Protein Function Prediction by an ARTMAP Neural Network

Vassilis Cutsuridis, George Efstathiou, Michael Kokkinidis
Institute of Molecular Biology & Biotechnology
Foundation for Research & Technology – Hellas (FORTH)
Heraklion, Crete, Greece
vcutsuridis@gmail.com, {george_efstathiou, kokkinid}@imbb.forth.gr

Abstract

Accurate prediction of protein functions solely from its amino acid sequence is of paramount importance, particularly in the development of new drugs. An ARTMAP neural network (NN) is employed to predict a protein's function based only on its amino-acid (AA) sequence. For our protein database, a Gene Ontology-based search against the UniProt/SwissProt database for "DNA sequence-specific binding proteins". The search complement set was also retrieved. For training and testing, various size datasets were generated. Datasets were generated either by random sampling from the existing categories or by classifying the proteins first into sub-groups based on a similarity measure and then randomly sampling from each sub-group. Our NN's performance with the latter method performed better than with the former method in every size dataset. Our NN has been successful in predicting the function of a protein from its AA sequence by extracting a shared sequence-specific feature that is linked to specific DNA binding proteins. This result is of major importance in structural biology and biomedicine as it can provide a basis of the development of highly specific tools for genome modification and gene therapy.

1 Introduction

In recent years we have experienced a dramatic growth of genomic and proteomic data. Making sense of millions of protein sequences as well as their evolutionary and functional relationships is of out-most importance for the development of highly specific tools for genome modification and gene therapy.

Various statistical and machine learning techniques including neural networks have been employed in recent years to understand the proteins sequence-structure-function relationship and uncover the mechanisms of their evolution. Backpropagation neural networks in particular have been used to predict protein secondary and tertiary structure [1, 2] and to distinguish ribosomal binding sites from non-binding sites [3] and encoding regions from non-coding sequences [4]. Similarly, Adaptive Resonance Theory (ART) family neural networks have been used for the probabilistic motif discovery in biological sequences [5].

In this paper we employ an Predictive ART (ARTMAP) neural network [6] to predict the function of proteins based only on their AA sequence.

2 Methods

2.1 Dataset and protein coding

For our protein database, a Gene Ontology-based search against the UniProt/SwissProt database for "DNA sequence-specific binding proteins" (see Fig. 1 for an example of such protein) retrieved 6492 sequences of amino acids. The search complement set comprising of 524406 sequences was also retrieved. All sequences less than 50 amino acids in length were thrown out, whereas the remaining ones were made equal-in-length by padding them with "Xs" till their length

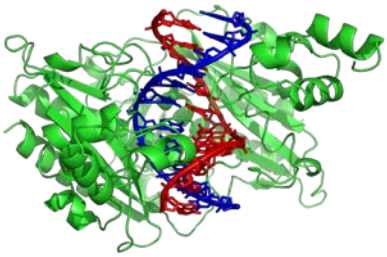

Figure 1: Restriction enzyme EcoRV (green) in a complex with its substrate DNA.

was equal to 1000. Every amino acid in each sequence was then converted into its corresponding 7-bit binary number (see Table 1) generating a

Table 1: Amino acid abbreviations and their corresponding binary codes

Amino acids								
Name	**Symbol**	**Binary code**	**Name**	**Symbol**	**Binary code**	**Name**	**Symbol**	**Binary code**
Isoleucine	I	1001001	Glycine	G	1000111	Glutamine	Q	1010001
Valine	V	1010110	Threonine	T	1010100	Asparagine	N	1001110
Leucine	L	1001100	Serine	S	1010011	Glutamic acid	E	1000101
Phenylalanine	F	1000110	Tryptophan	W	1010111	Aspartic acid	D	1000100
Cysteine	C	1000011	Tyrosine	Y	1011001	Lysine	K	1001011
Methionine	M	1001101	Proline	P	1010000	Arginine	R	1010010
Alanine	A	1000001	Histidine	H	1001000		X	0000000

sequence of length 7000 (see Fig. 2). For training and testing, various size datasets (Small dataset: 2600 proteins; Medium dataset: 4900 proteins; Large dataset: 6800 proteins) were generated. 90% of each dataset was used for training and 10% for testing.

2.2 ARTMAP system

For protein function prediction, the ARTMAP neural network was used. ARTMAP is a supervised learning system consisting of a pair of ART modules [6]. During training, an ART*a* receives a

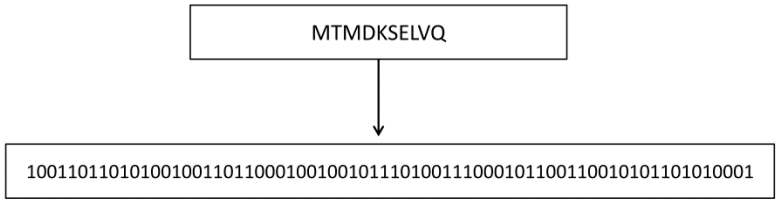

Figure 2: Example of a sequence of 10 amino acids and its binary number encoding. Each amino acid letter (e.g. 'M') in the sequence is converted to its corresponding from Table 1 binary code ('M' = 1001101) resulting into a new sequence of binary numbers of length 70.

stream of input patterns {**A**(*n*)} and an ART*b* a stream of input patterns {**B**(*n*)}, where **B**(*n*) is the correct prediction given **A**(*n*). Associative learning and a baseline vigilance parameter ρ representing a minimum matching criterion link these ART modules to enable ARTMAP to learn quickly and accurately by minimizing predictive error. High values of the vigilance parameter ensure the formation of fine categories, whereas low values the formation of coarse categories. Predictive failure at ART*b* increases ρ just enough to trigger a match tracking search by focusing attention on a different cluster of input features and checking on whether these features better predict the correct outcome. This way ARTMAP teaches itself to make a different prediction for a rare event

embedded in a cloud of similar frequent events.

3 Results

3.1 Random sampling

We first trained and tested ARTMAP's performance on predicting the function of a protein on three different size datasets (small, medium, large) created by randomly sampling the extracted UniProt/SwissProt database "DNA sequence-specific binding proteins" and "non DNA sequence binding proteins" datasets. From figure 3 we can see that when $\rho = 0.3$ (coarse categories) and as the size of the dataset increased, then the percentage of misclassified proteins ("DNA binding" vs "non-DNA binding" classes) increased from 15% to 40%. As ρ increased (fine categories) and a test input did not match any of the two learned classes, then the input was placed in the "I don't know" class. At $\rho = 0.7$ the error rate was roughly 30% regardless of the dataset size. The percent "I don't know" predictions were less 10%. At $\rho = 0.9$, the error rate dropped to less than 10% as the dataset size increased, but the percent "I don't know" predictions increased to almost 60% (large size dataset).

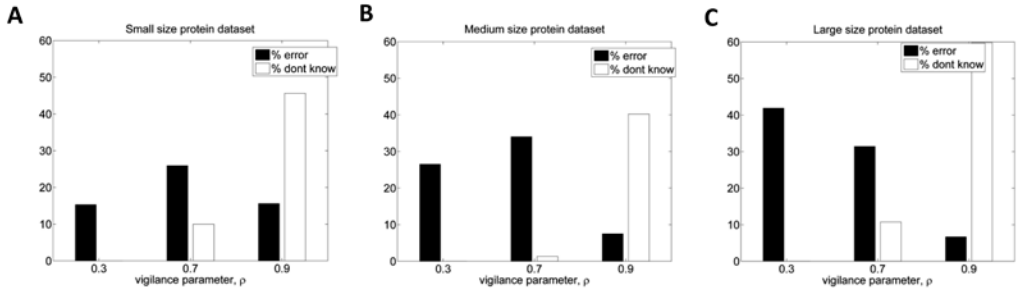

Figure 3: ARTMAP's performance using the "random sampling" methodology on three different size (small, medium, large) protein datasets as function of the vigilance parameter, ρ.

3.2 First similarity-based clustering, then random sampling

We then trained and tested ARTMAP's performance by classifying the proteins first into sub-groups based on a 40% similarity between its members and then randomly sampling 90% members from each sub-group for training and 10% for testing. This ensured that our sample was a representative one. From figure 4 we can see that when $\rho = 0.3$ and as the size of testing datasets increased, so did the error rate. When $\rho = 0.7$, the error rate fluctuated from 6% (small dataset) to 17% (large dataset). When $\rho = 0.9$, the error rate dropped to ~ 13% for the large dataset, but the number of "I don't know" predictions increased (~40%).

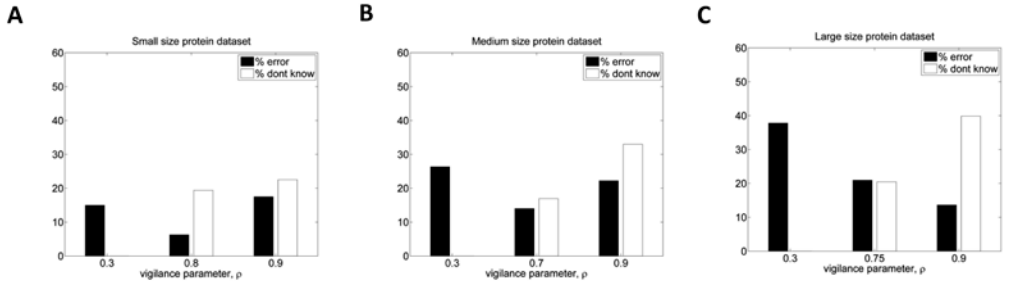

Figure 4: ARTMAP's performance using the "clustering first, then random sampling" methodology on three different size (small, medium, large) protein datasets as function of the vigilance parameter, ρ.

3.3 DNA bindingness feature

We then examined whether ARTMAP was able to extract a shared sequence-specific feature that is linked to all specific DNA binding proteins. As before, we first classified all DNA binding proteins into sub-groups based on 40% similarity and then we randomly selected N (10 or 30) sub-groups for testing and the remaining 90 sub-groups for training. The protein numbers varied in

each sub-group. From figure 5 we can see that for certain range of ρ values ($0.1 < \rho < 0.7$), ARTMAP can recognize correctly unseen during training proteins as DNA binding. As ρ increases, the ARTMAP's predictive success decreases, as it makes many more "I don't know" predictions and less correct ones.

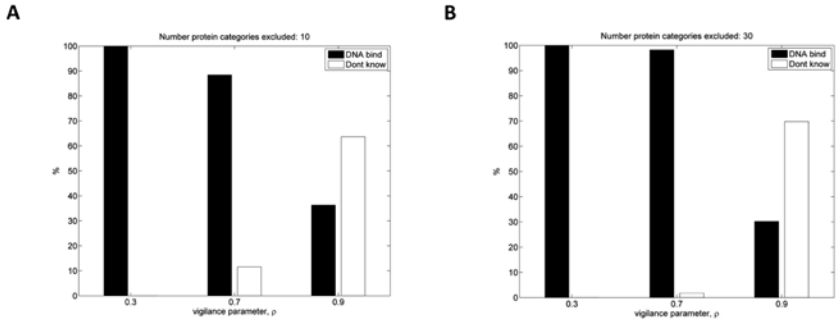

Figure 5: ARTMAP's predictive success when tested against N unseen during training DNA binding sub-groups of proteins. (A) N = 10. (B) N= 30. Protein members in each excluded sub-group varied.

4 Conclusions

In summary, we employed an ARTMAP neural network to predict the function ("DNA binding" vs "non-DNA binding") of a protein solely from its AA sequence. ARTMAP using the "clustering first, then random sampling" methodology performs better than using the "random sampling" method in all datasets and vigilance parameter values. The total number of "mis-classified" proteins and "I don't know" predictions was found to be less using the former method than with the latter method particularly in the large size dataset.

Also, ARTMAP has been successful in predicting the function of a protein from its AA sequence by extracting a shared sequence-specific feature ("DNA bindingness" feature) that seems to be linked to specific DNA binding proteins. This shared sequence-specific feature is imprinted in the weight matrix between the input (comparison) and output (recognition) layers of the ART*a* module of ARTMAP. Future research will attempt to decipher to what protein structural parameters these weight values correspond to.

Acknowledgments

This work was supported by the Regpot – InnovCrete 316223 project. The authors declare that they have no competing financial interests.

References

[1] Qian, N., & Sejnowski, T. J. (1988) Predicting the secondary structure of globular proteins using neural network models. *Journal of Molecular Biology* **202**: 865-884

[2] Bohr, H., Bohr, J., Brunak, S., Cotterill, R. M. J., Fredholm, H., Lautrup, B. and Peterson, S. B. 1990. A novel approach to prediction of the 3-dimensional structures of protein backbones by neural networks. *FEBS Letters* **261**: 43-46.

[3] Stormo, G. D., Schneider, T. D., Gold, L., & Ehrenfeucht, A. (1982) Use of the 'Perceptron' algorithm to distinguish translation initiation sites in E-coli. *Nucleic Acids Research* **10**: 2997-301

[4] Uberbacher, E., Mural, R. (1991) Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *PNAS* **88**: 11261-11265

[5] Hemalatha, M., Ranjit Jeba Thangaiah, P., & Vivekanandan K. (2009) FART Neural Network based Probabilistic Motif Discovery in Unaligned Biological Sequences. In S. I. Ao, O. Castillo, C. Douglas, D. Dagan Feng, and J.A. Lee (eds.), *Proceedings of the International Multi Conference of Engineers and Computer Scientists* Vol I, pp. 716-720. Newswood Limited.

[6] Carpenter, G.A., Grossberg, S., & Reynolds, J.H. (1991) ARTMAP: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network. *Neural Networks* **4**: 565-588.