Weightless Neural Network Array for Protein Classification

28434

Martin Chew Wooi Keat[1], Rosni Abdullah[2], Rosalina Abdul Salam[3],
and Aishah Abdul Latif[4]

28434 PLFSI

[1, 2, 3] Faculty of Computer Science, Universiti Sains Malaysia, Penang, M
martin.wooi.keat.chew@intel.com
[4] Doping Control Center, Universiti Sains Malaysia, Penang, Malaysia
aishah@dcc.usm.my

Abstract. Proteins are classified into superfamilies based on structural or functional similarities. Neural networks have been used before to abstract the properties of protein superfamilies. One approach is to use a single conventional neural network to abstract the properties of different protein superfamilies. Since the number of protein superfamilies is in the thousands, we propose another approach – one network *attuned* to one protein superfamily. Furthermore, we propose to use weightless neural networks, coupled with Hidden Markov Models (HMM). The advantages of weightless neural networks are: (a) the ability to learn with only one presentation of training patterns – thus improving performance, (b) ease of implementation, and (c) ease of parallelization – thus improving scalability.

1 Introduction

This concept paper relates to the field of protein classification, for the purpose of functional determination, in order to assist the process of drug target discovery. Given an unlabeled protein sequence S and a known superfamily F, we wish to determine whether or not S belongs to F. We refer to F as the target class and the set of sequences not in F as the non-target class. In general, a superfamily is a group of proteins that share similarities in structure and/or function [1]. If the unlabeled sequence S is detected to belong to F, then one can infer the function of S. Neural networks have been used to classify proteins before [2]. However, our proposed approach will use an array of weightless neural networks. Weightless neural networks have been used for image recognition before [5]. We modified the concept of a weightless neural network to suit the purpose of protein classification. A particular weightless network is attuned to a particular protein superfamily. An unknown protein sequence submitted to the network array is deemed to belong to the protein superfamily represented by the network with the most positive output (i.e. the *resonant* network). We may also be able to deduce the degree of relationship of the sequence to other protein superfamilies by comparing the outputs of the other networks, relative to the resonant network.

K.-M. Liew et al. (Eds.): PDCAT 2004, LNCS 3320, pp. 168–171, 2004.

2 System Description

The first step is to determine a particular transformation/encoding function, for the purpose of deriving from protein sequences, an array of real values to serve as inputs into the weightless network. The intention of the transformation function is to bring to surface (i.e. make explicit) the implicit feature(s) of protein sequences, for the purpose of abstraction. The transformation function plays a major role in the accuracy of the system. A particular protein superfamily is paired to the particular transformation function most suited to its properties. Each protein superfamily will be abstracted by its own weightless network. For the purpose of this paper, we will use a simple transformation function called 2-gram encoding for every protein family. The frequency of unique character *pairs* in a protein sequence is counted. Since there are 23 different amino acids, 2-gram encoding will give rise to 23x23 possible pairings. Therefore, the neural network will have 23x23 input units. Each input unit represents a particular *feature* of the protein. Different encoding techniques will give rise to different input features. For example, 3-gram encoding will give rise to 23x23x23 features. Each feature can be represented by an address. Since we have 23x23 possible pairings for 2-gram encoding, we need at least 23x23 addresses. The first feature will be mapped to address 000000000, the second feature to 000000001, etc. The content of each address will be initialized to zero. As each training sequence is presented to the weightless network, the value for a particular feature is accumulated in the address mapped to that feature.

Encoding systems such as 2-gram highlights global similarities at the expense of local similarities (i.e. local similarities such as motifs are lost during the transformation process). In order to factor in local similarities, a Hidden Markov Model (HMM) is used to abstract the motif indicative of a particular protein superfamily. HMMs have been used before, either singularly or in conjunction with conventional neural networks [3], [4]. Different superfamilies would most probably have different indicator motifs. There could also be more than one motif for a particular superfamily. First, the cluster of motifs has to be abstracted into a HMM. When a particular substring – window-scanned from a protein sequence - is submitted to the HMM, the model is able to return a probability value, indicating the consensus of the given substring, with respect to the cluster of motifs the model represents. Once these steps are done, the system is ready to be used in a predictive mode. Only addresses with final contents over a pre-determined threshold (we call this threshold the "weightless threshold") will be selected, the rest being ignored. Given an unknown protein sequence, the 2-gram encoding method is applied to extract an array of values (i.e. an integer count for each 2-gram pair). These values will serve as inputs (I) to a particular weightless network. To calculate the final output (R) of the weightless network, a mapping is done between cells of the input array and the selected addresses of the weightless network. Only cells with a count of more than 1 will be selected for the mapping, the rest being ignored. For each matching input cell and weightless address, a point will be scored. The final output (R) is the percentage of matches with respect to the number of active addresses.

28748

PERS I

3 Experimental Results

For experimental data, we relied on the Superfamily 1.65 website (http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/cgi-bin/align.cgi). We pulled data for three protein families (acid proteases, cytochrome b5, and cytochrome c) from the SCOP 1.63 Protein DataBase. Table 1 below shows the settings of our experiment, and Table 2 shows the results we obtained. (sample F1-1 refers to sample #1 from family F1, sample F3-3 refers to sample #3 from family F3).

Table 1. Weightless network settings

Network	Sequences	Average seq. length	Weightless threshold	Active addresses
Acid proteases	118	183	120	44
Cytochrome b5	28	81	40	7
Cytochrome c	129	120	95	14

Table 2. Results

Sample	Acid proteases (F1)	Cytochrome b5 (F2)	Cytochrome c (F3)	Correct
Sample F1-1	**65%**	0%	2%	Yes
Sample F2-2	**65%**	0%	4%	Yes
Sample F3-3	**65%**	2%	2%	Yes
Sample F2-1	28%	**57%**	28%	Yes
Sample F2-2	28%	**57%**	0%	Yes
Sample F2-3	28%	**100%**	14%	Yes
Sample F3-1	**21%**	**21%**	**21%**	No
Sample F3-2	**21%**	14%	**21%**	No
Sample F3-3	21%	21%	**50%**	Yes

We chose three random samples from each family (for a total of 9 samples), and we fed each sample into every network. Every sample scored the highest in its correct network, with the exception being sample F3-1 and sample F3-3. In other words, family 1 (acid proteases) and family 2 (cytochrome b5) was well abstracted, but not family 3 (cytochrome c). We attribute this to the loss of local similarities when the families when 2-gram encoding was applied. As we mentioned before, HMMs (each unique to a particular family) are necessary for further differentiation. In the case of the two samples of exception (sample F3-1 and sample F3-3), the HMM trained on local similarities of family 3 (cytochrome c) should return a high probability value for sample F3-1 and sample F3-3, and a low probability value for samples from other families. This will help to boost the score of sample F3-1 and sample F3-3, and push down the score of samples from the other families. The quality of the prediction will then be further improved. One drawback of this system is that the training data sets for each protein family must have sufficient members, and each member must be of sufficient length. This is necessary in order to enable the network to fully abstract the properties of the protein family. The training data sets should be filtered to exclude sequences which are too short. Another drawback is that each weightless threshold must be individually adjusted to provide the optimum results.

4 Potential Contributions to the Field of Bioinformatics

Currently, neural networks require numerous iterations over the training data set to reach convergence. Furthermore, an increase in the number of input units (e.g. from 2-gram encoding to 3-gram) could lead to an increase in training time as well. We intend to explore the use of weightless neural networks to help overcome this problem. The ease with which weightless neural networks may be implemented will help make parallelization easier. Our system requires one weightless network for one protein family. If there are a thousand protein families, we will have a thousand weightless networks. Parallelization will enable different protein families to be abstracted, and repeatedly reabstracted (in the event of a new encoding formula) in parallel. Finally, when an unknown sequence is submitted to the array of weightless networks, the outputs from the array will help us decide, not only from which protein family the unknown sequence may be from, but also, its degree of relationship to other families. Even if the array has several high outputs, this could help us narrow down the possibilities of which family an unknown sequence belongs to, by helping us to focus on a few most likely candidate families.

References

1. Pandit, Shashi B., et. al. : SUPFAM – A Database of Potential Protein Superfamily Relationships. Indian Institute of Science, Bangalore (2001).
2. Wang, J., Ma, Q., Shasha, D., Wu, C. : Application of Neural Networks to Biological Data Mining : A Case Study in Protein Sequence Classification. Department of Computer Science, New Jersey Institute of Technology (2001).
3. Krogh, A : An Introduction to Hidden Markov Models for Biological Sequence. Technical University of Denmark (1998).
4. Ohler, U., Stemmer, G., Niemann, H. : A Hybrid Markov Chain - Neural Network System for the Exact Prediction of Eukaryotic Transcription Start Sites. University Erlangen, Nuremberg, Germany. (2000).
5. Burattini, E., DeGregorio, M., Tamburrini, G. : Generating and Classifying Recall Images by Neurosymbolic Computation. Cybernectics Institute, Italy (1998).