

# A Visual Analytics Approach for Protein Disorder Prediction

Jaegul Choo and Fuxin Li and Keehyoung Joo and Haesun Park

**Abstract** In this chapter, we present a case study of performing visual analytics to the protein disorder prediction problem. Protein disorder is one of the most important characteristics in understanding many biological functions and interactions. Due to the high cost to perform lab experiments, machine learning algorithms such as neural networks and support vector machines have been used for its identification. Rather than applying these generic methods, we show in this chapter that more insights can be found using visual analytics. Visualizations using linear discriminant analysis reveal that the disorder within each protein is usually well separated linearly. However, if various proteins are integrated together, there does not exist a clear linear separation rule in general. Based on this observation, we perform another visualization on the linear discriminant vector for each protein and confirm that the proteins are clearly clustered into several groups. Inspired by such findings, we apply  $k$ -means clustering on the proteins and construct a different classifier on each group, which leads us to a significant improvement of disorder prediction performance. Moreover, within the identified protein subgroups, the separation accuracy topped 99%, a clear indicator for further biological investigations on these subgroups.

---

Jaegul Choo

Computational Science and Engineering, Georgia Institute of Technology e-mail: jaegul.choo@cc.gatech.edu

Fuxin Li

Computational Science and Engineering, Georgia Institute of Technology e-mail: fli@cc.gatech.edu

Keehyoung Joo

Center for Advanced Computation, Korea Institute for Advanced Study e-mail: newton.joo@gmail.com

Haesun Park

Computational Science and Engineering, Georgia Institute of Technology e-mail: hpark@cc.gatech.edu

## 1 Introduction

Today's internet era has bombarded analysts in many research fields with an excess amount of information. Extremely complicated structures in data have unearthed new challenges to the statistics and machine learning field. In the past 50 years, many excellent methods have been architected to handle *flat* data – data with simple problem structures. Good examples are the traditional classification and regression problems: given some training input and output, attempt to build a predictive model for the data that can accurately predict future inputs [15]. However, real-life data are often not flat, requiring that a certain structure unique to each problem be utilized in order to obtain good results. Since the structure may differ so much among the datasets, it is extremely hard to design automated methods to capture each and every one of the particular problem structures.

Because of this difficulty, visual analytics has drawn a lot of interest. Humans are much better than computers in gaining structural insights. However, a significant portion of their analytic ability comes visually, whereas even the fastest computers have yet to achieve a human's capability to quickly summarize fairly complicated pictures and plots. Certainly it would be extremely beneficial to combine the strengths of both humans and computers in order to make better sense of our data reservoir, but the question is exactly how this can be practically done. Solving practical problems of interest tends to be much more difficult than boasting about the accomplishments of theoretical principles.

The diverse nature of data indicates that there exists no simple answer. In general, pattern recognition techniques can be utilized to reduce the data into a form people can read and look at. Nonetheless, people can handle a limited number of objects [4]: it is well-known that usually a human can simultaneously handle at most a handful of objects [25, 3], an embarrassingly small number given that the data may contain millions of instances that contain thousands of features (dimensions). It is hard to believe that there exists any panacean algorithm that can reduce every kind of data to a much smaller number of objects of interest.

Therefore, in visual analytics, a lot of creativity and interaction with data are needed to analyze a problem. This does not necessarily mean producing beautiful renderings and eye-catching animations, but putting more problem-specific efforts to reveal the underlying structure in the data. We argue that in the current stage of visual analytics research, having a lot of use cases of applying visual analytics to a variety of problems is important, since these solid advices can potentially help people to draw more general guidelines in the future. Therefore, this chapter focuses on just one particular problem and shows how we apply visual analytic principles, combined with simple classic pattern recognition methods, to obtain some structural insights and enhance our knowledge and predictive ability about the problem. It is our hope that our analysis described in this chapter can give some inspiration to more and better visual analytics use cases in the future.

The rest of this chapter is organized as follows. Section 2 introduces the problem of protein disorder prediction and describes the dataset and features. Section 3 briefly discusses the variant of discriminant analysis algorithm that we use to visual-

ize the data. Section 4 presents the details of our visual analysis. Section 5 shows the experimental results of our visualization-driven approach. Finally, Section 6 draws conclusions and suggests possible future work.

## 2 Protein Disorder Prediction

Proteins are fundamental biochemical compounds in which a linear chain of amino acids (or residues) are formed via polypeptide bonds and folded into complex three-dimensional structures. In most cases, the complex structures of proteins are stable, but some proteins may contain some unstable sub-sequences within its amino acid chains, which we call intrinsically disordered regions.

These intrinsically disordered regions play important biological roles by facilitating flexible couplings and bindings with other proteins. Thus, the identification of disorder regions has been a crucial task within biology domains [11]. This problem has also continuously been one of the main focuses in the biannual world-wide experiment called critical assessment of methods of protein structure prediction, i.e., CASP [1].

This task is typically done by experimental methods such as X-ray scattering and nuclear magnetic resonance spectroscopy, which cost nontrivial amounts of time and money. Alternatively, a lot of effort has been spent in developing computational methods that statistically predict the disorder region of a given protein using a set of training proteins whose disorder labels are known. From a computational perspective, protein disorder prediction can be viewed as a binary classification problem, which determines whether each amino acid in a given protein is disordered or not. Until recently, numerous methods have been proposed [12], and some of them adopt popular classification techniques such as neural network [7, 16] and support vector machines (SVM) [26, 28].

The protein disorder prediction data in this study is a standard database [6]. It contains the amino acid sequences of 723 proteins, which has, in total, 215,612 residues as well as their labels that describe whether or not a residue is disordered. Approximately 6.4% of them are classified as disordered.

To apply classification techniques to the dataset, the data items that need to be classified are typically encoded as high-dimensional vectors. We have used one of the standard encoding schemes to represent each residue in a particular protein, which takes into account the neighborhood residues within a particular window size [21, 22]. To be specific, for a window size of  $(2w + 1)$ , a residue is encoded using itself as well as the previous  $w$  and the next  $w$  residues. For these  $(2w + 1)$  residues, their PSI-BLAST profiles [2], the secondary structure, the solvent accessibility, and the hydrophobicity features are concatenated as a high-dimensional vector to represent the residue at the center. The details of these features are as follows.

**PSI-BLAST profile** In the first part of the features, each of the  $(2w + 1)$  residues in the window is represented as a 20-dimensional PSI-BLAST vector. This 20-dimensional vector is then normalized so that it sums up to one. However, the first

and the last  $w$  residues at N- or C-termini do not have all the valid  $(2w + 1)$  residues in their windows. In order to allow a window to extend over N- and C-termini, an additional 21st dimension is appended. For those positions that extend out of the protein, no amino acid exists in which case the 20-dimensional vector is set to all zero, but the 21st dimensional value is set to one. Otherwise, the 21st dimensional value stays zero. Additionally, we put another dimension representing the entropy value of a PSI-BLAST vector. In the end, the PSI-BLAST profile takes up  $22 \times (2w + 1)$  dimensions.

**Secondary structure profile** The secondary structure of a protein refers to certain types of the three-dimensional local structure. Although it originally has 8 different types, we use a simpler categorization of three states of helix, sheet, and coil. To obtain the secondary structure profile, we utilize one of the popular secondary structure prediction methods called PSIPRED [19, 5]. Assigning one dimension for the resulting probability or likelihood of each state, the secondary structure of each residue is encoded as a three-dimensional vector, which has  $3 \times (2w + 1)$  dimensions in total.

**Solvent accessibility profile** The solvent accessibility is another important characteristic associated with residues in a protein. For this profile, we use a recent method based on the  $k$ -nearest neighbor classifier by Joo et al. [20] and encode it as a scalar value for each residue. Additionally, we add another dimension to represent the average solvent accessibility within the window. Thus,  $(2w + 1) + 1$  dimensions are used in total.

**Hydrophobicity profile** The hydrophobicity also plays an important role in disorder prediction, and in practice, hydrophilic residues are frequently shown in disorder regions. We encode the hydrophobicity of each residue as a scalar value by using the Kyte-Doolittle hydrophobicity values [23], and similar to the solvent accessibility profile, we include an additional dimension of its average value within the window. Furthermore, considering its significant influence on prediction, we put another additional dimension of the average hydrophobicity throughout the entire residue sequence within a certain protein. Finally,  $(2w + 1) + 1 + 1$  dimensions are used for this profile.

In our experiments, we set  $w$  to 7 since it resulted in providing a higher classification performance in a reasonable computation time. Finally, the total number of dimensions in the data is 408.

### 3 Discriminant Analysis for Visualization

Discriminant analysis transforms high-dimensional data into a low-dimensional space so that different classes of data are well separated from each other. One of the most popular methods is linear discriminant analysis (LDA)[14], and it has been successfully applied to the visualization of clustered high-dimensional data by reducing the data dimension to two or three [8, 9].

Let us briefly describe LDA by introducing the notion of scatter matrices used to define the cluster quality and the criteria of LDA. Suppose a given data matrix  $A = [a_1, a_2, \dots, a_n] \in \mathbb{R}^{m \times n}$  whose columns are data items and let  $\mathcal{N}_i$  denote the set of data item indices belonging to class  $i$ . Assuming that the number of classes is  $r$ , the within-class scatter matrix  $S_w$  and the between-class scatter matrix  $S_b$  are defined, respectively, as

$$S_w = \sum_{i=1}^r \sum_{j \in \mathcal{N}_i} (a_j - c^{(i)}) (a_j - c^{(i)})^T \text{ and}$$

$$S_b = \sum_{i=1}^r |\mathcal{N}_i| (c^{(i)} - c) (c^{(i)} - c)^T,$$

where  $c^{(i)}$  is the centroid of class  $i$ , and  $c$  is the global centroid. The traces of these matrices are expressed as

$$\text{trace}(S_w) = \sum_{i=1}^r \sum_{j \in \mathcal{N}_i} \|a_j - c^{(i)}\|^2 \text{ and}$$

$$\text{trace}(S_b) = \sum_{i=1}^r \left\| |\mathcal{N}_i| (c^{(i)} - c) \right\|^2,$$

respectively. A large  $\text{trace}(S_w)$  and a small  $\text{trace}(S_b)$  corresponds to a stronger discrimination between classes.

In the reduced dimensional space generated by a linear transformation  $G^T \in \mathbb{R}^{l \times m}$  ( $m > l$ ), a data item  $a_j$  is represented as  $G^T a_j$ , and accordingly, the scatter matrices  $S_w$  and  $S_b$  become  $G^T S_w G$  and  $G^T S_b G$ , respectively. LDA solves  $G$  such that it maximizes  $\text{trace}(G^T S_b G)$  while minimizing  $\text{trace}(G^T S_w G)$  by solving a single approximated criterion,

$$\max_G \text{trace} \left( (G^T S_w G)^{-1} (G^T S_b G) \right),$$

whose solution is obtained by generalized eigendecomposition [14] or generalized singular value decomposition [17].

In this vanilla version of LDA, the rank of  $G$  is at most  $k - 1$  due to the rank of  $S_b$ , and in the current binary classification problem only one-dimensional output can be generated, which is too restricted for visualization. To avoid this issue, we modify the centroid terms used in  $S_w$  and  $S_b$  to nearest neighbor points [13, 27] such that

$$S_w^{NN} = \sum_{j=1}^n \sum_{k=1}^K (a_j - NN_w(a_j, k)) (a_j - NN_w(a_j, k))^T \text{ and}$$

$$S_b^{NN} = \sum_{j=1}^n \sum_{k=1}^K (a_j - NN_b(a_j, k)) (a_j - NN_b(a_j, k))^T,$$

where  $NN_w(a_j, k)$  is the  $k$ -th nearest neighbor point of  $a_j$  among the data in the same class of  $a_j$ , and  $NN_b(a_j, k)$  is the one among the data that belongs to a class different than that of  $a_j$ . With such modifications, the rank of the matrix  $G$  is no longer restricted to the number of classes  $k$ , and one can visualize the data by using the two or three most significant dimensions of the solution obtained from generalized eigendecomposition/singular value decomposition.

## 4 Visualization of Protein Disorder Data

Although the protein disorder prediction problem can be described as a *flat* binary classification problem of individual amino acids, there is one more layer in its structure – the protein level. If this structure is used, each amino acid would not be treated separately, but rather grouped together by their respective proteins and some protein-level clue would be used. This is no longer trivial and needs both a motivation why it is needed, and a strategy to perform it. We will detail our visual approach in this section that gives both a motivation and a strategy.

### 4.1 Knowledge Discovery from Visualization

The first idea is to visualize a simpler subproblem: the amino acids within each protein. By using the neighborhood-based discriminant analysis described in Section 3, we have generated the 2D scatter plot of residues along with their disorder labels using different colors. Figure 1 shows several visualizations of different proteins. As can be seen in Figure 1 (a)-(c), the two classes are clearly separated from each other in almost all the proteins. Especially, the non-disorder amino acids almost form a Gaussian distribution in every protein, hinting that discriminant analysis methods are suitable for this problem (LDA is the optimal classifier when both classes are from Gaussians with equal covariance. If the covariance is not equal, quadratic discriminant analysis (QDA) generates the optimal classifier [10]).

Knowing that within each protein, linear separability is achievable, a natural next question is whether this extends when multiple proteins are analyzed together. Interestingly though, only a few proteins need to be put together to lose the separability: when performing discriminant analysis on residues from several proteins, the two classes almost always have significant overlap (Figure 1(d)). The structural knowledge we have gained through this visualization is that the non-disorder amino acids approximate a Gaussian distribution within each protein, but these Gaussians differ for different proteins.

These structural observations motivate us to carefully design the disorder predictor depending on the proteins. To this end, instead of visualizing individual residues, we now visualize the proteins as individual data items. For this case, however, we need a high-dimensional vector representation for each protein. Our second impor-

tant idea in the paper is to use the first basis (or discriminant) vector computed from discriminant analysis to represent each protein.

The justification of using it is as follows. As shown in Figure 1, in most of the proteins, the discrimination between the two clusters are achieved in the first dimension, i.e., along the horizontal axis, and in this sense, using the first basis vector is sufficient to characterize how the two clusters are separated in each protein. In addition, since discriminant vectors differ among proteins, visualizing them could reveal if some proteins have similar discriminant vectors, or if there is no special patterns there.

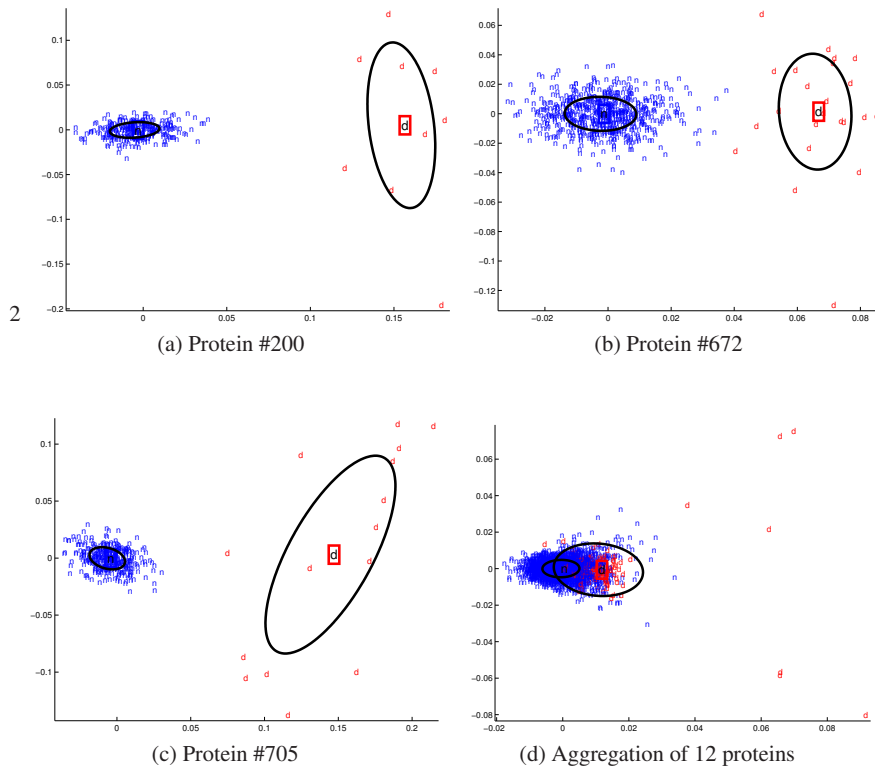


Fig. 1: Visualization examples of randomly chosen proteins from the 723 proteins database [6]. (a)-(c) are for individual proteins, and (d) for 12 proteins including the proteins used in (a)-(c). The blue and red color correspond to the non-disorder and the disorder clusters, respectively. The sample mean and covariance for each cluster are also shown as a letter in a rectangle and a black ellipse, respectively.

## 4.2 Visualizing the Discriminants

The 408-dimensional discriminant vector from each protein is used for the visual analysis in the next step. Here we first perform a simple principal component analysis (PCA) [18] to reduce the vector dimensions to three. Figure 2(a) shows the 3D scatter plot of these protein-level discriminant vectors. Unlike LDA, PCA does not directly take into account the cluster structures. Nevertheless, the visualization clearly shows a cluster structure in which there are four clusters among 723 proteins. The observed data invites us to use a clustering algorithm. Therefore, we have applied  $k$ -means clustering on the basis vectors by setting  $k$  to 4. It resulted in four clusters with 48, 61, 64, and 550 proteins, respectively, and this clustering result from the original 408-dimensional space matches our visual findings in the 3D space as shown in Figure 2(b). Although it is not clear in Figure 2(b), as we rotate the 3D scatter plot, the majority cluster with the orange color containing 550 proteins is shown to have a relatively high variance. The visualization of only the majority cluster, as shown in Figure 2(c), reveals that it is due to a heavy tail at the left side, and therefore, we further divided the majority cluster into two clusters by using  $k$ -means with  $k = 2$ . Consequently, the 48 proteins in the tail has been identified, which correspond to the brown cluster in Figure 2(d). Finally, the cluster distribution is summarized in Table 1.

To further confirm the protein clusters we found, we propose a stratified classification approach. In this approach, we train one classifier on each protein cluster. Ideally, when given a test protein we will first determine which cluster it belongs to and then use the respective classifier to predict its disorder regions. A Bayesian approach can also be taken such that the final decision is made by

$$\Pr(P_a \text{ is disordered}) = \sum_i \Pr(P \in G_i) \Pr(P_a \text{ is disordered} | G_i)$$

where  $P_a \in P$  is an amino acid in protein  $P$ , and  $G_i$  defines the protein groups found from the cluster analysis. This equation marks a difference from previous approaches: here we factorize the desired probability that an amino acid  $a$  is disordered into two distinctive parts. The first is the probability that the protein  $P$  belongs to a specific protein subgroup. Then, given this subgroup and the amino acid, the final decision is made.

In this chapter we will train the classifier to give  $\Pr(\text{disorder} | a, G_i)$  and leave the protein grouping as a future research topic. Solving the protein grouping problem would require the support of biologists, who already have a variety of tools and databases to select homologous proteins and similar proteins.



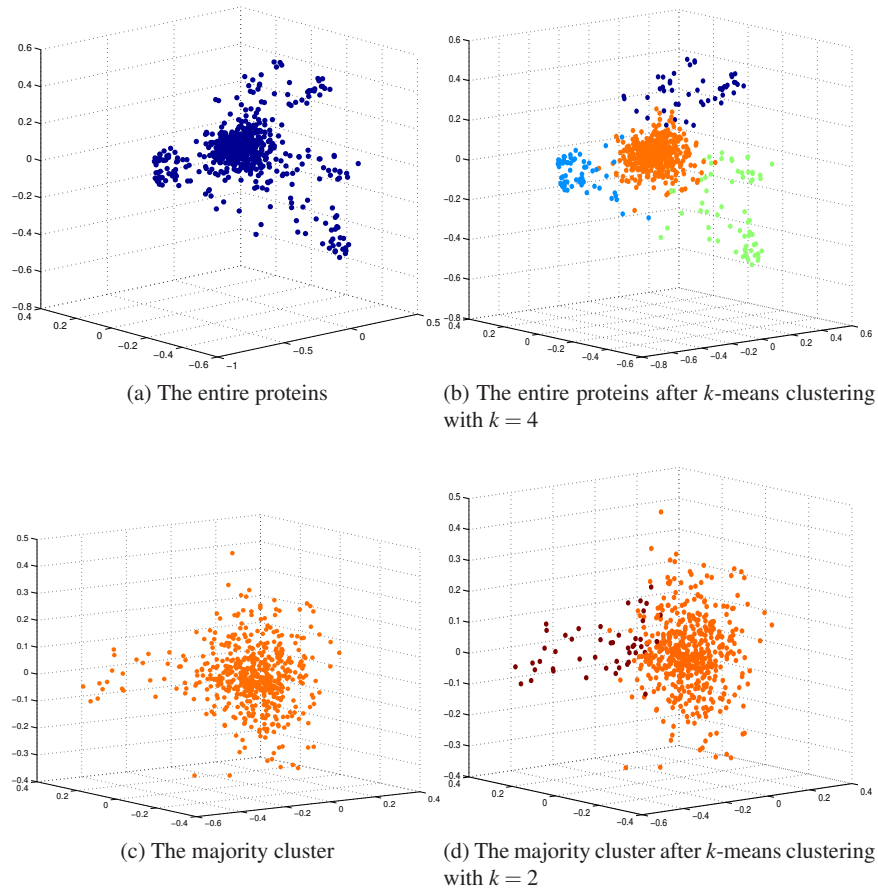


Fig. 2: 3D scatter plots of the first bases of discriminant analysis applied to each of the 723 proteins. PCA has been used to generated the scatter plots, and the different colors indicate the cluster labels obtained from  $k$ -means clustering.

Main group	Group 2	Group 3	Group 4	Group 5
Orange	Dark blue	Light blue	Green	Brown
502	48	61	64	48

Table 1: Cluster distribution of 723 proteins shown in Figures 2(b) and (d).

## 5 Classification Evaluation and Discussion

To evaluate disorder prediction performance, we adopt a standard procedure,  $K$ -fold cross-validation, where  $K$  is set to 10. The five different random cross-validation splits are used to assess the standard deviation of the methods. The split is done on the proteins so that each time the test prediction is performed on the protein data points that are not part of the training set. It is also independent of the stratification, which means we do not control the number of training proteins for each protein cluster. For stratified classification, we initially put the training proteins into their correspondent clusters and train one classifier per each cluster. Then for the test proteins, we identify their clusters and apply the corresponding classifier to predict whether the residues within the proteins are disordered or not. This setting is not realistic because for a new protein we do not know its true cluster, however, the purpose of this study is to verify that constructing the stratified classifier makes sense and improves the prediction accuracy significantly. This issue can be dealt with by learning another classifier that classifies the test protein into its proper cluster in the future research.

We compare the results from both literature and standard algorithms such as linear ridge regression and linear SVM applied to our feature representation described in Section 2. Linear SVM is computed with the LIBLINEAR package [24] with an  $L_2$ -loss and  $L_2$ -regularization. The ridge parameter of ridge regression is fixed as 500 and the linear SVM  $C$  parameter is fixed as 5. Area under ROC curve (AUC) is used as a performance measure since the dataset is highly imbalanced. The results are shown in Table 2. It can be seen that the performance greatly increases for stratified classifiers. With linear SVM, the performance shows the highest results, 91%, which are significantly better than the best known result on the dataset.

Moreover, depending on the identified protein clusters, the performance can be extremely good as shown in Table 3. On each of the protein groups 2, 3, and 4, the performances are more than 99%, which is almost perfect. Even on group 5, the performance is better, e.g., 95.66%, than the main group. In other words, if one determines that the protein is different from the main group, a very confident prediction of the protein disorder can be made. This finding makes the identification of protein groups a very interesting problem and should also shed some light on the biological side.

We want to re-emphasize again that this finding is a direct result of the visual analytics approach we take. Previous studies on this dataset have mostly emphasized performance improvements without careful investigation of the data themselves. Contrary to the previous studies, our study, which employed visualization techniques, has been able to pinpoint the structure in the protein disorder problem: 1) In each protein, ordered and disordered residues are well separated, but the separation rule is different for each protein. 2) Separation rules of each protein can be naturally clustered among various proteins. These two transitions are important for obtaining deep insight into the problem, further opening up an interesting direction in the biology and bioinformatics domains.

Method	AUC
Linear ridge regression	88.07 $\pm$ 0.14
Linear SVM	88.59 $\pm$ 0.12
Linear ridge regression on stratified data	89.74 $\pm$ 0.07
Linear SVM on stratified data	90.88 $\pm$ 0.08
DisProt [7]	87.8
SVMPrat [26]	87.7

Table 2: Comparison of classification performance between different methods

Protein group	Number of proteins	AUC
Main group	502	87.62 $\pm$ 0.16
Group 2	48	99.41 $\pm$ 0.07
Group 3	61	99.68 $\pm$ 0.07
Group 4	64	99.06 $\pm$ 0.06
Group 5	48	95.66 $\pm$ 0.47

Table 3: Classification performance on the protein clusters

## 6 Conclusion

In this chapter, we have studied the application of visual analysis principles in the protein disorder prediction problem. With simple techniques such as linear discriminant analysis and  $k$ -means clustering, we were able to unveil the special structure within the data that the disorder in each protein can be linearly separated while the separation rule is different between the proteins. Based on this visual observation, we grouped the proteins into five different groups and learned classifiers on each group. This turns out to perform better than many existing methods which have grouped all the proteins together. Especially, in three subgroups, we were able to obtain more than 99% accuracy, urging biological studies in these groups.

Of course, the reason we are obtaining some degree of success is because the structure in the data is still relatively simple. Instead of the linear structure, in bigger and more complicated datasets the inherent structure may be nonlinear manifolds, which are much harder to identify. Extension to such areas would be interesting for the future work.

**Acknowledgements** This research is partially supported by NSF grant CCF-0808863. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

## References

1. Protein structure prediction center. <http://predictioncenter.org/>.

2. Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.
3. Alan Baddeley. The magical number seven: Still magic after all these years? *The Psychological Review*, 101:353–356, 1994.
4. Christopher M. Bishop and Michael E. Tipping. A hierarchical latent variable model for data visualization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:281–293, 1998.
5. Kevin Bryson, Liam J. McGuffin, Russell L. Marsden, Jonathan J. Ward, Jaspreet S. Sodhi, and David T. Jones. Protein structure prediction servers at university college london. *Nucleic Acids Research*, 33(suppl 2):W36–W38, 2005.
6. Jianlin Cheng. Protein disorder dataset: Disorder723, June 2004. <http://casp.rnet.missouri.edu/download/disorder.dataset>.
7. Jianlin Cheng, Michael Sweredoski, and Pierre Baldi. Accurate prediction of protein disordered regions by mining protein structure data. *Data Mining and Knowledge Discovery*, 11:213–222, 2005. 10.1007/s10618-005-0001-y.
8. Jaegul Choo, Shawn Bohn, and Haesun Park. Two-stage framework for visualization of clustered high dimensional data. In *IEEE Symposium on Visual Analytics Science and Technology, 2009. VAST 2009.*, pages 67–74, oct. 2009.
9. Jaegul Choo, Hansung Lee, Jaeyeon Kihm, and Haesun Park. iVisClassifier: An interactive visual analytics system for classification based on supervised dimension reduction. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Conference on*, pages 27–34, oct. 2010.
10. Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2 edition, 2001.
11. A. Keith Dunker, Celeste J. Brown, J. David Lawson, Lilia M. Iakoucheva, and Zoran Obradovic. Intrinsic disorder and protein function. *Biochemistry*, 41(21):6573–6582, 2002. PMID: 12022860.
12. François Ferron, Sonia Longhi, Bruno Canard, and David Karlin. A practical overview of protein disorder prediction methods. *Proteins: Structure, Function, and Bioinformatics*, 65(1):1–14, 2006.
13. K. Fukunaga and J. M. Mantock. Nonparametric discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5:671–678, 1983.
14. Keinosuke Fukunaga. *Introduction to Statistical Pattern Recognition, second edition*. Academic Press, Boston, 1990.
15. Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2001.
16. Joshua Hecker, Jack Yang, and Jianlin Cheng. Protein disorder prediction at multiple levels of sensitivity and specificity. *BMC Genomics*, 9(Suppl 1):S9, 2008.
17. Peg Howland and Haesun Park. Generalizing discriminant analysis using the generalized singular value decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(8):995–1006, aug. 2004.
18. Ian T. Jolliffe. *Principal component analysis*. Springer, 2002.
19. David T Jones. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, 292(2):195–202, 1999.
20. Keehyoung Joo, Sung Jong Lee, and Jooyoung Lee. Sann: Solvent accessibility prediction of proteins by nearest neighbor method. *In submission*, 2011.
21. Hyunsoo Kim and Haesun Park. Protein secondary structure prediction based on an improved support vector machines approach. *Protein Engineering*, 16(8):553–560, 2003.
22. Hyunsoo Kim and Haesun Park. Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor. *Proteins: Structure, Function, and Bioinformatics*, 54(3):557–562, 2004.
23. Jack Kyte and Russell F. Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157(1):105–132, 1982.
24. Chih-Jen Lin. Liblinear – a library for large linear classification. <http://www.csie.ntu.edu.tw/~cjlin/liblinear/>.

25. George A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, 63:81–97, 1956.
26. Huzefa Rangwala, Christopher Kauffman, and George Karypis. svmprat: Svm-based protein residue annotation toolkit. *BMC Bioinformatics*, 10(1):439, 2009.
27. Fei Wang, Jimeng Sun, Tao Li, and N. Anerousis. Two heads better than one: Metric+active learning and its applications for it service classification. In *Data Mining, 2009. ICDM '09. Ninth IEEE International Conference on*, pages 1022–1027, dec. 2009.
28. Jonathan J. Ward, Jaspreet S. Sodhi, Liam J. McGuffin, Bernard F. Buxton, and David T. Jones. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *Journal of Molecular Biology*, 337(3):635–645, 2004.