PROTEIN FUNCTION PREDICTION BY AN ARTMAP NEURAL NETWORK

Vassilis Cutsuridis[1], Michael Kokkinidis[2]

[1]School of Computer Science, University of Lincoln, Lincoln, U.K.
[2]Department of Biology, University of Crete & IMBB-FORTH, Heraklion, Crete, Greece
Email: vcutsuridis@lincoln.ac.uk

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.

A Predictive ART (ARTMAP) neural network was exployed to predict a protein's function solely from its amino-acid (AA) sequence.

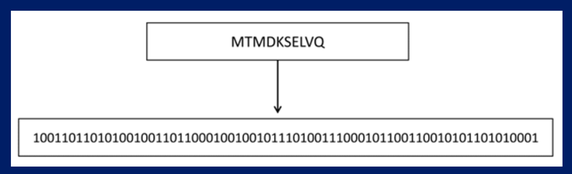
Dataset and Protein Coding

- Gene Ontology-based search against the UniProt/SwissProt database for "DNA sequence-specific binding proteins" retrieved 6492 sequences of amino acids.
- Complement search set returned 524406 sequences.
- 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 was equal to 1000.

Table 1: Amino acid abbreviations and their corresponding binary codes

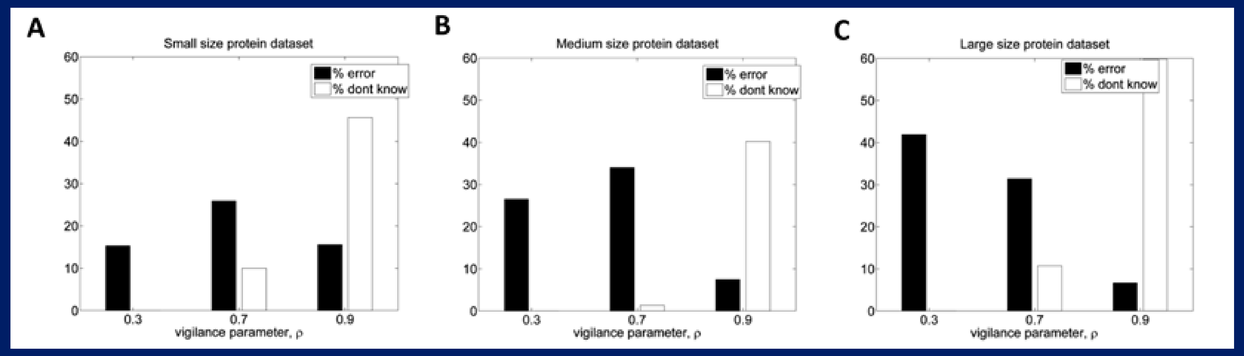
Amino acids								
Name	**Symbol**	**Binary code**	**Name**	**Symbol**	**Binary code**	**Name**	**Symbol**	**Binary code**
Isoleucine	I	1001001	Glycine	G	1000111	Glutamine	Q	1010001
Valine	V	1010110	Threonine	T	1010100	Asparagine	N	1001110
Leucine	L	1001100	Serine	S	1010011	Glutamic acid	E	1000101
Phenylalanine	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

- Every amino acid in each sequence was then converted into its corresponding 7-bit binary number generating a sequence of length 7000.



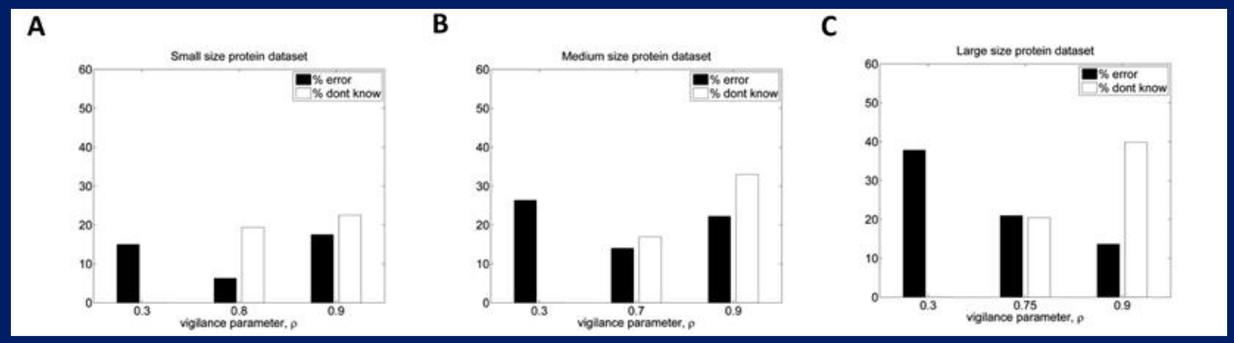
- For training and testing, various size datasets (Small dataset: 2600 proteins; Medium dataset: 4900 proteins; Large dataset: 6800 proteins) were generated.
- 90% of each dataset was used for training and 10% for testing.

Random Sampling



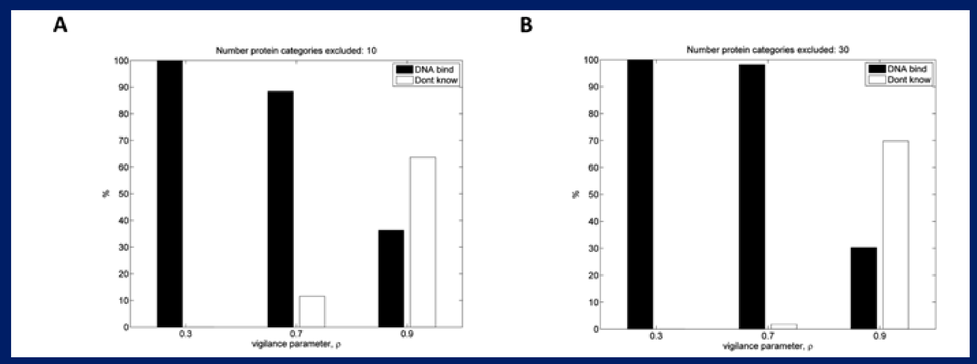
First Similarity-Based Clustering, Then Random Sampling

Classified proteins first into sub-groups based on a 40% similarity between its members and then randomly sampled 90% members from each sub-group for training and 10% for testing.

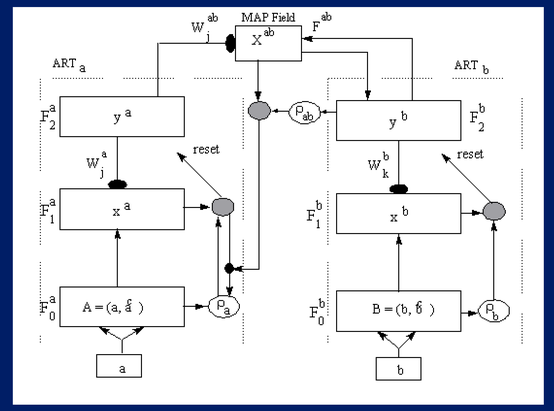


DNA Bindingness Feature

Examined whether a shared sequence-specific feature is linked to all specific DNA binding proteins. Classified all DNA binding proteins into sub-groups based on 40% similarity and then 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.



Predictive ART (ARTMAP) System



- A supervised learning system consisting of a pair of ART modules [1].
- During training, an ART*a* receives a 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 enables 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.

Conclusions

- "Clustering first, then random sampling" methodology outperformed "random sampling" method in all datasets and vigilance parameter values.
- 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.

References

[1] 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.