Intelligent Data Analysis for Protein Disorder Prediction

Pedro Romero [1] promero@eecs.wsu.edu
Zoran Obradovic [1] zoran@eecs.wsu.edu
A. Keith Dunker [2] dunker@mail.wsu.edu

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

Abstract

Although an ordered 3D structure is generally considered to be a necessary pre-condition for protein functionality, there are disordered counter examples found to have biological activity. The objectives of our data mining project are: (1) to generalize from the limited set of counter examples and then apply this knowledge to large data bases of amino acid sequence in order to estimate commonness of disordered protein regions in nature, and (2) to determine whether there are different types of protein disorder. For general disorder estimation, a neural network based predictor was designed and tested on data built from several public domain data banks through a nontrivial search, statistical analysis and data dimensionality reduction. In addition, predictors for identification of family-specific disorder were developed by extracting knowledge from databases generated through multiple sequence alignments of a known disordered sequence to other highly related proteins. Family-specific predictors were also integrated to test quality of general protein disorder identification from such hybrid prediction systems. Out of sample cross validation performance of several predictors was computed first, followed by tests on an unrelated database of proteins with long disordered regions, and the application of few selected predictors to two large protein data banks: Nrl_3D, currently containing more than 10,000 protein fragments of known 3D structure, and Swiss Protein, having almost 60,000 protein sequences. The obtained results provide evidence that long disordered regions are common in nature, with an estimate that 11% of all the residues in the Swiss Protein data bank belong to disordered regions of length 40 or greater. The hypothesis that different protein disorder types exist is supported by high specificity/low sensitivity results of two family-specific predictors, by hybrid systems outperforming general models on a two-family test, and by existence of significant gaps in Swiss Protein vs. Nrl_3D disorder frequency estimates for both families. These findings prompt the need for a revision in the current understanding of protein structure and function, as well as for the developing of improved disorder predictors that should have important uses in biotechnology applications.

1 Introduction

Proteins represent one of the most important and versatile classes of biological macromolecules. From catalyzing reactions to transmitting information within and between cells to providing the building blocks of biological structures, proteins carry out many important functions indispensable for life. One basic property of these molecules is their shape or 3-dimensional structure, as protein structure and function are intimately intertwined. Thus, the study of protein structure is of paramount importance for deepening our understanding of many biological processes.

When a protein is in its functional state, it is called *native*; upon loss of function, it is called *denatured* **(Through the remaining of this paper, terms in italics are explained in the glossary, unless completely defined in the text)**. The native form of a protein is assumed to

have a specific *3D structure*, and the loss of function, or *denaturation*, is assumed to be associated with unfolding or loss of the specific 3D structure. The 3D structure of a protein is generally taken to be a prerequisite for function. Indeed, the current "central dogma" of molecular biology states that information flows in the following manner:

$$DNA \rightarrow mRNA \rightarrow amino\ acid\ sequence \rightarrow 3D\ structure \rightarrow function,$$

where the step from *amino acid sequence* to 3D structure is considered so important that it has been called "the second half of the genetic code", and so difficult to understand from basic principles that this "protein folding problem" is considered to be one of the "Grand Challenges of Computer Science" [12]. In fact, a simplified model of the protein folding problem using lattices to discretize the space of conformations that the protein can assume is known to be an NP-hard problem [15].

The usual approach to the protein folding problem is to start with a collection of proteins of known 3D structures and their associated amino acid sequences. The sequence-3D structure relationships among these "knowns" are used to develop rules, energy functions, or other means to predict the 3D structure from amino acid sequence for a set of test proteins.

X-ray diffraction from protein crystals has provided the most information about the structures of proteins. In such studies, many proteins are found to be nonuniform, having both structured and disordered regions. The structured regions scatter x-rays coherently and so are observed. The disordered regions, however, fail to form fixed structures and so scatter x-rays incoherently. Because of the lack of coherent scattering, such disordered regions fail to contribute intensity in the resulting *electron density maps* [16], and so are invisible in the structure.

In previous attempts to predict 3D structure from amino acid sequence, these disordered or unfolded parts identified in protein crystal structures have been completely ignored. However, over the years, several such disordered regions have been discovered to be required for function. Selected examples include triose phosphate isomerase it binding with triose phosphate [2], avidin with biotin [24], the S-peptide with RNAse S [18], myosin with actin [32], tobacco mosaic virus (TMV) coat protein with its RNA [7], tyrosyl tRNA synthetase with its tRNA [14], and the trp repressor [28], the lac repressor [25] and Bam H1 [27] with their respective DNAs. Also, determination of protein structure by nuclear magnetic resonance (NMR) and structural characterization by other methods have uncovered additional proteins with functional disordered regions, including some that are apparently entirely disordered or unfolded [8, 11, 17, 22, 30, 43]. The terms *natively unfolded* [43] and *natively disordered* [36] have been suggested for describing proteins that are partially or entirely unfolded in their functional states.

A particularly interesting example of a natively disordered protein is calcineurin (CaN), a calcium/calmodulin (CaM) regulated serine/threonine phosphatase [20]. This protein contains three unobserved or disordered regions, the longest of which spans 95 consecutive amino acids [19]. This longest disordered region contains the binding site for CaM. This site had previously been shown to be disordered by its sensitivity to protease and to become resistant to protease upon CaM binding [26]. Protease digestion is one way of estimating order or disorder in a given region of a protein molecule, so these experiments imply that the calmodulin binding region of CaN is normally disordered and becomes ordered upon binding to CaM. In general, for this and other natively disordered proteins or proteins with natively disordered regions involved in binding, *molecular recognition* depends on disorder-to-order transitions as the natively unfolded proteins form *complexes* with their cognate partners [2, 5, 40, 41].

Given that disordered or unfolded regions frequently have function, the omission of such regions is a serious deficiency in previous attempts to solve the protein folding problem. The goal of the present

study is to estimate the importance of including the prediction of order and disorder as part of the protein folding problem.

2 The Protein Disorder Prediction Problem

Despite many reported examples of proteins with natively disordered regions, their commonness and function in nature is unknown. Another interesting unresolved question is whether there are distinct classes or types of disordered regions.

Solutions to these problems can not be found by searching an existing database of proteins with known 3D structure. First, information about protein disorder is not included in the current protein structure databases in any organized way. Second, the current structural databases are likely to be strongly biased against natively disordered proteins, if they are common, because such proteins are potentially difficult to isolate and purify, and because they would be refractory to crystallization, which is a necessary precondition for the determination of 3D structure by x-ray analysis.

Here we propose an alternative, data mining approach to determine whether disordered regions are common and whether distinct classes of disorder exist. Our approach depends on the development of a method to estimate the likelihood that a given amino acid sequence is natively folded or natively unfolded and the application of this method to a large database of amino acid sequences whose 3D structure is still unknown.

This represents a completely new problem in the protein structure prediction domain, having all the efforts in this area been directed towards the prediction of protein 3D structure, rather than lack of structure. In this sense, this paper shows how data mining techniques can help address a problem that has important implications in this domain, yet it is difficult to tackle due to scarce available data.

Amino acid sequence determines protein 3D structure [3], and so we reasoned that it should also determine lack of structure as well [35]. To test this hypothesis, a data engineering process for creating a labeled data set of ordered and disordered protein segments is proposed in Section 3, followed by data analysis, *attributes* generation, and data dimensionality reduction, as described in section 4., and a predictive model design through machine learning, its analysis and application methodology, explained in Section 5. Our explorations with neural network (NN) based disorder predictors reported in Section 6 indeed demonstrated strong relationships between amino acid sequence attributes and lack of foldability as discussed in Section 7.

3 Target Databases Construction

The development, testing and application of protein disorder predictors required prior construction of several databases, each containing a sufficiently large balanced set of two-class examples (ordered and disordered amino-acids). Several public domain protein data banks were used for this purposes as summarized in Figure 1 and as explained in more details in this section.

3.1 Identifying Confirmed Disordered Proteins

The Protein Data Bank (PDB) [1] at the Brookhaven National Laboratory is a public domain archive containing information on more than 6,000 experimentally determined three-dimensional structures of proteins, most of them by x-ray crystallography. The PDB is comprised of a series of text files, one for

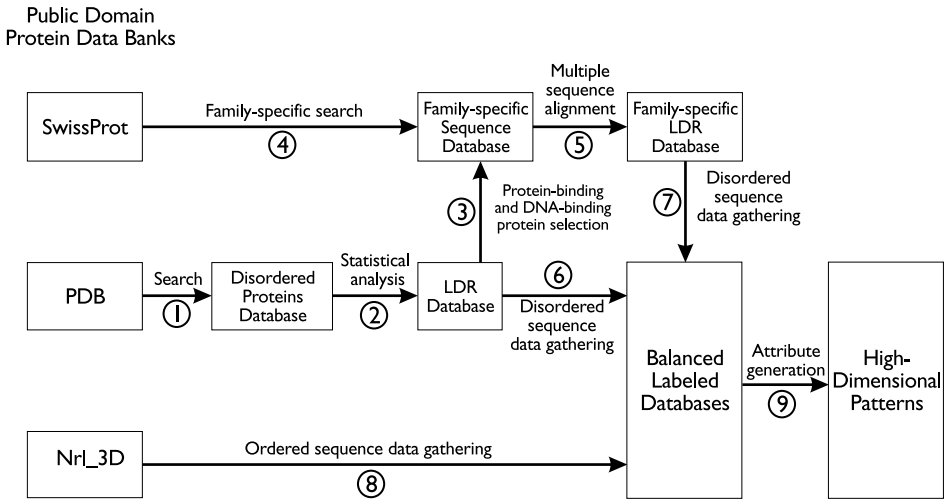

Figure 1: Target databases construction

each protein, containing the three dimensional coordinates of the studied molecule, various information about it, and the determination experiment, including bibliographical references, method of structural determination, other relevant experimental data, *secondary structure* assignments and the like. Most of the extra information is labeled as "remark." In principle, this source can be searched to identify all proteins reported to contain missing coordinate data, which would identify the disordered atoms. However, in practice, finding this information automatically (without reading the loosely structured "remarks" in PDB files) is a non-trivial task due to:

- lack of a standard format for reporting disordered regions in the PDB files (e.g. although invisible parts of the electron density maps are usually reported as "possibly disordered," reports of "weak" or "no electron density" are also used; sometimes coordinates from given *residues* are simply missing, with no more explanation about the reason why they are omitted);
- inappropriate continuous 3D map labeling for disordered regions in some proteins (although coordinates for the atoms of a disordered region cannot be determined experimentally, coordinates for such regions are sometimes inserted in the dataset through use of various methods, such as model building).
- gaps in the 3D map labeling that do not correspond to disordered regions (that is, disordered regions lead to gaps in the 3D map labeling due to missing coordinate data; however, gaps in the 3D map labeling do not necessarily correspond to disordered regions, because non standard numbering is often used).
- the existence of multiple entries for the same protein or nearly identical proteins. When multiple entries occur, all versions must be found and checked to confirm the presence of the same disordered regions in all of them. This search is not a trivial task, because the IDs or the headers of multiple versions of a protein are not necessarily related.
- existing complexes of two different molecules (these should be avoided since disordered regions tend to become ordered upon binding to a *substrate*).

- no easy way to interface the database with a search program (the files can be transferred via ftp from the Brookhaven National Laboratory, which requires frequent updates to keep the data bank current and either a large storage capacity or a multiple phase search on a local system.

Given these difficulties, no effort was made in this study to identify all proteins with confirmed disordered regions in the PDB. The idea was to find a sufficiently large set of proteins with confirmed disordered regions as needed for the design of a neural network predictor. This was achieved by using the following procedure (denoted as step 1 in Figure 1):

1. **Keyword search on PDB files**. A preliminary exploration of the PDB suggested that "disorder" and "disordered" were the most common descriptors for unobserved regions. Therefore, the search space was reduced by using the 3DB browser at the PDB world wide web page to generate names of files containing the keyword "DISORDER". This way, the number of files to download from the PDB FTP site was a fraction of the total contents of the data bank.

2. **Gaps identification in 3D coordinate entries**. A computer program was devised to read the 3D coordinate entries in the downloaded PDB files, searching for gaps, that is, regions of contiguous residues missing from the 3D coordinate entries. The program selected any file having gaps greater than 8 residues in length. This procedure misses those disordered regions that are included in continuous 3D labeling as explained previously, but a preliminary exploration indicates that those regions represent a minority of the reported disordered regions in PDB.

3. **Elimination of unwanted sequences**. The selected files were scanned to remove:

   - complexes (discarded due to their tendency to become ordered upon binding);
   - theoretical models (discarded for lack of experimental confirmation);
   - NMR structures (discarded due to existence of multiple 3D models in some files, which complicated the search process).
   - repeated or almost identical sequences (if a disordered region did not appear in all repeated or almost identical sequences, then all of them were discarded for inconsistencies; otherwise, a protein containing the longest common disordered segment was selected and the remaining are deleted);
   - sequences with unconfirmed disordered regions (only PDB entries that clearly confirmed that the gaps found in the 3D structure by our program are disordered regions were accepted).

A disorder sequence might have no fixed structure whatsoever or be partially folded, having secondary structural elements but with substantial *flexibility*, where the latter case occurs only in sufficiently long regions. This is one of the reasons that prompted us to focus our study to long disordered regions only. The other reason is to profit from the "biased coin effect", meaning that, when predicting long disordered regions, even a weak residue-by-residue predictor can provide strong region-by-region predictions.

Statistical analysis (step 2 in Figure 1) of the constructed database of proteins having confirmed disordered regions was used to determine a size threshold t for long disordered regions in our database. A subset of the disordered regions database, denoted as LDR, was built using only those disordered regions of size t or longer.

3.2 Creating Family-Specific Disorder Databases

Given the small size of the LDR database, the straightforward approach of splitting it and comparing prediction models designed on disjoint data subsets to find out if there are different protein disordered "flavors" was impossible. An alternative method proposed in this study is to design family-specific predictors developed on data generated from multiple alignments of a disordered protein from the LDR database (step 3 in Figure 1) with a family of highly similar (*homologous*) protein sequences (steps 4 and 5 in Figure 1). This disorder database enlargement technique is based on the fact that functionally similar proteins have highly similar 3D structures. Thus, we can assume that the homologues have highly similar ordered regions as in the original protein from LDR; by aligning them, it is possible to identify the approximate location of disordered regions in all the proteins in the alignment. In this study, a protein binding and a DNA binding protein from LDR were selected for this data-enlargement process using versions of the same molecule found in different living beings and obtained by searching the Swiss Protein database (SwissProt), which is a large collection of close to 60,000 protein sequences whose structure is mostly unknown. This yielded substantially larger family-specific databases of proteins with long disordered regions, as will be discussed in the Results section.

3.3 Generating Balanced Databases

Nrl_3D [29] is a structural database containing more than 10,000 protein fragments derived from the data contained in PDB, developed to facilitate searching and cross referencing. A Nrl_3D characteristic important for our study is that it contains only ordered residues from PDB, that is, any residues that do not appear in the 3D coordinate entries of a PDB file are omitted from Nrl_3D, which explains why the database is mostly comprised of protein fragments instead of full sequences. This means that all the residues in this database can be considered ordered and so a random subset corresponding in size to the number of disordered residues gathered from LDR and from family-specific LDR (steps 6 and 7 in Figure 1) can be used to generate balanced training and testing sets in all experiments and to reduce homology effects in family-specific experiments (step 8 in Figure 1).

4 Knowledge Representation and Dimensionality Reduction

Once labeled and balanced datasets are created, it is critical to focus on the construction of training *patterns* by identifying an appropriate knowledge representation, generating *attributes* (step 9 in Figure 1) and reducing initial patters dimensionality (step 1 in Figure 2a, and steps 1-3 in Figure 2b). Methodology used to achieve these steps is described in this section.

4.1 Knowledge Representation

All structural prediction techniques applied on proteins need to take into account that groups of amino acids act together to determine structure [38]. Thus, some information about neighboring residues is always included in such a predicting model. For example, in protein secondary structure prediction experiments [31], information for a given position in a protein sequence is usually represented by an ordered list of residues surrounding each sequence position and measured over a *window* of some fixed size k (typically $9 \leq k \leq 21$). To represent such a categorical list, each residue has to be represented by a vector of 20 binary values (one for each possible amino acid) resulting in $20k$-dimensional patterns. Such a large number of *features* requires huge training datasets for a reasonable predictor development (due to the "curse of dimensionality" that requires

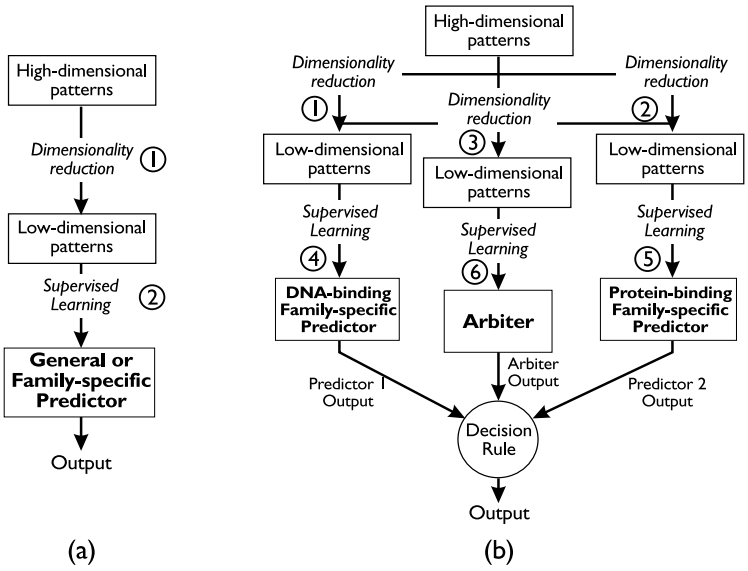

Figure 2: Dimensionality reduction and prediction: (a) a single model (general or family-specific); (b) a hybrid system (3 dimensionality reduction transformations are data specific and each is different)

an exponential number of patterns with respect to the number of features in order to design a reliable predictor for a given non-linear phenomenon [6]).

The small size of the data sets with confirmed protein disorder regions available for this study precludes such a sparse representation. So, in order to extract information from protein sequences, in our project a series of numerical (real-valued) valued attributes are calculated for each sequence position. An attribute associated with a given sequence position is averaged over a window of contiguous residues at and around that position. **So, for example, if we have a protein amino acid sequence like:** $a_1, a_2, \ldots, a_n$ **and we want to use as an attribute a numerical property** $P(a_i)$ **associated with each amino acid** a_i**, the value of** P **for a given position** m **in the sequence will be given by**

$$P^m = \frac{1}{k} \sum_{i=m-\frac{k}{2}}^{i=m+\frac{k}{2}} P(a_i)$$

where k **is the number of residues in a window centered at sequence position** m**.**

In this way, each attribute is represented by a single numerical value rather than 20, and so this representation results in much smaller patterns. Also, pattern size in this case is independent of the size of the window, whereas in the list of residues representation, window size directly determines pattern size.

After n attributes are calculated for a given position in the protein sequence, they are labeled with the class associated with that position (the class meaning either "ordered" or "disordered") to create an n- dimensional pattern suitable for data mining processing.

4.2 Pattern Dimensionality Reduction

Pattern dimensionality reduction is carried out due to: (a) a large number of potentially useful attributes subject to the "curse of dimensionality" constraints; (b) attributes that may have little or no

relation with the characteristic to be predicted; and/or (c) redundancy produced by highly correlated attributes. Item (a) is of major concern here as the number of available LDR patterns is not very large.

The feature selection and feature extraction methods for pattern dimensionality reduction considered in this study are summarized in this section.

4.2.1 Feature Selection

The feature selection approach to reducing the data dimensionality consists of eliminating a number of attributes from the original set in such a way as to minimize the information loss. Feature selection consists of: (1) a *technique* to search the space of all candidate feature subsets to find the optimal one; and (2) a *criterion* to compare among different subsets.

Due to combinatorial explosion, an exhaustive search of all possible feature subsets to find the optimal choice is impractical for all but the smallest number of features. Fortunately, p-feature selection, that is, identifying the optimal p features for our problem is possible by performing so called *branch and bound search* [34] and employing a monotone selection criterion.

The branch and bound search process consists of exploring the candidate feature subsets in a tree-like fashion, starting from the original feature set –including all attributes, and going down the tree by removing one feature at a time. Thus, the first level of the tree consists of all feature subsets of size $d-1$, where d is the number of features in the original set, the second level contains the subsets of size $d-2$ and so on. If a given subset has a criterion value that is smaller than that of a subset located at a lower level in the tree, then all the nodes below it are eliminated because, by the monotonicity property, their criterion values can not be larger.

The monotone selection criterion used in this study is the *Mahalanobis distance,* which is inversely proportional to the minimum error probability and measures the overlapping of class distributions. In a two-class order/disorder problem patterns can be grouped into two clusters, depending on their class. The Mahalanobis distance Δ between the two data clusters with mean vectors μ_1 and μ_2 is computed as

$$\Delta = \sqrt{(\mu_{\mathbf{2}} - \mu_{\mathbf{1}})^{\mathrm{T}} \mathbf{S}^{-\mathbf{1}} (\mu_{\mathbf{2}} - \mu_{\mathbf{1}})}$$

where

$$\mathbf{S} = \frac{(n_1 - 1)\Sigma_1 + (n_2 - 1)\Sigma_2}{n_1 + n_2 - 2}$$

and n_1 and n_2 are the number of patterns in each cluster, while Σ_1 and Σ_2 are the clusters' covariance matrices.

4.2.2 Feature Extraction

In *feature extraction* methods, attributes in the d-dimensional patterns are combined to produce smaller-dimensional patterns of p features. To achieve this reduction with a minimal information loss our study employed a linear transformation technique called *principal components analysis* [6], or *PCA*. This method relies only on the original attributes, without considering the respective class information as in the p-feature selection method.

The PCA algorithm maps d-dimensional patterns $\mathbf{x}_j$ to p-dimensional vectors $\mathbf{z}_j$, where $p < d$. Vectors $\mathbf{x}_j$ can be represented as linear combinations of d orthonormal vectors $\mathbf{u}_i$ as

$$\mathbf{x}_j = \sum_{i=1}^{d} z_i \mathbf{u}_i$$

and can be approximated by

$$\tilde{\mathbf{x}}_j = \sum_{i=1}^{p} z_i \mathbf{u}_i + \sum_{i=p+1}^{d} b_i \mathbf{u}_i$$

where all b_i are constants.

The minimum approximation error occurs when the basis vectors $\mathbf{u}$ satisfy

$$\mathbf{\Sigma u} = \lambda \mathbf{u}$$

meaning that they are the eigenvectors of the data set covariance matrix $\mathbf{\Sigma}$. The b_is of this minimum error approximation correspond to the eigenvalues, λ_i. So, the minimum error of a p-dimensional approximation is achievable by discarding the $d - p$ smallest eigenvalues and their corresponding eigenvectors.

In practice, the data set is first normalized so that each feature has zero mean and unit variance, and then the eigenvalues and eigenvectors of the covariance matrix are calculated. To generate a reduced p-dimensional data set, the original patterns are projected onto the eigenvectors corresponding to the p largest eigenvalues. This modified data set is used to develop the neural network predictor.

5 Predictors Development and Performance Analysis

All predictors considered in this study were based on feed forward neural networks with one hidden layer employing supervised learning from examples through the backpropagation learning algorithm [44] (step 2 in Figure 2a, steps 4-6 in Figure 2b). For each data set, an appropriate architecture is decided by a trial and error process, where the network size is bounded according to the size of the available training data set.

Although all predictors were neural network based, they can be categorized as:

- **general** - single models trained using all available disordered examples (Figure 2a);
- **family-specific** - single models trained to predict a particular type of disorder; and
- **hybrid** - systems that combine family-specific predictors into more general disorder predicting systems by using an *arbiter* neural network decision when the base predictors disagree (Figure 2b).

5.1 Cross-Validation Analysis

Neural network predictors developed on data sets obtained using PCA and p-feature selection were tested through a 5-cross validation process.

In experiments using feature selection-based data reduction, features were identified through an analysis performed on all available data and reduced dimensional data is randomly partitioned into five disjoint subsets each balanced as to have the same number of ordered and disordered patterns.

In contrast, when using PCA-based feature extraction, the data is partitioned before preprocessing, PCA is carried out separately on each "raw" training set, and the resulting eigenvector matrix is used to transform the originating training set and the corresponding validation and testing sets to a lower-dimensional space.

When analyzing a specific single neural network-based predictor, 4 of these data subsets are merged and then randomly partitioned in the following manner: 80% of the data points is used to train a neural network, while the remaining 20% is used to measure quality of the predictor during training as to decide when to stop the training process. Generalization is reported by testing the trained predictor's accuracy on the remaining fifth data subset. This process is repeated thrice per experiment, each time starting from different initializations of the neural network parameters, and the results are averaged. A total of five experiments are performed, each leaving out a different data subset for *out-of-sample* testing. The averaged testing result of these 15 neural networks is used to compare among different feature reduction techniques.

In a hybrid prediction system separate feature reductions were performed for each of two base family-specific models, as well as for the arbiter network. In such a system, and assuming feature selection-based dimensionality reduction, 5-cross validation process steps consisted of:

1. selecting features for each of two family-specific data sets denoted as set_1 and set_2;
2. generating two groups of patterns from each original family-specific data set:
   - $\text{set}_{1,1}$ consisting of Family-1 data with Family-1 features;
   - $\text{set}_{1,2}$ consisting of Family-1 data with Family-2 features;
   - $\text{set}_{2,1}$ consisting of Family-2 data with Family-1 features; and
   - $\text{set}_{2,2}$ consisting of Family-2 data with Family-2 features;
3. partitioning each of these data sets into 5 balanced subsets as in the single predictor case, but preserving the ordering of the patterns within the sets (e.g. i-th pattern on $\text{set}_{1,1}$ and on $\text{set}_{1,2}$ should both come from the i-th pattern on set_1);
4. generating training, validation and testing sets for each family-specific predictor and training them as usual;
5. generating 15 arbiter training data sets (one for each 5-cross validation experiment) by:
   - testing family-specific predictors 1 and 2 on training data from $\text{set}_{1,1}$ and $\text{set}_{1,2}$, respectively; adding the i-th pattern from set_1 to the arbiter's training data if there is a testing disagreement on the i-th pattern;
   - repeating the testing/adding process with family-specific predictors 1 and 2 being tested on $\text{set}_{2,1}$ and $\text{set}_{2,2}$ training data, respectively.
6. merging all 15 arbiter data sets, eliminating duplicate patterns (due to overlapping), and performing feature selection on the consolidated 2-family disagreement data set.
7. Training neural networks on the 15 original arbiter training data sets, after partitioning each set into training and validation subsets as earlier.

As a consequence, for each of the 15 experiments, there are two trained family-specific predictors and a 2-family arbiter. Three sets, one for each neural network, with corresponding patterns are used for testing. When testing, each family-specific predictor is applied to the pattern i from its corresponding testing set and if they agree the system's prediction is computed as the average of both experts' outputs. If they disagree, the arbiter's prediction on the i-th pattern from its test set is used as the system's output. Reported testing results are averaged same as in a single prediction model.

5.2 Performance Analysis on an Unrelated LDR Database

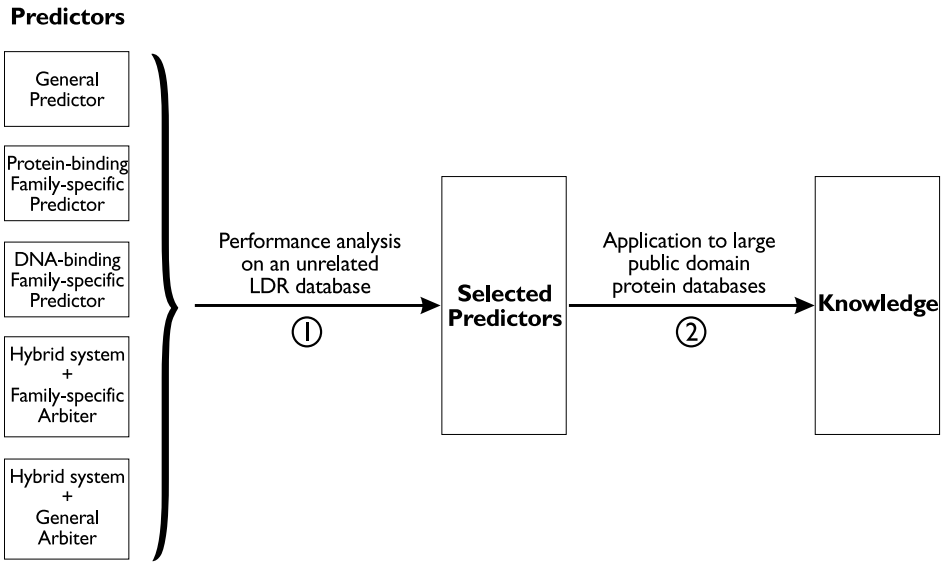

Figure 3: Performance analysis and application

All data used in 5-cross validation training for both single and hybrid predictors was regrouped into complete data sets which were used to train the final neural network predictors. These were then tested on a database of proteins with long disordered regions assembled through literature searches (step 1 in Figure 3). All these proteins are unrelated to those in the LDR database.

The statistics generated from these test complemented with previous cross validation results helped determine which predictors are worth applying to large public domain protein databases.

5.3 Application to Large Protein Databases

The Nrl_3D data bank was used to estimate the false positive error rate of

predictors on a large data set. Then, the predictors were applied to the SwissProt database to assess the commonness of disordered regions in nature and existence of various disorder flavors (step 2 in Figure 3).

6 Results

Results on general and family-specific databases construction and pattern generation are reported in Sections 6.1 and 6.2, followed by window size selection and dimensionality reduction analysis in

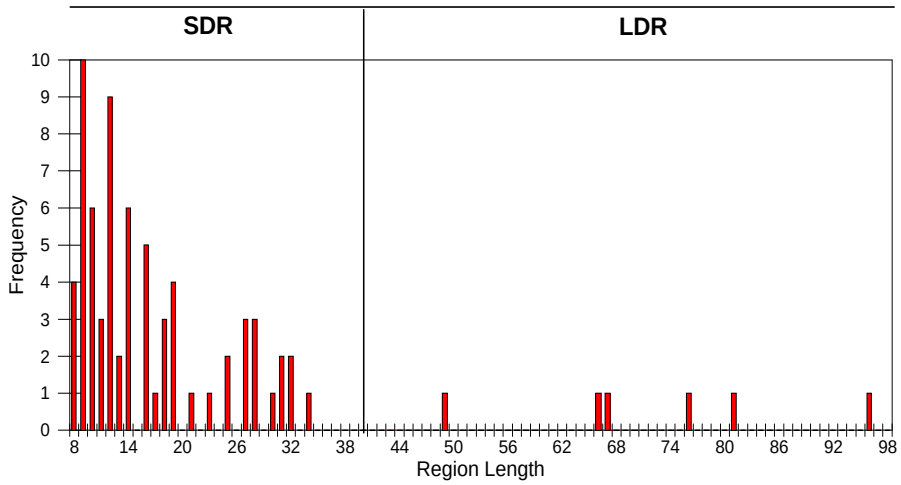

Figure 4: Disordered regions length histogram

Sections 6.3, 6.4 and 6.5. Cross-validation out of sample performance of developed neural networks is summarized in Section 6.6 and their test results on an independent LDR database are shown in Section 6.7. Finally, outcomes of large scale application to two real-life protein databases are presented in Section 6.8.

6.1 Constructed LDR Database and Generated Pattern

The search and analysis on the PDB data bank described in Section 3.1 produced a database of 53 proteins having confirmed disordered regions of different sizes. Study of the lengths of the disordered regions in this database (Figure 4) resulted in separation into disjoint subsets based on their length:

- Short Disordered Regions (SDR) having 8-40 consecutive residues, and
- Long Disordered Regions (LDR) consisting of more than 40 residues.

The LDR database was used in this study as a source of disorder data for predictor development. It consists of the sequences of 6 proteins: DNA topoisomerase II, elongation factor G, lactose operon repressor, tomato bushy stunt virus coat, apoptosis regulator Bcl-X and calcineurin. The database contains 1,453 ordered and 410 disordered residues, for a total of 1,863..

The disordered parts of the LDR database were used to generate patterns corresponding to disordered residues. To obtain data sets with a representative sample of ordered residues found in nature, all ordered patterns were generated from a random sample of sequences from Nrl_3D. Several balanced data sets were generated using attributes information averaged over various size windows. For each sequence position, calculated over windows of a specific size, following attributes were generated:

- 20 attributes corresponding to the composition of each of the 20 possible amino acids;
- average *hydropathy*, calculated using the Kyte-Doolittle scale [23];
- average *flexibility* [42]; and
- two attributes corresponding to the α and β *hydrophobic moments* [10].

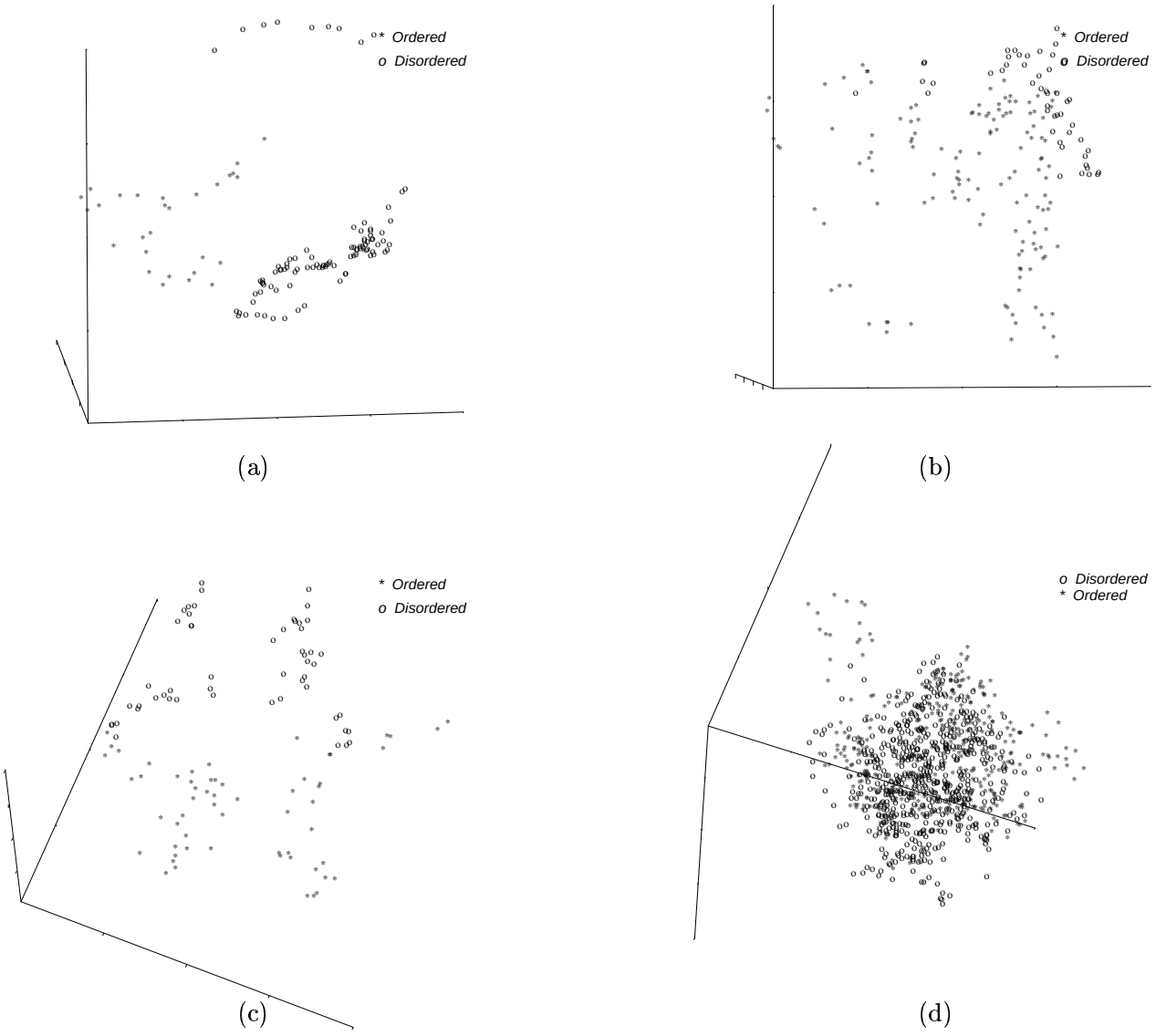

Figure 5: PCA projections of ordered and disordered patterns: (a) CaN, (b) topoisomerase II, (c) Bcl-X, (d) all molecules in the LDR database.

This produces a total of 24 attributes, representing each sequence position as a point in a 24-dimensional space. The resulting data sets were balanced to have the same number of ordered and disordered patterns.

6.2 Constructed Family-Specific LDR

For a distribution analysis of ordered and disordered points in the 24-dimensional attribute space, patterns were generated for all sequence positions (ordered and disordered) in the LDR database. Figures 5a through 5d show three dimensional projections of these points for three of the LDR database's

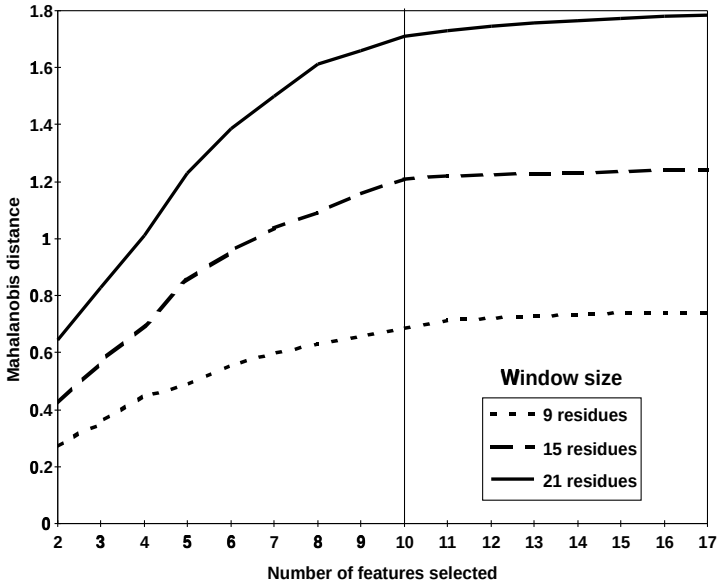

Figure 6: p-feature selection results on three data sets for $2 \leq p \leq 17$

molecules and for the whole database. These graphs were obtained using Principal Components Analysis of the 24-dimensional data sets and then projecting each point onto the first three principal components. The graphs have been rotated to see the separation between classes better.

Notice how, in the case of individual proteins, the points corresponding to ordered and disordered sequence positions seem to fall into well differentiated regions within the attribute space. This differentiation is not as clear when the points from all molecules in the LDR database are presented in a single graph (Fig. 5d). This supports the proposal of not only developing a general predictor trained on all the LDR data, but also producing family-specific predictors using multiple sequence alignments on a family of proteins, in order to take advantage of an apparently easier differentiation between family-specific ordered and disordered regions.

The graphs shown in Figure 5 also suggest the existence of different types of disorder, implying that the family-specific predictors would perhaps be very good at detecting their own type of disorder but is likely to miss other types. Thus, in order to predict long disorder in general (as the LDR predictor attempts to do), an integration of family-specific predictors trained on fairly different kinds of disorder, as explained in Section 5, should be considered. This would take advantage of possibly superior, albeit localized, performance of the family-specific predictors.

CaN and topoisomerase II were selected as prototype proteins for family-specific labeled databases construction. This decision was based on their functions involving different kinds of molecular recognition: CaN binds to proteins, whereas topoisomerase II binds to DNA. 13 proteins homologous to CaN and 12 homologous to topoisomerase II were selected from the Swiss Protein data bank to perform multiple sequence alignments, from which family-specific labeled databases were produced, as explained in Section 3.2. Attributes generation from these databases is carried out same as for the LDR database, producing balanced data sets having ordered patterns obtained from Nrl_3D and disordered patterns generated from the family-specific databases.

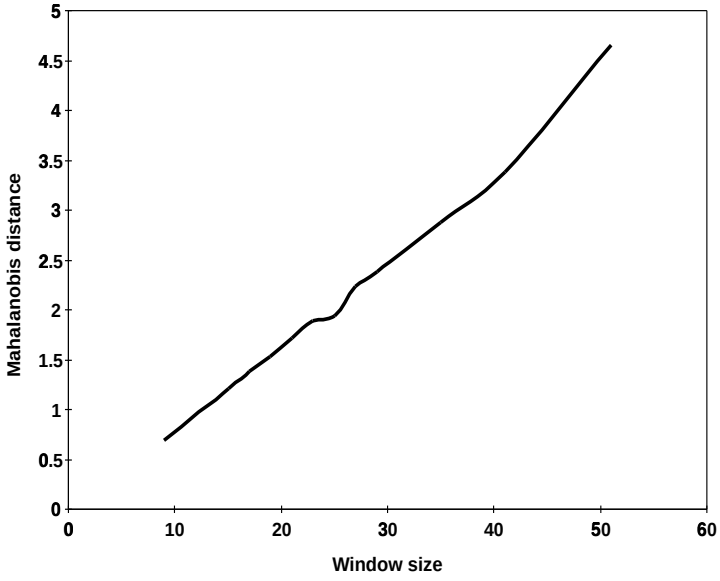

Figure 7: 10-feature selection results on data generated using window sizes ranging from 9 to 51 residues

6.3 Window Size Selection Experiments

To select a suitable window size for attributes generation, feature selection experiments were performed on data sets produced using window sizes ranging from 9 to 21 residues. Branch and bound search was used to find the p optimal features according to the Mahalanobis distance criterion. This procedure was done for both the LDR and the family-specific data sets. Figure 6 shows the results obtained on three CaN data sets using attributes computed over windows of 9, 15 and 21 residues for p ranging from 2 to 17 features.

As expected, the Mahalanobis distance grows with the number of selected features, but its growth slows down at about $p = 10$, implying that there is a smaller improvement in separability by adding more features beyond 10 to the selected subset. This result was consistently repeated for both the LDR and the topoisomerase data sets.

Figure 6 also shows how the Mahalanobis distance grows with the number of residues included in the window, so additional experiments were carried out to determine how the selection criterion value scaled with window size. After determining that 10 features were an appropriate size for the feature set, the 10-feature selection process was performed for data sets with attributes computed over windows ranging from 9 to 51 residues. The Mahalanobis distance increased linearly with the window size, as shown in Figure 7 for the CaN data set. This means that, within the range of window sizes studied, larger window size results in better class separation. Again, this behavior was also observed for the LDR and topoisomerase -specific data sets. Thus, an appropriate window size had to be selected based on its effect on the prediction range, as explained in Section 4. A window size of 21 residues was selected for the generation of all data sets used in this study, since it produces a calculation region big enough to adequately capture the presence of rare amino acids in the vicinity of the studied position without being so large as to limit severely the prediction range on an average-sized protein sequence.

6.4 Feature Selection Experiments

Table 1 show the features selected for all the data sets generated in this study. The data set for the family-specific arbiter is generated from the CaN and the topoisomerase data sets when combining the family-specific predictors, as explained in Section 5. Each 3-letter code corresponds to the composition of the amino acid represented by that code. The other selected attributes are the α and β hydrophobic moments.

Data set	Selected features									
CaN	Tyr	His	Ser	Trp	Phe	Val	Cys	Glu	Arg	β-moment
topoisomerase	Tyr	Gly	Ser	Ala	Asn	Lys	Cys	Pro	Asp	α-moment
LDR	Tyr	Met	Ser	Trp	Phe	Lys	Cys	Glu	Arg	Ala
2-family arbiter	Tyr	His	Ser	Ala	Thr	Lys	Cys	Glu	Arg	β-moment

Table 1: Selected features. 3-letter codes represent composition of the corresponding amino acids

6.5 Feature Extraction Experiments

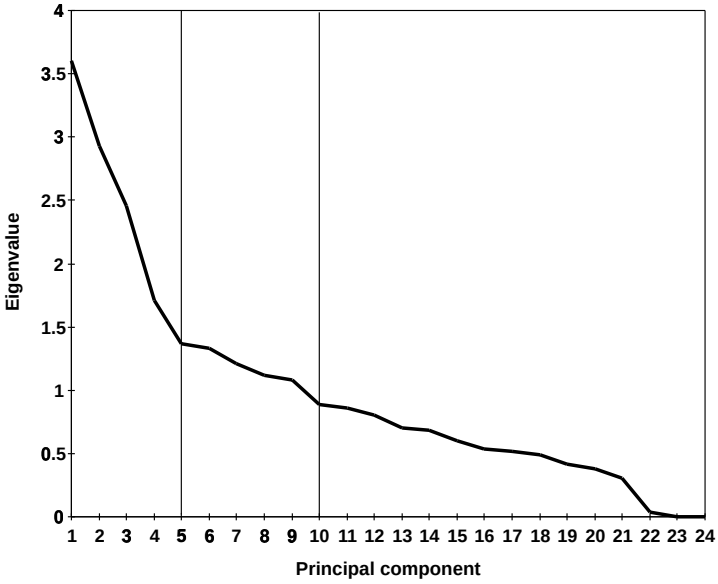

Figure 8: Eigenvalues comparison for deciding on an appropriate number of principal components

Principal components analysis was carried out on the LDR and family-specific data sets, which were normalized to zero mean and a standard deviation of 1 before preprocessing. The eigenvalues obtained for the CaN data set are shown in Figure 8. The relative high value of the first five eigenvalues suggested that a reduction to a 5-dimensional space spanned by the corresponding eigenvectors could produce a reasonable predictor, but the available data also allowed for a reduction to 10 dimensions as needed for a fair comparison to 10-feature selection. Again, similar results were obtained for the topoisomerase and LDR data sets.

Even though this analysis was performed on the whole data sets, feature extraction was later carried out independently on each of the five training subsets used for cross validation, as explained in Section 5. The eigenvectors obtained through this process were used to extract 10 features from the original data set.

6.6 5-Cross Validation Experiments

Data subset	Out-of-sample prediction accuracy		
	CaN	*Topoisomerase*	*LDR*
a	81%	79%	80%
b	80%	79%	84%
c	80%	80%	78%
d	81%	73%	82%
e	81%	77%	81%
Average	80%	78%	81%
Standard deviation	±2%	±3%	±2%
NN Architecture	10-10-1	10-10-1	10-6-1

Table 2: 5-cross validation, out-of-sample accuracy for single neural network predictors on data pre-processed with principal components analysis (PCA).

To select an appropriate dimensionality reduction technique, 5-cross validation experiments using data obtained through both feature selection and feature extraction techniques were carried out on all single-network predictors. Tables 2 and 3 show the average generalization on out of sample data for three runs on each of the five disjoint test sets, represented with the letters a-e. The last two rows show the average generalization over all 15 experiments (5 sets, 3 runs each), along with its standard deviation. Neural network architecture used for a general LDR predictor had six hidden neurons versus ten used in family specific neural networks (due to larger number of training examples available in alignment-based family specific disorder dataset).

The results indicate that the feature selection technique employed in this study produces equally good or better data sets for predictor development on protein disorder data. Thus, feature selection data was used for all the remaining neural network development and prediction experiments.

As explained in Section 5, the final predictors were trained on the entire corresponding data sets, those that were originally partitioned to generate 5-cross validation data. These final versions generated the results reported in the following two sections.

The results for the combination of family-specific data sets (CaN + topoisomerase) are contained in Table 4. It is important to observe lower generalization accuracy for the CaN and topoisomerase -specific predictors when they are applied to a 2-family data sets. This results support the hypothesis that CaN and topoisomerase have different flavors of disorder, as the accuracy drop occurs mainly when predicting on data coming from the other family.

5-cross validation was also carried out for a neural network trained on the whole CaN + topoisomerase dataset and called "CaN+topo" in Table 4. Even though this global predictor results are a big improvement over those of the CaN and topoisomerase predictors, its performance is significantly

Data subset	Out-of-sample prediction accuracy		
	CaN	*Topoisomerase*	*LDR*
a	91%	87%	81%
b	89%	87%	75%
c	85%	87%	84%
d	84%	86%	83%
e	90%	86%	81%
Average	88%	87%	81%
Standard deviation	±3%	±1%	±3%
NN Architecture	10-10-1	10-10-1	10-6-1

Table 3: 5-cross validation, out-of-sample accuracy for single neural network predictors on data preprocessed with the feature selection technique

Data subset	Out-of-sample prediction accuracy			
	Single NN predictors			Hybrid system
	CaN	*Topoisomerase.*	*CaN+topo.*	*2-family arbiter*
a	69%	69%	81%	84%
b	67%	68%	81%	81%
c	66%	71%	78%	82%
d	65%	67%	76%	81%
e	69%	67%	76%	83%
Average	67%	68%	78%	82%
Standard deviation	±2%	±2%	±2%	±1%
NN Architecture	10-10-1	10-10-1	10-10-1	10-10-1 each

Table 4: 5-cross validation, out-of-sample accuracy for single and hybrid neural network predictors on union of CaN and Topoisomerase family-specific datasets.

below that achieved by the hybrid system constructed with a 2-family arbiter as explained in Section 5. Thus, an integration of family-specific predictors can produce better results than a single neural network trained on the global data set, a result that further supports the idea of the existence of disorder flavors.

6.7 Testing on an Unrelated LDR Database

An unrelated database of proteins with known long disordered regions was assembled by gathering examples from protein literature. This set consisted of the following proteins: FlgM, Histone H5, prion, HmgY and glycyl-tRNA synthetase, for a total of 675 ordered and 465 disordered residues.

Several statistical measures of the predictor's performance on this data set are shown in Table 5. In these experiments, in addition to testing previously used single predictors and a hybrid with 2-family

specific arbiter, another hybrid system with LDR arbiter was also used. In Table 5, *predictive value* "disorder" and "order" are accuracy measures computed as

$$\frac{TP}{R_+} \text{ and } \frac{TN}{R_-}$$

where TP and TN are the number of correct disorder (true positives) and correct order (true negatives) predictions, and R_+ and R_- represent the total number of disorder and order predictions, respectively. A predictor's *sensitivity* and *specificity* are

$$\frac{TP}{C_+} \text{ and } \frac{TN}{C_-}$$

where C_+ and C_- represent the total number of disordered and ordered residues in the data set, respectively. Finally, the *correlation coefficient* is defined as

$$\frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TN + FN)(TN + FP)(TP + FN)(TP + FP)}}$$

where FP and FN are the number of false positive and false negative predictions.

Predictor	Performance statistics							
	Errors		Predictive value		Sensitivity	Specificity	Overall accuracy	Corr. coeff.
	false positive	false negative	Disorder	Order				
CaN	7%	18%	73%	75%	53%	88%	75%	0.45
topo	12%	31%	39%	62%	19%	81%	57%	0.00
LDR	21%	11%	56%	78%	70%	66%	67%	0.35
CaN + topo + 2-family arbiter	11%	19%	64%	72%	50%	83%	70%	0.34
CaN + topo + LDR arbiter	9%	18%	69%	74%	53%	85%	73%	0.41

Table 5: Performance on an unrelated set of proteins with long disordered regions

As explained in Section 5, both family-specific predictors were applied to the large protein data banks. The LDR predictor was selected as the general predictor to be used on those databases mainly because it has the smallest false negative error and the highest sensitivity of all predictors. This is important because we can use predictions of disorder on Nrl_3d as an estimate of false positive error on large databases, but there is no comparably accurate way of estimating false negative errors.

6.8 Disorder Predictions on Nrl_3D and SwissProt Databanks

Table 6 shows the cumulative results of applying the selected predictors on Nrl_3D and Swiss Protein. For each length i, the table shows the fraction of the total number of residues in the database that belong to predicted disordered regions of length i or longer. This is shown along with the total fraction of residues that were predicted as disordered, regardless of the size of the disordered region.

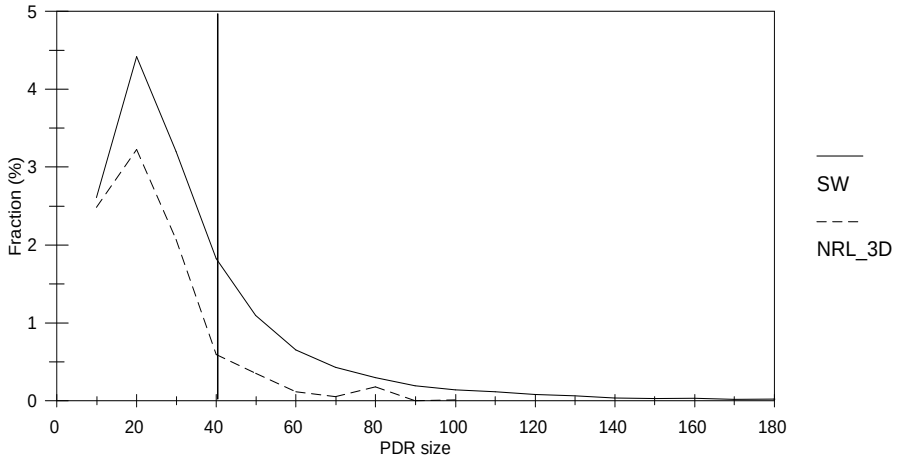

Figure 9: CaN family-specific predictions on large protein data banks.

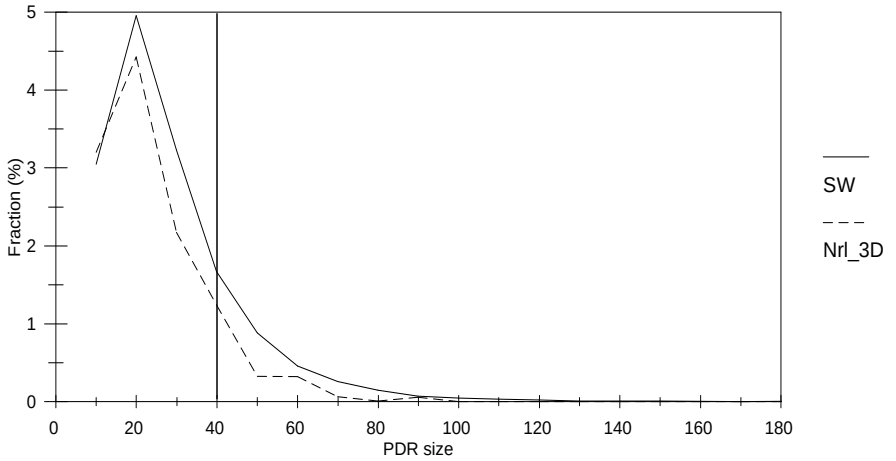

Figure 10: Topoisomerase family-specific predictions on large protein data banks.

Figures 9, 10 and 11 show histograms of fraction of the total number of residues in the database that belong to disordered regions (PDR) of length i. These histograms and Table 6 suggest that long disordered regions are common in nature. Indeed, using the general predictor's (LDR) results on Tables 5 and 6, the fraction of disordered residues belonging to regions of length 40 or greater on Swiss Protein is estimated to about 11%. Here, the SwissProt false positive error rate is estimated

Predictor	Total		$i \geq 20$		$i \geq 40$		$i \geq 60$	
	Nrl_3D	SW	Nrl_3D	SW	Nrl_3D	SW	Nrl_3D	SW
CaN	9%	15%	4%	9%	0.8%	4%	0.2%	2%
topoisomerase	12%	15%	5%	7%	0.9%	2%	0.2%	0.7%
LDR	28%	34%	18%	26%	7%	15%	3%	9%

Table 6: Fraction of Nrl_3D and Swiss Protein (SW) residues belonging to predicted disordered regions of length i or longer

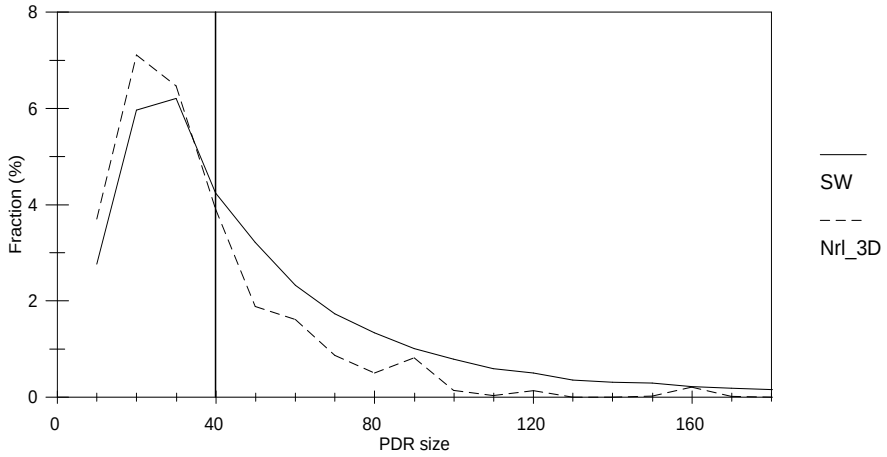

Figure 11: LDR general predictions on large protein data banks.

from Nrl_3D, while it is assumed that the negative error estimate obtained in Table 5 is representative of those on the large databases, as it is the case with the false positive errors shown in Tables 5 and 6. We can see in Table 6 that the estimated false positive error rate for the LDR predictor drops from 28% for all sizes of disordered regions to 7% for disordered regions of size 40 or longer; so for long disorders the false positive error rate drops to a fourth of the total value. Assuming a similar drop for the false negative error rate, we get $11/4 = 2.75\%$ false negative error for long disordered regions. So, fraction of disordered residues in SwissProt for region lengths ≥ 40 is estimated as:

$$f_{40} = \text{fraction predicted} - \text{false positive error} + \text{false negative error}$$

which in this case would be:

$$f_{40} \approx 15\% - 7\% + 3\% = 11\%$$

7 Conclusions and Discussion

Several conclusions can be drawn from the obtained experimental results:

- Disorder seems to come in several distinctive flavors, which makes it difficult for a general disorder predictor to achieve higher prediction accuracy. **Also, generalization on an unrelated set of proteins makes the accuracy drop for all predictors developed in this study, suggesting the possibility of more disorder flavors not represented in our training sets.** Different types of disorder seem to be related to different protein functions.

- When predicting their own kind of disorder, family specific predictors do achieve better out of sample accuracy as compared to a general predictor, but have insufficient sensitivity to other disorder types, which suggest that differences between family-specific disorder types can be deep. Combinations of family-specific predictors look promising as general disorder predictors, providing that an appropriate integration method is devised.

- Long disordered regions seem to represent a small but significant portion of nature's protein sequences. Indeed, the false positive prediction error gap between Nrl_3D and SwissProt and

a false negative error estimate for the LDR predictor suggest that about 11% of all residues in SwissProt belong to disordered regions of 40 or more residues in length. This represents more than two million residues, which is equivalent to about 7,000 average-size proteins.

Biological implications of these findings are discussed in the rest of this section.

7.1 Protein Function and the Commonness and Flavors of Disorder

Several protein functions have been identified that require disordered protein, including at least the following three: 1. regulation of enzymatic activity and/or lifetimes by proteolysis [33]; 2. mechanical uncoupling [13]; and 3. molecular recognition via disorder-to-order transitions upon binding [2, 5, 7, 8, 11, 14, 17, 18, 19, 22, 24, 25, 27, 28, 30, 32, 40, 41, 43]. The wide-spread occurrence of protein disorder indicated by the data presented herein and elsewhere [35, 36, 37] support the commonness and importance of disordered protein for these and possibly other protein functions.

Our data suggest that disordered regions in different proteins can have significantly different sequence characteristics. Thus, there are evidently flavors of protein disorder. We rationalize this observation in terms of the different functions of disordered regions in different proteins. Consider the pair of proteins in this study. A significant portion of the disordered region in calcineurin forms a helix. The protein that binds to calcineurin wraps around this helix upon complex formation. A substantial portion of the disordered region in topoisomerase binds to DNA; evidently, the topoisomerase wraps around the DNA. Formation of the complex in each case requires the respective disordered amino acids to adopt a specific structure when in the ordered state. The structural requirements for the complexed, ordered states almost certainly have consequences for the evolutionary selection of amino acids in the respective disordered regions. For these reasons and given the large differences in the complexes formed with their respective partners, we would expect the sequence characteristics of the disordered region of calcineurin to be very different from the characteristics of the disordered region of topoisomerase. Our findings are certainly concordant with such a difference.

Although different sequence characteristics for different disordered regions naturally follows from differences in their functions, the current and previous results [35, 36, 37] also suggest that disordered regions with diverse functions also have some features in common. Developing an appropriate classification of the flavors of disorder will be a challenging problem indeed.

7.2 Induced Fit Versus Induced Folding

There is potential confusion between induced folding [41] and the well-recognized induced fit [21] As commonly understood, induced fit refers to a protein that is essentially folded, but for which there is substantial *backbone* or side chain movement upon complex formation. **That is, although the protein has a specific 3D structure, it can undergo some changes in its spatial conformation. This way, it can better match the shape of the molecule it binds to in order to generate a complex.** Often the backbone movement is accomplished by the shifts of two relatively rigid domains via the deformation of a connecting hinge-like region. The usual analogy for induced fit is a glove changing shape to fit the hand.

Induced folding, which involves a disorder-to-order transition upon complex formation, is clearly distinct from the version of induced fit described above. **Here the protein is disorder in th unbound state, and acquires a specific 3D conformation upon forming a complex.** The problem is that the term "induced fit" has been stretched to cover virtually any type of conformational

change that occurs upon complex formation. If the stretched version of induced fit is accepted, then induced folding reactions are a subset of induced fit reactions.

7.3 Prior Folding Versus Induced Folding

The important features of any given molecular recognition event are its thermodynamics (e.g., the *affinity* **and** *specificity*) and its kinetics (e.g. the *on*-**and** *off-rates*). Using the ideas originally developed by Schulz [40], we explored the likely differences over evolutionary time between prior folding (ordered proteins) and induced folding with regard to molecular recognition [9].

With regard to binding thermodynamics, our explorations suggest that, for prior folding, affinity and specificity are likely to be strongly linked over time. That is, according to our analysis of prior folding, mutations are likely to increase both the affinity and specificity or decrease both. This linkage of specificity and affinity for prior folding is not generally appreciated. **It makes sense, though, that for a protein to be extremely specific, its 3D conformation should match that of its binding pattern in exceptional detail. This precludes binding to any other molecule, unless it is almost identical, structurally speaking, to the preferred binding partner. Such a detailed structural match produces very strong binding between protein and substrate, which implies a high affinity.** On the other hand, for induced folding, affinity and specificity become completely unlinked. That is, according to our analysis of induced folding, mutations can lead to high affinity with low specificity or low affinity with high specificity or anywhere in between.

With regard to binding kinetics, our explorations suggest that, for prior folding, on-rates should be diffusion controlled with possibly high orientation factors. Point mutations would be expected to be able to exert kinetic effects largely by changing the off-rates, although small effects on the on-rates could result from changing the orientation factors. On the other hand, for induced folding, a wide range of possibilities exist for mutational effects on the on- and off- rates, depending on the mechanistic details. For example, if a flexible region occluded a *binding site*, mutation-determined increases in flexibility should increase the on-rate; on the other hand, if the formation of structure were the rate-limiting step, then mutation- determined increases in the flexibility should decrease the on-rate.

Given the greater potential for variability for induced folding as compared to prior folding suggested by these studies, induced folding would be expected to be especially common in nature. This expectation is supported by the findings in this paper. Indeed, the commonness of disordered sequences suggests the need to consider revision of the prior folding paradigm and hence also a significant revision of the current Central Dogma of Molecular Biology.

8 Limitations and Further Research

The existence of a sizable fraction of disordered residues in nature implies that disordered regions can and do play important roles in many biochemical processes. This has important repercussions on many areas of biochemical and molecular biology research. Being able to accurately predict protein disorder would provide immediate benefits to crystallographers and drug designers, to mention just a couple of applications. However, to obtain better predictors some hurdles have to be cleared first. In fact, the data domain on which this study of protein disorder has been carried out has several inherent limitations that influence the results of data mining and knowledge discovery efforts in this area:

Limited availability of disorder data. Disordered regions are difficult to find both in the struc-

tural databases and in literature. Indeed, the structural databases must be biased against proteins with long disordered regions, mainly because of the difficulty to crystallize such molecules. On the other hand, literature reports of disordered regions are scarce and usually not very precise as to the exact location of those regions within the protein sequences. This makes it hard to obtain sizable training and testing sets which in turn limits the size and power of disorder predictors.

Ambiguity in x-ray structures. Being in a crystal is not the normal state of a protein, which are in solution under physiological conditions. This can affect the molecule structurally, specially due to contacts within a crystal that can provide order to a natively disorder fragment. This means that these disorder residues can be labeled as ordered, introducing noise in the data sets derived from such information.

Border effects. The areas where a disordered region starts or ends can be difficult to determine exactly from either x-ray or NMR information. For example, better resolution in one x-ray experiment can make observable a couple of such border residues that were invisible in a previous, lower resolution, experiment. This adds to the ambiguity problem explained above.

Redundancy. Public protein databases have redundant sequence information. This is specially true for structure databases like PDB and, consequently, Nrl_3D, where it is common to check the effects of one-point mutations (i.e., changing just one residue in the whole sequence) on the 3D structure of proteins.

Further research is being carried out to minimize or overcome these effects. To increase size of a training database for more accurate machine learning we currently study several alternative approaches for increasing the disordered regions database. Also, to provide less biased testing, low redundancy versions of the protein databases are under construction. Once a larger database is obtained, we plan to extend this research to include short disordered regions since there are many examples where they carry out important functions.

To exploit the increased accuracy of family-specific predictors, alternative integration approaches are being considered. Indeed, evolution seems to have separated disorder into distinctive flavors. Assuming existence of specific predictors for all distinctive protein disorder flavors, our results suggest that it might be possible to design an accurate hybrid protein disorder identification system. However, further research is needed in order to discover if protein disorder can be categorized into a limited number of classes and to find out statistical characteristics of these classes.

Biological systems contain networks of molecular recognition events, both to control the flow of matter (metabolic networks) and information (signaling networks). We speculate that the enhancement of the overall efficiency of these networks by natural selection would be greatly facilitated if the thermodynamic and kinetic parameters of the individual steps could be unlinked, allowing natural selection to operate with some degree of separability regarding specificity, affinity, on-rate and off-rate.

Biological systems also contain what we can term "end-point" molecular recognition events. Examples of these include the numerous hydrolytic enzymes involved in processes like digestion, inactivation of competing organisms, etc. For such events, enhancement of the efficiency of the specific, individual reaction would seem to be likely to confer an evolutionary advantage.

¿From the above two paragraphs, we have developed the hypothesis that end-point proteins will mostly utilize prior folding, whereas proteins within signaling, metabolic, or other networks will mostly utilize induced folding. We intend to use prediction of disorder from amino acid sequence as the starting point to test this hypothesis.

References

[1] 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. 107-132, 1987.

[2] 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 F. M. Richards, *Ciba Foundation Symposium,* 93:4-24, 1982.

[3] Anfinsen, C. B., "Principles That Govern the Folding of Protein Chains," *Science* 181:223-230, 1973.

[4] Arnold, G. E., Dunker, A. K., Johns, S. J. and Douthart, R. J. "Use of Conditional Probabilities for Determining Relationships Between Amino Acid Sequence and Protein Secondary Structure," *Proteins: Structure, Function and Genetics*, 12:382-399, 1992.

[5] Ayala, Y. M., Vindigni, A., Nayal, M.,Spolar, R. S., Record, Jr. M. T. and Di Cera, E., "Thermodynamic Investigation of Hirudin binding to the Slow and Fast Forms of Thrombin: Evidence for Folding Transitions in the Inhibitor and Protease Coupled to Binding," *J. Mol. Biol.,* 253:787-798, 1995.

[6] Bishop, C. M., *Neural Networks for Pattern Recognition,* Oxford University Press, Oxford, UK, 1995.

[7] Bloomer, A. C., Champness, J. N., Bricogne, G., Staden, R., and Klug, A. "Protein Disk of Tobacco Mosaic Virus at 2.8 ÅResolution Showing the Interactions Within and Between Subunits," *Nature*, 276:362-368, 1978.

[8] 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," *Nature Struct. Biol.,* 4:285-291, 1997.

[9] Dunker, A.K., Garner E., Guilliot S., Romero P., Albrecht K., Hart J., Obradovic Z., Kissinger C., and Villafranca, J.E., "Protein Disorder and the Evolution of Molecular Recognition: Theory, Predictions and Observations," *Pacific Symposium on Biocomputing,* 3:471-482, 1998.

[10] Eisenberg, D., Weiss, R.M., and Terwilliger, T.C. "The Helical Hydrophobic Moment: A Measure of the Amphiphilicity of a Helix," *Nature,* 299:371-374, 1982.

[11] Frankel, A. D. and Kim, P. S. "Modular Structure of Transcription Factors: Implications for Gene Regulation," *Cell,* 65:717-719, 1991.

[12] *Grand Challenges 1993: High Performance Computing and Communications, A Report by the Committee on Physical, Mathematical, and Engineering Sciences*, Federal Coordinating Council for Science and Technology, 1993.

[13] Gray, C. W., Brown, R. S., and Marvin, D. A. "Adsorption Couples of Filamentous fd Virus," *J. Mol. Biol.*, 146:621-627, 1981.

[14] 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.

[15] Hart, W., Istrail, S. "Robust Proofs of NP-Hardness for Protein Folding: General Lattices and Energy Potentials," *Journal of Computational Biology*, 4(2):1-20, 1997.

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

[17] Huth, J. R., Bewley, C. A., Nissen, M. S., Evans, J. N. S., 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," *Nature Struct. Biol.,* 4:657-665, 1997.

[18] 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.

[19] 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., Moornaw, E., Gastinel, L. N., Habuka, N., Chen, X., Maldonado, F., Barker, J. E., Bacquet, R., and Villafranca, J. E., "Crystal Structures of Human Calcineurin and the Human FKBP12-FK506-Calcineurin Complex," *Nature,* 378:641-644, 1995.

[20] Klee, C. B., Draetta, G. F., and Hubbard, M. J. "Calcineurin," *Advances in Enzymology.* ed. by A. Meister. 61:149-200, 1988.

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

[22] 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. Natl. Acad. Sci. USA,* 93:11,504-11,509, 1996.

[23] Kyte, J. and Doolittle, R. F. "A Simple Method for Displaying the Hydropathic Character of a Protein," *J. Mol. Biol,* 157:105-132, 1982.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] Pattabiramaan, 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.

[30] Plaxco, K. W. and Gross, M. "On the Importance of Being Unfolded," *Nature,* 386:657-658, 1997.

[31] Qian, N. and Sejnowski, T. J. "Predicting the Secondary Structure of Globular Proteins Using Neural Network Models," *J. Mol. Biol*, 202:865-884, 1988.

[32] 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.

[33] Rechsteiner, M. and Rogers, S. W. "PEST Sequences and Regulation by Proteolysis," *TIBS*, 21:267-271, 1996.

[34] Ripley, B. D., *Pattern Recognition and Neural Networks,* Cambridge University Press, Cambridge, UK, 1996.

[35] Romero, P., Obradovic, Z., Kissinger, C., Villafranca, J.E. and Dunker, A.K. "Identifying Disordered Regions in Proteins from Amino Acid Sequence," *Proc. IEEE Int. Conf. on Neural Networks,* Houston, TX, 1:90-95, 1997.

[36] Romero, P., Obradovic, Z., and Dunker, A.K. "Sequence Data Analysis for Long Disordered Regions Prediction in the Calcineurin Family," *The Eighth Workshop on Genome Informatics*, December 12-13, 1997, Tokyo, Japan, in press, 1997.

[37] Romero, P., Obradovic, Z., Kissinger, C., Villafranca, J.E., Garner, E., Guilliot, S. and Dunker, A.K. "Thousands of Proteins Likely to Have Long Disordered Regions," *Pacific Symposium on Biocomputing,* 3:435- 446, 1998.

[38] Rose, G. D., "Prediction of chain turns in globular proteins on a hydrophobic basis," *Nature*, 272: 586-590, 1978.

[39] Schweers, O., Schonbrunn-Hanebeck, E., Marx, A. and Mandelkow, E. "Structural studies of tau protein and Alzheimer paired helical filaments show no evidence for b-structure," *J Biol Chem*, 269:24290-242907, 1994.

[40] Schulz, G. E., "Nucleotide Binding Proteins," *Molecular Mechanism of Biological Recognition,* in Elsevier/North- Holland Biomedical Press, ed. by M. Balaban, 79-94, 1979.

[41] Spolar, R. S. and Record, Jr. M. T., "Coupling of Folding to Site-Specific Binding of Proteins to DNA," *Science,* 263:77-784, 1994.

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

[43] Weinreb, P. H., Zhen, W., Poon, A. W., Conway, K. A., and Lansbury, P. T. "NACP, A Protein Implicated in Alzheimer's Disease and Learning, Is Natively Unfolded," *Biochemistry,* 35:13709-13715, 1996.

[44] Werbos, P., *Beyond Regression: New Tools for Predicting and Analysis in the Behavioral Sciences,* Harvard University, Ph.D. thesis, 1974. Reprinted by Willey & Sons, 1995.

Glossary

3D structure (Tertiary structure)
The complete 3-dimensional shape of a protein represents its *tertiary structure*, which we refer to as simply 3D structure. It includes not only the secondary structure elements, but also their spatial position and orientation and the location of each atom. The function of a protein is thought to derive from its 3D shape.

Affinity
The binding strength of a protein-substrate complex. The tighter the binding, the higher the affinity of the complex.

Amino acids
The building blocks of proteins, amino acids are small organic molecules that bond with one another in a chainlike fashion to form proteins. If chemical modifications are ignored, there are basically twenty different amino acids in nature. All of them share a common core structure; they are differentiated by side chains having different chemical properties.

Attribute
A numerical value calculated over a specified number of consecutive amino acids often called a window [4]; examples include hydropathy [23], hydrophobic moment [10], flexibility [42] or simply amino acid composition [37].

Backbone
A continuous linear chain formed as water is removed during bond formation between the core structures of adjacent amino acids.

Binding
Proteins bind or attach to other molecules in order to carry out their biochemical functions. This process is sometimes referred to as *molecular recognition*. Two important aspects of binding are the affinity (tightness) and the specificity (binding to specific molecules rather than other similar ones).

Binding site
The region in a protein molecule where binding takes place. It is important to the binding site to be accesible to the protein's substrates.

Complex
Two or more molecules bound together form a complex.

Electron density map
The information resulting from the scattering of x-rays through the molecules in a crystal. The 3D positions of individual atoms in a molecule can be determined from its electron density map.

Feature
A product of preprocessing applied on a set of attributes; it can either be one of or a combination of the original attributes.

Flexibility
A measure of the capacity of a region of protein to undergo local motions. This capacity is affected by steric interactions among the nearby amino acid residues.

Homologous protein sequences or homologues

When two or more proteins have similar sequences and are related by evolutionary descent from a common ancestor, they are said to be homologous to one another and are called *homologues*. When there is no reason to believe that they descend from a common ancestor, they are just called *similar*. The hemoglobins from various animals are examples of homologous proteins.

Hydropathy

The tendency of the side chain of a residue to dissolve in water. When a residue is insoluble in water, it is called *hydrophobic*, while water soluble residues are called *hydrophilic*. The average hydropathy of a protein fragment is therefore related to its solubility in water.

Hydrophobic moment

When a protein fragment assumes a secondary structure shape, like an α-helix or a β-sheet, the distribution of hydrophobic and hydrophilic residues along the structure can determine certain properties. For example, if most of the hydrophobic residues in an α-helix fall in one side of the structure, with most of the hydrophilic on the other side, the helix has the property of being water soluble on one side (which will tend to be exposed on the surface of the protein) and insoluble on the other (which will face to the inside of the molecule). The quantities that measure this hydrophobicity imbalance on α-helices and β-sheets are called *α- and β-hydrophobic moment*, respectively.

Molecular recognition

See **Binding**.

Natively disordered or natively unfolded sequence

A sequence that does not fold into a single unique 3D structure under physiological conditions. Such sequence might have no fixed structure whatsoever (*random coil*), or be partially folded, having secondary structural elements but with substantial flexibility (*molten globule*).

Off-rate

The rate at which a complex separates into its constituent molecules.

On-rate

The rate at which the binding of two separate molecules into a complex takes place.

Out-of-sample

Testing using a data set that contains none of the examples from the training set.

Pattern

A tuple of attributes or features associated with a given sequence position, augmented with the actual class of that position (in this case "ordered" or "disordered").

Residue

When bonding with one another, a pair of amino acids undergo a reaction where they loose a water molecule. Such bonded amino acids are called *residues*.

Secondary structure

The residues in a protein sequence can rotate around the protein chain, or backbone, causing the protein molecule to fold into complex 3-dimensional shapes. A given protein fragment can

take one of several basic shapes, mainly *helix*, *sheet* or *loop*. The *secondary structure* of a protein consists of the type and location of these basic shapes within the molecule's sequence.

⟨*Protein* **or** *amino acid*⟩ **Sequence (primary structure)**
The sequentially ordered list of residues in a protein chain. This list specifies the order of the side chains that dangle from the core backbone structure.

Specificity
The ability of a protein to bind only to selected substrates. A protein that only accepts one or two binding patterns is said to be highly specific.

Substrate
The molecule to which a protein binds to carry out some function.