Self- Organizing Maps and Protein Analysis

Abstract

Self-Organizing Maps have been widely used in analyzing proteins. Proteins consist of amino acids; the sequence of amino acids usually determines the function of the protein. Therefore if two proteins are similar in sequence they are usually similar in function as well. Self-organizing maps have been used mainly to classify proteins in families. Another application is to use self-organizing maps to predict the secondary structure of a protein based on the primary structure, which is the amino acid sequence. Self-organizing maps have been used in conjunction with other methods such as spectral analysis of the protein in order to accurately depict the structure of the protein. In addition to analyzing patterns in proteins, gene expression patterns have been clustered using a modified SOM. The main architecture is a hierarchical tree that clusters the enormous amount of data in connection with the structure and pattern of the protein. Many of the experiments using SOM in order to predict the secondary structure and simply show similarities in proteins have been proven successful but not entirely accurate. As more becomes known about proteins through the use of unsupervised neural networking more research has been conducted about genes and drug discovery and other biomedical information.

Introduction

Proteins make up one of the four groups of macromolecules. There are four structures in a protein. The primary structure of a protein is its amino acid sequence. This sequence can range anywhere from around twenty to more than forty thousand amino acids. The secondary structure of a protein consists of how these amino acids fold, for example alpha helixes and beta sheets. The tertiary structure of a protein consists of subunits and how they are arranged in three-dimensional space. The quaternary structure of a protein consists of multiple subunits. The secondary, tertiary and quaternary structures of protein are all determined by the primary structure. An important task in biology is to predict the secondary structure of the protein [1]. Protein structure determines function and proteins basically control everything about living organisms. Therefore if every structure of a protein can be predicted simply by the amino acid sequence then the function can be known and this gives rise to what information is encoded in the genome. This will allow scientists to create better treatments for certain protein deficiencies and genetic disorders.

Self-Organizing maps have been used in order to classify proteins into families. These families depend upon the amino acid sequence of the protein, which in turn describe their function. The performance of this neural network relies hierarchical clustering. The results are then analyzed from a tree that is produced by linking the clusters [2]. In another aspect SOMs are used in order to recognize certain pattern in protein structure. For example some proteins that are found in bacteria contain helical domains that are distinct to these proteins. The SOM trains twelve sequences using sixteen neurons [3]. Other algorithms being used to analyze proteins involve training the SOM with six thousand secondary structures [4] and this data is then clustered into about two hundred classes. Some problems encountered with using SOM are when the proteins are of different sizes. Though they could be almost identical in shape the SOM does not prove accurate in this case.

Another way of analyzing proteins accurately is using SOM with another method. There are many protein databases available already. These databases allow the user to enter a sequence and then similar sequences from proteins are returned. Many of the SOMs that rely on pattern recognition use these databases in efforts to check the accuracy of the network [5]. In pattern recognition the SOM breaks the sequence into segments [6]. The SOM for most of the protein analysis programs uses Euclidean distance as measurement. A different way of analyzing structure is using a circular dichroism spectrum to train the SOM map. The spectrum measures optical activity of the protein; this activity is dependent upon the secondary structure. The SOM arranges the spectra.

The encoding of the amino acid sequence differs for each algorithm. However, the idea of analyzing proteins relies on two properties: the primary structure, which is the amino acid sequence and the secondary structure, which is the folding of these amino acids.

Methods

Self-Organizing Maps are used to cluster the protein sequences according to similarities in the amino acid sequence. The sequences cannot be compared to directly because the lengths of the protein sequences vary. Therefore two proteins can be very similar in pattern but be of different sizes. The SOM algorithm for comparing amino acid sequences breaks converts the amino acids into a twenty by twenty matrix. This data is used as the input vector. The SOM is trained with multiple protein sequences to allow proteins of different sizes to be analyzed. In general related proteins are mapped to nearby locations on the map. This allows the protein sequences to be classified into families [7].

The most important part of the SOM algorithm for protein analysis is to properly encode the amino acid sequences. If any part is encoded incorrectly this will lead to erroneous results [8]. Each amino acid sequence is coded using binary numbers. Each of the twenty amino acids is coded for. Corresponding amino acids are denoted by a one while a zero represents non-similar sequences. The map itself is a two-dimensional layer, each part of the grid represents one of the sequence vectors. The sequence vectors on the grid are known as slot vectors. During the training process, the slot vectors are set to values of the mean of the sequence vectors. The Euclidean distance from a slot vector is calculated and the winning slot vector is chosen an updated according to the Kohonen training algorithm. The map is obtained when the slot vectors in the winning neighborhood are also updated. During the training process the neighborhood decreases to simply one slot vector. The clustering properties of the map are used to classify the proteins into families. Some problems arise since the SOM slot vector cannot accurately represent the entire family. However these problems are resolved using the distance of the sequences in each cluster to determine how large each SOM slot vector should be [9]. The tree is formed by branching the clusters that contain the same sequences at consecutive levels. Therefore the clusters continue to split into subgroups. The map makes clusters of similar size. The slot vectors are compared linearly and then trained to allow for a better representation of the protein families. Each sample protein is assigned to the slot vector. Groups of similar sequences tend to stay in the same or nearby slots [10]. This algorithm allows for classification of protein families.

Other algorithms have been used to functionally classify proteins based on their secondary structure and amino acid sequence. As before the encoding is highly important in addressing this problem. Not only must the input data contain the sequence it must also contain information about the size of the protein and arrangement of amino acids in order to obtain an accurate depiction of the secondary structure [11]. The structure is represented in a twenty- element vector. The sequence is broken down in half in order to contain information about both the alpha helix and beta sheet; the most common secondary structures. However, in order for the network to compare proteins of different sizes, the network elongates shorter sequences. The length of the protein is set to the mean of all the proteins presented to the network [12]. The data is placed in a matrix and divided into two groups and instead of training every single sequence; only every other entry is taken and placed in the training data. The SOM architecture is hexagonal and two-dimensional. The network was trained using six thousand secondary structures. The output map was two hundred and twenty-five classes so it can be represented using a fifteen by fifteen map. The map was designed using MATLAB. The weights of the units were calculated using Euclidean distance. This algorithm proved useful in clustering the proteins based on structure.

The next algorithms differs from the others in the sense that the SOM is taking spectrums of the proteins and organizing them according to their corresponding sequences. The training set consists of twenty-four circular dichroism spectra of the proteins. The size of the training set then determines the size of the output layer. In order to achieve better performance the training set data increased to forty-four spectra. The weights are calculated using Euclidean distance as usual [13]. The map in this case is sixteen-by sixteen. Presenting the spectra to the network numerous times performs training. The amino acid sequences are labeled according to the shape they acquire in the protein. Each neuron is labeled with a three- letter sequence on a hexagonal map [14]. After training if self-organization has taken place accurately the spectra will be mapped on the grid. This algorithm is used to make a prediction on the secondary structure of a protein. Its efficacy is tested against a technique used to observe the structure of a protein known as X-ray crystallography [15].

A more general algorithm is used to simply self-organize all of the known protein sequences. One of the problems with this is that encoding long sequences of amino acids is very difficult. This new algorithm uses Euclidean metric space. The reason behind this mathematical measurement is that metric space encodes far more information than any other unit of measurement. The algorithm takes fifty amino acid segments of the protein and performs the Euclidean calculations [16]. The major part of this algorithm is its ability to cluster the data. The parameters of the clusters are tested against independent data. This is done in order to ensure that the data is clustered to the map accordingly. The algorithm splits the data into two clusters. Each data point is associated with a point in the center of the data. To evaluate the efficacy of this algorithm the output is tested against various protein databases.

Another SOM algorithm known as Self-Organizing Tree Algorithm (SOTA) is used to cluster gene expression patterns [17]. The output is a binary tree with a cluster of subgroups. Gene expression patterns tie in closely with protein sequences because one gene codes for one protein. Therefore if all known proteins are classified according to sequence the genes that code for these proteins can be classified more easily and vice versa.

Results

In the classification of protein families the tree that is generated represents a protein evolution tree. All of the families branch out from one another. The tree has a root and each branch comes from the root or a subgroup [18]. This tree can in fact be used to study the evolutionary relationship between proteins, considering there is relevance. A fascinating note of the tree obtained is that when the map is small certain protein families are grouped together but separated when the map is larger [19]. This is due to the fact that though the proteins are similar, when they are trained on a larger map with larger slot vectors there is more accuracy, and the entire sequence can be presented to the network. These results were checked against true evolutionary charts of proteins and the SOM proved efficient [20].

In the experiment where the classification was made on the functional aspect of the protein the SOM proved to be inaccurate in certain cases [21]. The reason for this is though a protein's structure and function are contingent upon each other sometimes proteins with similar structure have completely different functions.

In another experiment involving the circular dichroism spectra and secondary structure of the protein the SOM mapped certain structures to parts of the map. The beta sheet was mapped to the upper left of the region [22]. The map shows that weight vector arrangement is related to structure. Therefore combining the spectra with the amino acid sequences and SOM showed promising results.

The organization of all known protein sequences obtained one hundred and six clusters [23]. The tree obtained varied from clusters with many to few splits. An interesting note is that the splits that are present occur very early in the clustering process. According to the tree there are two hundred families per cluster. The SOM did an excellent job clustering the secondary structure of the proteins. For example, the alpha helixes and beta sheets are in separate clusters, as they should be. The map also clusters proteins that contain the same amino acids. In other words if the sequence does not contain one amino acid however, it is similar in sequence to another protein that does contain that specific amino acid these proteins are then mapped into different clusters. Another interesting biological aspect is that the SOM has mapped proteins that contain certain amino acid characteristics together. For instance some proteins contain low amounts of hydrophobic amino acids while others contain high amount. These proteins are clustered distinctly [24].

Discussion

Each algorithm uses SOM to classify and/or predict the structure of the protein. It seems safe to say that SOMs are better at classifying and clustering the proteins into families than they are at predicting the structure. One reason could be that Self-Organizing Maps are designed to cluster they are not specifically designed to generate helical and parallel sheets. A consensus can be made that though its predicting ability is not as accurate as scientists would like it to be, the recognition and clustering ability is phenomenal. For instance, the Self-Organizing Maps does in fact map certain structural proteins together [25]. Sequences that form a certain biological structure known as a motif were mapped into a certain cluster. This is proof that certain amino acid sequences form a structure and this structure corresponds to the function of that protein. For example, a motif known as a zinc finger is characteristic of proteins that bind DNA [26]. Proteins with this specific sequence were recognized by the SOM and clustered together. Another example known as homeobox domain is characteristic of helical proteins that also bind DNA. Proteins that fit this pattern were also clustered together. All of these proteins function to regulate transcription in the cycle of DNA replication. Other proteins function to maintain cellular activity and these proteins contain the same amino acids. Therefore they too are mapped in a separate cluster.

When it comes to predicting the secondary structure of a protein, SOM is used with spectral devices. This method of protein analysis uses the clustering of the SOM to deduce the structural representation of the protein [27]. However, it was not accurate at predicting certain structures of proteins that the SOM had not seen before. If the spectra of that particular structure was presented to the network in training the estimation of a protein’s secondary structure was not correct.

Conclusion

In experiments using protein sequences and Self-organizing maps the neural network is quite useful in classifying proteins to certain families. This information is highly important in today’s biological world because of the new advances being made in genetic research. As the information about proteins continues to grow scientists will be able to predict the structure of unknown proteins using SOMs and this will enable them to determine their function in the human body. With this information new medications will arise that can treat people that have protein disorders. Technology will improve the accuracy of Self-Organizing Maps and this will greatly impact the fields of molecular biology and medicine.

References

1 .Gerstein, Mark, Jansen, Ronald, The current excitement in bioinformatics analysis of whole-genome expression data: How does it relate to protein structure and function? Current Opinion in Structural Biology 10 (2000) p.1

2. Andrade, Miguel A, Casari, Georg, Sander, Chris, Valencia, Alfonso Classification of protein families and detection of the determinant residues with an improved self-organizing map Biological Cybernetics, 76, 441-450 (1997) p.445

3. Hanke, Jens, Beckmann, Georg, Bork, Peer, Reich, Jens Self-Organizing hierarchic network for pattern recognition in protein sequence Protein Science 5(1) 72-82 (1996) p.72

4. Pollock, Robert, Lane, Toby, Watts, Michael A Kohonen Self-Organizing Map for the functional classification of proteins based on one-dimensional sequence information University of Otago, New Zealand
http://divcom.otago.ac.nz/infosci/kel/CBIIS/pubs/pdf/ssSOM_paper2.1/pdf
p.1

5. Linial, Michal, Linial, Nathan, Tishby, Naftali, Yona, Golan Global Self-Organization of all known protein sequences reveal inherent biological signatures Molecular Biology 268 539-556 (1997) p.7

6. Linial, Michal, Linial, Nathan, Tishby, Naftali, Yona, Golan Global Self-Organization of all known protein sequences reveal inherent biological signatures Molecular Biology 268 539-556 (1997) p.1

7. Kangas, Jari On the Analysis of Pattern Sequences by Self-Organizing Maps Helsinki University of Technology (1994) p.57

8. Andrade, Miguel A, Casari, Georg, Sander, Chris, Valencia, Alfonso Classification of protein families and detection of the determinant residues with an improved self-organizing map Biological Cybernetics, 76, 441-450 (1997) p.447

9. Andrade, Miguel A, Casari, Georg, Sander, Chris, Valencia, Alfonso Classification of protein families and detection of the determinant residues with an improved self-organizing map Biological Cybernetics, 76, 441-450 (1997) p.444

10. Andrade, Miguel A, Casari, Georg, Sander, Chris, Valencia, Alfonso Classification of protein families and detection of the determinant residues with an improved self-organizing map Biological Cybernetics, 76, 441-450 (1997) p.448

11. Pollock, Robert, Lane, Toby, Watts, Michael A Kohonen Self-Organizing Map for the functional classification of proteins based on one-dimensional sequence information University of Otago, New Zealand

http://divcom.otago.ac.nz/infosci/kel/CBIIS/pubs/pdf/ssSOM_paper2.1/pdf
p.2

12. Pollock, Robert, Lane, Toby, Watts, Michael A Kohonen Self-Organizing Map for the functional classification of proteins based on one-dimensional sequence information University of Otago, New Zealand
http://divcom.otago.ac.nz/infosci/kel/CBIIS/pubs/pdf/ssSOM_paper2.1/pdf
p.4

13. Per Unneberg, Juan Merelo, Pablo Chacon, Federico Moran SOMCD: Method for Evaluating Protein Secondary Structure from UV Circular Dichroism Spectra
Proteins: Structure, Function, and Genetics 42:460-470 (2001) p.462

14. Per Unneberg, Juan Merelo, Pablo Chacon, Federico Moran SOMCD: Method for Evaluating Protein Secondary Structure from UV Circular Dichroism Spectra
Proteins: Structure, Function, and Genetics 42:460-470 (2001) p.463

15. Per Unneberg, Juan Merelo, Pablo Chacon, Federico Moran SOMCD: Method for Evaluating Protein Secondary Structure from UV Circular Dichroism Spectra
Proteins: Structure, Function, and Genetics 42:460-470 (2001) p.466

16. Linial, Michal, Linial, Nathan, Tishby, Naftali, Yona, Golan Global Self-Organization of all known protein sequences reveal inherent biological signatures Molecular Biology 268 539-556 (1997) p.7

17. Herrero, Javier, Valencia, Alfonso, Dopazo, Joaquin A hierarchic unsupervised growing neural network for clustering gene expression patterns Bioinformatics Vol.17 No.2 (2001) p.1

18. Andrade, Miguel A, Casari, Georg, Sander, Chris, Valencia, Alfonso Classification of protein families and detection of the determinant residues with an improved self-organizing map Biological Cybernetics, 76, 441-450 (1997) p.445

19. Andrade, Miguel A, Casari, Georg, Sander, Chris, Valencia, Alfonso Classification of protein families and detection of the determinant residues with an improved self-organizing map Biological Cybernetics, 76, 441-450 (1997) p.447

20. Andrade, Miguel A, Casari, Georg, Sander, Chris, Valencia, Alfonso Classification of protein families and detection of the determinant residues with an improved self-organizing map Biological Cybernetics, 76, 441-450 (1997) p.448

21. Pollock, Robert, Lane, Toby, Watts, Michael A Kohonen Self-Organizing Map for the functional classification of proteins based on one-dimensional sequence information University of Otago, New Zealand
http://divcom.otago.ac.nz/infosci/kel/CBIIS/pubs/pdf/ssSOM_paper2.1/pdf
p.3

22. Per Unneberg, Juan Merelo, Pablo Chacon, Federico Moran SOMCD: Method for Evaluating Protein Secondary Structure from UV Circular Dichroism Spectra
Proteins: Structure, Function, and Genetics 42:460-470 (2001) p.463

23. Linial, Michal, Linial, Nathan, Tishby, Naftali, Yona, Golan Global Self-Organization of all known protein sequences reveal inherent biological signatures Molecular Biology 268 539-556 (1997) p.7

24. Linial, Michal, Linial, Nathan, Tishby, Naftali, Yona, Golan Global Self-Organization of all known protein sequences reveal inherent biological signatures Molecular Biology 268 539-556 (1997) p.11

25. Linial, Michal, Linial, Nathan, Tishby, Naftali, Yona, Golan Global Self-Organization of all known protein sequences reveal inherent biological signatures Molecular Biology 268 539-556 (1997) p.9

26. Linial, Michal, Linial, Nathan, Tishby, Naftali, Yona, Golan Global Self-Organization of all known protein sequences reveal inherent biological signatures Molecular Biology 268 539-556 (1997) p.10

27. Per Unneberg, Juan Merelo, Pablo Chacon, Federico Moran SOMCD: Method for Evaluating Protein Secondary Structure from UV Circular Dichroism Spectra
Proteins: Structure, Function, and Genetics 42:460-470 (2001) p.470