Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Honing the *in silico* toolkit for detecting protein disorder

Robert M. Esnouf,[a]* R. Hamer,[a] J. L. Sussman,[b,c] I. Silman,[b,d] D. Trudgian,[e] Z.-R. Yang[e] and Jaime Prilusky[b,f]*

[a]Division of Structural Biology, University of Oxford, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, England, [b]The Israel Structural Proteomics Center, Weizmann Institute of Science, Rehovot 76100, Israel, [c]Department of Structural Biology, Weizmann Institute of Science, Rehovot 76100, Israel, [d]Department of Neurobiology, Weizmann Institute of Science, Rehovot 76100, Israel, [e]Department of Computer Science, University of Exeter, Exeter EX4 4QF, England, and [f]Department of Biological Services, Weizmann Institute of Science, Rehovot 76100, Israel

Correspondence e-mail: robert@strubi.ox.ac.uk, jaime.prilusky@weizmann.ac.il

Received 9 February 2006
Accepted 21 August 2006

Not all proteins form well defined three-dimensional structures in their native states. Some amino-acid sequences appear to strongly favour the disordered state, whereas some can apparently transition between disordered and ordered states under the influence of changes in the biological environment, thereby playing an important role in processes such as signalling. Although important biologically, for the structural biologist disordered regions of proteins can be disastrous even preventing successful structure determination. The accurate prediction of disorder is therefore important, not least for directing the design of expression constructs so as to maximize the chances of successful structure determination. Such design criteria have become integral to the construct-design strategies of laboratories within the Structural Proteomics In Europe (SPINE) consortium. This paper assesses the current state of the art in disorder prediction in terms of prediction reliability and considers how best to use these methods to guide construct design. Finally, it presents a brief discussion as to how methods of prediction might be improved in the future.

1. Introduction

The three-dimensional structure of a protein is primarily determined by its amino-acid sequence. However, the local and global environment in which the protein finds itself can also exert a large effect and protein-folding machinery may be required in order to drive the protein towards its target conformation. Under extreme conditions most proteins lose any specific three-dimensional structure and it has also become apparent that even under physiological conditions many protein sequences are partially or even totally disordered (Dunker *et al.*, 2000). Furthermore, the transition between ordered and disordered states can sometimes be effected by changing the protein's environment, even under physiological conditions. Computational studies of whole genomes have suggested varying numbers for the percentage of protein sequences that contain significant regions of disorder: one study predicted that 52–67% of eukaryotic proteins contain disordered regions longer than 40 amino acids (Vucetic *et al.*, 2003), while another predicted that 33% of eukaryotic proteins contain regions of disorder of longer than 30 residues (Ward *et al.*, 2004). Strikingly, the percentage of disordered regions in prokaryotic proteins is predicted to be substantially lower (Vucetic *et al.*, 2003). This may reflect the existence of entire families of eukaryotic proteins that are absent in prokaryotes, *e.g.* synaptic proteins in the nervous system.

When a protein is heated with concentrated urea it usually undergoes an order-to-disorder transition; when a protein is

Table 1
Selected disorder-prediction programs and their URLs.

Method	URL	Reference
DisEMBL (465)	http://dis.embl.de/	Linding, Jensen *et al.* (2003)
DisEMBL (coils)	http://dis.embl.de/	Linding, Jensen *et al.* (2003)
DisEMBL (hot)	http://dis.embl.de/	Linding, Jensen *et al.* (2003)
DISOPRED2	http://bioinf.cs.ucl.ac.uk/disopred/disopred.html	Ward *et al.* (2004)
FoldIndex	http://bip.weizmann.ac.il/fldbin/findex	Prilusky *et al.* (2005)
GlobPlot	http://globplot.embl.de/	Linding, Russell *et al.* (2003)
IST-ZORAN/*VSL*-1	http://www.ist.temple.edu/disprot/predictorVSL1.php	Obradovic *et al.* (2005)
IUPRED	http://iupred.enzim.hu/	Dosztanyi *et al.* (2005*a*,*b*)
PONDR	http://www.pondr.com/	Romero *et al.* (1997, 2001), Li *et al.* (1999)
PreLink	http://genomics.eu.org/spip/PreLink	Coeytaux & Poupon (2005)
RONN	http://www.strubi.ox.ac.uk/RONN	Yang *et al.* (2005)

successfully refolded it undergoes a disorder-to-order transition. Such transitions may also occur *in situ*, either resulting from a change in the environment or in response to an interaction with a specific binding partner. This latter case appears to be particularly significant biologically since it allows for extremely specific but reversible binding between protein partners (Oldfield, Cheng, Cortese, Romero *et al.*, 2005; Tompa, 2005; Wright & Dyson, 1999). Examples of such interactions include enzyme–substrate, receptor–ligand, protein–protein, protein–RNA and protein–DNA interactions. Amongst other uses, such controlled transitions appear to play a crucial role in cell signalling pathways.

It was hypothesized that if amino-acid sequence primarily determines the structure of a protein, then it also primarily determines which regions are unstructured. In support of this, early studies revealed that disordered regions often contain significant stretches of low-complexity sequence and that certain amino acids (charged, polar and flexible ones) are significantly more likely to be found in disordered regions (Garner *et al.*, 1998). For example, Glu, Asp and Lys are charged and Ser enhances solubility and provides flexibility, while low-complexity Pro-rich and/or Gly-rich sequences rarely form stable structures. Conversely, aromatic amino acids (Trp, Tyr and Phe) are primarily associated with ordered regions (Kissinger *et al.*, 1995) since they have a strong interaction capability which helps to develop structure (Burley & Petsko, 1985). The aliphatic amino acids (Leu, Ile and Val) are also similarly associated with ordered regions.

The importance of disorder in the study of molecular recognition has already been described. However, there is also a more practical and urgent need for accurate detection of disordered regions as a general tool in structural biology. Both X-ray crystallography and NMR spectroscopy rely on ensembles of almost identical structures to amplify the experimental signal. At best, spectroscopic methods can supply some information on the conformations of disordered sequences (Bernado *et al.*, 2005), while X-ray crystallography is inapplicable for such sequences (Oldfield, Chen, Cortese, Brown *et al.*, 2005). Furthermore, disordered regions can prevent structure determination entirely by affecting solubility and/or crystallizability.

The proliferation of disorder-prediction algorithms in recent years (for examples, see Table 1) is reflected in the inclusion of disorder prediction in both the CASP5 and CASP6 (http://predictioncenter.org/casp6/Casp6.html) trials. It may be noted that in CASP5, one of the submitted fully disordered proteins, Target 145 (Melamud & Moult, 2003), was in fact the cytoplasmic domain of the *Drosophila* adhesion protein gliotactin (Zeev-Ben-Mordehai *et al.*, 2003), one of the targets included in SPINE workpackage 10 (Human Proteins of Biomedical Relevance; Banci *et al.*, 2006). In CASP5, six methods were tested, whereas in 2004, for the CASP6 trial, 20 methods were evaluated. While *PONDR* (Li *et al.*, 1999; Romero *et al.*, 1997, 2001) remains perhaps the best known of these methods, we estimate that upwards of 40 algorithms have been developed. Two of the newer methods, developed since the CASP6 trial and very different in philosophy and applicability, have been developed by SPINE partners. *FoldIndex* (Prilusky *et al.*, 2005), developed at the Weizmann Institute, implements the algorithm described by Uversky *et al.* (2000) to make a calculation based on average net charge and average hydrophibicity of the sequence, thereby giving a single prediction of whether that sequence (or subsequence) is ordered or disordered. In contrast, *RONN* (Yang *et al.*, 2005), developed jointly by the Universities of Oxford and Exeter, uses a neural network technique to predict whether any given residue is likely to be ordered or disordered in the context of the surrounding amino-acid sequence. Both methods are freely accessible *via* their respective URLs.

This paper reviews the approaches to the disorder-prediction problem encapsulated in *FoldIndex* and *RONN* and considers in general terms the difficulties in objectively assessing the performance of such algorithms. Ways in which disorder can be combined with other bioinformatics analyses to guide the design of expression constructs are then considered. The discussion finishes by considering possibilities for improving the algorithms by recognizing differences between types of protein disorder.

2. Methods

2.1. Collecting data for disordered sequences

Central to any disorder-prediction approach is the collation of a database of sequences which are known to be either ordered or disordered. Ordered sequences can be easily

Table 2
Disorder information extracted from the PDB.

The table summarizes the parts of sequences of proteins which are assumed to be present in the crystallized/analysed entity but for which no structure was built. The analysis for 18 October 2005 was limited to structures determined by X-ray crystallography.

	PDB, 29 April 2004	PDB, 18 October 2005
Entries in PDB	25931	34347
Entries containing only short disordered regions	5754 (5–20 residues)	6105 (5–18 residues)
No. of disordered regions in these entries after filtering†	1925	2866
Entries containing at least one long disordered region	1573 (>20 residues)	1841 (>18 residues)
No. of long disordered regions in these entries after filtering†	530	687
No. of long ordered regions in these entries after filtering†	891	1358

† Filtering removed entries for heteromultimeric complexes and highly redundant sequences. Redundancy was addressed using *CD-HIT* (Li *et al.*, 2001, 2002) to remove sequences which were more than 70% identical to other sequences in the set.

extracted from the crystallographic and NMR structures found in the Protein Data Bank (PDB; Sussman *et al.*, 1998; Berman *et al.*, 2000). Proteins for which the structure has not yet been determined may also be known to be largely ordered from use of such techniques as heteronuclear single-quantum correlation (HSQC) spectroscopy, although it is not possible to know which residues are ordered and which are disordered. In contrast, it is difficult to gather data for disordered sequences. To confirm the presence of disorder it is necessary to have a soluble protein and in many cases this is not possible for reasons compounded by the presence of the disordered regions themselves. It is impossible to determine the structure of a completely disordered protein by crystallographic methods, so spectroscopic methods, including NMR, must be used for such proteins (Wright & Dyson, 1999). For partially ordered proteins it may be possible to determine a partial structure and thereby to infer that the missing regions are disordered, *e.g.* in the case of the human prion protein (Zahn *et al.*, 2000). Since these partial structures are usually deposited, the PDB is also the largest repository of disordered sequences associated with partially ordered structures. Uversky *et al.* (2000) tabulated fully disordered sequences characterized by spectroscopic methods, while Yang *et al.* (2005) trawled the 29 April 2004 release of the PDB, looking for disordered sequences using the Macromolecular Structure Database (MSD; Boutselakis *et al.*, 2003) at the European Bioinformatics Institute. Although only a small fraction of known proteins have had their structures determined and this method has some bias since only partly structured proteins can be included, this represents by far the largest resource of disordered sequences. For this paper, the analysis has been repeated and updated to reflect the state of the PDB as of 18 October 2005 (Table 2; detailed data are available from the *RONN* website, http://www.strubi.ox.ac.uk/RONN). The data are expected to contain some errors, *e.g.* crystallographic structure determinations where terminal regions of the protein may be incorrectly classified as disordered. Such regions may be missing (i) because they are genuinely disordered, (ii) because they are attached to the rest of the molecule by a flexible linker and thus have no fixed orientation with respect to the rest of the molecule or (iii) because the region may be absent from the crystallized entity owing to unexpected proteolysis. Another resource of experimentally measured disorder sequences was created so as to assess the reliability of *FoldIndex* (Prilusky *et al.*, 2005) and can be accessed at http://www.weizmann.ac.il/sb/faculty_pages/Sussman/papers/suppl/Prilusky_2005.

2.2. Different approaches to the prediction problem

It is well established that the amino-acid composition of disordered sequences is different from that of ordered sequences (Garner *et al.*, 1998) owing to the different physical properties of amino-acid side chains. Thus, most disorder-prediction methods focus on the properties of the individual amino acids, either based on experimentally measured parameters or calculated based on a statistical analysis of sequences known to be ordered or disordered (Wright & Dyson, 1999; Dyson & Wright, 2004) such as the Database of Protein Disorder (http://www.disprot.org/). Many different methods of parameterization have been tried and embodied in rule-based neural network methods such as *PONDR*, *GlobPlot*, *DisEMBL* and *DISOPRED*2 (see Table 1). However, this approach is most simply encapsulated in *FoldIndex* (Prilusky *et al.*, 2005), which codes the rules derived by Uversky *et al.* (2000). The program was originally designed to give a single overall prediction of 'ordered' or 'disordered' for any given sequence, but has since been adapted to give per-residue disorder propensity since this is useful for construct design. Although the physical properties of amino acids are clearly fundamental in the determination of disorder, the neural network used in *RONN* (Yang *et al.*, 2005) deliberately avoids explicit parameterization of amino acids in this way. Instead it uses non-gapped sequence alignment to measure 'distances' between windows of sequence for the unknown protein and windowed sequences for proteins of known folding state derived from the PDB analysis. Thus, in one sense, *FoldIndex* and *RONN* represent two extreme ways of approaching the disorder-prediction problem and, while both methods have their strengths and weaknesses, they perform well in comparison to other disorder-prediction methods. Fig. 1 compares both methods with the other most widely used disorder predictors using the data from the 29 April 2004 release of the PDB. Unsurprisingly, given the discussion above, *FoldIndex* performs particularly well for fully ordered

or fully disordered sequences, while *RONN* is more successful in identifying partially disordered sequences.

2.3. Measuring accuracies of prediction

At first glance, it seems trivial to assess the value of an algorithm for predicting whether a given amino acid (or amino-acid sequence) is ordered or disordered. However, it is not possible to devise a universally useful measure since the usefulness of any algorithm depends critically on the significance and consequences of correct and incorrect predictions, which in turn depend on the application to which the algorithm is being put. Many measures of such binary classifiers have been defined in computer science and an accessible discussion relating to text categorization has been posted on the web (http://www.islanddata.com/downloads/irt_whitepaper_perfmeasure.pdf). A second problem is that descriptions of algorithms often quote cross-validation results rather than the outcomes of true blind tests, which give more realistic results. Finally, since test sets tend to contain relatively few disordered amino acids and in the 'real world' of experimental construct design they are even less frequent, good measures of algorithm performance must not be overly affected by these differences in relative class frequency. What the experimental scientist engaged in construct design really wants is a realistic idea of how much better an algorithm is than simply guessing. Measures such as the 'probability excess' (Yang *et al.*, 2005) and the related scoring function used in the CASP trials attempt to provide such a measure and suggest that the best algorithms available today are, very roughly, 50% better than guessing for blind tests on partially ordered structures. Fig. 1 presents probability excess assessment of the common methods including *RONN* and *FoldIndex*. However, if the question is simply 'does a given sequence form an ordered structure or not?' then current methods, especially global predictors such as *FoldIndex*, can provide answers with a much higher degree of certainty.

One final complication with measuring algorithm performance is that disordered sequences themselves do not constitute a homogeneous class: some are of low complexity, for example containing runs of 20 or more consecutive glutamate residues, some are long sequences that appear to be reasonably complex and some are short regions in otherwise ordered sequences which clearly have a strong preference for disorder. These observations are considered in more detail below. Different algorithms are better tuned to different sorts of disorder, so that relative performance of algorithms varies depending upon the context of disorder prediction. For example, using disorder prediction to detect linker regions between domains is essentially a requirement for detecting short(ish) regions of disorder and methods which employ long prediction windows, such as the current version of *RONN*, may thus exhibit reduced sensitivity.

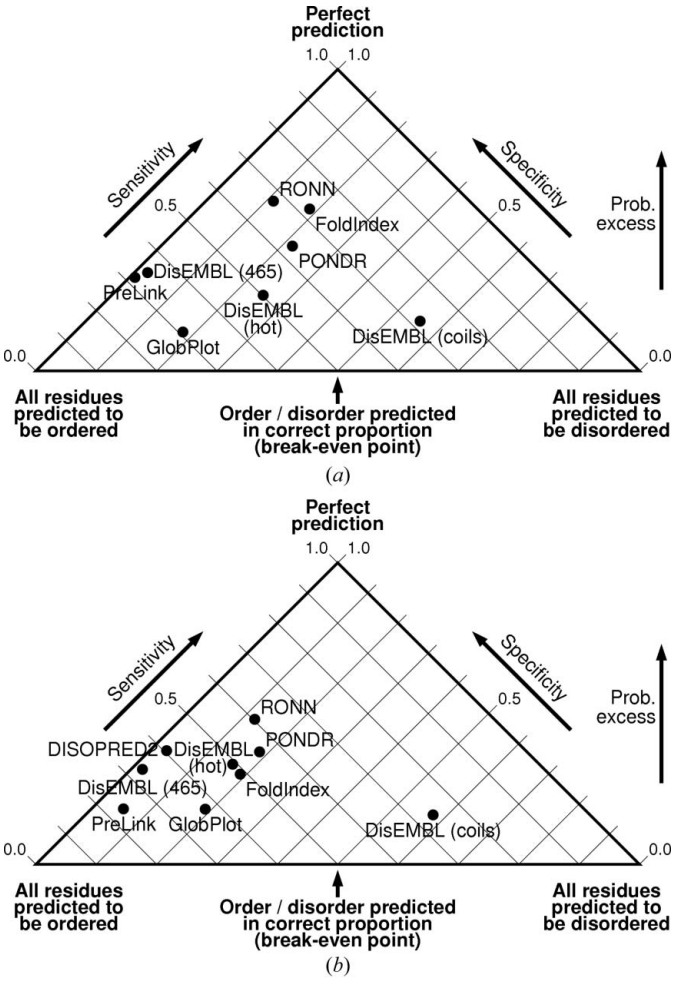

Figure 1
Assessment of prediction performance for nine different disorder-prediction methods. (*a*) The results of blind tests for a balanced mixture of fully ordered and fully disordered sequences. (*b*) The results of blind tests for partially ordered sequences extracted from the PDB. See §2.3 for a discussion of terms. Reproduced from Yang *et al.* (2005) by permission of Oxford University Press.

3. Results and discussion

3.1. Analysis of SPINE structures

To assess the robustness of prediction methods, an initial set of SPINE structures deposited with the PDB was used as a realistic (but not strictly 'blind') test set. At the time of analysis, 139 deposited SPINE structures had been fully annotated by the MSD, of which 15 corresponded to protein–protein complexes, which were excluded from further analysis; four others were excluded for other reasons. For the remaining 120 structures, 47 contained no disorder and 73 were partially disordered, giving a total of 25 230 ordered residues and 1476 disordered residues. Consistent with 'real-world' usage, the data set was not filtered in any way to remove similar sequences nor sequences that might have been used in training sets. Furthermore, this set contained disproportionately more structures from eukaryotic sources (45% of structures) than the PDB as a whole (Table 3). Thus, this data set would be expected to be difficult for prediction algorithms, especially *FoldIndex*, which was not designed for nor trained against partially ordered structures. The prediction results for *RONN*

Table 3
Disorder predictions for deposited SPINE structures.

Proteins are partitioned according to the domain of life to which they belong. The probability excess represents the relative improvement compared with simple guessing.

		Probability excess (%)	
	Structures	*RONN*	*FoldIndex*
Prokaryotic	59	58.4	28.5
Eukaryotic	54	49.9	14.2
Virus	7	7.4	−2.9
All	120	48.4	17.0

(Table 3) show that the performance is in line with previously published results for blind trials (Yang *et al.*, 2005), even though the proportion of disordered residues in this trial is substantially lower (5.5% compared with 10.9%). However, the performance is noticeably better with prokaryotic proteins, suggesting that training is biased toward these proteins and that there exists a significant difference between the proteins from the prokaryotic and eukaryotic domains of life. Somewhat surprisingly, the eukaryotic set contains proportionately less disorder than the prokaryotic set, although genomic studies using disorder-prediction tools suggest that disorder is more prevalent in eukaryotes. This apparent anomaly may reflect the difficulty in working with eukaryotic proteins: either target selection is more conservative or only very well ordered eukaryotic proteins lead to successful structure determinations. In sharp contrast, for the small set of viral proteins the predictions of both *RONN* and *FoldIndex* are virtually useless. While the scale of this trial makes it difficult to attach much significance to this observation, it is tempting to suggest that the determinants of disorder in viral proteins, particularly structural proteins, may be somewhat different from those of other proteins. As anticipated above, *FoldIndex* finds all the predictions more difficult (Table 3), although predictions for prokaryotic proteins are still useful.

3.2. Use of disorder prediction in construct design

Within the SPINE remit of developing methods for structural proteomics, the single most important use of disorder prediction is as an aid to construct design. Many biomedically important proteins contain regions of disorder and these regions are often implicated in difficulties with structure determination, *e.g.* because they can encourage aggregation and reduce solubility or inhibit growth of crystals. However, since many structures in the PDB are partly disordered, the PDB is the largest source of disorder data and the basis for most analyses. In many cases where disorder is suspected, the best strategy for crystal production is to work with multiple constructs in parallel (see, for example, Banci *et al.*, 2006), and disorder prediction has become an essential tool in the construct design process.

Disorder prediction is just one of an array of tools that can inform construct design. Others include comparison with domain definitions based either on sequence or structure, detection of signal peptides and nuclear localization sequences, detection of hydrophobic and low complexity regions, alignments with structures already deposited in the PDB and any known functional data or mutational data relating to the protein of interest. A significant goal of bioinformatics is to present all this information in a simple, preferably visual, way that can be conveniently accessed and assessed by the researcher (see Albeck *et al.*, 2006).

The output of most disorder predictors is a graph of per-residue disorder probabilities, with 50% probability being taken as the decision threshold. However, this is somewhat unrealistic and mainly driven by convenience since, for example, whether a residue is ordered or not also depends on the total sequence length, whether the complete structural domain is in the expression construct and, of course, the eventual protein environment. Nevertheless, these graphs are useful and our experience, primarily based on *RONN* and *FoldIndex*, suggests that more careful analysis of the output is justified. Firstly, regions predicted to have very low disorder probability tend to correspond to hydrophobic regions and transmembrane sequences and these regions can make constructs as difficult to work with as disordered regions. Secondly, although a single prediction threshold of 50% is used, it may be better to define the precise residue of transition as that which is half way between the plateaux of probabilities for ordered residues on one side and disordered residues on the other. Thirdly, for methods that rely on prediction windows the user should be aware of the smoothing effects they introduce. Although a short disordered linker between domains may not be long enough to force the individual prediction for any residue above 50%, a lower peak may still be clearly visible in the plot.

One final consideration is that the goal of construct design is to be able to express part (or all) of the protein of interest in a stable well behaved soluble form. This can require considerable accuracy in deciding start and end points. Cutting more than one or two residues into an ordered domain risks disrupting the folding of the entire domain, whereas being too conservative and leaving a long disordered tail on the expression construct risks an adverse effect on solubility, homogeneity and crystallizability. Results for exhaustive expression screening of all possible N- and C-terminal truncations (Hart & Tarendeau, 2006) suggest that the acceptable window can be quite narrow, perhaps no more than five amino acids. While many disorder predictors can give results largely in agreement with observed disorder, reaching this level of accuracy demands further improvements in the algorithms.

3.3. Different sorts of disorder?

It is likely that proteins do not have disordered regions by accident. As the study of disorder has progressed it has become clear that not all disorder is equivalent and it seems likely that if there are different sorts of disorder then they must fulfill different purposes. Low-complexity sequences bearing many charged amino acids have long been recognized as disordered and some of them may play a role in anchoring

proteins to charged surfaces such as membranes. Other low-complexity sequences bearing proline and glycine repeats also exist which may act as spacers or 'stalks'. Some long regions of disorder have more complex sequences, similar to structured regions, and may even include 'structure-favouring' amino acids such as Trp, Tyr and Phe. One possibility is that these regions may transition to an ordered state when binding to cognate partners, and the 'structure-favouring' amino acids may then play a key role in complex formation. Short regions of disorder are frequently found in proteins that are largely ordered; here one possibility is that they somehow reflect the evolutionary history of the parent gene and another is that they add a point where variety is desirable on a protein surface. The distribution of amino acids in these flexible loops is quite characteristic, with Gly predominating. Finally, one of the most common motifs among disordered regions is entirely man-made: purification His tags are the most easily detectable of all groups of disordered sequences.

The natural question that arises from the discussion above is whether these different types of disordered sequences cluster together in any way that might be useful for disorder-prediction algorithms. This in turn depends on whether different types of disorder can be classified from sequence alone. Three approaches suggest themselves: (i) do disordered regions divide into subgroups which can be characterized separately, (ii) do disordered regions of different lengths, especially very short regions, have characteristic sequences and (iii) is the transition between order and disorder characteristic? Based on an initial analysis of the disorder data extracted from the PDB (results presented in Table 2) the answer to the first two questions is not clear-cut. Except for obviously disordered sequences, producing clusters of sequences that divide ordered and disordered sequences requires very fine-grained clustering into many small clusters. This is in line with the Oxford group's experience with *RONN* that increasing the number of prototype sequences improved performance markedly. Detecting characteristic sequences for order-to-disorder and disorder-to-order transitions was, however, immediately more successful (Fig. 2). Although attempted before on a small scale (Radivojac *et al.*, 2003), this large-scale analysis showed that not only were certain amino acids more likely to be found on either side of the transition, but also that these frequencies are markedly higher than those observed for the middle of long regions of order or disorder. Furthermore, the distributions of amino acids for order-to-disorder transitions virtually mirror those of disorder-to-order transitions, suggesting that these effects are bidirectional in the primary sequence. We are attempting to exploit such findings in new disorder-prediction algorithms.

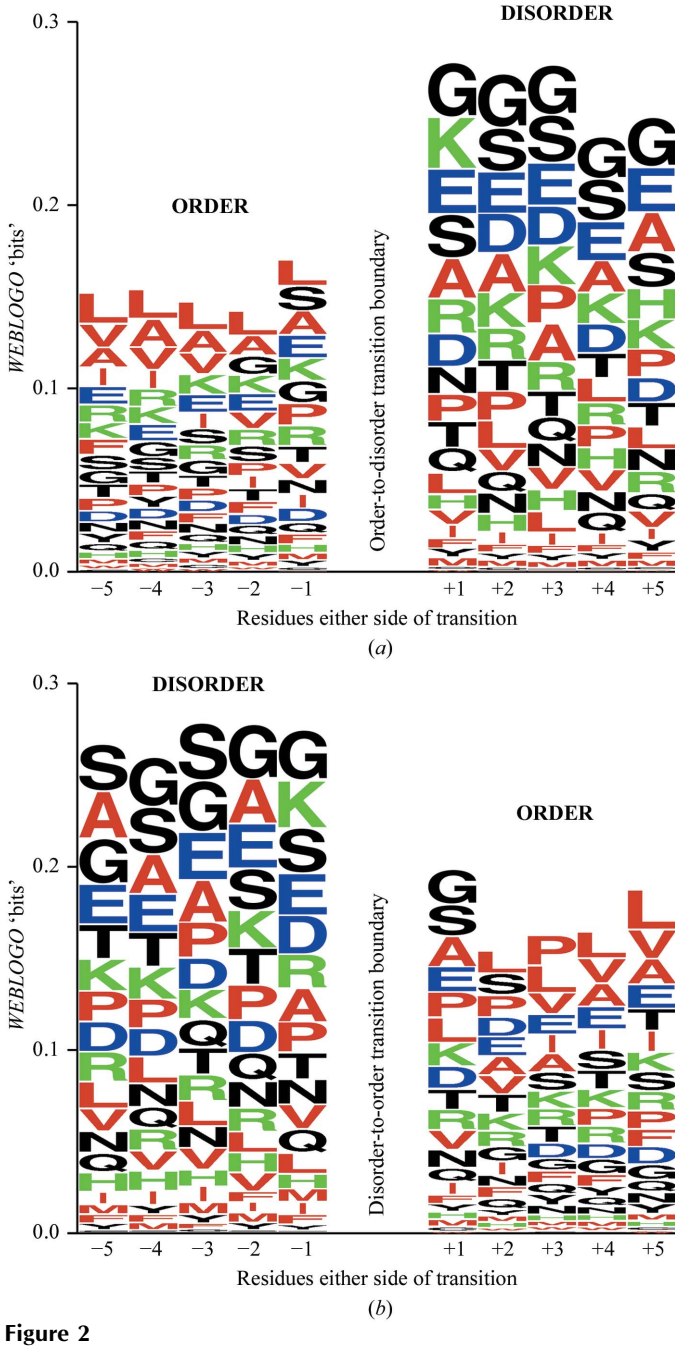

Figure 2
Amino-acid frequencies either side of order-state transition boundaries based on the analysis of the PDB (Table 2). Frequencies based (*a*) on 1709 order-to-disorder transitions and (*b*) on 1918 disorder-to-order transitions. A window of five amino acids on either side of the transition boundary was considered and amino-acid frequencies were analyzed with *WEBLOGO* (Crooks *et al.*, 2004). The height of each stack of single-letter amino-acid codes is proportional to the information content of (*i.e.* significance of the amino acid at) each residue position, while the height of each individual amino-acid code within a stack is proportional to its relative frequency at the position (see Crooks *et al.*, 2004 for more details). The colouring reflects the classes of amino acids (*i.e.* charged, aromatic *etc.*). The order-state transition boundary is indicated by the central gap in each figure.

3.4. Conclusions

For structural proteomics (and therefore for SPINE) the single most important application of disorder prediction is construct design for proteins of biomedical importance. This is an extremely demanding application in that many of these proteins are partially ordered (the most difficult disorder to predict) and that the boundaries between order and disorder regions need to be defined very accurately in order to be useful for defining stable soluble crystallizable constructs.

Many simple-to-use methods for the prediction of protein disorder are now available on the internet. Whilst different methods have different strengths and therefore different programs may be better suited to particular applications, the best tools appear to be about 50% better than guesswork. *FoldIndex* calculates the physical properties of all amino acids in a sequence to determine an overall probability of order or disorder. Many methods parameterize the physical properties of amino acids to train learning algorithms such as neural networks, while *RONN* uses a neural network in conjunction with non-gapped sequence alignment against prototype sequences to avoid having to define relevant physical properties explicitly.

As these methods have developed and more disordered sequences have been analyzed, it has become apparent that disorder displays many subtleties. The distribution of amino acids throughout disordered regions depends both on the length of the disordered region and on the proximity to an ordered region. This holds promise for a new generation of disorder predictors which will be more reliable and, in particular, better at defining the precise ends of disordered regions. Work on such improved algorithms is now under way.

The research described was supported by the European Commission as part of SPINE (Structural Proteomics In Europe) contract No. QLG2-CT-2002-00988 under the Integrated Programme 'Quality of Life and Management of Living Resources', the Israel Ministry of Science and Technology Grant for the ISPC, the Divadol Foundation, the MINERVA Foundation, the Bruce Rosen Foundation and by the Kimmelman Center. JLS is the Morton and Gladys Pickman Professor of Structural Biology. RME is supported by the UK Medical Research Council. RH and DT are supported by a UK MRC studentship and a UK EPSRC Doctoral Training Grant, respectively.

References

Albeck, S. *et al.* (2006). *Acta Cryst.* D**62**, 1184–1195.
Banci, L. *et al.* (2006). *Acta Cryst.* D**62**, 1208–1217.
Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.
Bernado, P., Blanchard, L., Timmins, P., Marion, D., Ruigrok, R. W. & Blackledge, M. (2005). *Proc. Natl Acad. Sci. USA*, **102**, 17002–17007.
Boutselakis, H. *et al.* (2003). *Nucleic Acids Res.* **31**, 458–462.
Burley, S. K. & Petsko, G. A. (1985). *Science*, **229**, 23–28.
Coeytaux, K. & Poupon, A. (2005). *Bioinformatics*, **21**, 1891–1900.
Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. (2004). *Genome. Res.* **14**, 1188–1190.
Dosztanyi, Z., Csizmok, V., Tompa, P. & Simon, I. (2005*a*). *Bioinformatics*, **21**, 3433–3434.
Dosztanyi, Z., Csizmok, V., Tompa, P. & Simon, I. (2005*b*). *J. Mol. Biol.* **347**, 827–839.
Dunker, A. K., Obradovic, Z., Romero, P., Garner, E. C. & Brown, C. J. (2000). *Genome Inform.* **11**, 161–171.
Dyson, H. J. & Wright, P. E. (2004). *Chem. Rev.* **104**, 3607–3622.
Garner, E., Cannon, P., Romero, P., Obradovic, Z. & Dunker, A. K. (1998). *Genome Inform.* **9**, 201–213.
Hart, D. J. & Tarendeau, F. (2006). *Acta Cryst.* D**62**, 19–26.
Kissinger, C. R., Parge, H. E., Knighton, D. R., Lewis, C. T., Pelletier, L. A., Tempczyk, A., Kalish, V. J., Tucker, K. D., Showalter, R. E., Moomaw, E. W., Gastinel, L. N., Habuka, N., Chen, X., Maldonado, F., Barker, J. E., Bacquet, R. & Villafranca, J. E. (1995). *Nature (London)*, **378**, 641–644.
Li, W., Jaroszewski, L. & Godzik, A. (2001). *Bioinformatics*, **17**, 282–283.
Li, W., Jaroszewski, L. & Godzik, A. (2002). *Bioinformatics*, **18**, 77–82.
Li, X., Romero, P., Rani, M., Dunker, A. K. & Obradovic, Z. (1999). *Genome Inform. Ser. Workshop Genome Inform.* **10**, 30–40.
Linding, R., Jensen, L. J., Diella, F., Bork, P., Gibson, T. J. & Russell, R. B. (2003). *Structure*, **11**, 1453–1459.
Linding, R., Russell, R. B., Neduva, V. & Gibson, T. J. (2003). *Nucleic Acids Res.* **31**, 3701–3708.
Melamud, E. & Moult, J. (2003). *Proteins*, **53**, *Suppl.* 6, 561–565.
Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P. & Dunker, A. K. (2005). *Proteins*, **61**, *Suppl.* 7, 176–182.
Oldfield, C. J., Cheng, Y., Cortese, M. S., Brown, C. J., Uversky, V. N. & Dunker, A. K. (2005). *Biochemistry*, **44**, 1989–2000.
Oldfield, C. J., Cheng, Y., Cortese, M. S., Romero, P., Uversky, V. N. & Dunker, A. K. (2005). *Biochemistry*, **44**, 12454–12470.
Prilusky, J., Felder, C. E., Zeev-Ben-Mordehai, T., Rydberg, E., Man, O., Beckmann, J. S., Silman, I. & Sussman, J. L. (2005). *Bioinformatics*, **21**, 3435–3438.
Radivojac, P., Obradovic, Z., Brown, C. J. & Dunker, A. K. (2003). *Pac. Symp. Biocomput.*, pp. 216–227.
Romero, P., Obradovic, Z. & Dunker, K. (1997). *Genome Inform. Ser. Workshop Genome Inform.* **8**, 110–124.
Romero, P., Obradovic, Z., Li, X., Garner, E. C., Brown, C. J. & Dunker, A. K. (2001). *Proteins*, **42**, 38–48.
Sussman, J. L., Lin, D., Jiang, J., Manning, N. O., Prilusky, J., Ritter, O. & Abola, E. E. (1998). *Acta Cryst.* D**54**, 1078–1084.
Tompa, P. (2005). *FEBS Lett.* **579**, 3346–3354.
Uversky, V. N., Gillespie, J. R. & Fink, A. L. (2000). *Proteins*, **41**, 415–427.
Vucetic, S., Brown, C. J., Dunker, A. K. & Obradovic, Z. (2003). *Proteins*, **52**, 573–584.
Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F. & Jones, D. T. (2004). *J. Mol. Biol.* **337**, 635–645.
Wright, P. E. & Dyson, H. J. (1999). *J. Mol. Biol.* **293**, 321–331.
Yang, Z. R., Thomson, R., McNeil, P. & Esnouf, R. M. (2005). *Bioinformatics*, **21**, 3369–3376.
Zahn, R., Liu, A., Luhrs, T., Riek, R., von Schroetter, C., Lopez Garcia, F., Billeter, M., Calzolai, L., Wider, G. & Wuthrich, K. (2000). *Proc. Natl Acad. Sci. USA*, **97**, 145–150.
Zeev-Ben-Mordehai, T., Rydberg, E. H., Solomon, A., Toker, L., Botti, S., Auld, V. J., Silman, I. & Sussman, J. L. (2003). *Proteins*, **53**, 758–767.