates, recent studies have pointed out that disorder predictors can been grossly classified into three categories.[43,44] These categories are by no means absolute, and are presented only for convenience, to allow a better interpretation of the results provided by the predictors and to optimize their use in combination.

Some predictors perform better on short disordered regions in the context of globally ordered proteins: Disopred2, Prelink, and Disembl (Remark465). Actually, they were specifically developed with that aim. These predictors also have a good specificity (i.e., they predict relatively few ordered residues to be disordered), but a moderate sensitivity (i.e., they miss a significant number of disordered residues). IUPred performs comparatively well for predicting long disordered segments, and has a good sensitivity. Finally, although no method has both a very high specificity and a very high sensitivity, some predictors are "polyvalent" (RONN, PONDR VSL1, Disembl "hot loops," FoldIndex and Globplot).

However, we would like to point out once again that these observations are only aimed at guiding the user, and in no way they are intended to provide a quantitative comparison. The section focused on practical examples (see below) will also give a qualitative idea of the respective sensitivities and specificities of these methods.

PREDICTING INDUCED FOLDING

The analysis of hydrophobic clusters and of secondary structures is of major interest for studying induced folding, because burial of hydrophobic residues provides the major driving force in protein folding. This force is, in turn, regulated by secondary structures that play a role in guiding the folding pathway. In some cases, hydrophobic clusters are found within secondary structure elements that are unstable in the native protein, but can stably fold upon binding a partner. Therefore, HCA can be very informative in highlighting potential induced folding. As an example, we suspected that the isolated hydrophobic cluster with a predicted α-helix within the disordered N_{TAIL} domain (Fig. 4) could correspond to a binding region for one of the partners of *N*, the viral phosphoprotein *P*, and that *N* would undergo induced folding upon binding *P*. Later, this hypothesis was proven experimentally[27,45–47] (see region highlighted by a bar in Fig. 4).

Molecular Recognition Elements (MoREs) are regions within an intrinsically disordered protein that have a propensity to bind to a partner and thereby to undergo induced folding. It has long been noticed that PONDR can highlight potential MoREs[48] (Table II). For instance, a fine analysis of PONDR plots led to the identification of segments of increased structural propensity (i.e., prone to induced folding) in the RNA degradosome-organizing domain of the *Escherichia coli* ribonuclease RNase E.[49] The group of Dunker recently developed a program to identify α-helix-forming MoREs (called α-MoREs) from the amino acid sequence.[50] For instance, the above-mentioned region within N_{TAIL} that undergoes an α-helical transition upon binding to *P*, was successfully identified. Linding also showed how neural network predictions of disorder can indicate the propensity of ELMs to undergo induced folding.[6]

PRACTICAL EXAMPLES SHOWING HOW COMBINING DIFFERENT METHODS IMPROVES DISORDER PREDICTION

Figure 1 illustrates a general sequence analysis scheme that integrates the peculiarities of each method to predict globular and disordered regions. As a first step, one should perform an analysis of sequence composition[51] and complexity,[24] a search for signal peptides, transmembrane regions,[52] leucine zippers,[53] and coiled-coil regions,[54–56] to premark regions of biased composition. This step is crucial in that it can avoid pitfalls that can lead to misspredictions, as exemplified in the next section.

It is also recommended to use DIpro[57] to identify possible disulfide bridges and to search for possible metal-binding regions by looking for conserved Cys_3–His or Cys_2–His_2 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.[58]

Then, ab initio methods, such as Globplot, Disembl, PONDR, Disopred2, IUPred, RONN, Prelink, Foldindex, and NORSp (Table I) can be combined to define a consensus on both globular and unstructured regions. Of course, any supplemental information, as for instance sequence similarity of a protein region to multidomain proteins, are precious in terms of domain boundary definition. 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 (see above).

An Example of a Pitfall: A Coiled-Coil

To illustrate a possible pitfall in disorder prediction, we have chosen the Heat-Shock Factor binding Protein 1. Biophysical and biochemical analyses have shown that it consists of a long, trimeric coiled-coil, with the N- and C-termini (respectively aa 1–8 and 58–76) being disordered.[59] Foldindex was not used, as it could not give reliable predictions due the small size of the protein (Table II). PONDR VSL1 predicts the whole protein as disordered (Fig. 2). IUPred and RONN predict borderline disorder for most of the protein (Fig. 2) and the C-terminus as disordered. SEG does not detect any long, nonglobular region (Fig. 2) (using the parameters shown in Table II). However when using more sensitive parameters (Table II), it detects a medium-length region of biased composition (aa 7–33). All other automated predictors predict the central region as ordered and the N- and C-termini as disordered (Fig. 2). A preliminary analysis using Multicoil[31] and HCA would have solved these discrepancies. Multicoil gives a high probability of coiled-coil over aa 30–60 (not shown), while the HCA plot is typical of a coiled-coil (a long and horizontally extended hydrophobic cluster encompassing aa 8–56) (Fig. 2). Furthermore, it is quite obvious from the plot that the protein has a biased composition, being rich in Q and D residues (noticeable thanks to their red color; see Fig. 2). In particular, the Q-rich region roughly corresponds to the low-complexity region detected by SEG (aa 7–33).

Thus, performing the preliminary analysis shown in Figure 1 would have allowed detection of a coiled-coil (which fooled some predictors into giving a wrong prediction of borderline disorder) and would have overcome this pitfall, while giving precious information on the protein (biased composition). Once the structural status of the region 8–60 has been established as a coiled-coil, the comparative analysis of the ensemble of the results gives a more accurate prediction. Indeed, almost all predictors correctly predict disordered N- and C-termini with reasonably accurate boundaries (Fig. 2). This example also illustrates the advantage of using predictors that rely on different principles: for instance, because Prelink is based on HCA, it is expected to correctly predicted coiled-coils as ordered. As another example, PONDR VL-XT, gives a correct prediction, whereas another version of PONDR, VSL1, optimized to detect short disordered regions (Table II), is completely fooled by the coiled-coil, and predicts it as disordered.

Domain Identification

Figures 3 and 4 illustrate the approach used to study the domain organization of the nucleoprotein (N) of measles virus, a protein that encapsidates the viral RNA. Experimental data available indicate that N is organized into two regions, N_{CORE} (aa 1–399) and N_{TAIL} (aa 400–525), respectively ordered[60] and disordered.[46,61] As shown in Figure 3, most ab initio methods converge to show the presence of a disordered region at its C-terminus (consensus is aa 437–484), and of a globular core (aa 145–344). Interestingly, Foldindex (Table I) highlights a very hydrophilic region (aa 100–150) that is also visible as a short plateau (aa 131–144) in the output of Disembl Remark 465 predictor and that is predicted by Disopred2 too (aa 131–149). Moreover, this region is hypervariable in sequence among *Morbillivirus* members (not shown). Finally, because changes in slope of Globplot often correspond to domain boundaries (see Table I), from this analysis one would suspect the following domain organization: a first domain or subdomain encompassing residues 1–130, that is not confidently predicted but might be ordered (cf. the negative slope of Globplot together with PONDR prediction); an exposed loop spanning aa 131–149; a second, more compact domain (aa 150–400, cf. steep negative slope), and a disordered domain encompassing aa 401–525. Finally, the charge/hydropathy method predicts that both suspected subdomains are ordered and confirms that the C-terminal domain is disordered (not shown).

HCA helps to refine these predictions. As shown in Figure 4, the density of hydrophobic clusters indicates without ambiguity that both subdomains identified by the combination of previous methods are ordered, and the lack of hydrophobic clusters within the 422–525 region indicates that it cannot be ordered by itself (the hydrophobic clusters in the 494–525 region are not long enough to lead to the formation of a compact domain).

Thus, no single method, nor even a combination of two predictors, could successfully unveil the organization of measles virus *N*, whereas the combined use of all predictors proved to be much more powerful in terms of domain boundary recognition. The hypervariable region (aa 131–149) is indeed accessible to antibodies and thus exposed to the solvent.[62] However, a wealth of mutational data (see Karlin et al.[60] and references therein) indicates that N_{CORE} cannot be divided into independent modules, but rather that the subdomains indicated above (aa 1–130 and aa 145–400) probably fold cooperatively. Thus, the exposed region (aa 131–149) is probably a loop and not a linker that would connect two mobile domains. Whether it is disordered or not is not known. However, as it is not sensitive to proteolysis[60] it is probably at least partially ordered. These unsolved issues nicely illustrate the present limits of disorder prediction.

Manual Refinement of Domain Boundaries using HCA

Figure 5 illustrates a frequently encountered case in disorder prediction, namely the occurrence of an extended region of intermediate length (20 > aa < 40) at one extremity of the protein followed by a globular domain (~70 aa). We have used the example of the Ubiquitin-like domain of hPLIC-2, whose structure has been solved by NMR.[63] As shown in Figure 5 (top), the region encompassing residues 1–31 is devoid of regular secondary structure elements and is extended in solution. All predictors detect a disordered region at the amino terminus (Fig. 5); however, the predicted boundaries of the disordered region vary from one predictor to another, with a predicted length ranging from 19 to 69 residues (see Fig. 5). Globplot predicts disorder for the 1–20 region, four predictors (Prelink, DisEMBL, VL-XT, and IUPred) define the C-terminal boundary of the disordered region around residue 28, two predictors (Disopred and Foldindex) predict a disordered region spanning residue 1–32, two predictors (VSL1 and RONN) extend the prediction of disorder to the region encompassing residues 1–60 (see Fig. 5), and SEG predicts a potential nonglobular region within aa 2–69.

Based on these results, the user can be confident than the N-terminal moiety of the protein is disordered, although the exact C-terminal boundary of the unstructured region remains uncertain (predictions vary from residue 19 to residue 69!) Use of HCA helps to reduce this uncertainty. The HCA plot clearly allows the identification of the 1–30 region as disordered, based on its almost total depletion in hydrophobic residues, and of the 53–103 region as ordered, given its high density in hydrophobic clusters. Thus, one can confidently predict that the protein is organized into two moieties, with HCA having narrowed the boundary between these two regions down to the 30–52 region. In the absence of functional or biochemical clues (such as limited proteolysis studies), the production of various truncated versions of each moiety is recommended in view of functional or structural studies. Such constructs should start or end at incremental positions between residues 30 and 52 (i.e., at the ends of predicted secondary structure elements; not shown). Indeed, the experience that we gained in the context of past and present structural genomics projects developed in our laboratory (see SPINE and VIZIER projects at http://www.afmb.univ-mrs.fr/-The-Spine-Program- and http://www.afmb.univ-mrs.fr/-VIZIER-) has shown that a critical factor in obtaining good-quality protein crystals is the number of constructs generated around the predicted boundary of the domain under study.

CONTRIBUTION OF DISORDER PREDICTIONS TO BIOINFORMATICS ANALYSES

As already mentioned, the identification of disorder can also avoid gross mistakes in protein sequence analysis. For instance, Iyer et al.[4] recently reported two examples in which the SEG program (Table I), in combination with multiple alignment and secondary structure prediction, invalidates previous functional assignments for two proteins, ATF-2 and PIF3, made on the basis of distant sequence similarity to two domains (respectively histone acetyltransferase (HAT) and PAS domain). The HAT and PAS domains are globular, whereas the similar regions of ATF-2 and PIF3 are confidently predicted to be unstructured, casting a strong doubt on their suspected homology.

Once regions of disorder have been identified, then further bioinformatics analyses, aimed at identifying related proteins, can be carried out avoiding spurious sequence similarity. For instance, Rabitsch et al. illustrated how to perform search for homologs of proteins composed mostly of unstructured regions (Sgo1 and Sgo2). They searched for candidate proteins having short stretches of sequence or structural similarity (i.e., presence of a coiled-coil) to Sgo1 and Sgo2, and distributed in a similar fashion as compared to the candidate protein sequence.[64]

Disorder prediction can also greatly help to identify short modules. For instance, it was instrumental in identifying five novel groups of Lsm domain proteins.[65] The architecture of these proteins, which guided the research for sequence similarities, was elucidated using the consensus of Globplot, Disembl, NORSp, and PONDR analyses. The authors identified conserved motifs described as “stable islands in a large sea of intrinsically unstructured sequence regions.” This is probably true of many large human proteins for which very short conserved motifs in the middle of long disordered regions remain to be discovered.[7] Another method that can be of great use in identifying short (50–70 aa), globular domains located within long, disordered regions (e.g., chromodomains) is HCA.[40]

CONCLUSION

As we have seen from two detailed examples, no single predictor can reveal the structural organization of a protein. However, in combination they provide relatively accurate results. Thus, there is obvious room for improvement of predictors by combining features of several programs (i.e., by including information on predicted secondary structure elements, or information derived from multiple sequence alignments). It would also be of major interest to check whether known regions of induced folding correlate well with isolated hydrophobic clusters, corresponding to predicted α-helices, within disordered regions.[45–47] Other improvements may arise from a better understanding of the different types, or "flavors" of disorder.[66]

As a last, optimistic note, one should never give up carrying out a crystallization experiment because of an order/disorder prediction: recently, Mavrakis et al. submitted the phosphoprotein of rabies virus to crystallization trials, despite the fact that the N-terminal moiety, which accounts for more than half of the protein, was predicted to be disordered. Crystals formed in the drop. In fact the N-terminus had been cleaved off by contaminating proteases and the crystals were made of the C-terminal part . . . whose structure was readily solved![67]

ACKNOWLEDGMENTS

We thank K. Dunker, P. Romero, J. Ward, R. Linding, V. Uversky, J. Wootton, I. Callebaut, A. Poupon, R. Esnouf, and Z. Dosztányi for their useful comments on their respective predictors. We are also grateful to V. Receveur-Bréchot and Sung Haeng Lee for critical reading of the manuscript.

REFERENCES

1. Tompa P. Intrinsically unstructured proteins. Trends Biochem Sci 2002;27:527–533.
2. Receveur-Bréchot V, Bourhis JM, Uversky VN, Canard B, Longhi S. Assessing protein disorder and induced folding. Proteins Struct Funct Bioinformat 2006;62:24–45.
3. Uversky VN. Natively unfolded proteins: a point where biology waits for physics. Protein Sci 2002;11:739–756.
4. Iyer LM, Aravind L, Bork P, Hofmann K, Mushegian AR, Zhulin IB, Koonin EV. Quoderat demonstrandum? The mystery of experimental validation of apparently erroneous computational analyses of protein sequences. Genome Biol 2001;2:RESEARCH0051.
5. Puntervoll P, Linding R, Gemund C, Chabanis-Davidson S, Mattingsdal M, Cameron S, Martin DM, Ausiello G, Brannetti B, Costantini A, Ferre F, Maselli V, Via A, Cesareni G, Diella F, Superti-Furga G, Wyrwicz L, Ramu C, McGuigan C, Gudavalli R, Letunic I, Bork P, Rychlewski L, Kuster B, Helmer-Citterich M, Hunter WN, Aasland R, Gibson TJ. ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. Nucleic Acids Res 2003;31:3625–3630.
6. Linding R. Linear functional modules. Implication for protein function. PhD Thesis, University of Heidelberg; 2004.
7. Neduva V, Linding R, Su-Angrand I, Stark A, Masi FD, Gibson TJ, Lewis J, Serrano L, Russell RB. Systematic discovery of new recognition peptides mediating protein interaction networks. PLoS Biol 2005;3:e405.
8. Friedberg I, Jaroszewski L, Ye Y, Godzik A. The interplay of fold recognition and experimental structure determination in structural genomics. Curr Opin Struct Biol 2004;14:307–312.
9. Melamud E, Moult J. Evaluation of disorder predictions in CASP5. Proteins 2003;53(Suppl 6):561–565.
10. Obradovic Z, Peng K, Vucetic S, Radivojac P, Dunker AK. Exploiting heterogeneous sequence properties improves prediction of protein disorder. Proteins 2005;61:166–182.
11. Dosztanyi Z, Csizmok V, Tompa P, Simon I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J Mol Biol 2005;347:827–839.
12. Karlin D, Ferron F, Canard B, Longhi S. Structural disorder and modular organization in Paramyxovirinae N and P. J Gen Virol 2003;84(Pt 12):3239–3252.
13. Ferron F, Rancurel C, Longhi S, Cambillau C, Henrissat B, Canard B. VaZyMolO: a tool to define and classify modularity in viral proteins. J Gen Virol 2005;86(Pt 3):743–749.
14. Severson W, Xu X, Kuhn M, Senutovitch N, Thokala M, Ferron F, Longhi S, Canard B, Jonsson CB. Essential amino acids of the hantaan virus N protein in its interaction with RNA. J Virol 2005;79:10032–10039.
15. Ferron FP. Approches bioinformatiques et structurales des réplicase virales. Marseille: Aix-Marseille II; 2005.
16. Llorente MT, Barreno-Garcia B, Calero M, Camafeita E, Lopez JA, Longhi S, Ferron F, Varela PF, Melero JA. Structural analysis of the human respiratory syncitial virus phosphoprotein: characterization of an a-helical domain involved in oligomerization. J Gen Virol 2006;87:159–169.
17. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW, Ausio J, Nissen MS, Reeves R, Kang C, Kissinger CR, Bailey RW, Griswold MD, Chiu W, Garner EC, Obradovic Z. Intrinsically disordered protein. J Mol Graph Model 2001;19:26–59.
18. Linding R, Russell RB, Neduva V, Gibson TJ. GlobPlot: Exploring protein sequences for globularity and disorder. Nucleic Acids Res 2003;31:3701–3708.
19. Egloff MP, Benarroch D, Selisko B, Romette JL, Canard B. An RNA cap (nucleoside-2′-O-)-methyltransferase in the flavivirus RNA polymerase NS5: crystal structure and functional characterization. EMBO J 2002;21:2757–2768.
20. Weathers EA, Paulaitis ME, Woolf TB, Hoh JH. Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein. FEBS Lett 2004;576:348–352.
21. Koonin E V, Galperin M. Sequence–evolution–function: computational approaches in comparative genomics. New York: Kluwer Academic Publishers; 2003.
22. Liu J, Tan H, Rost B. Loopy proteins appear conserved in evolution. J Mol Biol 2002;322:53–64.
23. Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK. Sequence complexity of disordered protein. Proteins 2001;42:38–48.
24. Wootton JC. Non-globular domains in protein sequences: automated segmentation using complexity measures. Comput Chem 1994;18:269–285.
25. Dyson MR, Shadbolt SP, Vincent KJ, Perera RL, McCafferty J. Production of soluble mammalian proteins in *Escherichia coli*: identification of protein features that correlate with successful expression. BMC Biotechnol 2004;4:32.
26. Karlin D, Longhi S, Receveur V, Canard B. The N-terminal domain of the phosphoprotein of morbilliviruses belongs to the natively unfolded class of proteins. Virology 2002;296:251–262.
27. Longhi S, Receveur-Brechot V, Karlin D, Johansson K, Darbon H, Bhella D, Yeo R, Finet S, Canard B. The C-terminal domain of the measles virus nucleoprotein is intrinsically disordered and folds upon binding to the C-terminal moiety of the phosphoprotein. J Biol Chem 2003;278:18638–18648.
28. Linding R, Schymkowitz J, Rousseau F, Diella F, Serrano L. A comparative study of the relationship between protein structure and beta-aggregation in globular and intrinsically disordered proteins. J Mol Biol 2004;342:345–353.
29. Tartaglia GG, Pellarin R, Cavalli A, Caflisch A. Organism complexity anti-correlates with proteomic beta-aggregation propensity. Protein Sci 2005;14:2735–2740.
30. Berger B, Wilson DB, Wolf E, Tonchev T, Milla M, Kim PS. Predicting coiled coils by use of pairwise residue correlations. Proc Natl Acad Sci USA 1995;92:8259–8263.
31. Wolf E, Kim PS, Berger B. MultiCoil: a program for predicting two- and three-stranded coiled coils. Protein Sci 1997;6:1179–1189.
32. Brown CJ, Takayama S, Campen AM, Vise P, Marshall TW, Oldfield CJ, Williams CJ, Keith Dunker A. Evolutionary rate heterogeneity in proteins with long disordered regions. J Mol Evol 2002;55:104–110.
33. Hurst LD. The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends Genet 2002;18:486.
34. Oldfield CJ, Cheng Y, Cortese MS, Brown CJ, Uversky VN, Dunker AK. Comparing and combining predictors of mostly disordered proteins. Biochemistry 2005;44:1989–2000.
35. Vucetic S, Obradovic Z, Vacic V, Radivojac P, Peng K, Iakoucheva LM, Cortese MS, Lawson JD, Brown CJ, Sikes JG, Newton CD, Dunker AK. DisProt: a database of protein disorder. Bioinformatics 2005;21:137–140.
36. Fernandez A, Berry RS. Molecular dimension explored in evolution to promote proteomic complexity. Proc Natl Acad Sci USA 2004;101:13460–13465.

37. Fernandez A, Scott R, Berry RS. The nonconserved wrapping of conserved protein folds reveals a trend toward increasing connectivity in proteomic networks. Proc Natl Acad Sci USA 2004;101: 2823–2827.
38. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
39. Garbuzynskiy SO, Lobanov MY, Galzitskaya OV. To be folded or to be unfolded? Protein Sci 2004;13:2871–2877.
40. Callebaut I, Courvalin JC, Worman HJ, Mornon JP. Hydrophobic cluster analysis reveals a third chromodomain in the Tetrahymena Pdd1p protein of the chromo superfamily. Biochem Biophys Res Commun 1997;235:103–107.
41. Jin Y, Dunbrack RL Jr. Assessment of disorder predictions in CASP6. Proteins 2005;61:167–175.
42. Coeytaux K, Poupon A. Prediction of unfolded segments in a protein sequence based on amino acid composition. Bioinformatics 2005;21:1891–1900.
43. Dosztanyi Z, Csizmok V, Tompa P, Simon I. IUPred: Web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 2005;21:3433–3434.
44. Yang ZR, Thomson R, McNeil P, Esnouf RM. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. Bioinformatics 2005;21: 3369–3376.
45. Johansson K, Bourhis JM, Campanacci V, Cambillau C, Canard B, Longhi S. Crystal structure of the measles virus phosphoprotein domain responsible for the induced folding of the C-terminal domain of the nucleoprotein. J Biol Chem 2003;278:44567–44573.
46. Bourhis JM, Johansson K, Receveur-Brechot V, Oldfield CJ, Dunker KA, Canard B, Longhi S. 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 Res 2004;99:157–167.
47. Kingston RL, Baase WA, Gay LS. Characterization of nucleocapsid binding by the measles virus and mumps virus phosphoproteins. J Virol 2004;78:8630–8640.
48. Garner E, Romero P, Dunker AK, Brown C, Obradovic Z. Predicting binding regions within disordered proteins. Genome Inform Ser Workshop Genome Inform 1999;10:41–50.
49. Callaghan AJ, Aurikko JP, Ilag LL, Gunter Grossmann J, Chandran V, Kuhnel K, Poljak L, Carpousis AJ, Robinson CV, Symmons MF, Luisi BF. Studies of the RNA degradosome-organizing domain of the *Escherichia coli* ribonuclease RNase E. J Mol Biol 2004;340:965–979.
50. Oldfield CJ, Cheng Y, Cortese MS, Romero P, Uversky VN, Dunker AK. Coupled folding and binding with alpha-helix-forming molecular recognition elements. Biochemistry 2005;44: 12454–12470.
51. Wilkins MR, Gasteiger E, Bairoch A, Sanchez JC, Williams KL, Appel RD, Hochstrasser DF. Protein identification and analysis tools in the ExPASy server. Methods Mol Biol 1999;112:531–552.
52. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res 2003;31:3784–3788.
53. Bornberg-Bauer E, Rivals E, Vingron M. Computational approaches to identify leucine zippers. Nucleic Acids Res 1998;26: 2740–2746.
54. Lupas A. Prediction and analysis of coiled-coil structures. Methods Enzymol 1996;266:513–525.
55. Lupas A, Van Dyke M, Stock J. Predicting coiled coils from protein sequences. Science 1991;252:1162–1164.
56. Lupas A. Predicting coiled-coil regions in proteins. Curr Opin Struct Biol 1997;7:388–393.
57. Baldi P, Cheng J, Vullo A. Large-scale prediction of disulphide bond connectivity. Adv Neural Inf Process Syst 2004;17:97–104.
58. Uversky VN, Gillespie JR, Fink AL. Why are “natively unfolded” proteins unstructured under physiologic conditions? Proteins 2000;41:415–427.
59. Tai LJ, McFall SM, Huang K, Demeler B, Fox SG, Brubaker K, Radhakrishnan I, Morimoto RI. Structure–function analysis of the heat shock factor-binding protein reveals a protein composed solely of a highly conserved and dynamic coiled-coil trimerization domain. J Biol Chem 2002;277:735–745.
60. Karlin D, Longhi S, Canard B. Substitution of two residues in the measles virus nucleoprotein results in an impaired self-association. Virology 2002;302:420–432.
61. Longhi S, Receveur-Brechot V, Karlin D, Johansson K, Darbon H, Bhella D, Yeo R, Finet S, Canard B. The C-terminal domain of the measles virus nucleoprotein is intrinsically disordered and folds upon binding to the C-terminal moiety of the phosphoprotein. J Biol Chem 2003;278:18638–18648.
62. Giraudon P, Jacquier MF, Wild TF. Antigenic analysis of African measles virus field isolates: identification and localisation of one conserved and two variable epitope sites on the NP protein. Virus Res 1988;10:137–152.
63. Walters KJ, Kleijnen MF, Goh AM, Wagner G, Howley PM. Structural studies of the interaction between ubiquitin family proteins and proteasome subunit S5a. Biochemistry 2002;41:1767–1777.
64. Rabitsch KP, Gregan J, Schleiffer A, Javerzat JP, Eisenhaber F, Nasmyth K. Two fission yeast homologs of *Drosophila* Mei-S332 are required for chromosome segregation during meiosis I and II. Curr Biol 2004;14:287–301.
65. Albrecht M, Lengauer T. Novel Sm-like proteins with long C-terminal tails and associated methyltransferases. FEBS Lett 2004;569:18–26.
66. Vucetic S, Brown C, Dunker K, Obradovic Z. Flavors of protein disorder. Proteins 2003;52:573–584.
67. Mavrakis M, McCarthy AA, Roche S, Blondel D, Ruigrok RW. Structure and function of the C-terminal domain of the polymerase cofactor of rabies virus. J Mol Biol 2004;343:819–831.
68. Romero P, Obradovic Z, Kissinger CR, Villafranca JE, Dunker AK. Identifying disordered regions in proteins from amino acid sequences. Proceedings of the IEEE International Conference on Neural Networks. 1997. p 90–95.
69. Li X, Romero P, Rani M. Dunker AK, Obradovic AZ. Predicting protein disorder for N-, C- and internal regions. Genome Informatics 1999;10:30–40.
70. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. Protein disorder prediction: implications for structural proteomics. Structure (Camb) 2003;11:1453–1459.
71. Liu J, Rost B. NORSp: predictions of long regions without regular secondary structure. Nucleic Acids Res 2003;31:3833–3835.
72. Zeev-Ben-Mordehai T, Rydberg EH, Solomon A, Toker L, Auld VJ, Silman I, Botti S, Sussman JL. The intracellular domain of the *Drosophila* cholinesterase-like neural adhesion protein, gliotactin, is natively unfolded. Proteins 2003;53:758–767.
73. Callebaut I, Labesse G, Durand P, Poupon A, Canard L, Chomilier J, Henrissat B, Mornon JP. Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. Cell Mol Life Sci 1997;53:621–645.
74. Coeytaux K, Poupon A. Prediction of unfolded segments in a protein sequence based on amino acid composition. Bioinformatics 2005;21:1891–1900.
75. Kingston RL, Hamel DJ, Gay LS, Dahlquist FW, Matthews BW. Structural basis for the attachment of a paramyxoviral polymerase to its template. Proc Natl Acad Sci USA 2004;101:8301–8306.