# THOUSANDS OF PROTEINS LIKELY TO HAVE LONG DISORDERED REGIONS

PEDRO ROMERO, ZORAN OBRADOVIC
*School of Electrical Engineering and Computer Science*
*Washington State University, Pullman, WA 99164*

CHARLES R. KISSINGER, J. ERNEST VILLAFRANCA
*Agouron Pharmaceutical, Inc., 3565 General Atomics Court, San Diego, CA 92121-1221*

ETHAN GARNER, STEPHEN GUILLIOT, A. KEITH DUNKER
*Department of Biochemistry & Biophysics*
*Washington State University, Pullman, WA 99164*
*e-mail: dunker@mail.wsu.edu*

Neural network predictors of protein disorder using primary sequence information were developed and applied to the Swiss Protein Database. More than 15,000 proteins were predicted to contain disordered regions of at least 40 consecutive amino acids, with more than 1,000 having especially high scores indicating disorder. These results support proposals that consideration of structure-activity relationships in proteins need to be broadened to include unfolded or disordered protein.

## 1. Introduction

Calcineurin (CaN) is a calcium / calmodulin (CaM) regulated serine / threonine phosphatase[1]. Using x-ray diffraction, two of us (CRK and JEV) along with additional collaborators determined the structure of human CaN[2]. This protein contains three unobserved or disordered[3] regions, the longest of which spanned 95 consecutive amino acids[2]. This longest disordered region contains the CaM binding site. This site had previously been shown to be disordered by its sensitivity to protease and to become resistant to protease upon CaM binding[4]. Given the central importance of the CaM binding site as the nexus between the calcium and phosphorylation / dephosphorylation signaling pathways, we wondered whether other proteins exhibited functional regions that were disordered.

Literature searches revealed many reports of x-ray diffraction experiments, some about 20 years old, describing disordered regions that become involved in binding. Selected examples include triose phosphate isomerase binding with triose phosphate[5], avidin with biotin[6], the S-peptide with RNAse S[7], myosin with actin[8], tobacco mosaic virus (TMV) coat protein with its RNA[9], tyrosyl tRNA synthetase with its tRNA[10], and the trp repressor[11], the lac repressor[12] and Bam H1[13] with their respective DNAs.

NMR has also been used to characterize disordered proteins. Involvement of disorder in the flagella-assembly related FlgM protein binding to the transcriptional activator, $\sigma^{28}$ [14] and cyclin-dependent kinase (Cdk) inhibitor, p21$^{Waf1/Cip1/Sdi1}$, binding to Cdk[15] are two especially interesting examples[16].

Given these many examples of functional disordered proteins, we have initiated studies on the relationship between sequence and disorder by testing whether neural network predictors (NNPs) can identify putative regions of disorder in proteins from their amino acid sequences[17]. Two further applications of these NNPs are presented here: (1) our overall predictions on a protein sequence database to estimate the commonness of disordered regions; and (2) the identification of a relatively small number of proteins that have especially strong predictions of disorder.

## 2. Materials and Methods

The databases used in these studies were SwissProtein (SW)[18], Nrl_3D[19], and the Protein Data Bank (PDB)[20]. The NNPs are described in more detail elsewhere[17]. To develop a training set of amino acids, 7 long disordered regions (LDRs) ranging in length from about 48 to about 100 were identified in the PDB entries. These 7 LDRs came from DNA topoisomerase II , elongation factor G, lactose operon repressor, tomato bushy stunt virus coat, tyrosyl-tRNA synthetase, apoptosis regulator Bcl-X, and calcineurin. The total number of amino acids in this training set was 930, half of them disordered. Another training set was developed using sequence data from a group of 13 homologous CaN molecules. Although the 3D structure of the 12 homologous CaN molecules was not known, the fact that highly similar proteins have similar 3D structures led to the assumption that those CaN molecules have the same disordered regions as the original CaN studied. To locate the disordered regions on the new CaN sequences, a multiple sequence alignment was performed using GCG[21], and all residues that aligned with disordered positions in the original CaN sequence were considered disordered. This procedure generated enough data (3,332 residues) for the training of a neural network-based, CaN-specific LDR predictor.

A feature selection process identified the 10 best attributes (or *features*) to be used as inputs to the predictor, as follows: information from statistical comparisons of ordered and disordered regions was used to pre-select 15 sequence-dependent attributes out of 22 considered. These attributes were then used as inputs for a modified[17] sequential forward search feature selection technique. A quadratic Gaussian classifier using different covariance matrices for ordered and disordered amino acids was used to calculate the minimal error probability during the search. Starting with a null set, features were selected incrementally, with the ($i$ + 1)th

feature being the one that most improved the discrimination between ordered and disordered amino acids when added to the i previously selected features. The 10 selected features include contents of W, Y, C, S, D, E, H, and K, a flexibility index, and hydropathy. This feature set was used in the development of neural network predictors (NNPs) for both the original LDR and the CaN-specific LDR data sets.

Balanced sets of the 10 dimensional feature values corresponding to ordered and disordered amino acids were used to carry out backpropagation-based supervised training[22] of feedforward neural networks having 10 inputs, one hidden layer with 6 units and a single output unit. The outputs of these NNPs were real numbers normalized to fall between 0 and 1, with values below a certain *prediction threshold* (*q*) indicating order and values above it indicating disorder[17]. Thus, a prediction threshold *q* of, say, 0.5 would mean that an NNP's output greater or equal to 0.5 corresponds to a prediction of disorder, and, conversely, an output smaller than 0.5 indicates order. For predicting regions of disorder, the amino acid residues are first predicted one-by-one to be ordered or disordered. Next, these predictions $\{\cdots, p_{i-1}, p_i, p_{i+1}, \cdots\}$ are smoothed by averaging over a window of nine to obtain $\{\cdots, s_{i-1}, s_i, s_{i+1}, \cdots\}$ where $s_i = (p_{i-4} + \cdots + p_{i+4})/9$. A region of length *m* is predicted to be disordered if m consecutive smoothed predictions exceed the specified prediction threshold, *q*, which was 0.5 in our initial work[17]. So, the region defined by sequence positions *i*+1 to *i*+*m* is predicted to be disordered if $s_j > q$ for $i < j \le i + m$.

## 3. Results

### *3.1. Predicting Disorder from Sequence Using Neural Networks*

If a protein structure has evolved to have a functional disordered state then a propensity for disorder might be predictable from its amino acid sequence and composition. To test this, we developed labeled data sets of disordered regions and used these to carry out supervised training of neural network predictors (NNPs). The 0 and 1 labels were used during the training to inform the NNPs which sequence positions belong to ordered and which to disordered regions.

The LDR NNP used here gave a cross-validated, out-of-sample, residue-by-residue prediction accuracy of 73 ± 4%, as described in more detail previously[17]; with exactly the same procedures, the CaN-based NNP exhibited an accuracy of 72 ± 5%. These results are comparable to those for predicting secondary structure[23], although we need to mention that our predictions are for two states –order or disorder– whereas the secondary structure predictions are for three states –helix, sheet, or other. On the other hand, our current NNPs are signficantly better than

secondary structure prediction methods that use only single sequence information. Adding multiple sequence alignment information as inputs to our NNPs will probably lead to a significant improvement just as for secondary structure prediction[23]. A further comparison between secondary structure and disorder prediction is that disordered regions can be much longer, and such LDRs should be easier to predict than the shorter segments of secondary structure. For this reason, we are initially focusing on those LDRs[17].

*3.2. Estimating the Commonness of Long Disordered Regions.*

To obtain an overall estimate of the commonness of LDRs in nature, one would apply the predictor to a sequence database, subtract an estimate *of false positive error* (e.g. prediction of disordered where the structure is actually ordered) and add an estimate of the *false negative error* (e.g. prediction of order where the structure is actually disordered). A summary of the LDR predictions on the SwissProtein[18] (SW) database, along with estimates of the false positive error rates, are shown in Table 1.

**Table 1: Estimated Numbers of Proteins with Long Disordered Regions.** The databases studied were Swiss Protein (SW) and the Naval Research Laboratory database, NRL_3D. The percentages of sequences estimated to have at least one prediction of a disordered segment of 40 or longer using the two different indicated predictors, with the prediction threshold *q* set at 0.5, is given in column 4, labeled as "% > 40". The predictor labeled "CaN or LDR NNP" accepts a protein as having a long disordered region if either the LDR NNP or the CaN NNP indicate it. The predictions on NRL_3D were taken to be measures of the false positive error rates for the two predictors, which were therefore subtracted from the corresponding estimates to give the net percentages of sequences predicted to contain long disordered regions as indicated in column 5, labeled as "Net". Finally, an estimate of the number of sequences in SW with disordered regions of 40 or longer were estimated by multiplying the net fractions times the total number of sequences.

| **Database** | **Number of Sequences** | **Predictor** | **% > 40** | **Net** | **Number Disordered** |
|---|---|---|---|---|---|
| SW | 59,021 | LDR NNP | 32% | 25% | 14,760 |
| | | CaN NNP | 14% | 12% | 7,080 |
| | | CaN or LDR NNP | 34% | 26% | 15,350 |
| NRL 3D | 6,063 | LDR NNP | 7% | — | — |
| | | CaN NNP | 2% | — | — |
| | | CaN or LDR NNP | 8% | — | — |

To estimate the false positive error rates, LDR NNP and CaN-based NNP were applied to the Naval Research Laboratory NRL_3D database[19], which contains only the ordered regions from a set of proteins of known 3D structure. These regions of ordered structure were taken from proteins in the PDB[20]. Because all the sequences in NRL_3D are ordered, any prediction of disorder on this database is probably a false positive. It should be kept in mind, however, that the process of crystallization itself has been shown to induce order; for example, a disordered region of the ras protein becomes ordered in some crystals when it is involved in crystal contacts, but not in other crystals and probably not in solution[24]. If such behavior is common, our current false positive error estimates may be too high.

We have not made estimates of false negative error rates so far. Omission of such errors means that the actual number of disordered proteins is even larger than the estimates given in Table 1.

The LDR NNP was trained using disordered regions from 7 different proteins whereas the CaN-based NNP used only calcineurin-type proteins. Thus, the CaN NNP would be expected to be biased for recognizing disordered regions more like those in CaN and so should find a smaller fraction of the disordered regions as compared to the NNP predictor, which is consistent with the data in Table 1.

### *3.3. Making a Few Strong Predictions*

Note that the rates of false positive prediction can be made arbitrarily small by elevating the prediction threshold *q*. In addition, since false positive prediction for a region of length *m* requires misclassification of *m consecutive* residues, it is easy to see that this error decreases when the region length *m* is increased. So, to find

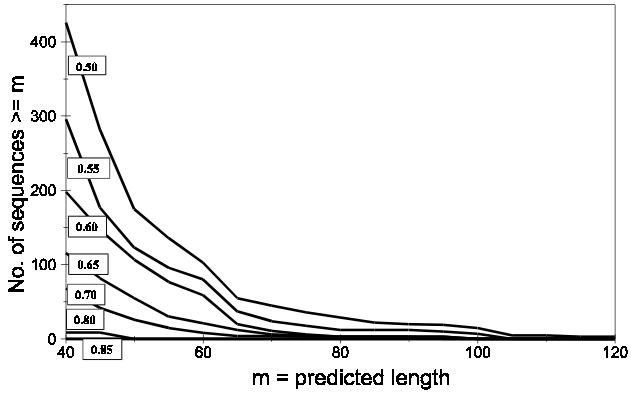


**Figure 1: Number of False Positive Predictions Versus Length for Several Thresholds.** The previously described LDR NNP[17] was repeatedly applied to NRL_3D with successively higher prediction thresholds each time as indicated. The false positive error rate is observed to drop as the prediction length, m, increases or as the prediction threshold, *q*, is elevated.

the strongest predictions with especially low false positive error rates, we studied the relationships among prediction threshold, predicted disordered region length and false positive prediction error rate (Figure 1).

For the prediction threshold *q* set at 0.85, not one sequence of 40 amino acids or longer in NRL_3D gave a false positive error using the LDR NNP and only 4 false positives were observed for the CaN-based predictor. Now, for a database with *i* total amino acids and *j* segments, the total number of prediction regions of length *m* is given by $i - j(k+m)$, where *k* is the number of sequence positions that are not assigned predictions on each segment. Our predictions use a windowing procedure that does not assign a prediction to the first and last 14 amino acids, for a total of $k = 28$ unpredicted positions per segment. Now, our version of NRL_3D has $j = 6{,}063$ segments and $i = 1{,}030{,}628$ amino acids; thus, the total number of predictions for $m = 40$ is equal to $(1{,}030{,}628 - [6{,}063] \times [40 + 28]) = 618{,}344$.

Since our NNP predicts on 618,314 regions of length 40 on NRL_3D, for the LDR NNP with *q* set at 0.85, the false positive error rate is *less than* 1 out of 618,314. For the CaN-based NNP, the false positive error rate is 76 out of 618,314. Applying these error estimates to predictions on SW assumes that the identified sequence features of the structured protein in NRL_3D represents an unbiased sampling of the sequence features for the structured protein in SW.

Winnowing the sequences in the SW database with the LDR NNP having the prediction threshold set at 0.85 left 1,069 proteins strongly predicted to contain disordered regions of 40 or longer; with the CaN-based NNP, 635 sequences remained (Figure 2). From the false positive error rate and the SW database characteristics, the expected number of false positives should be less than 30

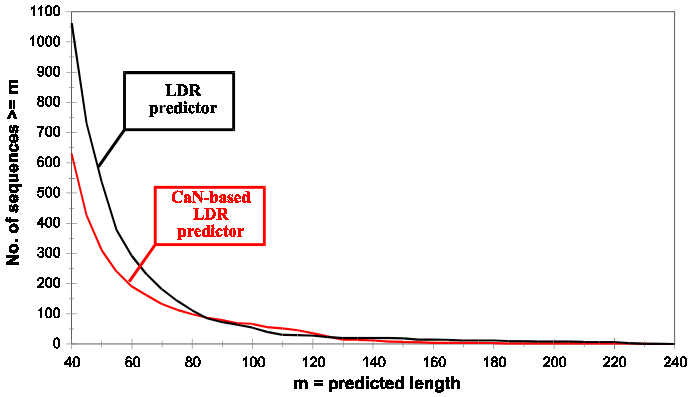


**Figure 2: Sequences Strongly Predicted to Contain LDRs.** The LDR and CaN-based NNPs were applied to the Swiss Protein (SW) database with the prediction threshold *q* set at 0.85. The cumulative number of sequences predicted to contain at least one disordered region of length *m* or longer is plotted versus *m*. A total of 1,069 sequences for the LDR NNP and 635 for the CaN-based NNP are strongly predicted to contain at least one disordered region of 40 or longer.

regions for the LDR NNP and on the order of 2,280 for the CaN-based NNP. Given that a prediction of disorder with a length, $m$, greater than 40 contains $m - 40$ predictions of length 40, the 1,069 and 635 sequences predicted by the LDR and CaN-based NNPs contain large numbers of predictions of length 40; indeed, the length distribution of the predicted LDRs was used to determine that the former set contains 22,595 predictions of length 40 and the latter set 16,007. Thus, the numbers of predicted disordered regions of length 40 are substantially greater than the expected number of false positive predictions.

**Table 2: Examples of Proteins Predicted to Contain Long Disordered Regions:** The 20 longest and/or strongest predictions of disorder for apparently nonhomologous proteins are presented here.

| No | SW ID | Description | Seq. length | Max. LDR | Location | Ave. Strength |
|---|---|---|---|---|---|---|
| 1 | SANT_PLAFW | s-antigen - plasmodium | 640 | 576 | 51-627 | 0.99 |
| 2 | SANT-PLAF7 | s-antigen - plasmodium | 593 | 490 | 68-557 | 0.93 |
| 3 | VG48_HSVSA | hypothetical gene 48 protein from herpes virus saimiri | 797 | 308 | 413-720 | 0.96 |
| 4 | MLH_TETTH | micronuclear linker histone poly protein from tetrahymena | 633 | 278 | 173-450 | 0.96 |
| 5 | CYCLI_HUMAN | human cyclin (fragment) | 598 | 255 | 238-492 | 0.98 |
| 6 | NFH_MOUSE | neurofilament triplet h protein | 1087 | 239 | 523-761 | 0.92 |
| 7 | RTOA_DICDI | slime mold rtoa protein | 400 | 227 | 77-303 | 0.96 |
| 8 | SR75_HUMAN | human pre-mRNA splicing factor, srp75 | 494 | 225 | 186-410 | 0.96 |
| 9 | RPB1_CRIGR | chinese hamster DNA-directed RNA polymerase: largest subunit | 1970 | 221 | 1610-1830 | 0.98 |
| 10 | LSTP_STAST | staphylococcus staphyloyticus lysostaphin precursor | 480 | 168 | 56-223 | 0.92 |
| 11 | YHFI_SALTY | salmonella typhimurium, hypothetical protein (orf3) | 416 | 154 | 57-210 | 0.99 |
| 12 | T2FA_DROME | drosophila melanogaster transcription factor iif, α subunit | 577 | 153 | 246-398 | 0.97 |
| 13 | H1_PEA | garden pea histone H1 | 265 | 152 | 100-252 | 0.97 |
| 14 | 110K_PLAKN | plasmodium knowlesi, 110 kd antigen (fragment) | 296 | 148 | 135-283 | 0.91 |
| 15 | XYNA_RUMFL | bifunctionall endo-1,4 β-xylanase precursor from ruminoccus flavefaciens | 954 | 129 | 248-376 | 0.90 |
| 16 | VIT2_CHICK | vitellogenin ii precursor from chick | 1850 | 125 | 1142-1266 | 0.96 |
| 17 | SNWA_DICDI | snwa protein from slime mold | 685 | 124 | 398-521 | 0.97 |
| 18 | FHL1_YEAST | pre-RNA processing protein fhl1 from baker's yeast | 936 | 123 | 800-923 | 0.95 |
| 19 | HYR1_CANAL | hyphally regulated protein from candida albicans (yeast) | 937 | 121 | 621-741 | 0.93 |
| 20 | ANKB_HUMAN | ankyrin, brain variant 1 from human | 3924 | 120 | 1778-1897 | 0.94 |

These 1,069 sequences from the LDR NNP were sorted by prediction length and average value of the prediction strength. The top 20 proteins from this sorting are listed in Table 2. For the full list of predictions see our website at *www.eecs.wsu.edu/~zoran/html/disorder.html*.

## 4. Discussion

A continuous stream of experiments, some about 20 years old, have suggested that disordered regions of proteins are involved in protein function via disorder-to-order transitions upon complex formation. As shown above, examples include enzyme binding with substrate, receptor with ligand, protein with protein, protein with RNA and protein with DNA.

Despite this long history and wide-spread list of examples, as recently as late 1996 the observation of a disorder-to-order transition for p21$^{Waf1/Cip1/Sdi1}$ upon binding to Cdk led to the following statement: "these observations challenge the generally accepted view that stable secondary and tertiary structure are prerequisites for biological activity and suggest that a broader view of protein structure should be considered in the context of structure-activity relationships"[15], and the discovery of a similar result for FlgM / $\sigma^{28}$ binding[14] led to a News and Views article[16] in Nature. Evidently, the many particular examples of important disorder-to-order transitions have failed to register within the molecular biology community as an important generality.

To determine the generality of functional disorder, it is reasonable to explore the use of bioinformatic approaches, for example, by the prediction of disorder from sequence information. Because amino acid sequence determines protein structure[25], it should also determine lack of structure[17]. Indeed, theoretical studies suggest that a very small fraction of sequence space corresponds to sequences that fold into unique 3D structures[26]. Furthermore, it is relatively simple to design sequences that collapse into compact, flexible globules, but so far nobody has identified novel sequences that fold into unique structures with rigid side chain packing, either by design[27] or by use of random sequence libraries for the exploration of local, promising regions of sequence space[28].

Overall, then, both theory and experiment suggest the notion that sequence determines lack of structure and that structured sequences comprise a small fraction of sequence space.

Our modestly successful first generation NNPs of protein disorder demonstrate that disordered regions share at least some common sequence features over many proteins and that more than 15,000 proteins in SW are identified as having long regions of sequence that share these same features. Although much work remains to determine what fraction of these predictions are actually correct and which of such

disordered regions are involved in function, our results nevertheless argue strongly that disordered regions deserve to be recognized as a *category* of protein structure every bit as much as a helix or sheet. Elevating disordered regions to the status of a category may be a necessary prerequisite to recognizing and understanding their general importance[29].

The biological importance of disorder-to-order transitions upon complex formation lies in two realms: energy and mechanism. Each of these will be considered briefly in turn.

With regard to energetic considerations, disorder-to-order transitions lead to low affinity[5,30,31] combined with high specificity[30]. Many biological recognition processes require high specificity, but they may also need relatively low affinity so that the binding undergoes reversal when appropriate. Thus, the combination of high specificity and low affinity is likely to be common; this is consistent with our findings that many particular examples have been identified and that predicted regions of disorder are common.

With regard to mechanism, disorder-to-order transitions have a significant use in one-to-many signaling processes. For example, the disorder of CaM allows it to recognize many different target helices[32,33]. The importance of disorder for one-to-many signaling was emphasized again more recently for p21$^{Waf1/Cip1/Sdi1}$, and it was suggested that this activity is a function of critical importance that only disordered regions of protein structure can provide[15].

Another critical function that can be carried out by disordered regions is the mechanical uncoupling of two domains. For example, the attachment protein of filamentous phage is evidently constructed of two domains connected by a flexible linker; this flexible connection permits the 8,600 Å long phage particle to sway without disrupting the phage's moorings to the cell surface[34] and may also facilitate phage attachment by allowing a local search of multiple orientations relative to the very long phage particle. There are likely to be many additional examples in which there would be an advantage to mechanically uncouple two domains.

We speculate that disordered regions of proteins or entirely disordered proteins will be found to have functions in addition to the few suggestions given here. We hope that our disorder regions predictors (and especially more accurate, next generation versions of them) will be helpful as the molecular biology community broadens its perspective of protein structure-activity relationships to include disordered proteins.

**Acknowledgements**

We would like to thank Gregory Petsko and Georg Schulz for helpful comments regarding protein disorder. This research was supported in part by the NSF grant

ESI-9254358. Special thanks are due to Partek Inc. and Dr. Radu Drossu, whose software were used for data analysis and neural network training, respectively.

**References**

1. Klee, C. B., Draetta, G. F., and Hubbard, M. J. Calcineurin, *Advances in Enzymology*. ed. by A. Meister. **61**, 149-200 (1988).
2. Kissinger, C. R., Parge, H. E., Knighton, D. R., Lewis, C. T., Pelletier, L. A., Tempczyk, A., Kalish, V. J., Tucker, K. D., Showalter, R. E., Moomaw, E., Gastinel, L. N., Habuka, N., Chen, X., Maldonado, F., Barker, J. E., Bacquet, R., and Villafranca, J. E. Crystal Structure of Human Calcineurin and the Human FKBP12-FK506-calcineurin Complex. *Nature*, **378**, 641-644 (1995).
3. Huber, R. Conformational Flexibility and its Functional Significance in Some Protein Molecules, *TIBS*, **4**, 271-276 (1979).
4. Manalan, A. S. and Klee, C. B. Activation of Calcineurin by Limited Proteolysis. *Proc. Nat'l. Acad. Sci.* U.S. A., **80**, 4291-4295 (1983).
5. Alber, T., Gilbert, W. A., Ponzi, C. R. and Petsko, G. A. The Role of Mobility in the Substrate Binding and Catalytic Machinery of Enzymes, *Mobility and function in proteins and nucleic acids*, ed. by Fred Richards. Ciba Founation Sympos. **93**, 4-24 (1982).
6. Livnah, O., Bayer, E. A., Wilchek, M., and Sussman, J. L. Three-dimensional Structures of Avidin and the Avidin-Biotin Complex, *Proc. Natl. Acad. Sci. USA*, **90**, 5076-5080 (1993).
7. Kim, E. E., Varadarajan, R., Wyckoff, H. W., and Richards, F. M. Refinement of the Crystal Structure of Ribonuclease S. Comparison with and between the Various Ribonuclease A Structures. *Biochemistry*, **31**, 12304-12314 (1992).
8. Rayment, I., Holden, H.M., Whittake, R. M., Yohn, C. B., Lorenz, M., Holmes, K. C., and R.A. Milligan, R. A., Structure of the Actin-myosin Complex and its Implications for Muscle Contraction, *Science*, **261**, 58-65 (1993).
9. Bloomer, A. C., Champness, J. N., Bricogne, G., Staden, R., and Klug, A. Protein Disk of Tobacco Mosaic Virus at 2.8 Å Resolution Showing the Interactions Within and Between Subunits. *Nature*, **276**, 362-368 (1978).
10. Guez-Ivanier, V., and Bedouelle, M. Disordered C-terminal Domain of Tyrosyl Transfer-RNA Synthetase: Evidence for a Folded State, *J. Mol. Biol*., **255**, 110-120 (1996) .
11. 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**, 321 (1988).
12. Lewis, M., Chang, G., Horton, N. C., Kercher, M. A., Pace, H. C., Schumacher, M. A., Brennan, R. G. and Lu, P. Crystal Structure of the Lactose Operon

Represssor and Its Complexes with DNA and Inducer. *Science*, **271,** 1247-1254 (1996).
13. 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**, 656-663 (1995).
14. Daughdrill, G. W., Chadsey, M. S., Karlinsey, J. E., Hughes, K. T., & Dalquist, F. W. The C-Terminal Half of the Anti-sigma Factor, FlgM, Becomes Structured When Bound to its Target. *Nature Struct. Biol.*, **4**, 285-291 (1997).
15. Kriwacki, R. W., Hengst, L., Tennant, L., Reed, S. I., and Wright, P. E. Structural Studies of $p21^{Waf1/Cip1/Sdi1}$ in the Free and Cdk2-bound state: Conformational Disorder Mediates Binding Diversity. *Proc. Nat'l Acad. Sci USA*, **93**, 11,504-11,509 (1996).
16. Plaxco, K. W., and Groß, M. The Importance of Being Unfolded. *Nature*, **386,** 657-658 (1997).
17. Romero, P., Obradovic, Z., Kissinger, C. R., Villafranca, J. E. and Dunker, A. K. Identifying Disordered Regions in Proteins from Amino Acid Sequence. *The 1997 IEEE International Conference on Neural Networks Proc*, Houston, TX. **1**, 90-95 (1997).
18. Bairoch, A. and Apweiler, R. The Swiss-Prot Protein Sequence Data Bank and its New Supplement TrEMBL. *Nucleic Acids Res.* **24**, 21-25 (1996).
19. 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 Analysis*, **3**, 387-405 (1990).
20. Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F., and Weng, J. Protein Data Bank, *Crystallographic Databases - Information Content, Software Systems, Scientific Applications*. ed by Allen, F. H., Bergerhoff, G., and Sievers, R. Data Commission of the International Union of Crystallography. Bonn/Cambridge/Chester pp. 107 - 132 (1987)
21. Wisconsin Package Version 9.0, Genetics Computer Group (GCG), Madison, Wisc.
22. Werbos, P. Beyond regression: New tools for predicting and analysis in the behavioral sciences. Harvard University, Ph.D. Thesis, 1974. Reprinted by Wiley and Sons, 1995.
23. Rost, R. and Sander, C. Combining Evolutionary Information and Neural Networks to Predict Protein Secondary Structure, *PROTEINS: Structure, Function, and Genetics*, **19**, 55-72 (1994).
24. Milburn, M. V., Tong, L., DeVos, A. M., Brunger, A., Yamaizumi, Z., Nishimura, S., Kim, S., Molecular Switch for Signal Transduction: Structural Differences Between Active and Inactive Forms of Protooncogenic ras Protein. *Science*, **247**, 939-945 (1990)

25. Anfinsen, C. B., Principles that Govern the Folding of Protein Chains, *Science*, **181**, 223-230 (1973).
26. Abkevich, V. I., Gutin, A. M., and Shakhnovich, E. I. How the First Biopolymers Could Have Evolved. *Proc. Natl. Acad. Sci. USA,* **93**, 839-844, (1996).
27. Betz, S. F., Bryson, J. W., and DeGrado, W. F., Native-like and Structurally Characterized Designed Alpha-helical Bundles, *Current.Opin. Struct. Biol.*, **5**, 457-463 (1995).
28. Davidson, A. R., Lumb, K. J., and Sauer, R. T., Cooperatively Folded Proteins in Random Sequence Libraries, *Nature Structural Biology*, **2**, 856-864 (1995).
29. Lakoff, G. Women, Fire and Dangerous Things: What Categories Reveal About the Human Mind. University of Chicago Press, Chicago, 1987.
30. Schulz, G. E. Nucleotide Binding Proteins, *Molecular Mechanism of Biological Recognition*, ed. by. M. Balaban, Elsevier/North-Holland Biomedical Press, 79-94 (1997).
31. Spolar, R. S. & Record, Jr., M. T. Coupling of Local Folding to Site-Specific Binding of Proteins to DNA. *Science* **263,** 777-784 (1994).
32. Meador, W. E., Means, A. R., and Quiocho, F. A. Modulation of Calmodulin Plasticity in Molecular Recognition on the Basis of X-ray Structures. *Science*, **262,** 1718-1721 (1993).
33. Crivici, A., and Ikura, M. Molecular and Structural Basis of Target Recognition by Calmodulin, *Annu. Rev. Biophys. Biomol. Struct*., **24**, 85-116 (1995).
34. Gray, C. W., Brown, R. S., and Marvin, D. A. Adsorption Comples of Filamentous fd Virus. *J. Mol. Biol.*, **146**, 621-627 (1981).