Detection of Underrepresented Biological Sequences Using Class-Conditional Distribution Models*

Slobodan Vucetic† Dragoljub Pokrajac‡ Hongbo Xie† Zoran Obradovic†

Abstract

A labeled sequence data set related to a certain biological property is often biased and, therefore, does not completely capture its diversity in nature. To reduce this sampling bias problem a data mining procedure is proposed for detecting underrepresented relevant sequences. The procedure is aimed at helping domain experts achieve a cost-effective qualitative enlargement of knowledge through an in-depth study of a small number of statistically underrepresented and functionally interesting sequences. Our procedure consists of: (i) learning a class-conditional distribution model on each class of labeled data; (ii) applying the models to select statistically underrepresented unlabeled sequences; and (iii) automatically evaluating their interestingness. An application of the proposed approach is illustrated on an important problem of increasing the data set of confirmed disordered proteins. The obtained results demonstrate the promise of the proposed approach for an efficient reduction of sampling bias in biological databases.

1 Introduction

Biological sequence data labeled to one of several classes with respect to a property of interest is often not representative of the underlying distribution due to the influence of (i) the evolutionary bias where some sequences are favored through the evolution, and (ii) an experimental bias where certain sequences are of larger interest to biologists, or are simply easier to analyze. Sampling bias should be taken into account when constructing and using the corresponding predictors, as well as when deriving biological conclusions from such data.

The most direct approach to resolving the sampling bias problem is enlarging labeled data set. However, considering the high costs of obtaining labeled biological data, a careful addition of new sequences to labeled data set is necessary. A successful selection should be helpful in directing use of available human resources towards detailed studies of the most promising sequences ranging from literature search to experiments in a "wet lab" in order to assign them a class label.

In our approach we concentrate on discovering sequences that are distributionally underrepresented in labeled sequence data. Such sequences are most likely to qualitatively improve the quality of labeled data and reduce the problem of sampling bias. Since distribution density estimation from high dimensional data is known as a difficult problem, we learn distribution implicitly by constructing autoassociator neural networks [4] on each class of labeled data.

Given a measure of deviation of a given sequence from each class obtained by autoassociators, a user could select distributionally underrepresented sequences from large repositories of unlabeled sequences. Since the selection is a highly subjective process, we propose an automatic procedure for its preliminary evaluation. This procedure is based on extracting information from well-annotated databases of biological sequences to summarize functional properties of selected underrepresented sequences. Based on the summary, a user could focus on the functionally most interesting sequences.

Steps of the proposed approach, are described in Section §2. In Section §3 the proposed approach is illustrated on the problem of enhancing a database used in our study of the protein disorder property [3].

2 Methodology

2.1 Learning class-conditional distributions

We consider a scenario where a set of labeled sequences is provided with each sequence position classified into one of several classes. For example, a given amino-acid within a protein sequence could be characterized by its secondary structure (e.g. helix vs. sheet vs. coil) or a specific function (e.g. binding sites). A special case of sequence labeling is when a single class label is assigned to an entire sequence. An example of this special case is classification of whole proteins by their participation within a specific cell process (e.g. energy metabolism). We are interested in learning distribution models that recognize sequence properties of each separate class.

*This study was supported in part by NSF grant CSE-IIS-0196237 to Z. Obradovic and A.K. Dunker and NSF grant IIS-0219736 to Z. Obradovic and S. Vucetic.

†Center for Information Science and Technology, Temple University, Philadelphia, PA 19122-6094

‡Computer and Information Science Department Delaware State University, Dover, DE 19901

Attribute construction. We represent each sequence position or a whole sequence with a set of attributes thought to be relevant to the studied property. For instance, attributes representing a given sequence position could be derived from statistics of a subsequence within a window centered at the position. Thus, the first step in distribution learning is choosing an appropriate knowledge representation and constructing a set of training examples from labeled sequences that are suitable for distribution learning.

More formally, given a labeled sequence $\mathbf{s} = \{s_i, i = 1, \dots L\}$ of length L, each position i is assigned a corresponding label $y_i \in \{1, \dots K\}$, where K is the number of classes. An appropriate M-dimensional attribute vector $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots x_{iM}]$ is constructed for each sequence position s_i. Finally, each sequence $\mathbf{s}$ is represented with a set of L examples $\{(\mathbf{x}_i, y_i), i = 1, \dots L\}$. A labeled set $\mathbf{S}$ is then constructed by repeating this procedure on all N available labeled sequences. Since our goal is developing 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.

Learning distributions. We avoid difficulties associated with explicitly learning probability distributions by applying models designed to detect examples that deviate from the underlying distribution. We use the so-called *autoassociators* constructed as multi-layer feedforward neural networks [4]

Given a sequence representation with M attributes the employed autoassociators have M-dimensional input layer $\mathbf{x}$, M-dimensional output layer $\mathbf{t}$, and three hidden layers with n_{13}, n_2, and n_{13} neurons, respectively. Each autoassociator is trained to reconstruct its input at the output and its parameters are optimized to minimize the Euclidian distance $||\mathbf{x} - \mathbf{t}||_2$ between its input $\mathbf{x}$ and output $\mathbf{t}$. Processing neurons in the autoassociator use logistic sigmoidal transfer functions.

The main property of autoassociators is that the number of neurons n_2 in the second hidden layer is smaller then the number of inputs M (where the most appropriate value n_2 is to be determined experimentally). To achieve an accurate reconstruction of an input example at its output 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. This requirement enforces specialization of the autoassociator to the distribution of training data with the consequence of making large reconstruction errors on the examples underrepresented in training data.

An important aspect of our approach is learning class-specific autoassociators A_j on each labeled set $\mathbf{S}_j$ instead of learning a global model on all labeled data $\mathbf{S}$. The main benefit of such decomposition is simplification of the learning task. Another important aspect is that we use class information in optimizing the topology of a class-specific autoassociator A_j by testing it both on in-distribution examples from $\mathbf{S}_j$ as well as on out-of-distribution examples $\mathbf{S} \backslash \mathbf{S}_j$. We describe this idea in more detail in the following subsection. Finally, the decomposition could facilitate selection of underrepresented sequences by providing estimates of their similarity to each labeled set $\mathbf{S}_j, j = 1, \dots K$.

Network topology optimization. To use autoassociators, we have to optimize the numbers of hidden neurons n_{13} and n_2, where the only constraint is $n_2 < M$. A properly chosen topology should result in autoassociators achieving a small reconstruction error on examples matching training data distribution and a large error elsewhere. Given a choice of the pair (n_{13}, n_2), class-specific autoassociators $A_j(n_{13}, n_2), j = 1, \dots K$, are trained first. The quality of the constructed autoassociators is then measured as their classification accuracy on a validation set. Given a validation example $(\mathbf{x}, y)$, autoassociators are combined into a classifier that predicts as $\hat{y} = \arg\min_j ||\mathbf{x} - A_j(\mathbf{x})||_2$ so that the autoassociator with the smallest reconstruction error determines the predicted class. The overall quality of autoassociators is thus measured simply as a fraction of correctly predicted examples from the validation data set. Given a set of candidate pairs (n_{13}, n_2), the best choice for topology of autoassociators is the one that provides the highest classification accuracy.

2.2 Selection and Evaluation of Underrepresented Sequences

Given a set of autoassociators and a database of unlabeled sequences one would like to select distributionally underrepresented unlabeled sequences that would be interesting for the subsequent in-depth biological analysis. However, choosing an appropriate measure of deviation from an underlying distribution and selecting the set of underrepresented sequences using a given measure is not a trivial task.

Measures of interestingness. Applying a class-specific autoassociator to an unlabeled sequence $\mathbf{s} = \{s_i, i = 1, \dots L\}$ of length L results in an L-dimensional vector $\mathbf{e} = \{e_i, i = 1, \dots L\}$ of position-by-position based reconstruction errors, where $e_i = ||\mathbf{x}_i - \mathbf{t}_i||_2$. Therefore, given K autoassociators, each unlabeled sequence is characterized by K such vectors $\{\mathbf{e}_1, \mathbf{e}_2, \dots \mathbf{e}_K\}$. Our goal is to summarize each L-dimensional vector $\mathbf{e}_j$ with a single number $f(\mathbf{e}_j)$, where the deviation measure $f(\cdot)$ can be some of the

sample statistics. This allows representing each sequence, regardless of its length, with the deviation vector $\mathbf{d} = [f(\mathbf{e}_1), \ldots f(\mathbf{e}_K)]$ of length K.

For the function $f(\cdot)$ we consider two measures: (1) $average(\mathbf{e})$ calculated as an average error of an autoassociator. This simple measure is aimed at detecting sequences that are overall the most different from an underlying sequence distribution; (2) $max(\mathbf{e})$ calculated as the maximum error of an autoassociator. This measure is aimed at identifying sequences containing the most underrepresented regions. Such regions in genes or proteins could indicate the existence of unusual biological domains with potentially interesting structural or functional properties.

Outlier Evaluation. It is highly desirable if evaluation of the selected underrepresented sequences is performed within the data mining environment. In our approach for outlier analysis we utilize the fact that some background knowledge often exists about unlabeled biological sequences. In the simplest form, a set of keywords is available that summarize the knowledge about each sequence. For example, in SwissProt, a high quality source of background information on more then $100,000$ proteins, about 840 keywords are in use to describe protein functional properties [2].

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. Therefore, in the framework of the evaluation of underrepresented sequences, one should be careful in designing the experiment and in interpreting the results of the evaluation.

We construct several summaries that compare the frequencies of the keywords in the whole database of unlabeled sequences with the selected set of underrepresented sequences. Such a summary potentially allows better understanding the properties of the identified sets of outliers in order to decide if their functional properties justify further in-depth analysis of the selection.

3 An Application: Deviations from Known Ordered and Disordered Proteins

We illustrate the use of the proposed system on an important biological problem related to protein disorder. One of the problems impeding the progress in understanding protein disorder is a relatively small data set of proteins with confirmed disorder property. Combining the proposed approach with an in-depth study of selected outliers could result in increasing the database of disordered proteins.

3.1 Background: Protein disorder property

Protein is a macromolecule defined by a sequence of amino acids corresponding to an alphabet of twenty symbols. In contrast to the standard paradigm that a fixed 3-D structure is necessary for a function, around 150 proteins have been shown to carry out functions by means of regions that are incompletely folded or that are even completely unfolded [3]. Annotating the biological functions of these disordered proteins, over 25 distinguishable functions have been identified ranging from ligand binding to flexible linkers to sites of post-translational modification. This important biological discovery is suggesting the need for a reassessment of the widely accepted protein sequence-to-structure-to-function paradigm [3].

While the evidence indicates the importance of protein disorder, the number of 150 confirmed disordered proteins is still small compared to several thousand completely ordered proteins from PDB structural database [6], or to more then $100,000$ protein sequences in the well-annotated SwissProt database. Just by considering these numbers one might conclude that protein disorder is more an exception than a rule in nature. However, such a small number of confirmed disordered proteins is a consequence of a strong experimental bias towards ordered proteins.

It is therefore very important to derive a target list of likely disordered proteins from a set of unlabeled proteins Such proteins could then be examined in-depth using time and financially more demanding methods (e.g. literature searches and "wet lab" structure determination experiments). Computational approaches to selecting the candidate list of proteins from unlabeled sequences include: (i) using predictors of disorder, and (ii) detecting sequences distributionally different from known examples of protein order and disorder. The second approach is more likely to qualitatively improve the set of disordered proteins by discovering examples substantially different from already known disordered proteins. The proposed methodology is aimed towards developing appropriate tools to facilitate the desired data enhancement using the second approach.

3.2 Data Sets

Labeled sequences. Our data set of labeled sequences consisted of 152 proteins containing disordered regions longer then 40 consecutive positions, and of 290 completely ordered proteins. All pairs of the labeled sequences have less than 25 percent sequence identity. Some disordered regions identified by NMR, circular dichroism or protease digestion were found starting by keyword searches of PubMed (http://www.ncbi.nlm.nih.gov). Additionally, disordered regions in X-ray crystal structures were identified by searching for residues having backbone atoms that are absent from the ATOMS lists in their PDB files. Among the 152 disordered proteins we identified 162 long disordered regions with the lengths of disordered regions ranging from 40 to more then 1000. The set of 290 nonredundant completely ordered proteins was extracted from PDB. In total, our labeled data set consists of $22,434$ disordered amino acids and $67,548$ ordered amino acids.

Unlabeled sequences. We used $101,602$ proteins listed in October 2001 release 40 of SwissProt database (http://us.expasy.org/sprot/) as a set of unlabeled proteins. From SwissProt we extracted amino acid sequences of each protein as well as the associated lists of keywords used in the evaluation of selected outliers. A total of 840 keywords are used in SwissProt to provide information about functional and structural properties of various proteins. SwissProt is a statistically redundant database as it contains a large number of homologous proteins with highly similar sequences. To reduce the redundancy that could bias the outlier analysis we used the ProtoMap database [7] that contains an automatically generated hierarchical classification of protein sequences based on the sequence similarities of all sequences in the SwissProt database. Using the ProtoMap we assigned each SwissProt protein to one of $17,676$ clusters obtained by using psi-blast similarity search [1] with the E-value threshold set to $1e^{-0}$.

3.3 Construction of Distribution Models for Ordered and Disordered Sequences

Attribute construction. The attributes used in autoassociators should be able to capture information relevant to the distributional properties of ordered and disordered proteins. For learning class-conditional distributions of disordered and ordered proteins we used only the following 7 attributes selected according to our earlier experiments and expert knowledge [5]:

Attribute 1. Real-valued disorder predictions obtained by our general-purpose linear disorder predictor;

Attributes 2-4. Real-valued predictions obtained by 3 linear disorder flavor predictors that capture different specific disorder properties better;

Attribute 5. Sequence complexity calculated as a nonlinear combination of amino acid frequencies over a window of length 41;

Attributes 6-7. Flexibility and hydropathy calculated as appropriate linear combinations of amino acid frequencies over a window of length 41.

Autoassociator topology optimization. Given the choice of attributes, we examined accuracy of autoassociators with the numbers of hidden neurons n_{13} ranging from 1 to 10 for the first and the third hidden layer, and n_2 ranging from 1 to 6 for the second hidden layer (upper bound being limited by a design choice of seven attributes at input and output layers).

To measure the overall quality of autoassociators as defined at Section §2.1 we used a 5–cross-validation procedure on balanced data with equal number of ordered and disordered proteins. The optimal distribution model structure was achieved with $n_{13} = 8$ neurons in the outer hidden layers and $n_2 = 1$ neuron in the inner layer with the achieved accuracy near 69.8 percent. While this accuracy is lower than the 80.5 percent accuracy achieved by our general-purpose disorder predictor, it is significantly higher then 50 percent accuracy of a random predictor. Considering the fact that class-conditional distribution autoassociators of disordered and ordered proteins were constructed in a completely unsupervised manner such that accuracy maximization was not explicitly enforced, near 70 percent accuracy should be considered an indicator of a satisfactory quality of distribution learning.

3.4 Illustration: Selection and Evaluation of Outliers

Data preparation. Starting from the $101,602$ SwissProt sequences, we removed all sequences not listed in ProtoMap database leaving $93,863$ sequences. We then applied one round of blastp algorithm (http://www.ncbi.nlm.nih.gov/BLAST/) to find all sequences among the remaining data similar to the 152 disorder sequences and 290 ordered sequences. We used the default parameters of the blastp software and an E-value threshold was set to 1. The procedure discovered a total of $18,965$ similar sequences that were removed from the further consideration. Since max and average interestingness measures described in Section §2.2 are sensitive to sequence length we retained only $34,233$ sequences with lengths between 200 and 500 amino acids. Finally, to remove the evolutionary bias in the remaining data, we used ProtoMap to select only a single representative sequence from each ProtoMap cluster in the remaining data. To retain the most well-studied sequences, each cluster representative was chosen as the

sequence with the most keywords in SwissProt database. This resulted in a total of $6,964$ sequences that we will call $SWISS$ data set.

Outlier selection. We applied the autoassociators of protein order and disorder on $SWISS$ sequences. and selected all sequences with $f(\mathbf{e}_{order}) + f(\mathbf{e}_{disorder}) > \theta$, for some threshold $\theta > 0$, where $f(\cdot)$ was max or average interestingness measure from Section §2.2. $OutMax$ with $2,582$ sequences was obtained using the max criterion, and $OutAvg$ with $2,079$ sequences was obtained using the average criterion.

Evaluation of the selection. To improve evaluation of $OutMax$ and $OutAvg$ selections we constructed an additional data sets $DisPred$ with $1,003$ sequences representing all $SWISS$ sequences predicted to have disorder longer than 40 consecutive amino acids by our linear disorder predictor [5]. For each of the 5 data sets we calculated the frequency of each of the 840 keywords listed in SwissProt. We list a summary for 10 of the most interesting keywords in the Table. Each of these keywords appears in at least 50 $OutMax$ proteins and their fraction in $OutMax$ is at least 50% larger than in $SWISS$.

It can be seen that selections based on max and average measures resulted in proteins with similar properties, with only major difference with respect to Transit Peptide proteins. Frequency of Glycoprotein, Inner Membrane, Membrane, Transmembrane and Transport proteins in $OutMax$ and $OutAvg$ is extremely high as compared to the other 3 data sets. They are all members of a large family of membrane proteins known to be structurally and functionally diverse and are highly underrepresented among our labeled order and disorder sequences. It is likely that most of the bias in our labeled data comes from membrane proteins. Although membrane proteins do not seem to be directly related to protein disorder, we plan to exploit this finding to improve accuracy of disorder prediction. Furhter analysis of the selections is in progress.

Table 1: Comparison of frequencies of 10 interesting keywords associated with proteins in 5 data sets

Keyword	Frequency of a keyword %			
	SWISS 6,965	DisPred 1,003	OutMax 2,582	OutAvg 2,079
Chloroplast	1.3	0.7	2.1	2.0
Glycoprotein	5.9	5.7	12.5	10.8
Inner Membrane	2.1	2.1	4.8	4.7
Membrane	21.1	13.2	44.0	46.7
Multigene Family	2.5	4.2	4.5	4.6
Receptor	1.1	0.8	2.7	2.5
Repeat	5.9	13.5	10.0	11.1
Transit Peptide	1.4	0.9	2.7	1.1
Transmembrane	17.7	8.9	39.7	43.4
Transport	5.6	4.0	10.6	10.8

4 Conclusions.

In this paper we proposed a data mining procedure for discovering statistically novel sequences with interesting biological properties. It is designed to reduce a common problem of sampling bias in bioinformatics. Our approach is enhancing the labeled sequence data sets by providing a domain expert with a set of sequences most likely to significantly improve the quality of labeled data. The domain expert is then expected to perform an in-depth evaluation of the selected sequences using traditional scientific procedures.

We provided an illustration of applying the proposed methodology to detection of novel examples of disordered proteins. By outlining the complete data mining process on this important real-life problem we pointed to a number of challenges and provided guidelines for solving them. Specific issues included deciding on an appropriate sequence data representation, optimizing autoassociator topology, measuring the interestingness of unlabeled sequences, and effectively using existing domain knowledge about unlabeled sequences.

Although the initial results seem promising, only extended use of the proposed methodology by domain experts will prove its usefulness. We are currently collaborating with bio-scientists to evaluate and tune the proposed data mining process.

References

[1] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*, Nucleic Acids Res., 25 (1997), pp. 3389-3402.

[2] A. Bairoch and R. Apweiler, *The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999*, Nucleic Acids Res., 27 (1999), pp. 49-54.

[3] A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva, and Z. Obradovic, *Intrinsic disorder and protein function*, Biochemistry, 41, 21(2002), pp. 6573-6582.

[4] N. Japkowicz, C. Myers and M. Gluck, *A novelty detection approach to classification*, Proc. Int'l Joint. Conf. Artificial Intelligence, , Montreal, (1995), pp. 518-523.

[5] S. Vucetic, C. J. Brown, C. J. Crosetto, Y. J. Van, A. K. Dunker and Z. Obradovic, *Flavors of disordered proteins*, Proteins (in press).

[6] J. Westbrook, Z. Feng, S. Jain, T. N. Bhat, N. Thanki, V. Ravichandran, G. L. Gilliland, W. Bluhm, H. Weissig, D. S. Greer, P. E. Bourne, and H. M. Berman, *The Protein Data Bank: Unifying the archive*, Nucleic Acids Res., 1, 30 (2002), pp. 245-8.

[7] G. Yona, N. Linial and M. Linial, *ProtoMap: Automatic classification of protein sequences, a hierarchy of protein families, and local maps of the protein space*, Proteins, 37 (1999), pp. 360-378.