Localized Neural Network Based Distributional Learning for Knowledge Discovery in Protein Databases

Dragoljub Pokrajac[1], Aleksandar Lazarevic[2], Teresa Singleton[1], Zoran Obradovic[3]
[1]Delaware State University, Dover DE, [2]University of Minnesota, Minneapolis MN, [3]Temple University, Philadelphia PA

Abstract. *In this paper, we investigate the application of localized neural network-based distributional learning techniques for characterizing interesting groups and potentially new types of disorder proteins. Instead of employing a single autoassociator model for learning global distributions of ordered and disordered classes, clustering-based partitioning techniques are first applied independently to both ordered and disordered labeled data set to identify regions of similar characteristics. Subsequently, local autoassociators are employed on labeled data to learn distribution of each cluster. These local autoassociators are used in testing phase to assign each tuple from the unlabeled data set to the cluster closest in distributional sense. Obtained partitions are analyzed for the presence and frequency of the expert-annotated keywords. Frequency comparison is applied to provide insight of keywords sensitive to the distribution heterogeneity and disorder/order labeling. Experimental results on a labeled database of confirmed order and disorder proteins and unlabeled data extracted from SWISS_PROT database are consistent with related literature and can provide further insight into relationship between protein similarity, keyword labeling and the disorder property.*

1. Introduction

Proteins are large complex molecules composed of amino acids that form the basic building blocks of life. Due to their ability to control the structure, function, and reproduction, proteins are referred as the "workhorse" of the cell. However, they can only function based on the accurate DNA blueprint found in the nucleus of each cell [17]. With the complete sequencing of the human genome, post-genomic era focuses on methods for estimating the structure and function of proteins. Emerging disciplines such as proteomics (the study of protein structure and their activity) and bioinformatics (a scientific discipline specifically aimed at using predicted biological functions from data in DNA sequencing) will be the focus of much research for years to come, and will help in solving many mysteries involved in the molecular basis of health and disease.

One of the greatest challenges in the proteomics is to identify proteins that are partially or wholly unstructured [12]. In studies of the protein disorder property [5], *disordered proteins* are characterized by long sequence regions that lack a fixed 3-D structure in their native states. To study this interesting property, known examples of *disordered proteins* as well as of *ordered proteins* with a fixed 3-D structure have been collected. However, due to experimental bias towards ordered proteins, examples of protein disorder are scarce in the existing databases of proteins with determined structure. Many of these databases also exhibit heterogeneity, which means that rules identified among the observed attributes in certain subsets do not necessary apply elsewhere. A heterogeneous data set can be partitioned into homogeneous subsets such that learning a local model separately on each of them results in improved overall prediction accuracy. For instance, our previous research work [19] has shown that disorder proteins may have several flavors (different behavior types). In addition, our previous results [13,20] have confirmed that attribute distributions in labeled protein datasets may be heterogeneous and related to the presence/absence of particular keywords assigned to the proteins. Therefore, partitioning the set of disorder proteins into more homogenous ones may help in identifying and characterizing new types of disorder proteins. Similar methodology has been demonstrated as successful in several data mining applications. For example, in spatial domain, DBSCAN clustering algorithm [15] was used to partition the spatial fields into several similar regions and then to build localized regression models on each of them in order to predict the wheat yield [9]. In addition, hierarchical partitioning followed by localized prediction was used to detect damages in large complex mechanical structures [11].

When different groups of disorder proteins are identified using a partitioning algorithm, each subset of disorder proteins needs to be characterized with a specific model. One of the standard techniques for characterizing data distributions is distributional learning, which is known as a difficult problem in machine learning [2]. In our approach we use autoassociator neural networks [2, 14] to learn class-conditional distributions for each class of labeled sequence data. While the autoassociators have been known as a promising approach to distribution learning, innovations in our approach include sequence representation by an appropriate set of attributes, as well as using class and cluster information to optimize the architecture of autoassociators. In our earlier work [20], multilayer autoassociators were used for qualitative enlargement of labeled data but without considering its potential distributional heterogeneity.

The approach proposed in this study consists of the following four steps: (1) partitioning labeled data using clustering algorithm, (2) learning class-conditional distributions from clusters of labeled sequence data, (3) partitioning of unlabeled data according to distributional similarity to the labeled data clusters, and (4) preliminary analysis of functional properties of clusters' learned distributions

based on frequency of keywords.

2. Methodology

Starting from a set of labeled sequences assigned to two classes (subsets) of ordered and disorder proteins, the first step of the proposed method involves partitioning, using a clustering algorithm, independently of the subset of ordered proteins and of the subset of disorder proteins. The main motivation behind this is identification of more homogenous subsets of ordered (disordered) proteins, which could be more successfully summarized by localized descriptive models than by the global ones. In the second step, we apply neural networks to learn distributions of data records in each cluster. The proposed distribution-learning approach consists of learning data distributions separately for each cluster that is discovered within each class. Using local distribution models, we assign each pattern from an unlabeled dataset to a labeled data cluster that is closest in distributional sense. Finally, we analyze the frequencies of keywords assigned to the corresponding proteins.

Attribute construction. In our approach we represent each sequence position or a whole sequence with a set of attributes shown relevant to the studied property [20]. In a particular application, the attributes should be able to capture information relevant to the distributional properties of each of K classes. For instance, such attributes corresponding to a given position in the sequence could be derived from statistics of a subsequence within a window centered at the position. More formally, given a labeled sequence $\mathbf{s} = \{s_i, i = 1,\ldots,L\}$ of length L, each position i is assigned a corresponding label $y_i \in \{1,\ldots,K\}$. An appropriate M-dimensional attribute vector $\mathbf{x}_i$ is constructed for each sequence position s_i. Finally, each sequence $\mathbf{s}$ is represented with a set of L examples $\{(\mathbf{x}_i, y_i), i = 1, \ldots L\}$. A labeled set $\mathbf{S}$ is then constructed by repeating this procedure on all N available labeled sequences. Since our goal is to develop a separate distribution model for each class, we construct training sets $\mathbf{S}_j = \{(\mathbf{x}_i, y_i), y_i = j\}$ composed of all examples from $\mathbf{S}$ labeled with the class j, j=1,..,K. For the considered proteomics application, we distinguish among ordered and disordered proteins, hence K=2.

Clustering. In this study, several clustering algorithms from the CLUTO clustering package [21] were employed to partition sets of ordered and disorder proteins into more homogenous ones. The applied clustering algorithms include k-way clustering algorithms with repeated bisections (*RB, RBR*), direct k-way clustering (*DIR*) and agglomerative clustering (*AGGL*) approach. In the k-way clustering with repeated bisections, clusters are obtained by performing sequence of k − 1 repeated bisections. Here, the data set is first clustered into two groups, with one of the groups selected and subsequently bisected further. This process continues until the desired number of clusters is found. During each step, the bisection is performed so that the resulting 2-way clustering solution optimizes a particular clustering criterion function. We have used two variants of the k-way clustering. In the first one (*RB*), the criterion function is locally optimized within each bisection, while in the second one (*RBR*) the overall solution is globally optimized. Basically, the *RBR* approach uses clusters obtained by the *RB* algorithm as initial cluster solution, and attempts to further optimize the clustering criterion function. However, the RBR is not necessarily better clustering solution due to this postprocessing step, since the global optimization does not necessarily give the best criterion for estimating the quality of obtained clusters. The direct k-way clustering algorithm, as a variant of k-means algorithm, simultaneously finds all k clusters. In general, computing a k-way clustering directly is slower than clustering via repeated bisections. In terms of quality, for reasonably small values of k (usually less than 10–20), the direct k-way clustering algorithm leads to better clusters than those obtained via repeated bisections. In agglomerative clustering approach, we first find the clusters by initially assigning each object to its own cluster and then repeatedly merging pairs of clusters until a certain stopping criterion is met. Here, the desired clusters are computed using the agglomerative paradigm which goal is to locally optimize a particular clustering criterion function. The algorithm starts with clusters assigned to each data object, and proceeds with merging two closest clusters and stopping when k clusters remain.

The criterion function that we use here in all clustering algorithms is based on maximizing the similarity between each data record and the centroid of the cluster the record is assigned to. Specifically, if we use the cosine function to measure the similarity between a data record and a centroid, then the criterion function may be defined to maximize I_2 criterion $\sum_{r=1}^{k} \sum_{d_i \in S_r} \cos(x_i, C_r)$, where x_i is a data record, S_r and C_r correspond to *the* data assigned to r-th cluster and the cluster centroid respectively, $r = 1, \ldots, k$.

Distribution learning. As the distribution models we use *autoassociators* — a special class of multi-layer feedforward neural networks with logistic sigmoidal transfer functions [2,14]. The number of inputs and outputs of autoassociators corresponds to the number M of attributes. In addition to M input and M output neurons, autoassociators have three hidden layers with n_{13}, n_2, and n_{13} neurons, respectively, where $n_2 < M$. Each autoassociator is trained to reconstruct its input $\mathbf{x}$ at the output $\mathbf{t}$ and its parameters are optimized to minimize the Euclidian distance $\|\mathbf{x}-\mathbf{t}\|_2$. To achieve an accurate reconstruction, the autoassociator is implicitly forced to discover an appropriate nonlinear mapping of the original M-dimensional attribute space into a smaller n_2-dimensional space that captures the properties of the underlying distribution [2].

An important aspect of our approach is learning *class-specific autoassociators* A_j for each cluster $\mathbf{S}_j$ of labeled set instead of learning a global model on all labeled data $\mathbf{S}$.

The main benefit of such decomposition is simplification of the learning task in the spirit of a mixture-of-experts approach to learning [8]. The autoassociators trained on labeled data are used to assign each tuple from the unlabeled data set to the cluster closest in distributional sense. For each tuple from unlabeled dataset, we compute the norm of difference between the tuple and the output of each autoassociator. The tuple is subsequently associated to the distribution corresponding to the autoassociator for which this norm is the smallest (Figure 1).

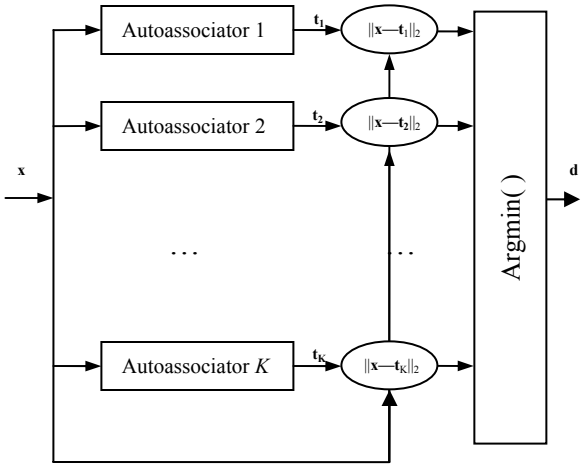

Figure 1. Using autoassociators to determine the closest distribution d to an unlabeled tuple **x**

Keyword Frequency Analysis. Tuples identified as belonging to particular partitions are analyzed for the presence and frequency of the expert-annotated keywords. (For example, in SWISS-PROT [18], about 840 keywords are used to describe protein functional properties). It should be emphasized that the number of keywords assigned to a given biological molecule is dependent on its biological activity (e.g. some proteins could be involved in a number of biological processes, while others are highly specific), and on the extent of experimental research performed on the molecule (e.g. very related proteins could differ largely in the number of associated keywords). Therefore, each existing database of keywords assigned to biological molecules is incomplete with a large fraction of missing information.

Another property of the existing biological databases is the presence of homologues – families of similar sequences with the common evolutionary origin. The presence of large homologous families could largely skew results of any analysis performed on these biological databases. Our goal in this paper is to compare frequencies of keywords in the whole SWISS-PROT and in its partitions, and not necessarily make conclusions about the SWISS-PROT as a "representative collection" of real-world proteins. Hence, the sampling bias present in SWISS-PROT equally propagates to its partitions, so we do not need construct its non-redundant subset [13, 20].

In our approach we first count the occurrences of the keywords in the whole database of unlabeled sequences, and in each partition of unlabeled sequences (identified by the clustering algorithm). Then, we compute the frequencies of the keywords in both cases and using statistical test for each keyword, determine whether the frequencies differ significantly. Subsequently, we determine the set of keywords with the frequency discrepancy that is significant for *each* cluster separately. The analysis of these keywords could provide insight of keywords sensitive to the distribution heterogeneity and disorder/order labeling. Particularly, this list of keywords could be compared to the keywords exemplified as associated with orders/disorders in other references [13,20].

To test whether the frequency of a keyword in a particular cluster differs from the frequency in the whole database, we use the Hypergeometric distribution [4]. Let N be the number of proteins in the whole database, P_i number of proteins having a specific *i-th* keyword in database, n_j number of proteins represented in a j-th cluster and $p_{i,j}$ number of proteins having the i-th keyword in a cluster j. Frequency of keywords in the database is computed as $F=P/N$, while the observed frequency of the i-th keyword in the j-th cluster is $f_{i,j}=p_{i,j}/n_i$. Under the null hypothesis H_0 that the distributions of the keywords in the cluster and in the whole database are actually the same, $p_{i,j}$ satisfies the Hypergeometric distribution $h(x)$ with parameters N, n_j and P_i [4]. It can be shown that the Type I error (of rejecting H_0 when it is actually true) is $\alpha = \sum_{x=0}^{p_{i,j}} h(x)$ if $f_{i,j} < F$ and $\alpha = \sum_{x=p_{i,j}}^{n_j} h(x)$ otherwise. The null hypothesis is rejected in favor of an alternative hypothesis (that the keyword frequency in the partition is *different from* the frequency in the whole database) if $\alpha<\alpha_{threshold}$ where $\alpha_{threshold}$ is a pre-specified significance level.

3. Experiments

In this section we illustrate our techniques when SWISS-PROT database [18] is used as the source of background information for discovering proteins underrepresented in a database of known ordered and disordered proteins.

3.1 Experimental Setup

Our data set of labeled sequences consists of 152 proteins containing disordered regions longer then 40 consecutive positions, and of 290 completely ordered proteins [6]. Some of disordered regions identified by NMR, circular dichroism or protease digestion were found starting from keyword searches of PubMed (www.ncbi.nlm.nih.gov) followed by a case by case confirmation based on detailed studies of relevant literature. Also, starting from a subset of the Protein Data Bank (PDB) called PDB_Select_25 (based on grouping PDB proteins into families having > 25% sequence identity), disordered regions in X-ray crys-

tal structures were identified by searching for residues having backbone atoms that are absent from the ATOMS lists in their PDB files. In total, our labeled data set consists of 22,434 disordered amino acids and 67,548 ordered amino acids.

We used 101,602 proteins listed in October 2001 release 40 of SWISS-PROT database [18] as a set of unlabeled proteins. From SWISS-PROT we extracted amino acid sequences of each protein as well as the associated lists of keywords used in the evaluation of selected outliers. Overall, 840 keywords are used in SWISS-PROT to provide information about functional and structural properties of various proteins.

To perform clustering and partitioning, we used seven attributes chosen in [20] according to our earlier experiments and expert knowledge. These attributes are shown as correlated with order/disorder property [19] and for each protein residual are constructed using statistics of its neighboring amino acids.

All seven constructed attributes were employed to perform various clustering algorithms from CLUTO package with different number of resulting clusters. When referring results obtained using a particular method, we will use the method abbreviation followed by the numbers of clusters in disordered and ordered class (e.g., RBR(7,8) means the repeated bisections algorithm by k-way refinement found 7 clusters in *disorder* and 8 clusters in the *order* labeled data). Table 1 illustrates applied methods. Since maximization of the I_2 criterion function was used to optimize clustering results, we started with a relatively large number of clusters (around 15) and we were observing the gain in the quality of clusters after merging selected clusters. When the gain significantly dropped, the clustering process was stopped. Due to several thresholds used in measuring gain, we have more than a single solution for an optimal number of discovered clusters.

Our major goal in this study was to demonstrate the potentials of the methods but not to investigate its final frontiers. Hence, although we tested several different configurations of the autoassociators (different numbers of neurons n_{13} and n_2) we did not perform systematic topology optimization. Here, we report results obtained using the networks with n_{13} = 8 and n_2 = 4 that are consistent to results obtained using other network topologies. The autoassociators were trained using the Levenberg-Marquart algorithm [3]. The optimal number of training epochs was determined empirically with training terminated when the improvements in training error were diminishing. We needed 100 epochs to achieve stable training error on labeled clustered data.

3.2 Results

Using the proposed technique, we first independently partitioned the unlabeled ordered and disordered datasets into subsets where each subset is the most similar to one of clusters identified on labeled data. Then, for each subset we identified SWISS-PROT keywords whose frequency is significantly different ($\alpha_{threshold}$=1e-6) from the corresponding frequency on the whole unlabeled dataset. For each method, we identified keywords that have frequencies significantly different for *all clusters*.

The set of identified keywords primarily depends on the clustering method and on the number of clusters in the set of *disordered* proteins. This means that there is really heterogeneity in disorders (and clear separation between orders and disorder property) while the set of ordered proteins seems to be fairly homogeneous (in contrast to our initial assertion that ordered protein set is also heterogeneous).

With the large number of clusters, the proposed techniques tend to be over-restrictive (e.g., using the RBR method that identified 11 disorder clusters, there were no keywords with significant frequency difference in all disordered clusters). Generally, the smaller number of clusters, the larger the number of discovered keywords. The larger number of clusters on labeled data leads to the smaller number of data for training each autoassociator, which can lead to overfitting and incorrect assignment of an unlabeled tuple to the partition that is actually *not* the closest in distributional sense.

Depending on the number of clusters and the clustering technique, we were able to detect the majority of keywords identified elsewhere as associated to the protein disorder property [13, 20]. For instance, *complete proteome*, *transmembrane*, *hypothetical protein*, and *inner membrane,* that were also identified as keywords associated with underrepresented sequences in PDB25 are detected as significant in all clusters, when RBR(7,8) clustering is used. Similarly, the keywords *repeat*, *nuclear protein*, *dna-binding*, *developmental protein*, *chromosomal protein* and *microtubules*, identified as disorder-correlated [20] appear associated to the disorder property using some of examined clustering methods.

Out of the keywords that are associated with disorders by our techniques, *alternative splicing*, *lyase*, *oxidoreductase*, *phosphorylation*, *transmembrane* and *transport* are exemplified by at least 50% of the examined clustering methods, as shown in Table 1. It is interesting to denote that, according to our best knowledge, for some of these keywords (e.g. *transport*) other methods could not demonstrate clear association with disorder property [20].

Observe that the keywords may be overrepresented or underrepresented in only *some* clusters. In fact, such keywords may be particularly interesting since their frequencies are associated with *particular* data regions (partitions), where the keywords have significant frequency mismatch. To further examine this important property of data, we extracted keywords that are underrepresented in at least one partition, for partitions corresponding to protein orders and to protein disorders. To reduce the number of presented keywords, in Table 2 we show only the keywords which frequency differs for more than 5% in com-

parison to the frequency in SWISS-PROT. These results suggest that heterogeneity in distribution of the examined dataset, efficiently discovered by the proposed method, is correlated to underlying properties of the data. For instance, Iakoucheva et al. [7] discovered that proteins related to regulatory function (and associated with the *transferase* keyword) have been linked with disorder property. It would be interesting to examine further biological meaning of our findings.

Using our technique, it is possible to look at keywords frequencies in *each* partition corresponding to disordered labeled data and to further analyze proteins corresponding to each partition. For each of 7 "disordered" partitions discovered by RBR(7,10), the keywords that have significantly higher (+) and significantly lower (-) frequencies in comparison to the frequencies in SWISS-PROT are shown in Table 3. Again, we show only the keywords which frequencies differ for at least 5% from the frequency in SWISS-PROT. As we can see, discovered partitions are correlated with underrepresentation or overrepresentation of particular keywords. For instance, sets of keywords over-represented in the partition 1 and the partition 2 are distinct. Furthermore, some of the keywords, (e.g., *dna-binding, nuclear protein, transmembrane, transport*) are overrepresented in one, while underrepresented in another partition. This clearly illustrates that our technique can distinguish not only keywords associated with disorders, but also the keywords associated with particular subclasses of unlabeled data, close in distributional sense. The follow-up examination of discovered subsets of keywords and related proteins may put new insight on better understanding different kinds of disorders and their relationship to the protein function. Here, none of the keywords from Table 3 appeared as underrepresented or overrepresented (with more than 5% frequency difference) in partitions 3 and 5. According to this example we can also observe that although keyword representation is correlated with obtained partitions, it is not necessary that keyword frequencies differ in *each* partition.

4. Conclusions

In this study we have presented a localized partitioning-based knowledge discovery technique that is applied to keyword-annotated protein databases to potentially help in discovering novel groups of proteins and their association with potentially unknown classes of protein disorders.

The principal goal of the study is to demonstrate the usefulness of the approach and not to explore its ultimate performance. Hence, we did not perform extensive research on technical details such as the choice of number of clusters, network topology (number of neurons in hidden layers), optimal learning algorithm, and comparison of the proposed autoassociator-based technique with other existing methods for distribution learning. Research on these important aspects is currently in progress. In addition, we are working on a follow-up expert analysis of discovered protein partitions and on application of the proposed method on other keyword annotated protein databases (e.g., PIR, GenPept/Gene [1]).

5. Acknowledgements

This work was partially supported by NIH (P20 RR16472), DoD Department of Army (45395-MA-ISP), NSF (#0310163, #0320991) Grants and by the Army High Performance Computing Research Center contract number DAAD19-01-2-0014. The content of the work does not necessarily reflect the position or policy of the government and no official endorsement should be inferred. Access to computing facilities was partially provided by AHPCRC.

6. References

[1] Baldi, P., Brunak, S. *Bioinformatics- The Machine Learning Approach*, 2edn. MIT Press, 2001.

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

[3] Demuth, H., Beale, M. *Neural Network Toolbox for use with MATLAB, Users Guide,* Version 3. The MathWorks, Inc., 1998.

[4] Devore, J.L. *Probability and Statistics for Engineering and the Sciences*, 4th edn. International Thomson Publishing Company, Belmont, CA, 1995.

[5] Dunker, A.K., et al."Intrinsically Disordered Proteins," *Journal of Molecular Graphics and Modeling*, 19, 2001.

[6] Dunker A.K., et al. "Intrinsic Disorder and Protein Function," *Biochemistry,* 41 (21) 2002, pp. 6573 – 6582.

[7] Iakoucheva, I., et al. "Intrinsic Disorder in Cell-signaling and Cancer-associated Proteins," *Journal of Molecular Biology*, 323 2002, pp. 573-584.

[8] Jordan, M.I., Jacobs, R.A. "Hierarchical Mixtures of Experts and the EM Algorithm", *Neural Computation*, 6 1994, pp. 181-214.

[9] Lazarevic, A., et al. "Clustering-Regression-Ordering Steps for Knowledge Discovery in Spatial Databases," *Proc. IEEE/INNS Int'l Joint Conf. on Neural Networks*, 1999, No. 345, Session 8.1B.

[10] Lazarevic, A., et al. "Distributed Clustering and Local Regression for Knowledge Discovery in Multiple Spatial Databases," *Proc. European Symposium on Artificial Neural Networks*, 2000, pp. 129-134

[11] Lazarevic, A., et al: "Localized Prediction of Multiple Target Variables Using Hierarchical Clustering", *Proc. IEEE International Conference on Data Mining*, Melbourne, FL, November 2003.

[12] Linding R, et al. "Protein Disorder Prediction: Implications For Structural Proteomics," *Structure (Camb)*. 11(11), 2003, pp. 1453-9

[13] Peng, K., et al. "Exploiting Unlabeled Data for Improving Accuracy of Predictive Data Mining," *Third IEEE Int'l Conf. on Data Mining*, 2003.

[14] Pokrajac D., et al. "Distribution Comparison for Site-specific Regression Modeling in Agriculture," *Proc.*

IEEE/INNS Int'l Joint Conf. on Neural Networks 1999, pp. 346-352.

[15] Sander J., et al. "Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and its Applications," *Data Mining and Knowledge Discovery, Kluwer Academic Publishers*, 2 (2) 1998.

[16] Sensen, C. *Essentials of Genomics and Bioinformatics*. Wiley-VCH, 2002.

[17] Snustad, P., Simmons, M. *Principles Of Genetics*, 2nd edn. John Wiley & Sons, 2002.

[18] SWIIS-PROT database, http://us.expasy.org/sprot.

[19] Vucetic, S., et al. "Flavors of Protein Disorder, "*Proteins: Structure, Function and Genetics*, 52 2003, pp. 573-584.

[20] Vucetic, S., et al. "Detection of Unusual Biological Sequences Using Class-Conditional Distribution Models," *Proc. 3rd SIAM Data Mining Conference*, 2003, pp. 279-283.

[21] Y. Zhao and G. Karypis, Criterion Functions for Document Clustering Experiments and Analysis, *Army High Performance Computing Research Center (AHPCRC) Technical Report* #01-40, 2002.

Clustering Method		**RBR**						**AGGL**									**RB**				**DIR**
Number of clusters	Disorder	**11**		**5**		**7**		**4**			**6**			**8**			**4**		**7**		**8**
	Order	10	8	10	8	10	8	11	7	8	11	7	8	11	7	8	6	8	6	8	8
Keyword	**alternative splicing**			*	*	*	*	*	*		*	*	*	*	*	*	*	*	*	*	*
	lyase			*	*	*	*		*	*					*		*	*	*	*	
	oxidoreductase			*	*	*	*	*	*		*	*	*		*		*				*
	phosphorylation			*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
	transmembrane			*		*	*	*	*	*			*				*	*	*		
	transport			*		*	*	*	*	*		*	*	*	*	*	*	*		*	

Table 1. Keywords with frequencies significantly different in each disordered cluster vs. the frequency in the whole dataset for each examined clustering algorithm. The keywords identified significant by at least 50% of examined methods are annotated as **bold.**

Clustering Method		**RBR**						**AGGL**									**RB**				**DIR**
Number of clusters	Disorder	**11**		**5**		**7**		**4**			**6**			**8**			**4**		**7**		**8**
	Order	10	8	10	8	10	8	11	7	8	11	7	8	11	7	8	6	8	6	8	8
Keyword	complete proteome			*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
	dna-binding					*															
	glycoprotein							*					*								
	hydrolase			*				*					*	*							*
	hypothetical protein				*			*	*	*	*	*	*	*	*	*	*	*	*	*	*
	nuclear protein					*															
	oxidoreductase			*	*	*	*	*	*		*	*	*	*	*	*	*	*	*	*	*
	signal												*								
	transferase			*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
	transmembrane	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
	transport	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*

Table 2. Keywords identified be underrepresented (having significantly lower frequency in at least one partition corresponding to protein disorders as compared to the whole SWISS-PROT database).

Keyword		**alternative splicing**	**chromosomal protein**	**complete proteome**	**dna-binding**	**g-protein coupled receptor**	**glycoprotein**	**inner membrane**	**membrane**	**mitochondrion**	**nuclear protein**	**oxidoreductase**	**tion**	**receptor**	**repeat**	**rna-binding**	**signal**	**transcription**	**transcription regulation**	**transferase**	**transmembrane**	**transport**
Partition	**1**	+	+	–	+				–		+	–	+		+	+				–	–	–
	2				–	+	+	+	+	+	–			+			+				+	+
	4			–	+						+	–			+			+	+			–
	6			–							+											
	7			–																		

Table 3. Keywords significantly overrepresented (+) or underrepresented (-) in partitions corresponding to disordered tuples obtained using RBR(7,10) clustering technique. Only the keywords significant with $\alpha_{threshold}$=1e-6 with frequencies at least 5% different from the frequencies on the SWISS-PROT are presented.