To appear in the Proc. IEEE International Conference on Neural Networks, Huston, TX, June 1997

Identifying Disordered Regions in Proteins from Amino Acid Sequence[1]

P. Romero[2], Z. Obradović[2], C. Kissinger[4], J. E. Villafranca[4], and A. K. Dunker[3]

[2]School of Electrical Engineering and Computer Science
[3]Department of Biochemistry and Biophysics
Washington State University, Pullman, Washington, 99164-2752
and
[4]Agouron Pharmaceuticals, Inc.
3565 General Atomics Ct., San Diego, CA 92121-1221

Abstract

A rule-based and several neural network predictors are developed for identifying disordered regions in proteins. The rule-based predictor was suitable only for very long disordered regions, whereas the neural network predictors were developed separately for short-, medium-, and long-disordered regions (S-, M-, and LDRs, respectively). The out-of-sample prediction accuracies on a residue-by-residue basis ranged from 69 to 74% for the neural network predictors when applied to the same length class, but fell to 59 to 67% when applied to different length classes. Application of both the rule-based and LDR neural network predictors to large databases of protein sequences provide strong evidence that disordered regions are very common in nature. These results are consistent with our recent proposal that disordered regions are crucial for the evolution of molecular recognition.

1 Introduction

X-ray diffraction from protein crystals can yield the structure of these molecules. Many proteins are nonuniform, having both structured and disordered regions. When crystallized, the structured regions scatter x-rays coherently and so are observed. The disordered regions, however, fail to crystallize into fixed structures and so scatter x-rays incoherently. These disordered regions are therefore invisible in the resulting electron density maps [8].

We have so far identified more than 15 proteins containing disordered regions that become ordered upon formation of a molecular complex with a partner. Homology searches extend these to hundreds of examples. These include the following types of molecular interactions: enzyme/substrate, receptor/ligand, protein/protein, protein/RNA and protein/DNA. Thus, disorder-to-order transitions upon binding span the biological domain [4]. Schulz proposed that loss of disorder upon binding leads to a biologically desirable pair of features, namely: high specificity coupled with modest binding affinities [13]. Affinities that are too high result in essentially irreversible binding, which is unsuitable for most biological processes.

For proteins whose x-ray structures are known, the existence of disordered stretches can be identified directly by looking for amino acids that are missing from the electron density maps. However, x-ray structures are known for just a small fraction of the set of proteins with known amino acid sequences. Consequently, to estimate the occurance of proteins with disordered regions in nature, alternative information-based approaches have to be taken. The approach considered in this study is to develop predictors that can identify disordered regions in proteins. It is well-established that amino acid sequence determines protein 3-D structure [2]; here we assume that amino acid sequence determines lack of fixed 3-D structure (e.g. *disorder*) as well.

Design of a rule-based predictor is summarized in Section 2. Construction of a labeled data set of proteins with disordered regions and its use to develop neural network predictors is discussed in Section 3. Finally, both approaches were tested on several databases as reported in Section 4.

[1]Corresponding E-mail: zoran@eecs.wsu.edu

2 Rule-Based Approach

Studies on calcineurin (CaN) [9] sparked the present work; we noticed that the long disordered region (LDR) in this protein has a low content of aromatic amino acids (Trp, Tyr, Phe). Several other disordered regions were found to have this same characteristic. This makes structural sense because the side chains of aromatic amino acids have strong and specific interactions [3] and so would be expected to induce structure and inhibit disorder. Using CaN as the prototype, the average fraction of aromatic amino acids was calculated over a window of 31 amino acids surrounding each sequence position. The aromatic content dropped significantly in both of the longer unobserved regions in CaN (Fig 1).

The following prediction rule was developed from these observations: (a) for a given protein, the average content of aromatic residues is calculated throughout the amino acid sequence, as explained above; (b) if there is a contiguous region of more than 80 sequence positions with an average content of aromatic residues below 6.5%, the protein is predicted to have an LDR. This predictor was intended for LDRs like the one in CaN; because of the large window sizes for this predictor, it is not suitable for predicting M- or SDRs.

3 Neural Network Approach

The rule-based predictor discussed in the previous section was developed based on information from a single protein, CaN, which served as the prototype for our studies. An alternative is to design feedforward neural network predictor trained using the backpropagation learning algorithm [15]. This predictor requires construction of a larger set of examples of disordered regions (DRs) and determination of appropriate features, as discussed in this section.

3.1 Disordered Regions Labeled Data Sets

A search for proteins with invisible regions, which are presumed to be locally disordered, was performed on the Protein Data Bank (PDB) at the Brookhaven National Laboratory. This is a public domain archive of more than 4,600 experimentally determined three-dimensional structures of proteins.

Searching PDB for proteins with DRs is a non-trivial task since no standard format for reporting such findings is imposed. In addition, several other problems like complexes and repeated sequences further complicated this search. In this study, no effort was made to identify all proteins with DRs in the PDB. The main objective was to find a sufficiently large set of proteins with confirmed DRs as needed for the design of a neural network predictor.

The PDB search supplied a set of proteins each having at least one DR longer than seven residues. These proteins from PDB were supplemented by two additional proteins with DRs (CaN [9] and Bcl [12]). A histogram of the lengths of the DRs suggested a partition into short, medium and long labeled data sets, denoted as SDR (7-21 amino acids), MDR (22-44 amino acids), and LDR (45 or more amino acids), respectively.

3.2 Feature Selection

The LDR labeled data set was analyzed to identify a pool of properties that discriminate between structured and disordered regions. The exploratory analysis considered several attributes measured by averaging over windows of consecutive amino acids. Considered attributes included individual amino-acid compositions, flexibility [14], hydropathy [11] and hydrophobic moments [5].

In addition to the lack of aromatics (in this case just Tyr and Trp) mentioned above, low amounts of Cys and His and high amounts of Glu, Asp, Ser, and Lys were also found to be associated with disorder. Cys can make special covalent bounds, so its absence in disordered regions is reasonable. Glu, Asp, and Lys are charged; charge imbalance would be expected to contribute to disorder. Ser increases solubility and provides a flexible locus. Finally, disordered regions would be expected to be soluble and flexible in a manner consistent with the findings on hydrophathy and flexibility. Thus, overall, the identified attributes seem reasonable for promoting disorder.

A set of m attributes was identified through this analysis as being more discriminative. For each identified attribute, n different features were generated by computing this attribute for n different window sizes, yielding

an $m \times n$ matrix of features, where each row corresponds to a different attribute and each column to a different window size.

A formal procedure was used to select the most appropriate feature from each row of this matrix. The method used here is an adaptation of the sequential forward search feature selection technique with the minimal error probability selection criterion [7]. A quadratic Gaussian classifier using different covariance matrices for each class was used to calculate minimal error probability during the search.

The standard sequential forward search selection technique is a greedy algorithm that begins with an empty feature set and adds features to it one at a time. The first feature added is the one deemed to be the best according to the selection criteria. The next feature added is the one which results in the largest improvement when considered in conjunction with the first feature. Similarly, the i-th feature added is the one that results in the largest improvement when considered in conjunction with the previous $i-1$ features.

In the method used here, when the i-th feature is added to the selected features set, its corresponding row is removed from the matrix and the search continues on the reduced $(m-i) \times n$ matrix. This prevents the same attribute from being selected with more than one window size; the resulting selected feature set contains the most appropriate window size for each attribute. A set of examples for neural network training was constructed from the LDR labeled data set using the m selected features. For convenience the same features set was used when training on examples from the SDR and MDR labeled data sets.

4 Results

The SDR labeled data set contained 38 disordered segments from 34 proteins with 411 disordered amino acids and 11,050 total amino acids; MDR set contained 22 disordered segments from 20 proteins with 464 disordered amino acids and 4,764 total amino acids; and finally, LDR set contained 7 regions from 7 proteins with 465 amino acids and 2,069 total amino acids. The result of the exploratory analysis on 24 considered attributes was the selection of 10, shown in Table 1. Shown here are the most appropriate windows for each of these attributes obtained through the feature selection process discussed in Section 3.2 by exploring odd-numbered values ranging from 9 to 21.

4.1 Prediction Accuracy Estimates

The rule-based predictor was designed using the CaN knowledge. When tested on a residue-by-residue prediction on the remaining 6 LDR proteins, it achieved 70% success rate. This result is surprisingly good for the simplicity of the rule and suggests that lack of aromatic amino acids is a strong determinant for the development of LDRs.

Balanced sets of the 10 dimensional feature values corresponding to the unobserved and observed amino acids were constructed from the S-, M-, and LDR labeled data sets. These feature sets were each randomly partitioned into 5 disjoint balanced subsets from observed and unobserved amino acids. A neural network architecture was determined through limited experimentation and a machine with 10 inputs, one hidden layer with 6 units, and a single output unit was used for in depth testing. 5-cross validation experiments starting from 3 different random initializations of neural network parameters were performed, resulting in a total of 15 runs each for the S-, M- and LDR labeled data sets (Table 2, rows a-e).

Averaging the results from the 5-cross validation experiments gave residue-by-residue prediction accuracies ranging from 66 to 74%. Lumping all length-classes together (ADR) and repeating the 5-cross validation experiment led to a drop in prediction accuracy, to about 60%. Finally, when testing each predictor on data sets of other length classes, the prediction accuracies dropped to 59-67%. The greatest drop was observed

Selected	composition								average	
Attribute	His	Glu	Lys	Ser	Asp	Cys	Trp	Tyr	Hydropathy	Flexibility
Window Size	9	9	9	9	13	21	21	21	9	15

Table 1: Selected Features

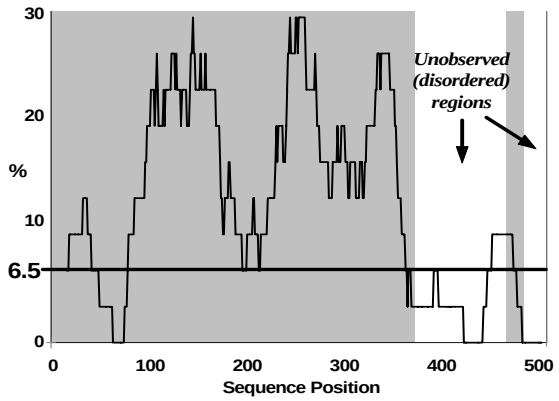

Figure 1: Fraction of aromatic residues on CaN

Test Set	Out of Sample Accuracy			
	SDR	MDR	LDR	ADR
a	71%	75%	70%	60%
b	71%	74%	77%	62%
c	69%	76%	77%	57%
d	67%	72%	72%	61%
e	66%	75%	70%	60%
Ave	69%	74%	73%	60%
SDR	-	63%	59%	-
MDR	63%	-	67%	-
LDR	59%	61%	-	-

Table 2: NN generalization results (5-cross validation averaged over 3 initializations)

for the LDR predictor applied to the SDR labeled data set or vice versa (both to 59%). The smallest drop was observed for the LDR predictor applied to the MDR labeled data set (67%).

4.2 Prediction of Disordered Regions in Proteins

The rule-based predictor produces binary outputs that are directly applicable for observability predictions for stretches of proteins due to the fact that the aromatic residues' content changes slowly when a window of 31 residues slides from one sequence position to the next. This is not true for the neural network predictor whose outputs are real numbers in (0,1) range that can vary significantly between adjacent positions. Consequently, neural network outputs are averaged over a window of 9 neighboring residues in order to smooth the signal as necessary for predicting disordered regions. Using a 0.5 threshold this signal is discretized to 0 or 1 corresponding to structured and disordered residue prediction, and the resulting binary signal is checked for regions of contiguous prediction of unobserved residues.

Estimating false positive error rates for the predictors is key to determining whether disordered regions are common. The presence of undetermined false negative error rates means only that our estimates represent lower bounds for the numbers of DRs. An estimate of false positive error rates was accomplished using the NRL_3D database, which contains the sequences of the observed parts of the proteins in PDB. Consequently, a perfect predictor should not accept any protein from NRL_3D.

The LDR predictor was applied to four data bases: NRL_3D (to estimate false positives), PDB (known to contain proteins with disordered regions) and also to two large sequence-only databases, SwissProt (SW) and the Protein Identification Resource (PIR). SW and PIR contain 59,031 and 89,926 sequences, respectively. Figure 2 and its accompanying Table 3 show the percentage of sequences predicted to have at least one predicted unobserved stretch longer than t. Due to the lower accuracy of the LDR predictor in the SDR region, data for $t < 20$ should be ignored.

The curve for the predicted DR rates in PDB is shifted to the right of the curve for NRL_3D. Examination of a few specific proteins indicates that this right-ward shift is due mostly to correct predictions.

From 20 to about 35 amino acids, the predicted rates of DRs in PDB, SW, and PIR are very similar. In contrast, the predicted DR rates in SW and PIR diverge from the predicted rates in PDB for disordered segments longer than 35. Disordered molecules rarely form crystals. Thus, this divergence of PDB from SW and PIR probably results from the inhibition of protein crystallization when LDRs are present, thus leading to a bias in PDB against proteins with LDRs.

The frequencies of predicted DRs in SW and PIR are well above estimates of false positives from the NRL_3D curves for all length classes, but especially in the LDR region. For example, for predicted DRs of 80 or longer, the LDR-based NN predictor gave 0.3% false positives error rates on NRL_3D (rounded to 0% in Table 2), whereas it gave 7% on SW and 6% on PIR. Application of the rule-based predictor to the same problem gave 3% false positives and 17% for SW and 14% for PIR. Perhaps our current implementation of the LDR neural network predictor is overly stringent and misses substantial numbers of LDRs. Overall,

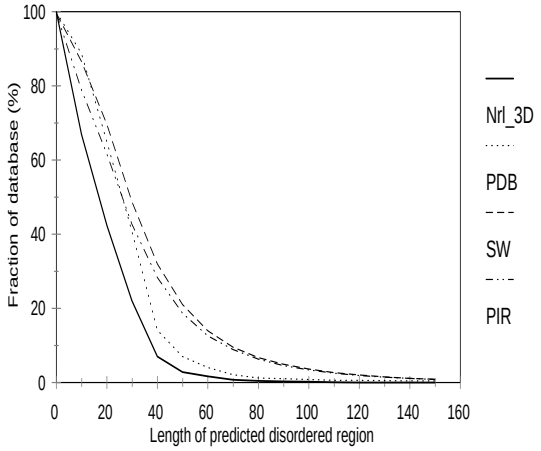

Figure 2: Comparison of LDR-based NN prediction on 4 major databases

Predicted DR longer than:	Fraction of proteins over whole database			
	NRL_3D	PDB	SW	PIR
20	42%	65%	70%	62%
40	7%	14%	32%	28%
60	2%	4%	14%	13%
80	0%	1%	7%	6%
100	0%	1%	4%	3%
120	0%	1%	2%	2%
140	0%	0%	1%	1%
160	0%	0%	1%	1%
180	0%	0%	0%	0%

Table 3: Fraction of proteins predicted to have DR longer than specific values.

these observations suggest that DRs, even LDRs, are very common in nature.

5 Conclusions

The LDR, MDR and SDR predictors were significantly more accurate when applied to the same length class. Even the process of grouping all disordered amino acids together reduced the prediction accuracy substantially (Table 2). These results strongly suggest that amino acid sequence characteristics leading to disorder are dependent on the length of the disordered region.

The current views of protein structure and function still seem to be dominated by the concepts of rigid organization and lock-and-key interactions [6], despite many examples of disorder-to-order transitions upon binding. As we point out, disorder-to-order transitions upon binding have been found for a diversity of molecular interactions that span the biological domain [4].

Koshland's induced-fit hypothesis introduced flexibility as an alternative to the lock and key [10]. Without reference to induced fit, Schulz [13] pointed out that the increases in free energy when flexible components solidify upon binding enable high specificity without excessively high affinities. Petsko and his collaborators [1] independently showed that loss of flexibility could help prevent excessively tight binding, but failed to note the coequally important feature of trading flexibility for specificity. We recently extended Schulz's proposal to show that disorder-to-order transitions allow natural selection to operate separately on affinity and specificity. We propose that such a separation is essential for the evolution of complex signaling and metabolic networks [4].

If disordered regions are required for the separation of affinity and specificity as we propose, then such regions should be very common. The commonness of such regions is fully supported by the findings presented herein. The next steps will be to determine whether the regions predicted to be disordered do indeed carry binding function and to determine whether the predictions are correct. Studies in these directions are underway.

Here we assume that all invisible regions in the x-ray structures are equivalent; however, three different causes have been identified for such regions, including crystal packing disorder, static disorder, and dynamic disorder [8]. Only the last of these involves the local disorder required by Schulz's proposal, so lumping all invisible regions together as we have done may be inappropriate. This is an acknowledeged weakness of the present study and might be adding noise to the data. We plan to improve our labeled data sets by distinguishing the 3 types of disorder for as many entries as possible, using either literature-based investigations or laboratory-based experiments.

Due to the curse of dimensionality and the small sizes of the DR labeled data sets, our current neural network predictors were very simple and limited to just a few features. Yet fairly good predictions are evidently being made despite this limitation. Increasing the sizes of DR labeled data sets and repeating the feature selection process for each length class will enable us to test a larger pool of candidate features and to design more complex predictors.

The more complex predictors will, hopefully, give more accurate predictions. This is especially important for short disordered segments since none of the current predictors do very well in this region. Since SDRs are found frequently to be involved in disorder-to-order transitions upon binding [4], improved predictions in this domain are certainly important.

Acknowledgments

Susan Johns and Steve Thompson of the Center for Visualization, Analysis and Design in the Molecular Sciences at WSU are acknowledged for their help with the various molecular biology data bases and with the use of the GCG package for calculating some sequence features. This work was sponsored in part by a grant from the American Heart Association, Washington State Chapter awarded to A. K. Dunker and Chul-Hee Kang. Molecular Kinetics is thanked for providing additional support.

References

[1] Alber, T. et al. [1982] "The role of mobility in the substrate binding and catalytic machinery of enzymes," *Mobility and function in proteins and nucleic acids,* Pitman, London, pp. 4-24.

[2] Anfinsen, C. B. [1973] "Principles that govern the folding of protein chains," *Science* vol. 181, pp. 223-230.

[3] Burley, S. K. and Petsko, G. A. [1985] "Aromatic-aromatic interaction: A mechanism of protein structure stabilization," *Science,* vol. 229, pp. 23-28.

[4] Dunker, A. K., Romero, P., Obradovic, Z., Kissinger, C. R., and Villafranca, J. E. [in preparation] "Role of protein disorder in the evolution of molecular recognition."

[5] Eisenberg, D., Weiss, R.M., and Terwilliger, T.C. [1982] "The helical hydrophobic moment: a measure of the amphiphilicity of a helix," *Nature,* vol. 299, pp. 371-374.

[6] Fischer, E. [1894] "Einfluss der configuration auf die wirkung derenzyme," *Ber. Dt. Chem. Ges.* vol. 27, pp. 2985-2993.

[7] Fukunaga, K. [1990] *Introduction to statistical pattern recognition,* Academic Press, San Diego, CA.

[8] Huber, R. [1979] "Conformational flexibility and its functional significance in some protein molecules," *TIBS,* vol. 4, pp. 271-276.

[9] Kissinger, C. R. et al. [1995] "Crystal structures of human calcineurin and the human FKBP12-FK506-calcineurin complex," *Nature,* vol. 378, pp. 641-644.

[10] Koshland, D. E. [1958] "Application of a theory of enzyme specificity to protein synthesis," *Proc. Nat'l. Acad. Sci. USA* vol. 44, pp. 98-104.

[11] Kyte, J. and Doolittle, R. F. [1982] "A simple method for displaying the hydropathic character of a protein," *J. Mol. Biol,* vol. 157, pp. 105-132.

[12] Muchmore, S. W. et al. [1996] "X-ray and NMR structure of human Bcl-xl, an inhibitor of programmed cell death," *Nature,* vol. 381, pp. 335-341.

[13] Schulz, G. E. [1979] "Nucleotide binding proteins," *Molecular Mechanism of Biological Recognition,* Elsevier/North-Holland Biomedical Press, pp. 79-94.

[14] Vihinen, M., Torkkila, E. and Riikonen, P. [1994] "Accuracy of Protein Flexibility Predictions," *Proteins: Structure, Function, and Genetics,* vol. 19, 1994, pp. 141-149.

[15] Werbos, P. [1974] "Beyond regression: New tools for predicting and analysis in the behavioral sciences," Harvard University, Ph.D. Thesis. Reprinted by Wiley and Sons, 1995.