

Identification of Intrinsically Unstructured Regions in Proteins Using Primary Structure

Mary Qu Yang

National Human Genome Research Institute
National Institutes of Health
U.S. Department of Health and Human Services
5625 Fishers Ln 5N01, Rockville, MD 20850 USA
yangma@mail.NIH.GOV

Jack Y. Yang

Harvard University, Harvard Medical School
and Massachusetts General Hospital
Department of Radiation Oncology
Boston, Massachusetts, 02114 USA
jyang@hadron.mgh.harvard.edu

Abstract

Most proteins function only when folded into a particular 3D configuration. Recently, a class of proteins has been discovered that do not fold into any particular configuration; these are known as Intrinsically Unstructured (IU) proteins. We construct a classifier to identify IU regions in proteins based on features derived from protein sequence information alone, and evaluate it on out-of-sample data. Our results indicate that the resulting classifier represents a viable alternative to existing IU classifiers.

1 Introduction

Most proteins function only when folded into a particular 3D configuration. Recently, a class of proteins has been discovered that do not fold into any particular configuration; rather, they exist as dynamic ensembles in their native state. These proteins have been variously called natively unfolded, natively disordered or Intrinsically Unstructured (IU) proteins [12, 11, 5, 13]. Unlike regular proteins, which unfold and lose their ability to function when subjected to environmental challenges such as detergents, urea, or heat [2], IU proteins continue to function under such conditions, as they do not have to be folded into a particular configuration in order to carry out their function.

IU proteins have been associated with a wide range of protein functions such as molecular recognition, molecular assembly/disassembly and protein modification [12, 11, 2]. They also play a central role in diseases characterized by protein misfolding and aggregation [2, 16, 12]. Furthermore, the identification of such proteins can aid both structure determination and sequence alignment, and may aid in drug design.

IU protein regions can be identified through anal-

ysis of a protein tertiary structure. Traditionally, the tertiary structure of proteins is determined using experimental methods such as X-ray crystallography, Overhauser-Effect Enhanced Nuclear Magnetic Resonance spectroscopy (NMR), and Circular Dichroism Spectra (CD). However, these experimental methods are usually time consuming and expensive, and often have their own limitations and problems. For instances, X-ray crystallography may run into difficulty because some proteins are difficult to crystallize, while the use of NMR to determine tertiary structure is limited to proteins with molecular weights of about 15,000 or less, due to the fact that for large molecules the NMR spectrum can be extremely complex.

It would be thus advantageous to develop computational approaches to identifying IU regions on the basis of protein primary structure alone. Several computational approaches to predicting IU regions from sequence information have already been developed, including PONDR [12, 11, 7, 14], disEMBL [8] and GlobPlot [9]. PONDR and disEMBL are mainly based on neural networks, while GlobPlot is based on a single attribute - disorder propensity. Our approach, discussed in Section 3, is based on a hierarchical classifier. We begin by describing how the features that we use in our classifier were selected.

2 Feature Extraction and Selection

In order to classify a given amino acid residue A as either belonging to or not belonging to an IU region, we center a window of length W at A ; included in this window are the $(W-1)/2$ residues preceding A in the sequence, as well as the $(W-1)/2$ residues following A . We then compute features over this window, and feed the calculated feature values into a classifier, which then outputs a decision (either IU or NOT IU).

Based on an in-depth analysis of the effectiveness of various features in discriminating IU from non-IU regions [18], we extracted the following features:

- **First-order statistics** - For each residue position j and each amino acid type t_i , we define the feature $P_j(t_i)$ as the fraction of amino acids of type t_i contained in a window of length W centered at the j^{th} residue, i.e.

$$P_j(t_i) = \frac{\sum_{k=k_{\min}}^{k_{\max}} 1_{\{a_k=t_i\}}}{k_{\max} - k_{\min} + 1} \quad i = 1, \dots, 20 \quad (1)$$

where a_1, \dots, a_L is the amino acid sequence, t_1, \dots, t_{20} are the 20 amino acid types, $1_{\{a_k=t_i\}}$ is 1 if the amino acid at the k^{th} position is t_i and 0 otherwise, and k_{\min}, k_{\max} respectively specify the indices corresponding to the first and last residue in the window, i.e.

$$\begin{aligned} k_{\min} &= \max\{1, j - (W - 1)/2\} \\ k_{\max} &= \min\{L, j + (W - 1)/2\} \end{aligned} \quad (2)$$

Since there are 20 amino acid types, for each residue position j , there are 20 features $P_j(t_i)$, corresponding to $t_i, i = 1, \dots, 20$.

- **Hydropathy** - a measure of the relative hydrophobicity of an amino acid. We consider 4 hydropathy scales: the Kyte-Doolittle [6], Eisenberg [3], Engelman [4], and Liu-Deber [10] scales. The average hydropathy \bar{H}_j associated to the j^{th} amino acid in the sequence is the average hydropathy of all residues in a window of width W centered at the j^{th} residue, that is

$$\bar{H}_j = \frac{\sum_{k=k_{\min}}^{k_{\max}} H(a_k)}{k_{\max} - k_{\min} + 1} \quad (3)$$

where a_1, \dots, a_L is the amino acid sequence, $H(a_k)$ is the hydropathy associated with amino acid a_k , and k_{\min}, k_{\max} are defined by (2).

- **Complexity** - essentially the entropy of the probability distribution over the different amino acid types contained in the window of length W . Thus the complexity value associated to the residue at the j^{th} position of the sequence is:

$$C_j = -\sum_{i=1}^{20} P_j(t_i) \log_2 P_j(t_i) \quad (4)$$

where t_1, \dots, t_{20} are the 20 amino acid types, and $P_j(t_i)$ is calculated by Equation (1). In [18], it was noted that IU regions tend to have low complexity values.

After some initial experiments, we decided to add several additional features:

- **IU Propensity** - a measure of how likely an amino acid is to be unfolded. We consider two different scales of IU Propensity, the Russell/Linding and Deleage/Roux scales [9], as shown in Table 1. The average IU Propensity \bar{R}_j associated to the j^{th} amino acid in the sequence is the average hydropathy of all residues in a window of width W centered at the j^{th} residue, that is

$$\bar{R}_j = \frac{\sum_{k=k_{\min}}^{k_{\max}} R(a_k)}{k_{\max} - k_{\min} + 1} \quad (5)$$

where a_k is the amino acid at the k^{th} position, $R(a_k)$ is the hydropathy associated with that amino acid, and where k_{\min}, k_{\max} are defined by (2).

Table 1: Russell/Linding and Deleage/Roux IU propensity scales.

Residue	IU Propensity	
	Russell/Linding	Deleage/Roux
A	-0.26154	-0.275
C	-0.015152	-0.1255
D	0.22763	0.4645
E	-0.20469	-0.2745
F	-0.22557	-0.497
G	0.43323	0.6675
H	-0.0012174	0.135
I	-0.42224	-0.515
K	-0.100092	-0.0495
L	-0.33793	-0.4385
M	-0.22590	-0.4765
N	0.22989	0.479
P	0.55232	1.117
Q	-0.187677	-0.055
R	-0.17659	-0.179
S	0.14288	0.2965
T	0.0088780	0.145
V	-0.38618	-0.7055
W	-0.243375	-0.257
Y	-0.20751	0.0825

- **Second order statistics** - For each residue position j and each amino acid types t_i, t_m , we define the feature $P_j(t_i, t_m)$ as the fraction of adjacent amino acid pairs in the window of length W centered at the j^{th} residue such that the first member of the pair is of type t_i and the second member

of the pair is of type t_m , i.e. for $i, m = 1, \dots, 20$

$$P_j(t_i, t_m) = \frac{\sum_{k=k_{\min}}^{k_{\max}-1} 1_{\{a_k=t_i\}} 1_{\{a_{k+1}=t_m\}}}{k_{\max} - k_{\min}} \quad (6)$$

where a_1, \dots, a_L is the amino acid sequence, t_1, \dots, t_{20} are the 20 amino acid types, $1_{\{a_k=t_i\}}$ is 1 if the amino acid at the k^{th} position is t_i and 0 otherwise, and where k_{\min}, k_{\max} are defined by (2).

The first and second order statistics of the 20 amino acids account for $20+20^2 = 400$ features. To compress the number of features to a more manageable level, we introduce a 9-ary encoding scheme in which each amino acid is classified to one of the 9 groups shown in Table 2 based on its physiochemical properties; this encoding is similar to the substitution matrix used in BLASTp.

Table 2: 9-ary encoding scheme for amino acids.

Group	Residues	Description
1	C	Highly conserved
2	M	Hydrophobic
3	N, Q	Amides, polar
4	D, E	Acids, positive, polar
5	S, T	Alcohols
6	P, A, G	Aliphatic, small
7	I, V, L	Aliphatic
8	F, Y, W	Aromatic
9	H, K, R	Bases, charged

One of the advantages of the 9-ary encoding scheme is that the resulting first and second order statistics comprise $9+9^2 = 90$ features, a considerable reduction as compared to 400 for the unencoded case. If as a result of this encoding scheme the number of features remaining after the feature selection step is reduced as compared to the unencoded case, the benefits include lower algorithmic complexity, and possibly a reduction in the generalization error.

The first and second order statistics for the 9-ary encoding are defined as follows:

- **First-order statistics** - For each residue position j and each of the 9-ary encoding groups g_i , we define the feature $P_j^G(g_i)$ as the fraction of amino acids belonging to the 9-ary encoding group g_i contained in a window of length W centered at the j^{th} residue, i.e.

$$P_j^G(g_i) = \frac{\sum_{k=k_{\min}}^{k_{\max}} 1_{\{G(a_k)=g_i\}}}{k_{\max} - k_{\min} + 1} \quad i = 1, \dots, 9 \quad (7)$$

where a_1, \dots, a_L is the amino acid sequence, g_1, \dots, g_9 are the 9-ary encoding groups, $G(a_k)$ is a function mapping residue a_k to one of the 9-ary encoding groups, and where k_{\min}, k_{\max} are defined by (2).

- **Second order statistics** - For each residue position j and 9-ary encoding groups g_i, g_m we define the feature $P_j^G(g_i, g_m)$ as the fraction of adjacent amino acid pairs in the window of length W centered at the j^{th} residue such that the first member of the pair is of type g_i and the second member of the pair is of type g_m , i.e. for $i, m = 1, \dots, 9$

$$P_j^G(g_i, g_m) = \frac{\sum_{k=k_{\min}}^{k_{\max}-1} 1_{\{G(a_k)=g_i\}} 1_{\{G(a_{k+1})=g_m\}}}{k_{\max} - k_{\min}} \quad (8)$$

where a_1, \dots, a_L is the amino acid sequence, g_1, \dots, g_9 are the 9-ary encoding groups, $G(a_k)$ is a function mapping residue a_k to one of the 9-ary encoding groups, and where k_{\min}, k_{\max} are defined by (2).

At this point, we generated a total of 517 features:

20	first order statistics (unencoded)
400	second order statistics (unencoded)
9	first order statistics (9-ary encoding)
81	second order statistics (9-ary encoding)
1	average complexity C_j
4	average hydrophathy \bar{H}_j (4 scales)
2	IU Propensity \bar{R}_j (2 scales)

As this is far too many features, we have incorporated a feature selection step, which consists of choosing the features x_{ij} that yield the largest values of $D(x_{ij})$ [15], where

$$D(x_{ij}) = \frac{|m_{\text{IU}} - m_{\text{IU}}|}{\sqrt{\sigma_{\text{IU}}^2 + \sigma_{\text{IU}}^2}} \quad (9)$$

where m_{IU} and σ_{IU}^2 are respectively the sample mean and variance of feature x_i , restricted to instances that are labeled as IU, whereas m_{IU} and σ_{IU}^2 are respectively the sample mean and variance of feature x_{ij} , restricted to instances that are labeled as NOT_IU.

Due to correlations between the 4 hydrophathy scale features, we decided to include a maximum of one hydrophathy scale in the feature set, specifically the Kyte-Doolittle scale, which tended to yield the largest values of (9). For similar reasons, we included a maximum of one IU propensity scale in the feature set, specifically the Russell/Linding scale, which tended to yield the largest values of (9). Subject to these restrictions, we

selected the features that yielded the largest values of (9) and ended up with a total of 59 features, including a number of first- and second-order statistics, average hydrophathy, average IU propensity, and complexity.

3 Recursive Maximum-Contrast Tree Classifier

The classifier that we use to discriminate IU from non-IU regions is based on the Recursive Maximum-Contrast Tree (RMCT) [17], a hierarchical top-down clustering algorithm. The output of this algorithm is a clustering tree in which the root node contain all the training instances, and the leaf nodes in general contain only a single training instance. Once the hierarchical clustering tree has been constructed from the training data, it can be used to classify an arbitrary test instance \vec{x}^{test} as follows:

1. Find the leaf node L of the tree containing the training instance \vec{x}_i that is closest to \vec{x}^{test} in the sense of minimizing the distance

$$d(\vec{x}^{\text{test}}, \vec{x}_i) = \sqrt{\sum_{j=1}^{59} (x_j^{\text{test}} - x_{ij})^2}$$

where \vec{x}^{test} and \vec{x}_i are vectors with elements x_j^{test} and x_{ij} , respectively.

2. Go up the tree from the leaf node L until a node N with at least K training instances is reached, where K is a parameter.
3. Of the training instances in node N , select the K training instances that are closest to \vec{x}^{test} , and assign \vec{x}^{test} the majority class of these K training instances.

The algorithm depends on a parameter, K , the number of neighbors used in the classification rule. It thus functions similarly to the K -Nearest-Neighbor classification algorithm [1], where the neighbors are determined by the tree structure.

4 Dataset

Our data, obtained from the Protein Data Bank (PDB), consists of 290 totally ordered protein sequences, which do not contain any IU regions, as well as 290 protein sequences having one or more IU regions. The total number of amino acid residues in the totally ordered sequences was 67,522. IU regions can be identified in X-ray crystallography data as regions having a missing backbone coordinate, indicating that no electron density was measured; only data with resolution better than 2 angstroms was used to ensure that

missing coordinates were not due to poor data quality. Furthermore, the selected sequences were constrained to have pair-wise sequence identities of less than 25%.

5 Choice of Parameters W , K

The algorithm as described has two parameters, the feature window length W and the number of neighbors K used by the classification rule. We ran a series of experiments to determine the optimal values of these parameters for our dataset. Table 3 and Figure 1 show the true positive rate (fraction of IU residues classified as IU), true negative rate (fraction of NOT_IU residues classified as NOT_IU), and average accuracy (defined as the average of the true positive rate and the true negative rate) as a function of the number of neighbors K used by the classification rule, for a fixed value of W (in this case, $W = 13$). We observe that the average accuracy reaches a maximum at $K = 21$, but does not vary much beyond $K = 17$. Figure 2 shows the average accuracy, true positive rate, and true negative rate of the classifier as a function of the feature window length W , for a fixed value of K (in this case, $K = 17$). We observe that the overall accuracy reaches a maximum at $W = 13$. On the basis of these results and others, we concluded that the highest values for accuracy and the true positive rate (which in many applications is considered more important than the true negative rate) were obtained with $W = 13$ and $K = 17$.

Table 3: Performance of IU classifier as a function of the number of neighbors K used in classifying out-of-sample data. The average accuracy, true positive, and true negative rates are shown (the true positive rate is the fraction of IU residues classified as IU, while the true negative rate is the fraction of non-IU residues classified as non-IU).

K	TP	TN	Average
3	0.8031	0.7183	0.7607
5	0.8014	0.7336	0.7675
7	0.804	0.7349	0.7694
9	0.8106	0.7414	0.776
11	0.8036	0.7453	0.7744
13	0.8015	0.749	0.7752
15	0.8023	0.7514	0.7768
17	0.8025	0.7536	0.778
19	0.8024	0.7575	0.7799
21	0.8005	0.7597	0.7801
23	0.7964	0.7624	0.7794
25	0.7947	0.7641	0.7794

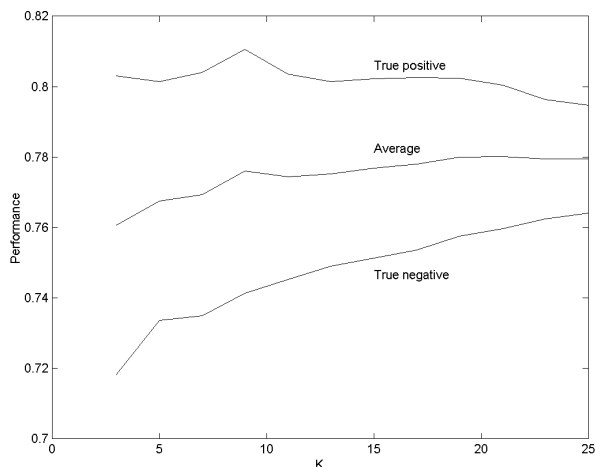


Figure 1: Performance of IU classifier as a function of the number of neighbors K used in classifying out-of-sample data, where W was fixed at 13. The true positive and true negative rates are shown, along with their average (the true positive rate is the fraction of IU residues classified as IU, while the true negative rate is the fraction of NON_IU residues classified as NON_IU).

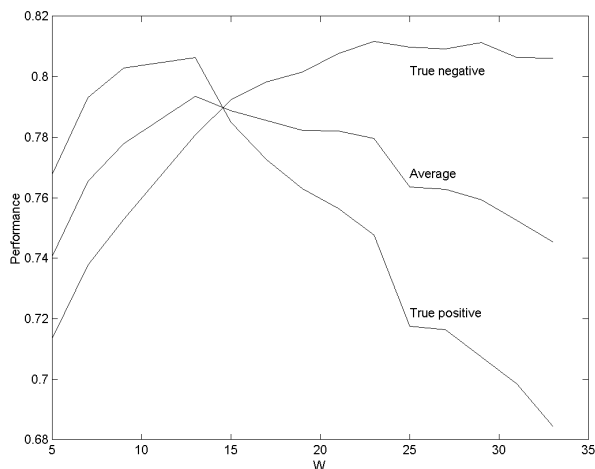


Figure 2: Performance of IU classifier as a function of the feature window length W , where the number of neighbors K was fixed at 17. The true positive and true negative rates are shown, along with their average (the true positive rate is the fraction of IU residues classified as IU, while the true negative rate is the fraction of NON_IU residues classified as NON_IU).

6 Results

We compared the RMCT classifier to PONDR [12, 11, 14], disEMBL [8] and GlobPlot [9]:

- The results of a comparison of the RMCT classifier to the PONDR VLXT classifier based on 255 out-of-sample proteins is reported in Table 4. While the PONDR classifier has a higher true positive rate than the RMCT classifier, the RMCT classifier has both a higher true negative rate and a higher overall accuracy.
- The results of a comparison of the RMCT classifier to the DisEMBL and GlobPlot classifiers for a number of individual out-of-sample proteins is reported in Table 5. These results show that the RMCT classifier compares favorably to both the DisEMBL and GlobPlot classifiers.

Table 4: A comparison of the RMCT IU classifier to the PONDR IU classifier based on 255 out-of-sample proteins. TP and TN are respectively the true positive and true negative rates, while Accuracy represents the overall accuracy.

Classifier	TP	TN	Accuracy
RMCT	73.63 ± 0.40	78.48 ± 1.40	74.85 ± 0.57
PONDR	77.72 ± 0.72	64.44 ± 2.18	72.92 ± 0.91

7 Conclusions

The RMCT classifier compares favorably to several existing classifiers that identify IU regions based on protein sequence information. In our experiments, we found that it attained true positive rates and true negative rates that generally exceeded 70%, and were often much higher.

Acknowledgments

We thank Dr. Laura L. Elnitski and Dr. Craig W. Codrington for many valuable discussions. This research was conducted while the corresponding author (Dr. Jack Yongsheng Yang) was an IU Medical School Post-Doctoral Fellow and Faculty Member of IUPUI (Indiana University Purdue University Indianapolis); and while the first author (Dr. Mary Qu-Xuanyuan Yang) was an interdisciplinary Bilsland Dissertation Fellow for biological physics and computer engineering dual-degree at Purdue University (Main Campus in West Lafayette) and NIH Post-Doctoral Fellow for National Human Genome Research.

Table 5: A comparison of the RMCT IU classifier to the DisEMBL and Globplot IU classifiers on out-of-sample data. TP and TN are respectively the true positive and true negative rates, while AVG represents the average of the two.

Proteins	RMCT			DisEMBL			GlobPlot		
	TP	TN	AVG	TP	TNR	AVG	TP	TN	AVG
PhosCarboxykinase	75%	95%	85%	0%	91%	46%	0%	100%	5%
Avi.Sar.Vir.Cat.Do	81%	75%	78%	0%	79%	40%	0%	100%	5%
Cyanide	100%	88%	94%	100%	85%	92%	100%	94%	9%
Rent.-Bind.Pig Plas	86%	99%	93%	0%	83%	42%	0%	95%	4%
Nat. Plas. Act.Inhib	100%	70%	85%	100%	78%	89%	0%	100%	5%
N-Ter.Lobe. Lactof	100%	70%	85%	100%	78%	89%	0%	100%	5%
Calp.D.ViCa.Boun	91%	96%	94%	100%	85%	93%	100%	98%	9%
Therm.Ser Protease	78%	77%	77%	0%	78%	39%	0%	100%	5%
Bov.Pur.Nuc.PhosC	86%	98%	92%	100%	67%	83%	0%	93%	4%
LFuc.PhosAldolase	100%	98%	99%	100%	83%	92%	88%	100%	9%
N-Ter.Dom.Sec18p	82%	81%	82%	40%	88%	64%	40%	98%	6%

References

- [1] T. Cover and P. E. Hart. Nearest neighbor pattern classification. *IEEE Trans. Information Theory*, 13(1):21–27, 1967.
- [2] A. K. Dunker and Z. Obradovic. The protein trinity–linking function and disorder. *Nature Biotechnology*, 19(9):805–806, September 2001.
- [3] D. Eisenberg, E. Schwarz, M. Komaromy, and R. Wall. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J. Mol. Biol.*, 179:125–142, 1984.
- [4] D. M. Engelman, T. A. Steitz, and A. Goldman. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu. Rev. Biophys. Biophys. Chem.*, 15:321–353, 1986.
- [5] L.M. Iakoucheva, P. Radivojac, C. J. Brown, T. R. O’Connor, J. G. Sikes, Z. Obradovic, and A. K. Dunker. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.*, 32(3):1037–1049, February 2004.
- [6] J. Kyte and R. Doolittle. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, 157:105–132, 1982.
- [7] X. Li, P. Romero, M. Rani, A. K. Dunker, and Z. Obradovic. Predicting protein disorder for N-, C-, and internal regions. *Genome Informatics*, 10:30–40, 1999.
- [8] R. Linding. Protein disorder prediction: Implications structural proteomics. *Protein Structure*, 11:1316–1317, 2003.
- [9] R. Linding, R. B. Russell, V. Neduva, and T. J. Gibson. Globplot: exploring protein sequences for globularity and disorder. *Nucleic Acid Res.*, 31:3701–8, 2003.
- [10] L.-P. Liu and C.M. Deber. Guidelines for membrane protein engineering derived from de novo designed model peptides. *Biopolymers (Peptide Science)*, 5(47):41–62, 1998.
- [11] Kang Peng, Slobodan Vucetic, Predrag Radivojac, Celeste J. Brown, A. Keith Dunker, and Zoran Obradovic. Optimizing long intrinsic disorder predictors with protein evolutionary information. *Journal of Bioinformatics and Computational Biology*, 3(1):35–60, February 2005.
- [12] P. Radivojac, Z. Obradovic, D. K. Smith, G. Zhu, S. Vucetic, C. J. Brown, J. D. Lawson, and A. K. Dunker. Protein flexibility and intrinsic disorder. *Protein Science*, 13(1):71–80, January 2004.
- [13] P. Romero and A. K. Dunker. Intelligent data analysis for protein disorder prediction. *Artificial Intelligence Review*, 14, 2000.
- [14] P. Romero, Z. Obradovic, X. Li, E. Garner, C. Brown, and A. K. Dunker. Sequence complexity of disordered protein. *Proteins: Struct. Funct. Gen.*, 42:38–48, 2001.
- [15] V. V. Solovyev and K. S. Makarova. Protein sequence classification based on oligopeptide frequency analysis and its application to search for functional sites and to domain localization. *Computer Applications in the Biosciences*, 9:17–24, 1993.
- [16] V. N. Uversky and A. L. Fink. *Protein Misfolding, Aggregation and Conformational Diseases*. Springer, 2005.

- [17] Mary Qu Yang, Okan K. Ersoy, and Jack Y. Yang. Sequential bifurcation approach to learning protein functional classes. In *Advances in Bioinformatics and its Applications*, volume 8 of *Series in Mathematical Biology and Medicine*, pages 264–275. World Scientific, 2005.
- [18] Mary Qu-Xuanyuan Yang. *Predicting Protein Structure and Function Using Machine Learning Methods*. PhD thesis, Purdue University, West Lafayette, Indiana, 2005.