Predicting Binding Regions within Disordered Proteins

Ethan Garner [1] egarner@disorder.chem.wsu.edu
Pedro Romero [2] promero@eecs.wsu.edu
A. Keith Dunker [1] dunker@disorder.chem.wsu.edu
Celeste Brown [1] celesteb@disorder.chem.wsu.edu
Zoran Obradovic [2] zoran@eecs.wsu.edu

[1] School of Molecular Biosciences
Washington State University, Pullman, WA 99164, U.S.A.
[2] School of Electrical Engineering and Computer Science
Washington State University, Pullman, WA 99164, U.S.A.

Abstract

Disordered regions are sequences within proteins that fail to fold into a fixed tertiary structure and have been shown to be involved in a variety of biological functions. We recently applied neural network predictors of disorder developed from X-ray data to several protein sequences characterized as disordered by NMR (Garner, Cannon, Romero, Obradovic and Dunker, Genome Informatics, 9:201–213, 1998). A few predictions on the NMR-characterized disordered regions were noted to contain "false" negative indications of order that correlated with regions of function. These and additional examples are examined in more detail here. Overall, 8 of 9 functional segments in 5 disordered proteins were identified or partially identified by this approach. The functions of these regions appear to involve binding to DNA, RNA, and proteins. These regions are known to undergo disorder-to-order transitions upon binding. This apparent ability of the predictors to identify functional regions in disordered proteins could be due to the existence of different flavors, or sub-classes of disorder, originating from the sequence of the disordered regions and perhaps owing to local inclinations toward order. These different flavors may be a characteristic that could be used to identify binding regions within proteins that are difficult to characterize structurally.

1 Introduction

As the amount of sequence data from various genomic projects increases, it becomes ever more important to predict structure from primary sequence. Predictions of structure from sequence, whether by *ab initio* approaches or homology-based methods, are then used to deduce function. This scheme assumes that defined structures are a prerequisite for function [10, 15, 17, 20].

Increasing attention has recently focused on proteins or regions of proteins lacking fixed tertiary structure, essentially being partially or fully unfolded [7, 11, 26, 39]. Such disordered regions (DRs) have been shown to be involved in a variety of functions, including DNA recognition [5, 16, 23, 24], modulation of specificity/affinity of protein binding [1, 8, 34, 28], molecular threading [6], activation by cleavage [4, 9], and control of protein lifetimes [18]. Although these DRs lack a defined 3-D structure in their native states, they frequently undergo disorder-to-order transitions upon binding to their partners.

As it is known that sequence determines structure [2], we assumed that sequence would determine lack of structure as well. To test this, we developed a series of neural network predictors (NNPs) that use amino acid sequence data to estimate the likelihood of disorder in a given region [30, 31]. The 70% accuracies shown from 5-cross validation and out-of-sample testing [13, 30, 33] support the hypothesis that DRs are encoded by local amino acid sequence. Further support for this hypothesis comes from observations that DRs have characteristics that are consistent with expectations for non-folding sequences, such as significant net charge, lack of aromatic residues, an excess of hydrophilic groups, or a combination of these and other appropriate sequence attributes [7, 40].

Proteins with DRs are largely unexplored, yet represent a significant percentage of proteins in nature. Previous studies have indicated that, at a threshold with a false positive rate less than 1 in 618,344 predictions, more than 1000 proteins out of the database SWISS-PROT contain DRs longer than 40 amino acids in length. At a threshold with a false positive rate of about 7% per protein chain, almost 25% of the SWISS-PROT sequences are predicted to have such long DRs [32].

Gerstein et al. [14] showed that the proteins from structural databases are related to only a small subset of those from genomic databases and that the compositional qualities of the unrepresented proteins are indeed quite different from those sequences represented in structural databases. The unrepresented proteins from the genomes had significantly more charged residues, less cysteine and less tryptophan than proteins in the Protein Data Bank (PDB). These sequence characteristics are some of those associated with native protein disorder [7, 40]. Thus, a reasonable suggestion is that the biases discovered by Gerstein et al. are due to a significant degree to the existence of protein disorder. Studies are in progress to estimate the relative amount of disorder in the current complete genomes.

In a previous study of regions known to be disordered through NMR characterization, we noted a couple of examples in which prediction errors corresponded to functionally important segments [13]. That is, apparent false negative errors (predictions of order within regions structurally characterized to be disordered) mapped to functional regions within the disordered protein sequences. Here we present a more complete examination of these previous examples and extend the study to new ones. Although the sample size is still too small to serve as the basis for generalization, these few examples demonstrate a novel use for our NNPs, namely, to perhaps identify binding segments within DRs.

2 Materials and Methods

2.1 Development of NNPs

Two of our disorder predictors are utilized in this study, the X-ray and calcineurin (CaN) NNPs. The X-ray NNP is the same as the long disordered region (LDR) predictor described previously [33] but renamed for clarity [13]. The NNPs use primary sequence information, within a sliding window of 21. Attributes, such as numbers of particular amino acids or hydropathy, are calculated over this window and used as inputs into the NNP. The NNP then assigns an output value to the central amino acid within the window. Any output value exceeding a threshold of 0.5 is considered disordered. Past studies have shown that as the length of the predicted disordered region increases, the false positive error rate of that given prediction decreases, and that errors frequently occur at boundary regions, perhaps due to the use of windows by the predictor [13].

The X-ray NNP was trained upon a disordered data set of 7 structures with missing coordinates in PDB. These were selected to have no associating subunits or bound cofactors within the crystal structure. An ordered set was obtained from randomly selected patterns from NRL_3D [25]. The X-ray NNP's attributes were selected using a method of sequential forward search. The resulting NNP utilized a feed forward architecture with 10 inputs, 7 fully connected nodes within the hidden layer, and a single output. The accuracy was 73% as judged by 5-cross validation on a residue-by-residue basis [31].

The calcineurin NNP, an example of a family-specific predictor, was developed based upon data from alignments between the disordered region of human calcineurin and those of calcineurins from other species identified in SWISS-PROT [30]. Any SWISS-PROT sequence segment that aligned with the disordered region of human CaN was considered disordered. Ordered data patterns were again selected randomly from NRL_3D. Branch and bound searching was utilized to determine the optimal features to be used as inputs. The CaN NNP used 10 inputs, 10 fully connected nodes, and one output, and had a 5-cross-prediction accuracy of 83% [30].

The features that serve as inputs for the two NNPs are similar; indeed, 6 of 10 selected attributes are the same. Thus, similar attributes were selected despite the different training sets used. A list

of selected features is presented in Table 1, where the individual amino acids correspond to their compositions in windows of 21. The hydropathy attribute used the Kyte and Doolittle scale [19], while flexibility values were based on backbone atom B-factor values, averaged for each amino acid type [37]. The β-moment was calculated as described previously [30]. A list of selected attributes is presented in Table 1.

Table 1: Features used by the NNPs.

X-ray NNP	H	C	S	W	Y	E	D	K	Hydropathy	Flexibility
CaN NNP	H	C	S	W	Y	E	V	F	R	β-moment

2.2 Application of NNPs to Disordered Proteins and Cross Validation

Both the LDR and the CaN NNP were applied to a set of structurally well-characterized proteins (see http://disorder.chem.wsu.edu/proteinlist.html), all of which have been shown through a variety of methods to be fully disordered or to contain long disordered regions. For the NMR-characterized proteins in the absence of binding partners, no tertiary structure was evident, and secondary structure, if present at all, was highly dynamic.

In the present study, the NNPs were applied only to the disordered parts of the proteins. First, we identified regions that showed predictions of order or large differences between the outputs of the two different predictors, with at least one of the predictors giving false negative indications of order. These predicted regions of order were then examined to see whether they corresponded to any functional sites within the sequence.

2.3 Proteins Used in This Study

The 4e binding protein 1 (4E-BP1) is a translational regulator, inhibiting the formation of the initiation factor complex eIF4F by binding to eIF4E. A 20 residue fragment of 4E-BP1 was shown to be sufficient to bind to the eIF4E protein and inhibit translation [11, 12]. 4E-BP1 has been shown by NMR to have little folded structure in solution. Only the binding site (residues 49-68) becomes structured upon binding, while the remainder of the protein remains disordered.

High mobility group factor (HMGI(Y)) is a protein involved in assembly of higher order transcription enhancer complexes. This protein contains 3 DNA binding domains that interact with the B DNA minor groove [16]. This protein has been shown to be fully disordered by NMR, and becomes structured upon binding to DNA.

The transcriptional activation protein N of phage lambda (APN) regulates genes expressed from phage promoters to allow the phage to transcribe through terminators. The protein includes a Box B RNA binding region, a Nus A binding region, and a carboxy-terminal region that interacts with RNA polymerase. NMR indicates that the protein is disordered in solution, and that the Box B RNA binding region undergoes a local disorder-to-order transition when interacting with the BoxB RNA, while the other regions remain disordered in the absence of interactions with their target proteins [22].

Flagellin is the sole component of bacterial flagella, and is a self-polymerizing protein that has unfolded N and C terminal ends in its monomeric form [38]. When these regions are proteolytically removed, the flagella still assemble, but the stability and polymorphic ability of the flagella are lost. The N terminal region is involved in the completion of outer-tube domain folding and in the regulation of the initiation and stabilization of these interactions between subunits. This region becomes ordered upon polymerization [21].

The N transcriptional activation domain of thyroid transcription factor-1 (TTF-1) is a unique DNA binding domain that falls outside of the proline-rich, glutamine rich and acidic domain classifications

previously suggested for transcription factors. TTF-1 functions as a tissue specific transcription factor, regulating transcription of thyroid and lung-specific genes. It has been shown by CD to exist in a random coil conformation and is quite susceptible to protease digestion [36]. Similar to acidic transcription domains, TTF-1 assumes a helical structure when binding to DNA. The formation of a helix within regions required for binding is supported by CD spectroscopy, by protection from protease digestion in solution with triflouroethanol (TFE), and by secondary structure predictions [36].

Table 2: Proteins Examined.

Protein name	Disordered Region Examined	Method of Structural Characterization	Functional Regions	Refs
4E-BP1	1-118 Fully disordered	NMR, CD	49-68 Minimal eIF4E binding region	[11, 12]
HMGI(Y)	1-106 Fully disordered	NMR	23-31 DNA BD I, 55-70 DNA BD II, 81-89 DNA BD III	[16, 29]
APN	1-107 Fully disordered	NMR	1-22 Box B RNA binding, 34-47 Nus A binding, 73-107 RNA polymerase binding	[22]
N term. of flagellin	1-65	Protease digestion, Adiabatic compressibility	40-65 Involved in outer domain folding and subunit stabilization	[21]
TTF-1	1-156 Fully disordered	CD, Protease Digestion	51- 102 Minimal activating region, 49-73 Predicted helix, 58-78 TFE + protease resistant	[36]

3 Results

3.1 Cross Predictions

We have developed several different NNPs based on different training sets [30, 31, 33]. Predictions on each other's training examples, i.e. cross predictions, revealed different types, or "flavors", of disorder [33]. Thus, cross prediction is a useful approach for characterizing similarities and differences between two predictors.

For the studies presented here, we applied several of our predictors to known regions of disorder to test for different "flavors". Although interesting details come from comparing the results of several of these predictors, the main points are evident from comparing just two: the CaN and X-ray NNPs.

The X-ray NNP included CaN within its training set; indeed, almost 1/4 of the disordered amino acids came from this protein. Despite this overlap, only 6 out of 10 of the attributes were common between the two training sets. To determine the extent to which the two predictors give concordant predictions, the CaN NNP was applied to the DRs of the X-ray NNP's training data and *vice versa*.

The X-ray NNP exhibited a very high success rate (97%) in cross-predictions on the CaN disorder data, whereas the CaN NNP gave a much lower accuracy (59%) on the X-ray disorder data, and an even lower value (40%) after the CaN disorder was removed from the X-ray training set. These results are consistent with other studies showing that the X-ray NNP does fairly well across a wide spectrum

of disordered proteins, whereas the CaN NNP exhibits specificity for CaN-like disordered regions [33]. The presence of a CaN sequence in the X-ray NNP's training set certainly contributes to its high success rate on the CaN disorder data, but the success is still unexpectedly high for reasons that we don't understand.

3.2 Predictions on Disordered Proteins

Fig. 1-Fig. 5 show the numerical NNP outputs of the predictions on the well-characterized DRs of the selected proteins. The vertical axis of each graph represents the NNP's output, while the horizontal axis represents the residue number. The outputs are given by solid (X-ray NNP) and dashed (CaN NNP) lines. Values above 0.5, indicated by the central line, are predictions of disorder; below this line, order. Shaded regions have been identified as important to function.

Application of the two NNPs to the 4E-BP1 protein (Fig. 1) gave concordant and correct predictions of disorder in the C-terminal region, concordant but false negative predictions of order in the central region, and discordant predictions in the N-terminal region. The concordant false negative predictions of order and the identified binding region of this protein are essentially coincident as indicated by the degree of overlap of the shaded area (49-68) with the two strong predictions of order.

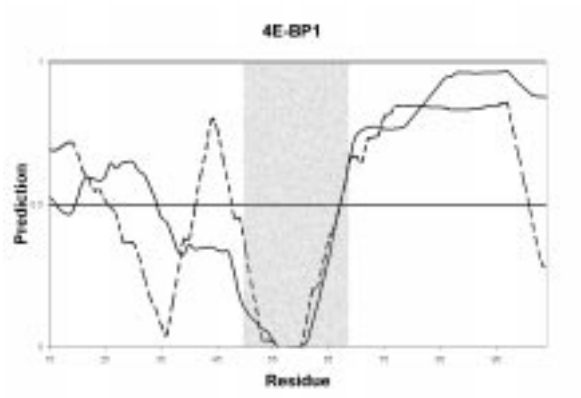

Figure 1: 4E-BP1.

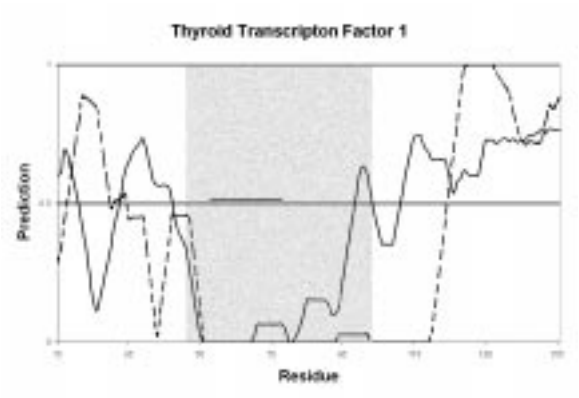

Figure 2: TTF-1.

As for TTF-1 (Fig. 2), the shaded area and the bar represent the suggested region essential for function. That is, the shaded area (residues 51-102) indicates the minimal activating region, while the black bar (residues 58-78) denotes the location of the protease-resistant, TFE-induced helix, which is a putative DNA binding site. Again, the false negative prediction of order and the functional region are mostly coincident. A very low score is predicted for the minimal activating region, but an absolute minimum (e.g. a neural net output of 0) is predicted in the location of the protease-resistant, TFE-induced helix.

For the remaining 3 proteins the two NNPs differ in their false negative predictions of order. In these examples, differences between the predictor outputs identify functionally important regions that undergo disorder-to-order transitions upon binding.

For flagellin (Fig. 3), the X-ray NNP (solid line) indicates that this segment of the protein is disordered, save for a small dip in the 45-55 residue region. The CaN NNP (dashed line) shows a near absolute minimum (e.g. an output near 0) for the residues 43-62, all falling within the shaded 40-65 segment involved in the intersubunit interactions.

The predictions upon HMGI(Y) (Fig. 4) show that the CaN NNP (dashed line) predicts minimas to be overlapping with, but slightly off-shifted from, the shaded DNA binding domains (residues 23-31, 55-70, and 81-89). In contrast, the X-ray NNP (solid line) considers this a fully unfolded protein as indicated by the high score throughout the sequence.

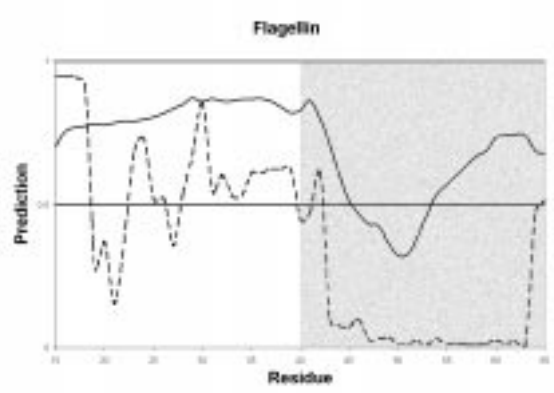

Figure 3: N terminus of Flagellin.

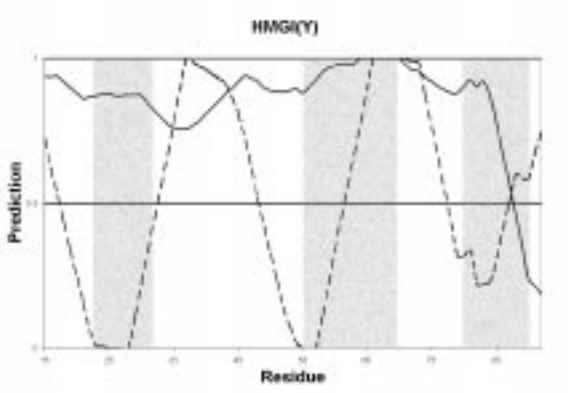

Figure 4: HMGI(Y).

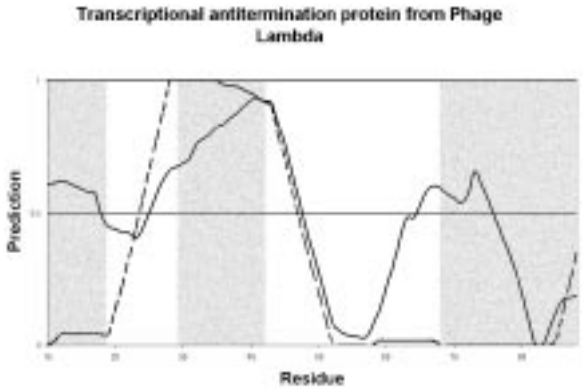

Figure 5: APN.

The predictions on APN (Fig. 5) show that the X-ray NNP (solid line) regards the Box B RNA binding site (first shaded region on the left) to be disordered while the CaN NNP (dashed line) gives an output indicating order. This pattern is again repeated in the RNAP binding site in the C-term (right shaded region). Here the CaN NNP considers the residues 55 and higher to be ordered, while the X-ray predicts disorder for residues 65-77. An additional 20 residues (55-75) are predicted to be ordered aside from the known functional region. We wonder whether these extra 20 residues are involved in some yet unknown function. The NUSA binding region, residues 34-47, is predicted as disordered rather than ordered, and is considered missed by the criteria of this study.

4 Discussion

4.1 Potential Binding Sites by Disorder Prediction

In the cases presented above, the NNPs are able to identify binding sites within disordered proteins by yielding apparent false negative predictions, e.g. predictions of order. The two predictors may give differing patterns of agreement in regards to the binding regions, but in all cases the functional regions were identified using sequence information alone.

The structural assignment of a protein as being "disordered" is difficult, and in many cases, somewhat arbitrary. While a protein may be disordered in its monomeric or unbound form, particular regions within the protein may become structured upon binding, as is the case for all of the aforementioned examples. Therefore, during predictor development, a judgement must be made whether

the disordered region is either consistently disordered or disordered until binding.

The X-ray NNP was trained upon a data set of unliganded, monomeric molecules, many of which undergo disorder-to-order transitions upon association with ligands. Since the molecules were unbound within the crystal structure, the binding regions of the X-ray NNP's training set were considered disordered during the training. This becomes manifest as shown by 3 examples above, where the unliganded binding regions are considered disordered by the X-ray NNP as it has been taught to recognize those regions as disordered.

The X-ray NNP, which is broad spectrum and has a fair out-of-sample accuracy, may miss functional information. The CaN NNP, on the other hand, has a very poor out-of-sample accuracy and it might therefore provide increased discrimination of the functional regions.

4.2 Different Types of Disorder

The results herein and the results of past work [30] have indicated the presence of different "flavors" of disorder. Evidence for these "flavors" has come from clustering analyses of our disordered and ordered data sets in attribute space. While random patterns from NRL-3D cluster in a given well-defined region, the attributes for disordered proteins lie in a much greater region of attribute space, and cluster within subgroups [30]. Indeed, the discovery of these differing inclinations toward disorder is what prompted the development of the CaN and other family-specific NNPs, so as to be able to differentiate possible disorder subtypes. Here we are investigating whether these "flavors" can provide information regarding protein function.

The presence of differing flavors of disorder may be responsible for the identification of the binding sites. The X-ray NNP evidently considers many of the binding domains disordered due to the larger "breadth" of its training set (and the inclusion of CaN). In contrast, the CaN NNP has not been trained upon as many regions of differing function as the X-ray NNP, and hence is more selective in its prediction of disorder. Similar arguments can be developed for the predictions on 4E-BP1 and TTF-1, where both predictors agree that the binding sites are ordered. A more globally recognized type of disorder may be seen in the linker regions between these binding sites, where both predictors consider the region disordered. These areas must contain attributes shared by both predictors and recognized by both as disordered.

4.3 Structural Implications

A different source of the false negative minimas can also be envisioned. The degree of order/disorder within proteins is by no means a binary state, as is assumed in the inputs to our predictors. That is, any given region could be in equilibrium between order and disorder. Different regions could have different values for the equilibrium constant, ranging from being disordered most of the time to being predominantly ordered. We would expect a collection of such sequences to exhibit a gradient of disorder tendencies rather than the two-state behavior assumed during training.

Perhaps the binding areas of the afore mentioned proteins, and of proteins in general that undergo disorder-to-ordered transitions upon binding, do indeed contain sequence compositions that are natively disordered, but are missing the nonlocal interactions required to drive them into the ordered state. The extreme false negative minimas witnessed on the binding regions within this study may contain an inclination toward, or transient state of, order, but again need the driving force provided by the interactions with the associating partner. This would allow for the increased rate of association of the interacting complex proposed in unstructured domains, through both an increased rate of association as well as modulation of the specificity of the binding [8, 27, 34]. Much evidence is growing to support this, including thermodynamic arguments which have been proposed stating that site-specific DNA-protein interactions are indeed disordered, and undergo disorder-to-order transitions upon binding [35].

This study shows that NNPs trained upon different disordered proteins can differentiate between functional DRs. This supports the view that DRs may contain differing compositional features characteristic of their function, which may prove to be a powerful tool for protein function identification. As the amount of both structural and functional data regarding disordered proteins increases, we hope to develop more NNPs based upon family and function-specific data, to perhaps discriminate other functional regions from sequence.

On the flip side of the disorder problem, many DRs have been identified without any known function. Are these truly "orphans" without function, or do they contain functions yet to be discovered. Using our predictors, systematic studies on "disorder orphans" may shed light upon both the locations and functions of such sequences.

4.4 Future Efforts

The results presented herein suggest that functional regions within disordered domains can be recognized by their local tendencies to form ordered structure. Of course, the number of examples is too small to know whether or not this is a general feature of such sequences. On the other hand, functional sites within disordered regions that are missed with a given set of predictors might be identified as more predictor flavors are added. Given the recent call for the identification and study of more disordered proteins [39], we look forward to a much larger sample size that will enable us to more completely test the proposals presented herein.

Acknowledgments

Support from NSF research grant NSF-CSE-IIS-9711532 to Z. O. and A. K. D. is gratefully acknowledged. We also thank Dr. R. Drossu, whose neural network simulator was used in the development of the NNPs.

References

[1] Alber, T., Gilbert, W.A., Ponzi, D.R., and Petsko, G.A., The role of mobility in the substrate binding and catalytic machinery of enzymes, *Ciba Found. Symp.*, 93:4–24, 1982.

[2] Anfinsen, C.B., Principles that govern the folding of protein chains, *Science*, 181(96):223–230, 1973.

[3] Bairoch, A. and Apweiler, R., The SWISS-PROT protein sequence data bank and its new supplement TREMBL, *Nucleic Acids Res.*, 24(1):21–25, 1996.

[4] Bennet, W.S. and Hubber, R., Structural and Functional Aspects of Domain Motions in Proteins, *CRC Critical Reviews on Biochemistry*, 15(4):291–369, 1984.

[5] Cho, H.S., Liu, C.W., Damberger, F.F., Pelton, J.G., Nelson, H.C., and Wemmer, D.E., Yeast heat shock transcription factor N-terminal activation domains are unstructured as probed by heteronuclear NMR spectroscopy, *Protein Sci.*, 5(2):262–269, 1996.

[6] Daughdrill, G.W., Chadsey, M.S., Karlinsey, J.E., Hughes, K.T., and Dahlquist, F.W., The C-terminal half of the anti-sigma factor, FlgM, becomes structured when bound to its target, sigma 28, *Nat. Struct. Biol.*, 4(4):285–291, 1997.

[7] Dunker, A., Obradovic, Z., Romero, P., Kissinger, C., and Villafranca, E., On the importance of being disordered, *PDB Newsletter*, 81:3–5, 1997.

[8] Dunker, A.K., Garner, E., Guilliot, S., Romero, P., Albrecht, K., Hart, J., Obradovic, Z., Kissinger, C., et al., Protein disorder and the evolution of molecular recognition: theory, predictions and observations, *Pacific Symposium on Biocomputing*, 3:471–782, 1998.

[9] Fehlhammer, H. and Bode, W., The refined crystal structure of bovine beta-trypsin at 1.8 A resolution. I. Crystallization, data collection and application of patterson search technique, *J. Mol. Biol.*, 98(4):683–692, 1975.

[10] Fetrow, J.S. and Skolnick, J., Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases, *J. Mol. Biol.*, 281(5):949–968, 1998.

[11] Fletcher, C.M., McGuire, A.M., Gingras, A.C., Li, H., Matsuo, H., Sonenberg, N., and Wagner, G., 4E binding proteins inhibit the translation factor eIF4E without folded structure, *Biochemistry*, 37(1):9–15, 1998.

[12] Fletcher, C.M. and Wagner, G., The interaction of eIF4E with 4E-BP1 is an induced fit to a completely disordered protein, *Protein Sci.*, 7(7):1639–1642, 1998.

[13] Garner, E., Cannon, P., Romero, P., Obradovic, Z., and Dunker, A., Predicting disordered regions from amino acid sequence: common theme despite differing structural characterization, *Genome Informatics*, 9:201–214, 1998.

[14] Gerstein, M., How representative are the known structures of the proteins in a complete genome? A comprehensive structural census, *Fold Des.*, 3(6):497–512, 1998.

[15] Hagerman, P.J.I., From sequence to structure to function, *Curr. Opin. Struct. Biol.*, 6(3):277–280, 1996.

[16] Huth, J.R., Bewley, C.A., Nissen, M.S., Evans, J.N., Reeves, R., Gronenborn, A.M., and Clore, G.M., The solution structure of an HMG-I(Y)-DNA complex defines a new architectural minor groove binding motif, *Nat. Struct. Biol.*, 4(8):657–665, 1997.

[17] Koonin, E.V., Tatusov, R.L., and Galperin, M.Y., Beyond complete genomes: from sequence to structure and function, *Curr. Opin. Struct. Biol.*, 8(3):355–363, 1998.

[18] Kriwacki, R.W., Hengst, L., Tennant, L., Reed, S.I., and Wright, P.E., Structural studies of p21Waf1/Cip1/Sdi1 in the free and Cdk2-bound state: conformational disorder mediates binding diversity, *Proc. Natl. Acad. Sci. USA*, 93(21):11504–11509, 1996.

[19] Kyte, J. and Doolittle, R.F., A simple method for displaying the hydropathic character of a protein, *J. Mol. Biol.*, 157(1):105–132, 1982.

[20] May, A.C., Johnson, M.S., Rufino, S.D., Wako, H., Zhu, Z.Y., Sowdhamini, R., Srinivasan, N., Rodionov, M.A., et al., The recognition of protein structure and function from sequence: adding value to genome data, *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, 344(1310):373–381, 1994.

[21] Mimori-Kiyosue, Y., Vonderviszt, F., and Namba, K., Locations of terminal segments of flagellin in the filament structure and their roles in polymerization and polymorphism, *J. Mol. Biol.*, 270(2):222–237, 1997.

[22] Mogridge, J., Legault, P., Li, J., Van Oene, M.D., Kay, L.E., and Greenblatt, J., Independent ligand-induced folding of the RNA-binding domain and two functionally distinct antitermination regions in the phage lambda N protein, *Mol. Cell*, 1(2):265–275, 1998.

[23] Newman, M., Strzelecka, T., Dorner, L.F., Schildkraut, I., and Aggarwal, A.K., Structure of Bam HI endonuclease bound to DNA: partial folding and unfolding on DNA binding, *Science*, 269(5224):656–663, 1995.

[24] Otwinowski, Z., Schevitz, R.W., Zhang, R.G., Lawson, C.L., Joachimiak, A., Marmorstein, R.Q., Luisi, B.F., and Sigler, P.B., Crystal structure of trp repressor/operator complex at atomic resolution, *Nature*, 335(6188):321–329, 1988.

[25] Pattabiraman, N., Namboodiri, K., Lowrey, A., and Gaber, B.P., NRL-3D: a sequence-structure database derived from the protein data bank (PDB) and searchable within the PIR environment, *Protein Seq. Data Anal.*, 3(5):387–405, 1990.

[26] Plaxco, K.W. and Gross, M., Cell biology. The importance of being unfolded, *Nature*, 386(6626):657, 659, 1997.

[27] Pontius, B.W., Close encounters: why unstructured, polymeric domains can increase rates of specific macromolecular association, *Trends Biochem. Sci.*, 18(5):181–186, 1993.

[28] Rader, S.D. and Agard, D.A., Conformational substates in enzyme mechanism: the 120 K structure of alpha-lytic protease at 1.5Åresolution, *Protein Science*, 6:1375–1386, 1997.

[29] Reeves, R. and Nissen, M.S., Interaction of high mobility group-I (Y) nonhistone proteins with nucleosome core particles, *J. Biol. Chem.*, 268(28):21137–21146, 1993.

[30] Romero, P., Obradovic, Z., and Dunker, A.K., Sequence data analysis for long disordered regions prediction in the calcineurin family, *Genome Informatics*, 8:110–124, 1997.

[31] Romero, P., Obradovic, Z., Kissinger, C.R., Villafranca, J.E., and Dunker, A.K., Identifying disordered regions in proteins from amino acid sequences, *Proc. IEEE International Conference on Neural Networks*, 1:90–95, 1997.

[32] Romero, P., Obradovic, Z., Kissinger, C.R., Villafranca, J.E., Guilliot, S., Garner, E., and Dunker, A.K., Thousands of proteins likely to have long disordered regions, *Pacific Symposium on Biocomputing* 3:435–446, 1998.

[33] Romero, P.Z., Obradovic, C., and Dunker, A.K., Intelligent data analysis for protein disorder prediction, *Arificial Intelligence Reviews*, (in press).

[34] Schulz, G.E., Nucleotide Binding Proteins, *Molecular Mechanism of Biological Recognition*, Elsevier/North-Holland Biomedical Press, 79–94, 1979.

[35] Spolar, R.S. and Record II, M.T., Coupling of local folding to site-specific binding of proteins to DNA, *Science*, 263:777–784, 1994.

[36] Tell, G., Perrone, L., Fabbro, D., Pellizzari, L., Pucillo, C., De Felice, M., Acquaviva, R., Formisano, S., et al., Structural and functional properties of the N transcriptional activation domain of thyroid transcription factor-1: similarities with the acidic activation domains, *Biochem. J.*, 329(Pt 2):395–403, 1998.

[37] Vihinen, M., Torkkila, E. and Riikonen, P. Accuracy of Protein Flexibility Predictions, *Proteins: Structure, Function, and Genetics*, 19:141–149, 1994.

[38] Vonderviszt, F., Uedaira, H., Kidokoro, S., Namba, K., Structural organization of flagellin, *J. Mol. Biol.*, 214(1):97–104, 1990.

[39] Wright, P.E. and Dyson, H.J., Intrinsically Unstructured Proteins: Re-assesing the Protein Structure-Function Paradigm, *J. Mol. Biol.*, 293(2):321–331, 1999.

[40] Xie, Q., Arnold, G.E., Romero, P., Obradovic, Z., Garner, E., Dunker, A.K., The sequence attribute method for determining relationships between sequence and protein disorder, *Genome Informatics*, 9:193–200, 1998.