Sequence Data Analysis for Long Disordered Regions Prediction in the Calcineurin Family

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

Our recently reported results [14, 29, 30] provide strong support for a hypothesis that some amino acid sequences code for disordered regions rather than structured ones and that such disordered regions are commonly involved in function. General and family-specific neural network predictors developed in those previous studies suggest that different classes of disordered regions exist. Here, family-specific data preprocessing for disorder prediction in the calcineurin (CaN) family is explored. The results show that prediction of order and disorder on CaN sequence data benefits significantly from the use of family-specific preprocessing, with feature extraction through principal components analysis (PCA) outperforming feature selection techniques, although all methods do a good job of discriminating CaN-specific disordered regions from CaN-specific ordered regions. On the other hand, for the discrimination of CaN-specific disordered regions from general (unrelated to CaN) ordered regions, feature selection approaches proved to be more appropriate than PCA. The results further support a hypothesis that different kinds of disordered regions exist, as all family-specific disorder predictors developed in this study significantly outperformed a previously reported general multi-family disorder predictor.

1 Introduction

The standard view of protein structure and function is that the amino acid sequence determines the 3D structure and the 3D structure is a prerequisite for function. However, there is a growing recognition that not all proteins fit this standard view. Some proteins contain local regions that fail to fold into particular 3D structures; other proteins are evidently entirely unfolded in their native states (Table 1).

Many of these "natively unfolded" proteins [34] or natively unfolded regions are, paradoxically, involved in molecular recognition. For these proteins, molecular recognition depends on disorder-to-order transitions as the natively unfolded proteins form complexes with their cognate partners [3, 6, 31, 32]. Increasing attention is being paid to a diverse collection of natively unfolded proteins or natively unfolded regions that participate in cell signaling pathways [11, 16, 19, 20, 21, 26, 34]. The prion protein, which was in the news recently due to the award of the Nobel Prize to S. Prusiner for his infectious protein hypothesis, also has a very long disordered region under physiological conditions; a part of this region apparently undergoes a disorder-to-order transition during the conversion of this protein into its infectious state [27].

Although many examples of proteins with natively disordered regions have been reported [3, 6, 11, 16, 19, 20, 21, 26, 27, 31, 32, 34], it is difficult to assess their commonness and overall importance. For example, natively disordered proteins might be more difficult to isolate and purify compared to their well-structured cousins. This would introduce a bias right from the beginning, leading to underestimation of the commonness and the importance of natively disordered protein.

Bioinformatics approaches provide a possible alternative means to carry out an assessment of the importance of native disorder. The overall strategy would be to develop a method to estimate the

Term	Description
natively unfolded sequence	a sequence that does not fold into a single unique 3D structure under physiological conditions; might be like a random coil, or partially folded like a molten globule [13]; "natively disordered" is an alternative descriptor.
attribute	a numerical value calculated over a specified number of consecutive amino acids [5] often called a window; examples include hydropathy [22], hydrophobic moment [15], flexibility [33] and amino acid composition [30].
feature	a product of preprocessing applied on a set of attributes; it can either be one of or a combination of the original attributes.
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").
out-of-sample	testing using a data set that contains none of the examples from the training set.
k-cross-validation	dividing the data at random into k disjoint subsets and repeating k times the process of training/verifying a neural network using data from $k-1$ of the subsets and using the remaining subset for out-of-sample testing; each experiment is leaving out a different subset for testing, and the test accuracy is averaged over all k experiments.

Table 1: Jargon

likelihood that a given amino acid sequence is natively folded or natively unfolded and then apply this method to databases of amino acid sequences. That is, since amino acid sequence determines protein 3D structure [4], we reasoned that it should also determine lack of structure as well. Our initial explorations with neural network based disorder predictors indeed demonstrated strong relationships between amino acid sequence attributes and lack of foldability [29, 30].

Following our initial studies on the relationship between sequence and foldability [29], we turned our attention to a single natively disordered region identified by alignment of a family of closely related sequences [30]. Our initial family-specific predictor was developed using features (defined in Table 1) chosen through a prior analysis of data from several different individual proteins having disordered regions [29]. The purpose of these prior studies was to ask whether a predictor trained on a set of disordered sequences from one protein family, in this case the calcineurin (CaN) family, would work as well as a predictor trained on several unrelated regions of disorder; for such a comparison, we needed to use the same set of features.

Application of the family-specific predictor and the general predictor to sequences known to be natively disordered often, but not always, yielded concordant predictions (manuscript in preparation). Prediction disparities, when they occurred, suggested that distinguishable classes of disordered regions probably exist. A reasonable extrapolation of this suggestion is that a collection of family-specific predictors might give better estimates of disorder as compared to estimates based on a single predictor formed from several classes of disordered regions. However, a family-specific predictor should be based on family-specific features rather than features selected from the study of a collection of different proteins.

The first goal of the present study was to learn whether better family-specific long disorder predictions can be achieved by identifying family-specific features. The second goal was to compare different data preprocessing methods for feature identification in this domain.

In order to achieve the first of the two specified goals, two different CaN-specific predictors were to be compared: one based on a set of features selected from different individual proteins with disordered regions [29, 30] and one based on CaN-specific feature selection. During the course of these studies,

we discovered that our original CaN-specific predictor's accuracy [30] could be improved. The new version of this predictor was used here. In addition, the accuracy of the CaN-specific predictors would be expected to be affected by the sequence homology in the ordered regions as well as by the feature patterns in the disordered regions. For this reason, all the experiments were repeated for new data in which the set of ordered regions of CaN was replaced by a "general" ordered data set developed from the ordered regions randomly selected from the Nrl_3D database [25].

The results demonstrate the benefits of using family-specific features when discriminating between ordered and disordered residues within the same family. In this case, the out-of-sample, cross-validated prediction accuracy shows an increase from about $86 \pm 2\%$ for the original but improved family-specific predictor, up to $92 \pm 1\%$ when using feature reduction methods for identification of an appropriate family-specific set of features. On the other hand, differentiating CaN-disorder from Nrl-order proved to be more difficult for some of the CaN-specific predictors employing family-specific-preprocessing (i.e., the ones based on PCA) while BBS(10) preprocessing still worked well. In any case, all the family-specific predictors studied here outperformed a multi-family predictor developed in [29], whose average out-of-sample prediction accuracy was $73 \pm 2\%$. These results strongly support the idea that different kinds of disordered regions exist, and suggest that further exploration of family-based predictors of protein disorder would be worthwhile.

Previous work relevant to this study is summarized in Section 2, the methodology is explained in Section 3, the results are reported in Section 4, and a discussion is presented in Section 5.

2 Previous Work

Our research on disordered regions prediction in proteins was prompted by structural studies performed on calcineurin (CaN) whose largest disordered region, containing 95 residues, plays an important part in this protein's function [20]. In our initial work [29], disordered regions from CaN and 52 other proteins from different families were analyzed, resulting in separation into three disjoint subsets based on their lengths:

- Short Disordered Regions (SDR) having 8-22 consecutive residues,
- Medium Disordered Regions (MDR) with 23-40 residues, and
- Long Disordered Regions (LDR) consisting of more than 40 residues.

A combination of sequence-based attribute comparisons on all subsets and a formal feature selection process performed on the LDR subset resulted in the identification of 10 relevant attributes out of 22 considered. A modified sequential forward search technique with a minimum error probability criterion was used for the feature selection algorithm, denoted as *MSFS(10)*. The 10 features selected by this technique were used to develop SDR, MDR and LDR neural network predictors.

In a further study [30] another LDR predictor was developed using sequence data from a group of 13 homologous CaN molecules, as opposed to the seven previously used LDR proteins, which belong to different families. Although the 2D structure of the 12 homologous CaN molecules were not known, the fact that highly similar proteins have similar 3D structures led to the assumption that all the 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, and all residues that aligned with disordered positions in the original CaN sequence were considered disordered. This procedure generated enough data for the training of a neural network-based, CaN-specific LDR predictor. The features used to train this neural network predictor were the same 10 features used in the original multi-family LDR predictor. Here, we are considering an improved version of that CaN-based predictor, called MSFS-LDR(10), that is based on features selected using MSFS on LDR data. Both this predictor and another one called PCA-LDR(10), based on features

obtained through principal components analysis (PCA) on LDR data, were used for comparison to the predictors developed using CaN family-specific preprocessing. The idea was to perform a thorough sequence data analysis on the 13 homologous CaN molecules to find out whether a more accurate predictor for CaN-like LDRs can be developed by feeding a neural network with CaN-specific features instead of those based on multi-family data, like the LDR database.

The present study involved two tasks. In the first one, a data set consisting of both CaN ordered and CaN disordered data, called CaN-CaN, was used for data analysis and predictor development. In task two, the possible effect of the CaN molecules' high homology was reduced by substituting ordered data obtained from a random sample of Nrl_3D sequences for the original CaN ordered samples. In fact, similarity analysis on the aligned CaN sequences [14] has determined that the disordered regions of calcineurin show far less identity across homologues than the ordered ones. Thus, using non-homologous ordered sequence data –abundantly available in Nrl_3D– helps reduce any bias that high homology may introduce in the predictions. Also, prediction results on these two data sets can hint at the possible differences between family-specific and general ordered and disordered regions.

3 Methods

3.1 Data Generation

To predict order or disorder on a residue-by-residue basis, a number of sequence-dependent attributes is calculated over a *window* of n residues surrounding each position in a protein sequence (see Table 1). These attributes include amino acid compositions, average hydropathy and flexibility, and hydrophobic moments.

The window size n over which attributes are calculated is a compromise between the granularity of the attribute values and the actual number of sequence positions that can be used for the predictions. A larger n is desirable since a very small n limits the range of possible attribute values and causes the composition of rare amino acids to be almost always evaluated as zero. However, a smaller n is also desirable since the prediction range is limited to only $m - 2\lfloor n/2 \rfloor$ sequence positions, where m is the total number of residues in a protein, because predictions can not be made for $\lfloor n/2 \rfloor$ residues at each end of a sequence.

Once a window size n is decided upon, the values of all attributes are calculated for each position within the prediction ranges of the selected protein sequences. The set of attribute values for each sequence position is augmented with the class value for that position (ordered or disordered) to make a pattern.

The number of attributes in a pattern can be large, producing a high-dimensional prediction problem, and prompting the use of feature reduction preprocessing methods.

3.2 Feature Reduction

Pattern dimensionality reduction is carried out to: (a) cope with the "curse of dimensionality" that requires 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 [8]; (b) to eliminate attributes that may have little or no relation with the characteristic to be predicted; and (c) to eliminate redundancy produced by highly correlated attributes.

Item (a) is of major concern as the number of available LDR patterns from the CaN sequences is not very large. In fact, we were motivated to study the utility of sets of homologous proteins for training neural network disordered regions predictors in the first place because this use of homologous proteins, if successful, would enable the very rapid enlargement of our data sets of disorder. That is, use of homologous sequences for training disordered regions predictors would potentially allow us to take advantage of the rapidly expanding sequence data bases.

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

3.2.1 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, identifying the optimal p features for our problem is possible by performing so called *branch and bound search* [28] 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

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

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.

In this paper, the branch and bound based selection of p out of d features employing the Mahalanobis distance criteria will be called the *BBS(p)* method.

3.2.2 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* [8], denoted here as *PCA(p)*. This method relies only on the original attributes, without considering the respective class information as in the *BBS(p)* method.

The *PCA(p)* 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_is 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.

3.3 Predictor Development

Predictors were developed following the same procedure for both the CaN-CaN and the CaN-Nrl_3D data sets. For feature selection approaches, the available data is used to select a number of features, which are then used to generate new patterns. Our preprocessed data were randomly partitioned into five disjoint subsets each balanced as to have the same number of ordered and disordered patterns. In the case of feature extraction, the data were partitioned *before* preprocessing. PCA was carried out separately on each "raw" training set, and the resulting eigenvector matrix was used to extract the desired number of features from the originating training set and the corresponding validation and testing sets.

For a specific experiment, 4 of these data subsets were merged and this 80% of the total data were then randomly partitioned in the following manner: 64% of the data points were used to train a feed forward neural network with one hidden layer employing the backpropagation learning algorithm [35], while an additional 16% were used to measure quality of the predictor during training in order to decide when to stop the training process.

Generalization was estimated by testing the trained predictor's accuracy on the remaining 20% of the data. This process was repeated three times starting from different initializations of the neural network parameters and the results were averaged. A total of five experiments were performed, each leaving out a different data subset for out-of-sample testing. The averaged testing result of these 15 neural networks was used to compare among the different feature reduction techniques.

4 Results

4.1 Data Generation and Cleaning

Data used for the reported experiments were generated using information from a group of 13 homologous CaN proteins. Twelve of these proteins were found in SwissProt [7] (P2BA_HUMAN, P2B1_YEAST, P2B2_YEAST, P2B_EMENI, P2B_NEUCR, P2BC_HUMAN, P2BC_MOUSE, P2BA_RAT, P2BB_RAT, P2BA_BOVIN, P2BB_MOUSE, and P2BB_HUMAN) while the thirteenth was found in PIR [17] (A36222). This group of proteins contains a total of 6,861 residues.

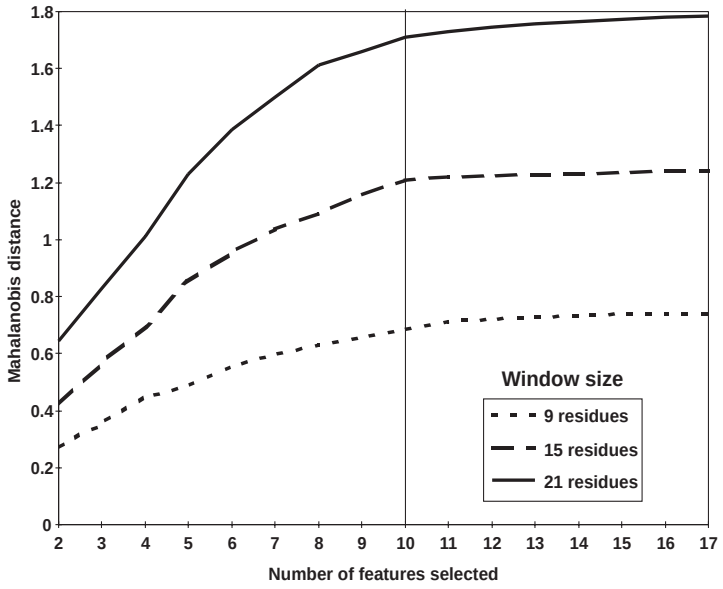

Figure 1: *BBS(p)* results on three data sets for $2 \leq p \leq 17$

For a given position within a CaN protein sequence, the following 24 attributes were calculated:

- the composition of each one of the 20 amino acids;
- the average hydropathy (using the Kite-Doolitle scale [22]);
- the average flexibility [33]; and
- the α and β hydrophobic moments [15].

These attribute statistics were calculated over a window consisting of the studied sequence position and $\lfloor n/2 \rfloor$ residues before and after it.

Due to the high similarity of the aligned sequences, some of the generated patterns were identical. To avoid a possible prediction bias, any repeated patterns were eliminated. The resulting data set was further reduced in order to have the same number of ordered and disordered patterns. When using a window size of 21 amino acids, the total number of remaining patterns after all these reductions was 1,720 (860 ordered and 860 disordered). This same number was maintained for the CaN-Nrl_3D data set by substituting 860 unique ordered patterns generated from the Nrl_3D database for the CaN ordered patterns. Several cleaned data sets computed using windows of various size ($9 \leq n \leq 51$) were analyzed to determine an appropriate window and to reduce the number of attributes as explained in the following sections.

4.2 BBS Preprocessing Experiments

To select an appropriate number of features from the original 24, a branch and bound optimal search was carried out using the Mahalanobis probabilistic distance criterion as explained in Section 3.2.1. Figure 1 shows the *BBS(p)* results on three 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.

Figure 1 also shows how the Mahalanobis distance grows with the size of a window over which attribute statistics are computed. Additional experiments were carried out to determine how the

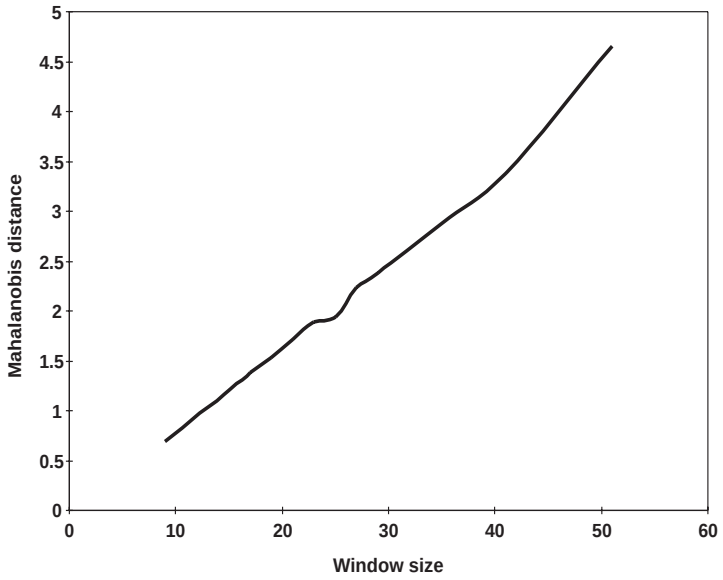

Figure 2: *BBS(10)* results on data generated using window sizes ranging from 9 to 51 residues

selection criterion value scales with window size. After determining that 10 features were an appropriate subset size, the *BBS(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 Fig. 2. This means that, within the range of window sizes studied, larger window size results in better class separation. Thus, an appropriate window size had to be selected based on its effect on the prediction range. A window size of 21 residues, the same one used in our previous work, was selected since it produces a 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.

Table 2 contains the 10 selected features for each of the three window sizes shown in Figure 1, along with the 10 features selected for the CaN-Nrl_3D data set using the chosen window size of 21, and the 10 features selected for the multi-family data set studied previously [29] (as reviewed in Section 2) and used to develop the MSFS-LDR(10) predictors.

Notice that, even though certain attributes were consistently selected (most notably the compositions of tyrosine (Y), histidine (H) and serine (S)), there are certain differences among the various groups of features, especially when they correspond to different data sets.

Data set	Window size	Selected features									
CaN-CaN	9	Y	H	S	A	F	D	C	N	P	Flexibility
	15	Y	H	S	A	F	K	T	G	R	E
	21	Y	H	S	A	D	K	T	G	P	Flexibility
CaN-Nrl_3D	21	Y	H	S	W	F	V	C	E	R	β-moment
LDR [29]	various	Y	H	S	W	D	K	C	E	Hydropathy	Flexibility

Table 2: Selected features with letters representing composition of the corresponding amino acids

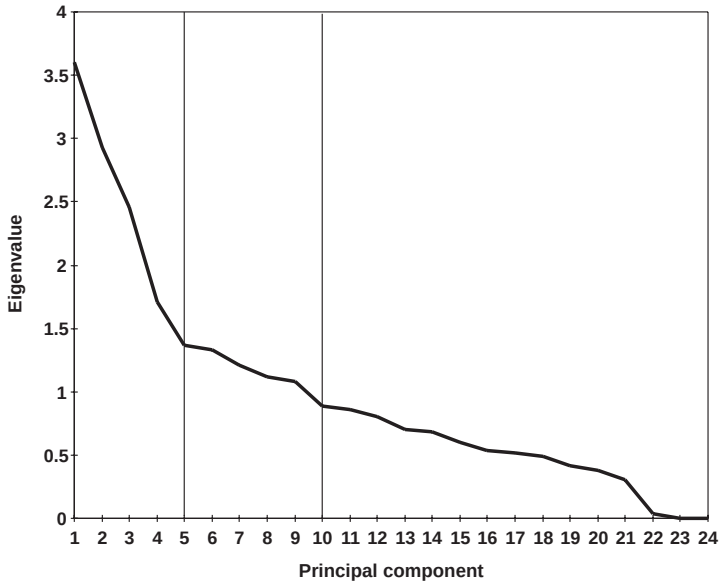

Figure 3: Eigenvalues comparison for deciding on an appropriate number of principal components

4.3 PCA Preprocessing Experiments

Principal components analysis was carried out on the CaN-CaN data set, obtaining the eigenvalues shown in Figure 3. 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. The available CaN data also allowed for a reduction to 10 dimensions as needed for fair comparison to *MSFS-LDR(10)* and to *BBS(10)* feature selection considered in this study. These choices appear reasonable since the first five and the first 10 principal components contribute 50% and 74% of the variance of the entire data set, respectively.

Even though this analysis was performed on the whole data set, feature extraction was later carried out independently on each of the five training subsets used for cross validation, as explained above (Section 3.3). Also, PCA was applied to the original LDR data set from which the MSFS-LDR(10) features were selected. The resulting eigenvector matrix was used to develop yet another predictor, called PCA-LDR(10), which was applied to the CaN-CaN and CaN-Nrl_3D data sets.

4.4 Neural Network Prediction Experiments

Cross-validation out-of-sample prediction results for 75 neural network predictors (15 for each set of features) trained on the CaN-CaN data set are summarized in Table 3. The corresponding experiments for data set CaN-Nrl_3D are shown in Table 4. The prediction results shown for each data set represent an average over three runs, each starting from a different random initialization of neural network parameters. The average and standard deviation values correspond to all 15 runs (five data sets and three runs per data set). The first column corresponds to the CaN-specific neural network based predictor developed using 10 features selected by the modified sequential forward search over patterns from seven different protein families [30]. The neural network (NN) architectures were selected by a trial and error process with u-v-w denoting a machine with u inputs, v units in a single hidden layer, and w output units. After the trial and error process, all the chosen neural network architectures were very similar, so the same number of hidden units (10) was used for all of them.

In the case of the CaN-CaN data set (Table 3), the PCA neural network predictors outperformed the predictors developed through feature selection techniques (MSFS-LDR(10) and BBS(10)). In general, family specific preprocessing produced predictors that were as good or better than their LDR-

CaN-CaN Data subset	Out-of-sample prediction accuracy				
	Multi-family preprocessing		Family-specific preprocessing		
	MSFS-LDR(10)	*PCA-LDR(10)*	*BBS(10)*	*PCA(5)*	*PCA(10)*
a	86%	90%	79%	89%	94%
b	86%	89%	78%	85%	92%
c	82%	89%	91%	87%	91%
d	88%	87%	90%	87%	91%
e	86%	90%	84%	85%	93%
Average	86%	89%	85%	87%	92%
Standard deviation	±2%	±1%	±5%	±2%	±1%
NN Architecture	10-10-1	10-10-1	10-10-1	5-10-1	10-10-1

Table 3: 5-cross validation, out-of-sample accuracy for predictors trained on the CaN-CaN data set

based counterparts. Notice how the PCA-LDR(10) predictor produces good results, although not as good as the PCA(10) neural network, which also supports the notion that family-specific feature reduction is significantly better in this case. This further suggests that different types of disorder depend on different relevant sequence attributes. The use of 10 principal components performed on CaN-specific patterns proved to be the best preprocessing technique of those applied to CaN-CaN data, producing a neural network that yielded an average out-of-sample prediction accuracy of about 92%.

On the other hand, the results shown in Table 4 for the CaN-Nrl_3D data set imply that features selected for a specific family are not quite as good for discriminating between family-specific disordered and general ordered residues. Indeed, the predictor developed using the MSFS-LDR(10) features had the least drop in performance when trained on the CaN-Nrl_3D data. Also, notice that all the PCA approaches, which gave the best results on the CaN-CaN data set, were the ones that suffered the greatest performance reduction when applied on the CaN-Nrl_3D set.

Even the lowest prediction accuracy reported in Tables 3 and 4 (e.g. 79 ± 2%) is significantly better than accuracy observed previously (e.g. 73 ± 2% [29]) when our predictors were applied to a

CaN-Nrl_3D Data subset	Out-of-sample prediction accuracy				
	Multi-family preprocessing		Family-specific preprocessing		
	MSFS-LDR(10)	*PCA-LDR(10)*	*BBS(10)*	*PCA(5)*	*PCA(10)*
a	83%	79%	83%	75%	81%
b	85%	77%	77%	72%	80%
c	88%	80%	88%	72%	80%
d	86%	78%	85%	65%	81%
e	83%	79%	84%	69%	81%
Average	85%	79%	83%	71%	80%
Standard deviation	±2%	±2%	±5%	±2%	±2%
NN Architecture	10-10-1	10-10-1	10-10-1	5-10-1	10-10-1

Table 4: 5-cross validation, out-of-sample accuracy for predictors trained on the CaN-Nrl_3D data set

collection of LDRs from a diverse set of proteins, and the best values in Tables 3 and 4 show very substantial improvement. This observation suggests that family-specific predictors can have better performance than general ones, probably because the feature sets of families of disordered regions are less diverse compared to feature sets from unrelated regions of disorder. This in turn suggests that the feature set of the LDR in CaN has remained relatively invariant over evolutionary time. The data further show that feature selection performed on multi-family data result in predictors that are better suited for discriminating between general order and family-specific disorder. Finally, the data show that it was very easy to differentiate CaN-specific disorder from CaN-specific disorder, regardless of the preprocessing method used.

5 Summary and Conclusions

5.1 Corroborating Studies

The concept that many amino acid sequences fail to fold completely on their own is supported not only by a series of specific examples from natural proteins [3, 6, 11, 16, 19, 20, 21, 26, 27, 31, 32, 34], but also by attempts to design rigidly folded proteins and by theoretical studies. With regard to protein design experiments, it is relatively simple to construct amino acid sequences that fold into ensembles of collapsed, but still flexible structures [9], but rather more difficult to design proteins that fold into rigidly packed, unique 3D structures, even when diversity synthesis is used to explore tens of thousands of sequences in locally promising regions of sequence space [12]. Thus, most if not all currently designed protein sequences fail to fold into unique 3D structures with fully rigid side chain packing. With regard to theoretical investigations, simple lattice models suggest that uniquely folding sequences represent a vanishing small fraction of sequence space [1, 2, 18]. Overall, the specific examples and the corroborating experimental and theoretical studies suggest that natively unfolded sequences are apt to be common.

The specific examples of natively unfolded protein given above were chosen because they were known to be involved in function. In each case, the disordered protein or the disordered region acquired order upon binding with its partner. For these examples, the partners range from being small molecules [3], to being other copies of the same protein [27], to being other proteins [6, 11, 20, 21, 34], to being nucleic acids [16, 19, 31, 32]. These selected specific examples suggest that disorder-to-order transitions upon complex formation, also called *induced folding* [32], encompass a broad range of biological specificities.

Given the importance of being unfolded [26], as illustrated by the many specific examples and various observations given above, studies to determine whether protein disorder can be predicted from amino acid sequence seem warranted. We previously designed several neural network based predictors trained using a collection of proteins whose disordered regions were identified by missing electron density in crystal structures [29]. One of these predictors, trained using long disordered regions (LDRs), was then compared with a second predictor trained on a set of LDRs from homologous proteins related to CaN rather than using different disordered proteins [30]. Comparison of these predictors, both trained using the same feature set, but from different disordered sequences, indicated similar overall 5-cross-validated, out-of-sample, residue-by-residue prediction accuracies and similar false positive error rates [30].

5.2 Testing the Predictors

An ongoing effort has been to find reports describing additional examples of natively unfolded protein and then to apply our predictors one-by-one to these out-of-sample test cases. In the course of such studies we noticed that the CaN-based predictor often failed for unfolded proteins containing nucleic acid binding regions whereas the LDR predictor gave better results and *vice versa* for some disordered

regions involved in protein binding (manuscript in preparation). Qualitatively such results seemed reasonable because the multi-family LDR training set contained disordered regions that bind nucleic acids, whereas the CaN unfolded regions are involved in protein binding.

The obvious extrapolation of these comparisons is that different types of disordered regions (having different functions) could exhibit different sets of underlying features determinant of the unfolding. If so, carrying out predictions on a single type of disorder should lead to significant improvements in the predictions. To test this, we carried out studies on the predictions of a single family of protein disorder, namely 95 amino acid LDR in CaN [20]).

To avoid a possible prediction bias when testing out-of-sample accuracy on homologous data, it was necessary to eliminate all repeating patterns (data cleaning step discussed in Section 4.1). Further homology reduction was accomplished by substituting ordered sequences from the Nrl_3D database for the ordered regions of CaN. The fact that the ordered parts of the CaN sequences showed far greater similarity than the disordered regions justified this substitution as a means of reducing homology effects. Still, the results of the current study imply that, although family-specific preprocessing is the best approach for family-specific prediction, the performance of such a predictor is limited to this type of disorder/order discrimination. To be able to predict a family-specific disordered region among generic ordered residues, a more robust predictor is needed. This study suggests that performing feature selection on multi-family sequence data and using general ordered patterns for training is likely to be the best way to obtain such a predictor.

5.3 Implications for Protein Structure and Function

Ordered regions are classified into structural subtypes: helix, sheet, loops, reverse turns, etc., based on visually obvious structural criteria. The results of the present paper suggest that disordered regions may likewise have subtypes, but it is unclear at this time how best to classify such disordered sequences. The feature selection and feature reduction results presented herein suggest that studies aimed to reveal the most appropriate mappings from feature space into structure space should be tried. Such an approach might be useful not only for disordered sequences, but perhaps also as a means to improve the classification of the ordered sequences as well.

Disordered regions can underlie numerous different protein functions [14, 29, 30]. For example, disorder facilitates protease digestion, which is often required for enzyme activation and which could also be used to regulate protein turn-over. Also, disorder evidently improves the ability of one protein to bind to many different targets, as suggested both for $p21^{Waf1/Cip1/Sdi1}$ [21] and earlier for calmodulin [10, 23, 24, 36]. In addition, as emphasized here, disorder can also play a role in specific binding. In this case, the key event is a disorder-to-order transitions upon complex formation.

It might seem to be a small matter whether a protein folds first and then forms a complex (e.g. prior folding) or whether folding and complex formation occur concomitantly (e.g. coupled folding). However, this apparently small distinction has enormous consequences. Coupled folding enables the biologically useful combination of high specificity and low affinity [3, 14, 31] or low specificity and high affinity [14], whereas, to the first order of approximation, prior folding does not allow such a separability of affinity and specificity [14]. That is, for prior folding, affinity and specificity are linked; both are high or both are low. On the other hand, for coupled folding, affinity and specificity are readily separable through evolution and natural selection [14]. Thus, the apparently small difference of prior folding versus coupled folding has a radical affect on the biological potential of a given complex formation event.

If a region of sequence undergoes a disorder-to-order transition upon complex formation, why should such a sequence be predictable as a region of disorder? Indeed, it seems to be a conundrum that we are evidently able to predict disorder for sequences that become ordered upon binding with partners.

The solution to this false conundrum is simple. Obviously, our predictors can only define likelihoods

or tendencies. In such a circumstance, the local tendency for a region of sequence to be disordered can be overcome when that local region interacts with something else. That something else could be another region in the same protein, in which case the prediction of disorder would lead to a false positive. Alternatively, that something else could be another protein or an entirely different class of molecules such as DNA or RNA, in which case a correct prediction would result.

Interactions between a given region of sequence and other regions that are well separated along the sequence are frequently called nonlocal interactions. Nonlocal interactions are believed to lower the predictability of secondary structure [5]. In the same way, such nonlocal interactions could lower the accuracies of the prediction of disorder. However, if a region of disorder were very long or if a region of disorder were to have the function of binding to nonprotein such as nucleic acid, then nonlocal interactions with other parts of the same protein might be less likely to induce order. In these circumstances, predictions of disorder are likely to be more accurate.

Our predictions of disorder may have practical uses as well. For example, the prediction that a protein is very likely to be entirely disordered would indicate that NMR would probably be a better approach than x-ray crystallography for determining its 3D structure. The prediction of substantial regions of disorder and other regions of order would indicate that protease digestion experiments would be informative. Researchers involved in studies on proteins that may contain regions of disorder are encouraged to contact us to arrange for the application of our predictors to their sequences. This may yield immediate practical benefit and may also, over time, provide information for improving our predictors.

Acknowledgments

We thank Mr. B.O. Oswalt-Rumsey and Mr. L. W. Becker, participants in the 1997 WSU/NSF Teacher Institute for Science/Mathematics Education through Engineering Experiences program, who helped running some of the feature selection experiments. We also thank the NSF grant ESI-9254358 for providing financial support for their participation at our project. Special thanks are due to Partek Inc. whose software product was used for data analysis and to Dr. R. Drossu whose neural network simulator was used for prediction.

References

[1] 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.

[2] Abkevich, V. I., Gutin, A. M., Shakhnovich, E. I., "Computer Simulations of Prebiotic Evolution," *Pacific Symposium on Biocomputing,* 2:27–34, 1997.

[3] 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," in *Mobility and Function in Proteins and Nucleic Acids,* ed. by F. M. Richards, *Ciba Foundation Symposium,* 93:4–24, 1982.

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

[5] 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.

[6] 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.

[7] Bairoch, A. and Apweiler, R., "The Swiss-Prot Protein Sequence Data Bank and its New Supplement TrEMBL," *Nucleic Acids Res.,* 24:21–25, 1996.

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

[9] Betz, S. F., Bryson, J. W. and DeGrado, W. F. "Native-like and Structurally Characterized Designed α-Helical Bundles," *Current Opin. Struct. Biol.,* 5:457–463, 1995.

[10] Crivici, A., and Ikura, M. "Molecular and Structural Basis of Target Recognition by Calmodulin," Annu. Rev. Biophys. Biomol. Struct. 24: 95–116, 1995.

[11] 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.

[12] Davidson, A. R., Lumb, K. J., and Sauer, R. T. "Cooperatively Folded Proteins in Random Sequence Libraries," *Nature Struct. Biol.,* 2:856–864, 1995.

[13] Dolgikh, D. A., Gilmanshin, R. I., Brazhnikov, E. V., Bychkova, V. E., Semisotnov, G. V., Yu, S., and Ptitsyn, O. B. "α-Lactalbumin: Compact State with Fluctuating Tertiary Structure?" *FEBS Lett.* 136:311–315, 1981.

[14] 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.

[15] 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.

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

[17] George, D. G., Dodson, R. J., Garavelli, J. S., Haft, D. H., Hunt, L. T., Marzec, C. R., Orcutt, B. C., Sidman, K. E., Srinivasarao, G. Y., Yeh, L. S-L., Arminski, L. M., Ledley, R. S., Tsugita, A., and Barker, W. C., "The Protein Information Resource (PIR) and the PIR-International Protein Sequence Database," *Nucleic Acids Res.,* 25:24–28, 1997.

[18] Gutin, A. M., Abkevich, V. I. and Shakhnovich, E. I. "Evolution-like Selection of Fast-Folding Model Proteins," *Proc. Natl. Acad. Sci. USA,* 92:1282–1286, 1995.

[19] 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.

[20] 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.

[21] 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,* 93:11,504–511,509, 1996.

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

[23] Meador, W. E., Means, A. R., and Quiochio, F. A., "Modulation of Calmodulin Plastcity in Molecular Recognition on the Basis of X-ray Diffraction Structures," *Science,* 262:1718–1721, 1993.

[24] O'Neil, K. T. and DeGrado, W. F. "How Calmodulin Binds its Targets: Sequence-Independent Recognition of Amphiphilic α-helices," *TIBS,* 15:59–64, 1990.

[25] 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.

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

[27] Riek, R., Hornemann, S., Wider, G., Glockshuber, R. and Wüthrich, K. "NMR Characterization of the Full-length Recombinant Murine Prion Protein mPrP(23-231)," *FEBS Lett* 413:282–288, 1997.

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

[29] 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.

[30] 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.

[31] Schulz, G. E., "Nucleotide Binding Proteins," *Molecular Mechanisms of Biological Recognition,* ed. by M. Balaban, in Elsevier/North-Holland Biomedical Press, 79–94, 1977.

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

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

[34] 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.

[35] 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.

[36] Zhang, M. and Tanake, T., "Calcium-induced Conformational Transition Revealed by the Solution Structure of apoCalmodulin," *Nature Struct. Biol.,* 2:758–767, 1995.