doi:10.1016/j.jmb.2004.02.002 J. Mol. Biol. (2004) 337, 635–645

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life

J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton and D. T. Jones*

Bioinformatics Unit Department of Computer Science, University College London, Gower Street, London WC1E 6BT, UK

*Corresponding author

An automatic method for recognizing natively disordered regions from amino acid sequence is described and benchmarked against predictors that were assessed at the latest critical assessment of techniques for protein structure prediction (CASP) experiment. The method attains a Wilcoxon score of 90.0, which represents a statistically significant improvement on the methods evaluated on the same targets at CASP. The classifier, DISOPRED2, was used to estimate the frequency of native disorder in several representative genomes from the three kingdoms of life. Putative, long (>30 residue) disordered segments are found to occur in 2.0% of archaean, 4.2% of eubacterial and 33.0% of eukaryotic proteins. The function of proteins with long predicted regions of disorder was investigated using the gene ontology annotations supplied with the *Saccharomyces* genome database. The analysis of the yeast proteome suggests that proteins containing disorder are often located in the cell nucleus and are involved in the regulation of transcription and cell signalling. The results also indicate that native disorder is associated with the molecular functions of kinase activity and nucleic acid binding.

Keywords: protein structure; native disorder; molecular recognition; functional genomics; *Saccharomyces*

Introduction

One of the central tenets of structural biology is that the function of a protein is determined by its three-dimensional structure. As a result, predicting protein structure has often been at the forefront of efforts to infer function. However, it appears that a large proportion of protein sequences do not form complete globular structures. The natively disordered regions within these proteins may adopt an ensemble of structural states with transitions between the states leading to dynamic flexibility of the protein structure[1] or have non-globular structures that are extended in the solvent.[2]

It has been shown experimentally that disordered regions are involved in DNA-binding[3] and several other types of molecular recognition. One of the advantages of disordered binding sites is that their multiple metastable conformations allow them to recognize several targets with high specificity and low affinity.[4] Transitions between the native unfolded state and a globular structure, induced by phosphorylation or some other type of interaction, may also provide thermodynamic regulation of binding. The prediction of disordered regions would therefore provide a first step in methods for identifying functionally relevant disordered regions and the flexible segments that hinder successful crystallization of the protein.

It has been shown in a series of papers[5–8] that there are clear patterns that characterize disordered regions such as low sequence complexity, amino acid compositional bias (e.g. towards aromatic residues) and high flexibility, and that disorder can be predicted successfully from amino acid sequence. We describe here the development of a new method for predicting native disorder. The classifier, DISOPRED2, is benchmarked on targets from the last critical assessment of structure prediction (CASP) experiment, which included an evaluation of the latest disorder prediction methods.[9]

The new method is also used to investigate

Abbreviations used: CASP, critical assessment of structure prediction; GO, gene ontology; PDB, protein data bank; SGD, *Saccharomyces* genome database; SVM, support vector machine; ROC, receiver operating characteristic.

E-mail address of the corresponding author: dtj@cs.ucl.ac.uk

disorder in several archaea, eubacteria and eukaryote genomes. Previous genome-wide analyses of disordered regions have been based on classifiers with high false positive rates (16% for disordered segments longer than 40 residues).[10,11] Although the results presented here cannot be interpreted as a lower bound on the proportion of proteins that contain disorder, they are intended to be very conservative with false positive rates estimated to be lower than 0.5% on long disordered segments.

The functions of potentially disordered proteins are also investigated using the gene ontology (GO) annotations[12] for the budding yeast *Saccharomyces cerevisiae*. The aim of the analysis was to investigate which processes rely directly on dynamic flexibility of the protein structure. This was achieved by mapping each long disordered segment to the GO annotations attached to its parent protein. The frequency with which each GO term occurred was then compared to its frequency of occurrence in random simulations. The random model corresponds to a null hypothesis whereby each protein's probability of containing a long disordered segment is proportional to its length. The replicates were then used to provide confidence estimates, under the null model, for GO terms that were over- or under-represented in the set of disorder predictions.

Results

Estimating error rates

The false positive rate for DISOPRED2 was established by classifying a set of 7169 ordered proteins with less than 95% sequence similarity to each other. (All residues for the protein set have atomic co-ordinates recorded in the Protein Data Bank (PDB).[13]) This threshold allows inclusion of a large proportion of the PDB but removes multiple models of the same structure or very close homologues. Although DISOPRED2 was developed with the aim of optimizing per residue accuracy, it is important to distinguish between long contiguous regions of disorder and short disordered segments, which are less likely to be functionally relevant. The per residue false positive rate was found to be 3.2% on the set of proteins from the PDB with a large fraction arising from short predictions of disorder that typically occur at the C and N termini.

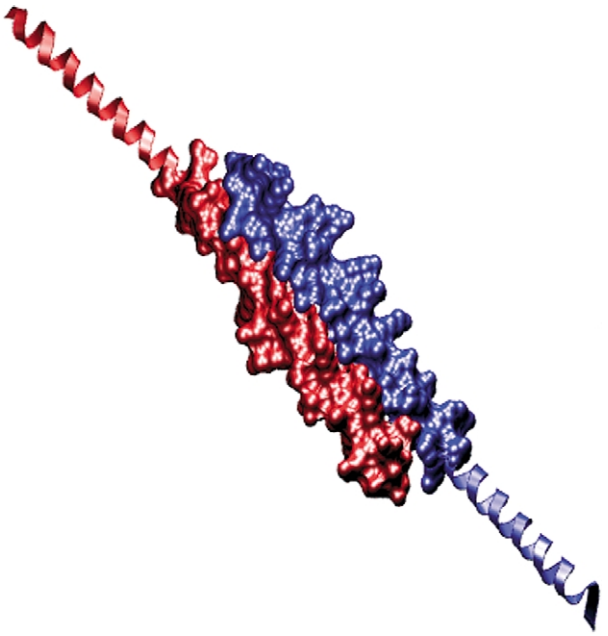

Figure 1. ATPase inhibitor (1gmj). The disordered predictions are highlighted by the surface map structure.

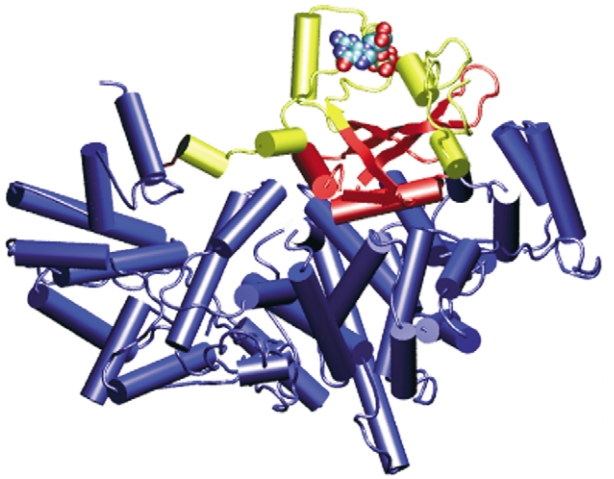

Figure 2. Nuclear cap-binding protein (1h2u). The region of the protein that is predicted to be disordered is colored yellow with the second chain colored in blue. The molecule in contact with the protein is the nucleotide GDP.

Only 37 (0.5%) of the ordered structures were predicted to contain long (>30 residue) regions of disorder. This value is likely to overestimate the false positive rate on this set, as many of the chains were crystallized as part of structural complexes and may be disordered prior to the formation of quaternary structure (see Figure 1). Some of the

Table 1. The number of ordered crystal structures with predicted regions of disorder longer than 30 residues

Bound to ligand	6
Bound to protein	14
Bound to DNA	5
No secondary structure	4
Ribosome	1
Domain linker	5
Surface of protein	5
Total	37

The results come from a set of 7169 proteins that are non-redundant at 95% sequence identity. In all cases where the protein is described as bound, the predicted disordered region is in contact with the ligand, DNA or other chain. The models with no secondary structure were determined using NMR and contained almost no visible helical or strand elements.

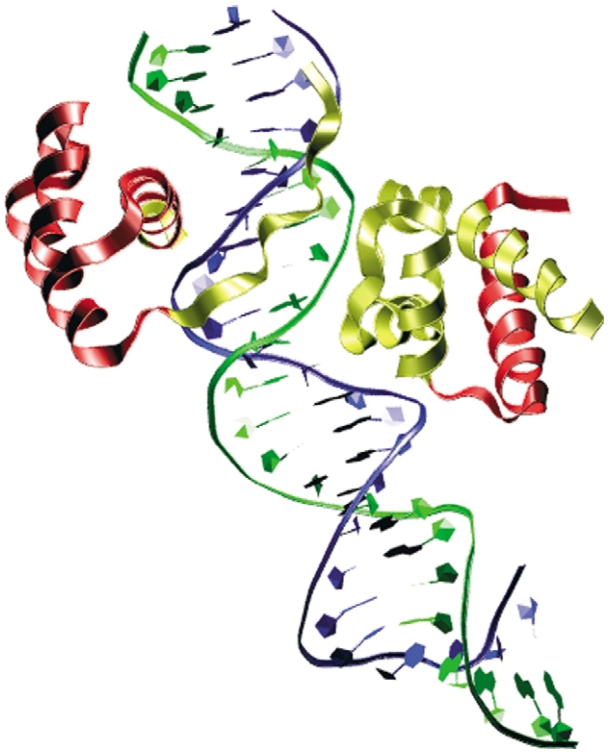

Figure 3. Transcription factor (1gt0) bound to DNA with predicted disordered regions colored in yellow. The protein structure is taken from a single chain with the apparent discontinuity caused by missing co-ordinates in the electron density map.

proteins, such as the nuclear cap-binding protein shown in Figure 2, undergo induced folding in the presence of a ligand[14] and others are stabilized by binding to DNA as shown in Figure 3. Other long predictions of disorder occur in domain linker regions, which may be unstructured in solution to allow structural uncoupling of two or more globular domains.[15] It therefore appears that only the five (0.07%) false positives that occur on the protein surface are certain, as shown in Table 1.

Disorder frequencies in complete genomes

Table 2 shows the estimated disorder frequencies for six archaean, 13 bacterial and five eukaryotic genomes, in addition to overall totals for each kingdom and predictions from a non-redundant set of resolved crystal structures in the PDB. An average of 2.0% of archaean, 4.2% of eubacteria and 33.0% of eukaryotic proteins are predicted to contain long regions of disorder.

Functional roles of disorder in *S. cerevisiae*

Table 2 shows a clear disparity between the predicted disorder rates in eukaryotic and prokaryotic genomes. The functional relevance of putative long regions of disorder was investigated in the budding yeast *S. cerevisiae* using the GO annotations[16] supplied with the *Saccharomyces* genome database (SGD).[17]

The proteins in the subset containing long

Table 2. Estimated disorder frequencies

Kingdom organism	Number of sequences	Disorder frequency	Length >30	Length >50
Archaea *Aeropyrum pernix*	1841	4.7	2.1	0.5
Archaea *Archaeoglobus fulgidis*	2409	2.8	0.9	0.2
Archaea *Halobacterium* sp.	2442	6.2	5.0	1.9
Archaea *Methanococcus jannaschi*	1784	2.8	1.0	0.3
Archaea *Pyrococcus abyssi*	1769	3.0	1.4	0.7
Archaea *Thermoplasma volcanium*	1497	3.2	1.0	0.3
Bacteria *Agrobacterium tumefaciens* C58	5288	6.4	5.7	2.0
Bacteria *Aquifex aeolicus* VF5	1557	3.3	1.9	0.4
Bacteria *Chlamydophila pneumoniae* AR39	1111	6.2	4.8	2.3
Bacteria *Chlorobium tepidum* TLS	2248	5.1	3.3	0.5
Bacteria *Escherichia coli* K12	4247	4.6	2.8	0.8
Bacteria *Haemophilus influenzae* Rd	1650	4.4	3.8	1.3
Bacteria *Mycobacterium tuberculosis* H37Rv	3890	9.1	7.0	3.3
Bacteria *Neisseria meningitides* MC58	2020	5.7	4.5	1.7
Bacteria *Salmonella typhi*	4714	4.9	2.7	0.9
Bacteria *Staphylococcus aureus*	2632	6.2	4.5	2.2
Bacteria *Synechocystis* species PCC 6803	3140	5.4	4.7	1.8
Bacteria *Thermotoga maritima*	1857	3.3	1.8	0.6
Bacteria *Treponema pallidum*	1035	6.1	6.4	2.6
Eukaryota *Arabidopsis thaliana*	21,482	16.8	33.8	19.0
Eukaryota *Caenorhabditis elegans*	20,506	15.9	27.5	15.6
Eukaryota *Drosophila melanogaster*	13,913	21.6	36.6	22.1
Eukaryota *Homo sapiens*	26,385	21.6	35.2	21.9
Eukaryota *S. cerevisiae*	6245	17.0	31.2	19.3
Archaea	11,742	3.8	2.0	0.7
Bacteria	35,389	5.7	4.2	1.6
Eukaryota	88,531	18.9	33.0	19.6
PDB (non-redundant at 95% sequence identity)	7169	3.2	0.5	0.1

The columns show the number of sequences, the percentage of residues predicted as being disordered and the percentage of chains with contiguous disordered segments of length greater than 30 and 50 residues, respectively.

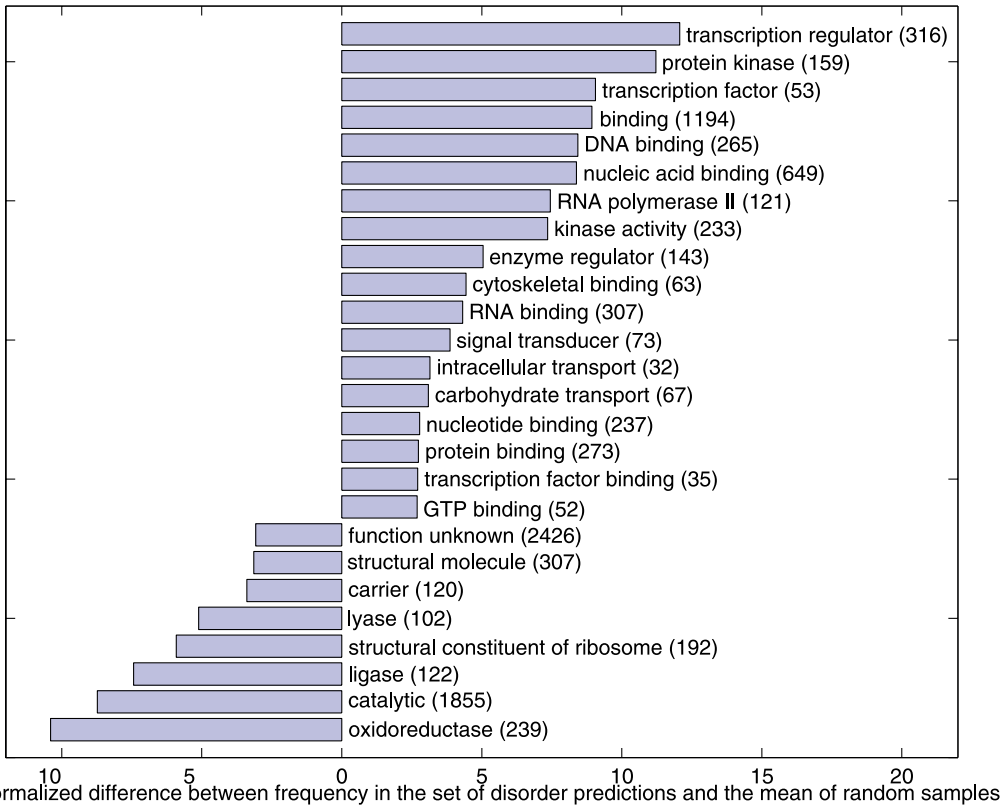

Figure 4. GO terms from the molecular function ontology that are significantly over- or under-represented in the set of proteins predicted to contain long regions of disorder. Each term is followed by the number of proteins in the yeast proteome that have been assigned this annotation. The terms are ordered by the normalized differences between the terms' frequency of occurrence in the random samples and the set of disordered predictions.

regions of predicted disorder were longer on average than the population (704.6 compared with 497.1 residues). This arises because large multi-domain proteins have more linker regions and a higher probability of incorporating a disordered domain if these are distributed uniformly across the proteome. The sampling method accounts for this by constructing a null model where the disordered segments with lengths greater than 30 residues are distributed randomly across the length of the proteome (see System and Methods for further details).

Selected GO terms that obtained a *p*-value lower than 0.2 and that describe more than 50 protein annotations† are listed in Figures 4–6. The Figures divide the results into the three separate ontologies representing molecular function, biological process and cellular component.

Discussion

The difficulty in investigating dynamically flexible polypeptide sequences is the main reason for the relative paucity of experimental data on native disorder compared with globular structures. This difficulty also extends to the identification of disordered regions for the purposes of pattern recognition. The definition of native disorder is also fairly heterogeneous as it applies to global structures such as collapsed molten globule proteins and extended random coil-like proteins, and to the localized disorder that can exist in flexible domain linkers and ligand binding sites.

In the training of DISOPRED2, residues with missing atomic co-ordinates are defined as disordered. Although this definition was also used in the CASP experiment, it is imperfect as missing residues can also arise as an artifact of the crystallization process such as rigid body wobble or crystal contacts. It is also possible that false prediction of order can be caused by the crystallized fragment being part of a structural complex *in vitro* or that tags added or regions removed from the sequence can alter the stability of the structure. However, this appears to be the most effective means of identifying disordered regions in the absence of further experimental characterization of the protein structure.

The development of DISOPRED2 has demonstrated that information from homologous sequences leads to a slight improvement in the prediction of native disorder. However, the

† A *p*-value of 0.2 corresponds to fewer than 100 out of the 10,000 resamplings receiving a more extreme Z-score. Full results can be found at http://bioinf.cs.ucl.ac.uk/disopred/suppInfo.html

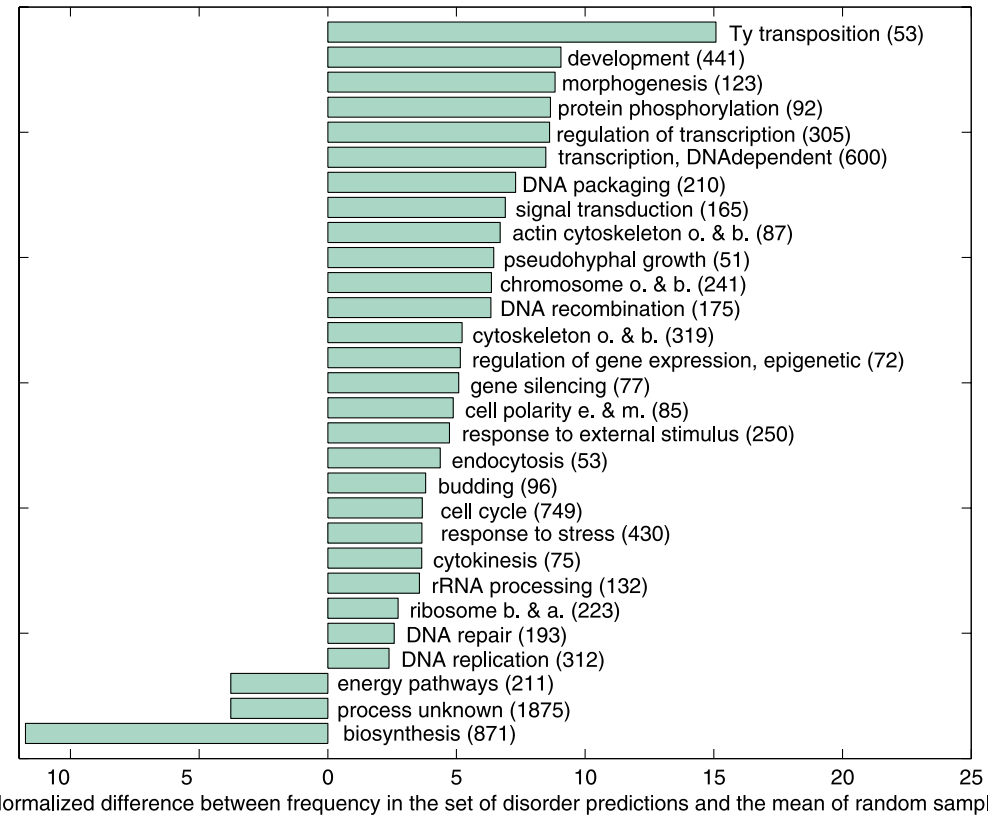

Figure 5. GO terms from the biological process ontology that are significantly over- or under- represented in the set of disordered predictions. The abbreviations used are: organization (o), biogenesis (b), establishment (e), maintenance (m) and assembly (a). Terms describing various types of metabolic and biosynthetic processes are omitted in the interests of space (native disorder is under-represented in these categories).

improvement is not as great as that observed in predicting the secondary structure of globular proteins.[18] It is believed that patterns of conservation improve the secondary structure prediction by implicitly encoding global constraints on the local structure.[19] This improvement is likely to be less effective in the prediction of natively disordered regions as, by definition, they are not constrained by the protein's tertiary structure.

There are likely to be several other reasons for the improved accuracy of DISOPRED2 compared with the other algorithms described in System and Methods. The main difference is that DISOPRED2 is trained directly on protein sequence rather than measures of amino acid composition, sequence complexity[7,10] or biophysical properties such as mean hydrophobicity.[20] This may allow the classifier to recognize sequence motifs that have been shown to be associated with disorder such as Pro-X-Pro-X-Pro or Lys-X-X-Lys-X-Lys. (S. Lise & D.T.J., unpublished results). The cascaded classifier also improves accuracy by increasing the confidence in long predicted segments of disorder at the expense of shorter predictions. Another factor may be the training set, which is taken exclusively from crystal structures and does not restrict the definition of disorder to long continuous regions.

The genomic analysis was designed to provide quantitative estimates for the abundance of native disorder, though this is complicated by the difficulty in establishing the true error rates for DISOPRED2. Most of the structures in the PDB come from proteins that have been successfully crystallized and it does not constitute a random sample. The subset used to estimate error rates does not therefore contain members of structural families in the same proportions as those in the population and there is also bias towards smaller, single domain proteins. However, the estimates are likely to be conservative because of the very low false positive rate and the likelihood of there being a significant number of disordered regions that are falsely predicted as ordered. It is possible, for example, that the under-representation of "unknown" annotations in the disordered set could be caused by DISOPRED2 failing to recognize other types of disorder that exist in the hypothetical proteins.

Although the most striking feature of Table 2 is the discrepancy between eukaryotes and prokaryotes, smaller differences are also observed between the archaea and the eubacteria. The scarcity of disordered regions in the thermophiles is perhaps caused by the strong evolutionary constraint on protein melting point in these organisms. Indeed, the only reference archaean

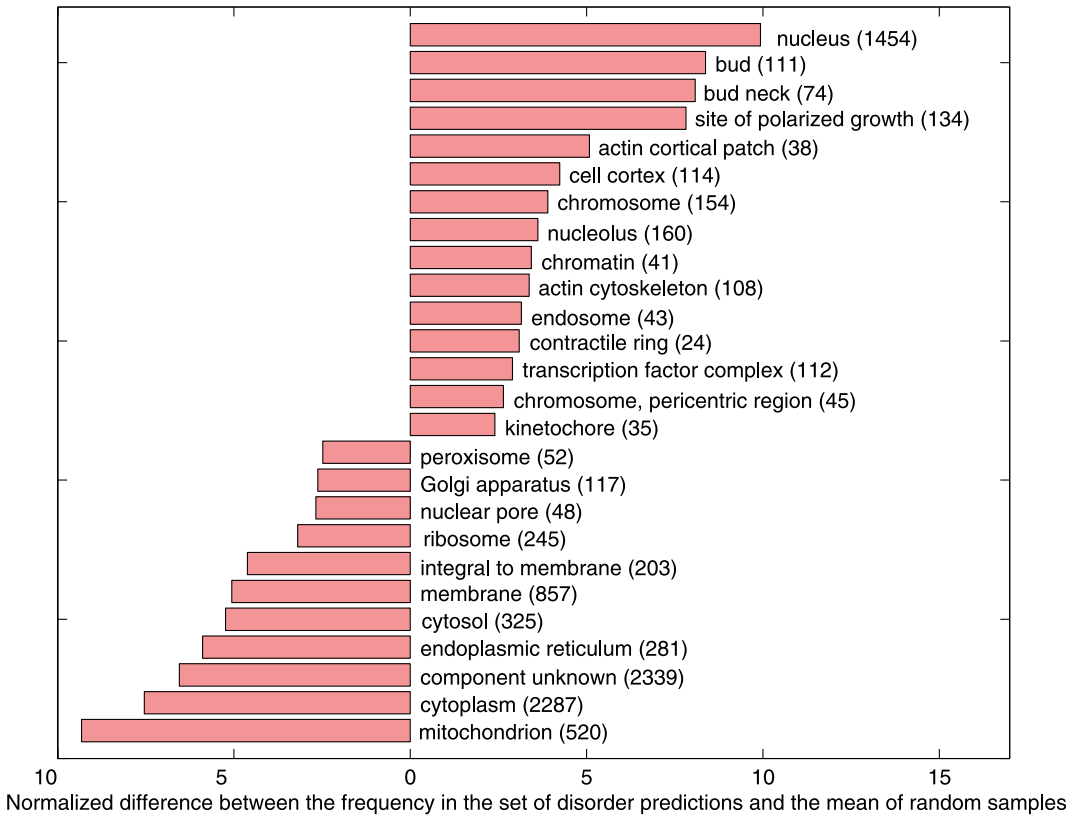

Figure 6. GO terms from the cellular component ontology that are significantly over- or under-represented in the set of disorder predictions.

organism with an optimum growth temperature below 60 °C is *Halobacterium* species, which is predicted to have a far larger proportion of long disordered segments. Amongst the eubacteria, the anomalously high disorder in *Mycobacterium tuberculosis* may be a result of its high G-C content and a raised propensity towards the amino acids Ala, Gly, Pro, Arg and Trp.[21]

Previous studies[10,11] have found disorder to be ubiquitous in all three kingdoms of life with around 60% of eukaryote, 28% of eubacteria and 36% of archaea proteins predicted to possess disordered regions longer than 40 residues. Here, the comparatively low over-prediction rates (0.5% cf 17%) mean that large systematic differences in the error rates between organisms are less likely. Table 2 therefore provides convincing evidence for disorder being common in eukaryotes but less so in prokaryotes. This confirms much of the experimental evidence to date, which has shown that dynamic flexibility of the protein structure is more often associated with eukaryotic protein function.[2] The results from the analysis of *Saccharomyces* also show that many of the functions associated with disordered regions are unique to eukaryotes such as the organization and biogenesis of the cytoskeleton.

There are several explanations for the lower occurrence of disorder in prokaryotes. Prokaryotes are subject to strong selective pressure on biochemical efficiency and do not have highly-regulated degradation pathways such as ubiquitination, so the cost of short protein lifetimes is likely to be far greater. The absence of cell compartments also reduces the ability of prokaryotic cells to physically protect unfolded structures from degradation. This is confirmed by Figure 6, which shows that the majority of putative disorder-containing proteins are located in cellular components that provide some protection from proteolysis such as the cell cortex and nucleus. The low levels of disorder in the mitochondrial proteins are likely to be a result of the organelle's propinquity to prokaryotes.

Many of the terms associated with disorder in Figures 4–6, such as DNA and cytoskeleton-binding, have been indicated by previous theoretical approaches[8,22] and numerous experiments.[2,20] The predominant molecular functions of long disordered segments appear to involve molecular recognition and, in particular, binding of DNA to facilitate processes such as transcription, transposition, packaging, repair and replication. The other processes that are linked to disorder include signalling, cell cycle, development and endocytosis.

Other results suggest that disorder is involved in signal transduction *via* the small GTPases and cell surface receptors, in addition to the protein kinases. These pathways also facilitate responses

to external stimuli, stress and the phases of cell cycle. It has been suggested that transformations from order to disorder allow the cell to rapidly and irreversibly reduce the concentration of signalling proteins in response to external or intracellular conditions.[23,24] The uniform proportion of disordered proteins across eukaryote proteomes also suggests that native disorder is involved in the multicellularity. An obvious candidate in higher organisms is the cell differentiation where disorder is likely to be present in proteins that modify the cytoskeleton and control gene expression.

Native disorder has been implicated in cancer-associated proteins present in the human genome.[22] Figure 5 provides further detail on the causal mechanisms of cancer that may involve disorder such as gene silencing, epigenetic regulation of expression[25] and DNA repair.[26] The presence of disorder in Ty transposable element proteins also suggests that it is a feature of retroviruses, because of Ty elements' origin as a retroviral infection that has been fixed in the yeast genome. It is likely that the use of disorder for reversible binding of DNA, and possibly transport through a small orifice,[1] is advantageous to retroviral infectivity in eukaryote cells.

The low occurrence of disorder in functions such as biosynthesis and metabolism has also been indicated.[22] This suggests that the rigid body model of molecular recognition applies fairly generally to the interactions between catalytic proteins and their substrates, and may also explain the preponderance of enzymes in the PDB.[27] Protein kinases do not conform to this general trend as they are strongly associated with disorder and are not readily crystallized. However, this is consistent with other functions that utilize disorder, since kinases are involved in regulatory processes, and are required to simultaneously bind a nucleotide (ATP) and the protein phosphorylation site.

In summary, native disorder is involved in some of the most important regulatory processes in eukaryotes; cell damage that renders some of these processes inactive is known to contribute to the development of cancer in humans. The abundant disorder in eukaryotes indicates that the folding of a water-soluble protein into compact structure is often incomplete in the absence of stabilizing proteins, ligands or DNA. This represents a limitation on the scope of structural genomics projects, and has implications for our understanding of structural biology and protein–protein interactions. The analysis using the GO also suggests that the presence of long disordered regions is linked to several locations, functions and processes, and may be of use in annotating protein function.

System and Methods

Recognition of native disorder

The training set for DISOPRED2 was the same as that used to train the original version of DISOPRED[28] and was composed of non-redundant chains with X-ray structures in the PDB[13] and less than 25% pair-wise sequence identity. Only structures with resolutions better than 2.0 Å were used to ensure that missing regions were not caused by poor model quality. Disordered residues were identified by aligning the sequence of the protein chain in the SEQRES records with the sequence as specified by the ATOM records (alpha-carbon coordinates). Residues, which were found in the SEQRES records but not in the ATOM records were classed as disordered. The final training set comprised of 715 protein chains, in which a total of 176,550 residues were classed as ordered and 4590 residues as disordered.

Discriminating between ordered and disordered regions is a binary classification problem that can be solved using a support vector machine (SVM). SVMs may improve generalization by controlling the classifier's capacity and the associated potential for overfitting. This is achieved by ensuring that the decision boundary separating two classes does so with a large margin.[29] In this case, the SVM also has the advantage that it can be trained more efficiently than back-propagation networks.

The *SVMlight* support vector machine package was used to train the classifiers.[30] The linear kernel was used, corresponding to a hyperplane in the input space, and the learning parameters were found by fourfold cross-validation. Unbalanced class frequencies can result in classifiers that output the majority class exclusively, since this optimizes overall accuracy. This behaviour is prevented in a formulation of the SVM that places asymmetric costs on points that violate the geometric margin.[31] This allows a greater cost to be placed on margin breaches by points from the minority (disordered) class than examples from the majority (ordered) class. Correctly setting the asymmetric cost parameter results in informative classifier outputs.

For each protein in the training set, a sequence profile was generated using three iterations of a PSI-BLAST[32] search against a non-redundant sequence database. Figure 7 shows receiver operating characteristic (ROC) curves for several classifiers trained using various combinations of binary-encoded amino acid sequence, secondary structure predictions from PSIPRED[33] and PSI-BLAST profiles for symmetric windows of 15 positions. The N and C termini were treated separately as it has been demonstrated that there are different patterns in disordered sequences at the terminal positions.[6]

The area under the ROC curve has similar properties to the non-parametric Wilcoxon statistic with a score of 50% representing random and 100% perfect classification. Table 3 shows estimates of the area under each ROC curve in addition to the results from a second smoothing classifier trained on the outputs of the profile SVM. The

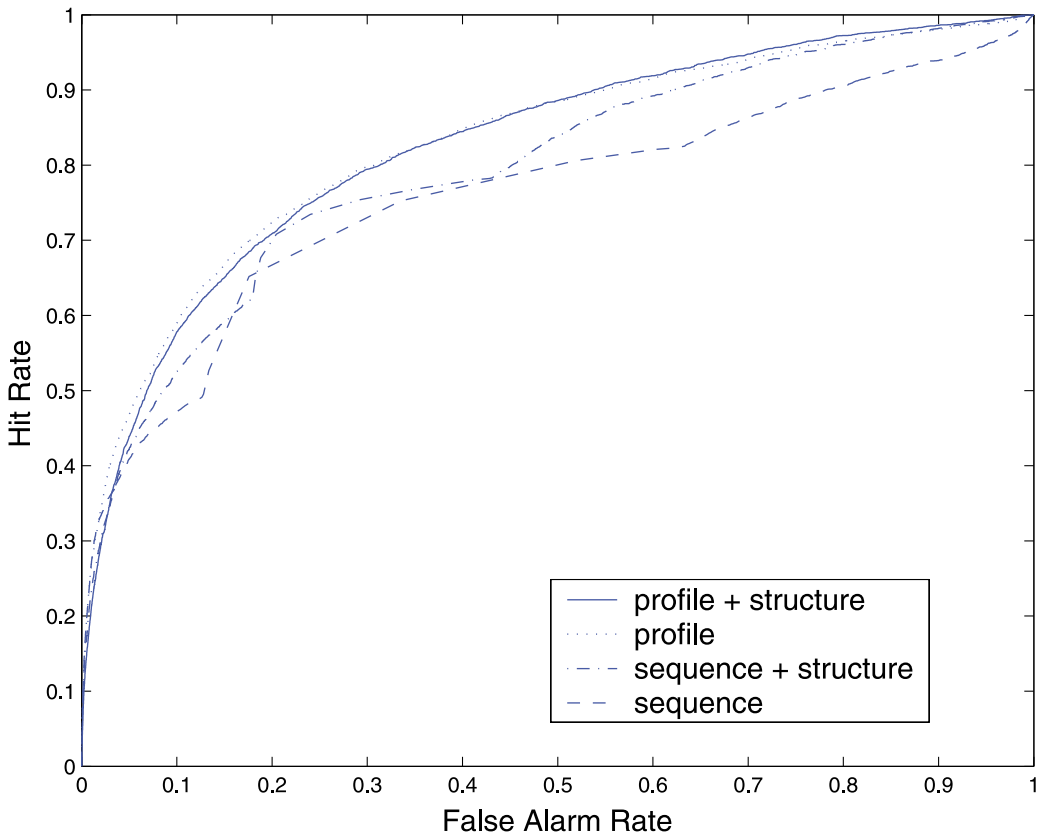

Figure 7. Receiver operator characteristic curves for linear SVM classifiers generated using fourfold cross-validation on the non-redundant set of proteins. The ROC curves were generated by varying the decision threshold of each SVM classifier.

Wilcoxon test statistics can be used to indicate whether the difference between two methods is statistically significant.[34] In this case, all the differences apart from the two profile-based classifiers are significant at the 95% level.[35]

Classifiers trained on PSI-BLAST profiles outperform those trained on single sequences across the range of error rate thresholds, indicating that evolutionary information improves prediction of disorder. Secondary structure predictions improve the accuracy of the sequence classifiers because they implicitly contain information from the position-specific scoring matrix but do not improve the profile classifiers. This contradicts the results from the first version of DISOPRED,[28] where structure predictions were used to improve accuracy.

The difference may be attributable to the different learning algorithms used in the two cases. Here, we use a linear SVM with a relatively low capacity[29] (i.e. will avoid overfitting the data) and a capacity-controlling maximal margin learning algorithm. On the other hand, the two-layer neural network used to train the original classifier had a relatively large number of hidden units and may have been prone to overfitting. It is possible that the improved generalization of the SVM prevents prediction of disorder in regions that are likely to form helical or strand elements in the core of a globular protein.

Further improvements in accuracy can be achieved by inputting the first set of predictions into a second smoothing network, which extends the effective length of the input window from 15 to 29 residues and increases prediction accuracies of longer (>15 residue) disordered segments. The cascaded classifier therefore has the familiar two-layer neural network topology with 15 hidden units but without full connectivity in the first layer of adaptive weights. DISOPRED2 is comprised of this cascaded classifier, trained on the full set of proteins.

Table 3. Cross-validated performance measures for several disorder classifiers

	C	Q_2	Wilcoxon	SE
Profiles + structure	0.26	93.74	82.70	0.34
Profiles	0.27	93.79	83.02	0.36
Sequence + structure	0.25	93.69	79.84	0.38
Sequence	0.24	93.63	76.16	0.46
Cascaded classifier	0.35	94.05	86.75	0.30

Columns show Matthews correlation coefficient (C) and two-state accuracy (Q_2) for a false hit rate of 0.05 and the Wilcoxon statistic with its standard error. The area under the ROC curves (Wilcoxon statistic) was calculated using trapezium rule numerical integration.

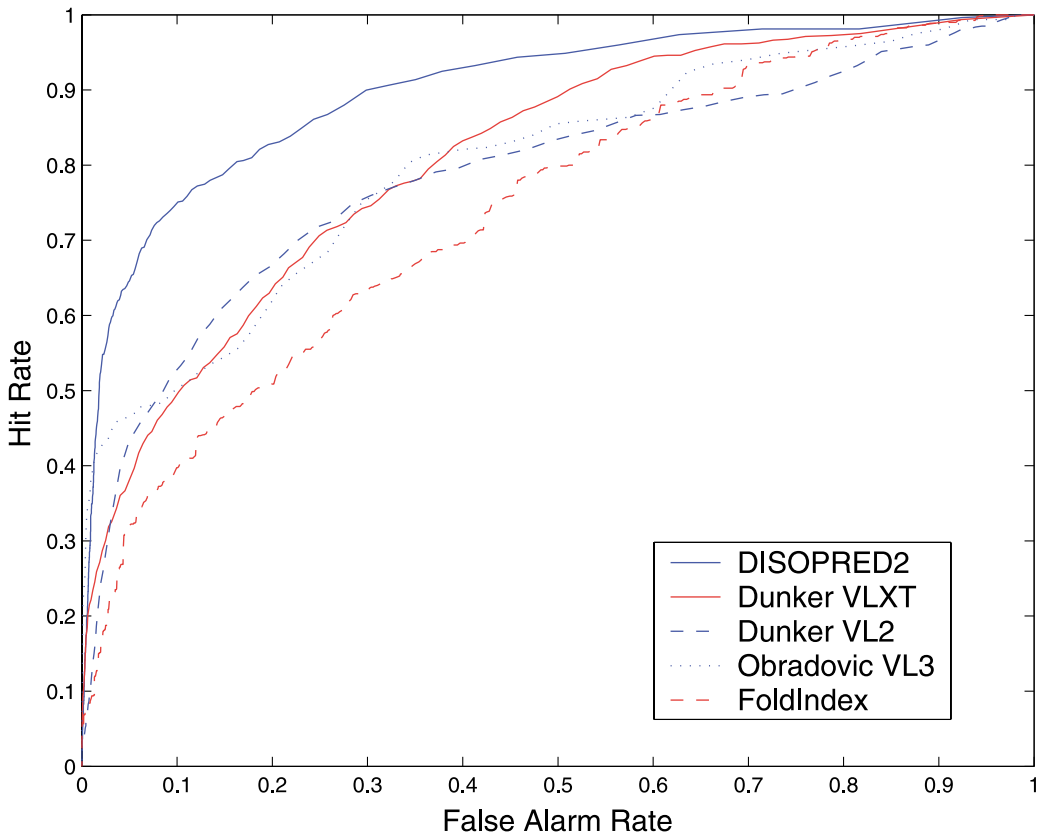

Figure 8. Receiver operator characteristic curves comparing the outputs of DISOPRED2 to four other methods evaluated on the CASP targets.

Benchmarking

An objective comparison was carried out between DISOPRED2 and several other disorder prediction methods evaluated on targets from the last CASP experiment. The PSI-BLAST search database and the training set for DISOPRED2 were compiled before the start of CASP so the test can be considered fair. Figure 8 and Table 4 show results from DISOPRED2 along with those from the Obradovic (VL3) and Dunker (VLXT) groups submitted in the model 1 category and the VL2 method from the Dunker group, which achieved highest accuracy according to their own assessment.[36]

The Obradovic & Dunker groups have established several disordered prediction methods in collaboration.[36] During the CASP experiment, the two groups submitted predictions independently. The predictions from VL2 and VLXT come from ensembles of neural networks trained on combinations of amino acid composition, flexibility and sequence complexity. The VL3 predictor was trained using ordinary least squares regression with partitioning of the training set to cluster various "flavors" of disorder.[10] The FoldIndex program† is based on the calculations developed by Uversky, Gillespie & Fink,[20] and predicts whether a sequence will fold by computing its mean net charge and hydrophobicity. The window parameter for the FoldIndex classifier was set to 31 residues as this value achieved highest accuracy on a validation set.

The DISOPRED2 predictor achieves higher accuracy than the other methods across the range of decision thresholds apart from a slight deficiency over VL3 at very low false positive rates (<1.27%). This is likely to be a result of VL3 being trained on only long regions of disorder, which can be predicted more accurately than shorter regions. Applying a simple rule to the outputs of DISOPRED2 that removes predictions for short disordered segments yields even higher accuracies at the low thresholds. The differences in the Wilcoxon scores between DISOPRED2 and

Table 4. Matthews correlation coefficient and two-state accuracy (Q_2) for a false hit rate of 0.05 and the Wilcoxon statistic for CASP targets

	C	Q_2	Wilcoxon	SE
DISOPRED2	0.51	93.1	90.0	0.64
Dunker VLXT	0.31	91.41	80.94	0.80
Dunker VL2	0.36	91.76	78.62	0.99
Obradovic VL3	0.38	92.05	80.07	0.91
FoldIndex	0.26	91.0	73.8	0.92

† http://bioportal.weizmann.ac.il/fldbin/findex

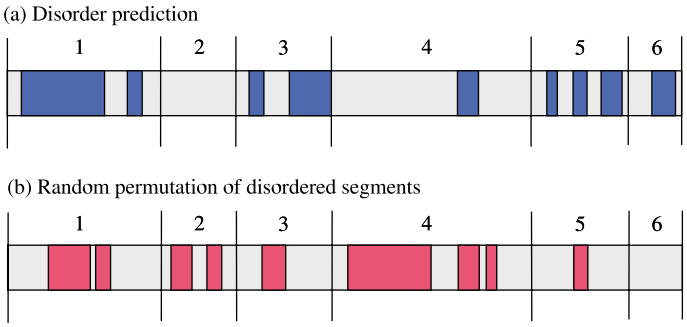

Figure 9. (a) Representation of the yeast proteome. The six proteins are separated by vertical lines with long predictions of disorder colored in blue. (b) Randomly shuffled segments are displayed in red.

the other three methods are all statistically significant at the 99% level.

Predicting disorder in complete genomes

The protein sequences for six archaea, 13 eubacteria and five eukaryote genomes were downloaded from the NCBI ftp server. These sequences were first filtered using the sequence masking program pfilt to remove coiled-coil and transmembrane regions.[37] The low sequence complexity and compositional bias filters were not used as these regions are often disordered. The PSI-BLAST jobs, used to calculate the inputs to DISOPRED2, were distributed across a Linux beowulf cluster of Intel Pentium and AMD Athlon processors and two associated SunFire 880 servers running Solaris. Sequences were submitted to the Sun Grid Engine scheduler using servlet technology.

The analysis of the annotations associated with predicted disorder was carried out using the July 2003 release of the SGD. The database contains 2337 unique GO terms attached to 5889 proteins in the set of translated open reading frames. The electronic annotations were excluded from the analysis to ensure reliability. The SGD annotates each protein with the most specific terms available in the GO. However, the hierarchical structure of GO means that all ancestral terms also provide a valid description in a more general sense. For example, the term "cell cycle" is part of "cell proliferation", which is a process of "cell growth and maintenance". Including the ancestral nodes in GO's directed, acyclic graph hierarchy expands the number of unique terms to 3299.

A diagram of the sampling method used to determine the functional significance of disorder is shown in Figure 9. Each disordered segment was mapped to the GO annotations for the protein in which it occurred. The number of times each GO term occurred was then counted across the entire set of disordered predictions. In each random simulation, segments with identical lengths to the disordered predictions were randomly distributed across the yeast proteome with the constraint that segments could not cross the boundaries separating each protein. The number of times each GO term occurred in 10,000 simulations was used to obtain p-values for the disorder predictions under the null model.

Acknowledgements

This work was supported by the Medical Research Council (J.J.W. and J.S.S.). Thanks to Stefano A. Street for assistance with the distributed computing, and David Corney and Kevin Bryson for useful discussions.

References

1. Daughdrill, G. W., Hanely, L. J. & Dahlquist, F. W. (1998). The C-terminal half of the anti-sigma factor FlgM contains a dynamic equilibrium solution structure favoring helical conformations. *Biochemistry*, **37**, 1076–1082.
2. Wright, P. E. & Dyson, H. J. (1999). Intrinsically unstructured proteins: re-assessing the protein structure–function paradigm. *J. Mol. Biol.* **293**, 321–331.
3. Weiss, M. A., Ellenberger, T., Wobbe, C. R., Lee, J. P., Harrison, S. C. & Struhl, K. (1990). Folding transition in the DNA-binding domain of GCN4 on specific binding to DNA. *Nature*, **347**, 575–578.
4. Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M. & Obradovic, Z. (2002). Intrinsic disorder and protein function. *Biochemistry*, **41**, 6573–6582.
5. Romero, P., Obradovic, Z., Kissinger, C. R., Villafranca, J. E. & Dunker, A. K. (1997). Identifying disordered regions in proteins from amino acid sequences. *Proc. IEEE Int. Conf. Neural Netw.*, 90–95.
6. Li, X., Romero, P., Rani, M., Dunker, A. K. & Obradovic, Z. (1999). Predicting protein disorder for N-, C-, and internal regions. *Genome Inform.* **10**, 30–40.
7. Romero, P., Obradovic, Z., Li, X., Garner, E., Brown, C. & Dunker, A. (2001). Sequence complexity and disordered proteins. *Proteins: Struct. Funct. Genet.* **42**, 38–48.
8. Dunker, A. & Obradovic, Z. (2001). The protein trinity—linking function and disorder. *Nature Biotechnol.* **19**, 805–806.
9. Melamud, E. & Moult, J. (2003). Evaluation of disorder predictions in CASP5. *Proteins: Struct. Funct. Genet.* **53**, 561–565.
10. Vucetic, S., Brown, C. J., Dunker, A. K. & Obradovic, Z. (2003). Flavors of protein disorder. *Proteins: Struct. Funct. Genet.* **52**, 573–584.

11. Dunker, A. K., Obradovic, Z., Romero, P., Garner, E. & Brown, C. (2000). Intrinsic protein disorder in complete genomes. *Genome Inform.* **11**, 161–171.
12. Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. *Nature Genet.* **25**, 25–29.
13. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al.* (2000). The protein data bank. *Nucl. Acids Res.* **28**, 235–242.
14. Mazza, C., Segref, A., Mattaj, I. W. & Cusack, S. (2002). Large-scale induced fit recognition of an m(7)GpppG cap analogue by the human nuclear cap-binding complex. *EMBO J.* **21**, 5548–5557.
15. Dyson, H. J. & Wright, P. E. (2002). Coupling of folding and binding for unstructured proteins. *Curr. Opin. Struct. Biol.* **12**, 54–60.
16. Gene Ontology Consortium (2001). Creating the gene ontology resource: design and implementation. *Genome Res.* **11**, 1425–1433.
17. Dwight, S. S., Harris, M. A., Dolinski, K., Ball, C. A., Binkley, G., Christie, K. R. *et al.* (2002). *Saccharomyces* genome database (SGD) provides secondary annotation using the gene ontology (GO). *Nucl. Acids Res.* **30**, 69–72.
18. Rost, B. & Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* **232**, 584–599.
19. Rost, B. & Sander, C. (2000). Third generation prediction of secondary structures. In *Methods in Molecular Biology* (Webster, D., ed.), pp. 71–96, Humana Press, Totowa, NJ chapt. 5.
20. Uversky, V., Gillespie, J. & Fink, A. (2000). Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins: Struct. Funct. Genet.* **41**, 415–427.
21. Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D. *et al.* (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, **393**, 537–544.
22. Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradovic, Z. & Dunker, K. A. (2002). Intrinsic disorder in cell-signalling and cancer-associated proteins. *J. Mol. Biol.* **323**, 573–584.
23. Nakayama, K., Hatakeyama, S. & Nakayama, K. (2001). Regulation of the cell cycle at the G1-S transition by proteolysis of cyclin E and p27Kip1. *Biochem. Biophys. Res. Commun.* **282**, 853–860.
24. Parker, D., Rivera, M., Zor, T., Henrion-Caude, A., Radhakrishnan, I., Kumar, A. *et al.* (1999). Role of secondary structure in discrimination between constitutive and inducible activators. *Mol. Cell. Biol.* **19**, 5601–5607.
25. Nephew, K. P. & Huang, T. H. (2003). Epigenetic gene silencing in cancer initiation and progression. *Cancer Letters*, **190**, 125–133.
26. Khanna, K. K. & Jackson, S. P. (2001). DNA double-strand breaks: signaling, repair and the cancer connection. *Nature Genet.* **27**, 247–254.
27. Hegyi, H. & Gerstein, M. (1999). The relationship between structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.* **288**, 147–164.
28. Jones, D. T. & Ward, J. J. (2003). Prediction of disordered regions in proteins from position specific scoring matrices. *Proteins: Struct. Funct. Genet.* **53**, 573–578.
29. Vapnik, V. (1998). *Statistical Learning Theory*, Wiley, New York.
30. Joachims, T. (1999). Making large-scale SVM learning practical. In *Advances in Kernel Methods—Support Vector Learning* (Schoelkopf, B., Burges, C. & Smola, A., eds). MIT Press, Cambridge, MA, USA.
31. Morik, K., Brockhausen, P. & Joachims, T. (1999). Combining statistical learning with a knowledge-based approach—a case study in intensive care monitoring. *Int. Conf. Mach. Learn. (ICML)*.
32. Altschul, S. F., Madden, T. L., Schaeffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.
33. Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 196–202.
34. Hanley, J. A. & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29–36.
35. Hanley, J. A. & McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, **148**, 839–843.
36. Obradvic, Z., Peng, K., Vucetic, S., Radivojac, P., Brown, C. J. & Dunker, A. K. (2003). Predicting intrinsic disorder from amino acid sequence. *Proteins: Struct. Funct. Genet.* **53**, 566–572.
37. Jones, D. T. & Swindells, M. B. (2002). Getting the most from PSI-BLAST. *Trends Biochem. Sci.* **27**, 161–164.

Edited by J. Thornton

(Received 20 October 2003; received in revised form 20 January 2004; accepted 3 February 2004)