Analysis of Relationship between Amino Acid Composition of Proteins and Environmental Features of Microorganisms using Evolutionary Algorithm and Self-Organizing Maps

Maciej Sobczyński
Department of Genomics
Faculty of Biotechnology, University of Wrocław
ul. Przybyszewskiego 63/77
51-148 Wrocław
macsebsob@poczta.onet.pl

Paweł Mackiewicz
Department of Genomics
Faculty of Biotechnology, University of Wrocław
ul. Przybyszewskiego 63/77
51-148 Wrocław
pamac@smorfland.uni.wroc.pl

Categories and Subject Descriptors
J.3.1 [**Computer Applications**]: **Life and Medical Sciences** – *Biology and Genetics*; I.5.1 [**Pattern Recognition**]: Models – *Neural nets*; I.2.6 [**Artificial Intelligence**]: Learning – *Connectionism and neural nets*

Keywords
amino acid composition, extremophiles, evolutionary algorithm, genomics, microbes, protein, proteome, Self-Organizing Maps

1. INTRODUCTION
Amino acid composition of proteins in a given microorganisms is its characteristic feature and results from selectional constraints on protein structure and function as well as mutations introduced into the genome sequences. It was shown that the composition is related with the taxonomic affiliation of microorganisms and their environmental conditions. The most characteristic usage of amino acid was found in hyperthermophiles and thermophiles, organisms that prefer habitats with high temperature; halophiles, which live in environments with very high concentrations of salt; aerobes, which require oxygen to optimal grow; and intracellular microbes (parasites and endosymbionts) living inside the cell of their hosts.

The former analyses of amino acid composition of proteomes (sets of proteins in a given organism) were based on simple calculations of global or averaged amino acid frequencies and standard multidimensional approaches, i.e. Principal Component Analysis or Correspondence Analysis to reduce the number of parameters and dimensions [5]. Therefore, it led to generalizations and loss of important information included in the data. Here we proposed an alternative approach using Self-Organizing Maps (SOM) [3] and evolutionary algorithm (EA) [1] to describe differences in amino acid composition of proteomes in various ecological groups of prokaryotic organisms. The combination of these methods turned out very useful and sensitive to extract even small differences in amino acid composition of studied proteomes.

2. METHODS
Self-Organizing Maps were used to classify 434,000 proteins from 194 prokaryotic proteomes according to their amino acid composition and calculate a distance between the proteomes on the map [4]. In the studies we selected five representatives of all three ecological groups of microorganisms: aerobes, anaerobes, hyperthermophiles, psychrophiles and intracellular micro-organisms. We have also included artificial proteomes that consisted of amino acid sequences with the same length as real sequences but with the amino acid composition generated based on the global nucleotide composition of corresponding genomes. The distances were related to differences in amino acid compositions between studied proteomes and measured by:

$$d(\mathbf{B}_1,\mathbf{B}_2)=\frac{1}{2}\sum_{k=1}^{K}\left|\frac{n_{k\mathbf{B}_1}}{m}-\frac{n_{k\mathbf{B}_2}}{N-m}\right|,$$

where: $n_{k\mathbf{B}_1}$ and $n_{k\mathbf{B}_2}$ are numbers of proteins classified to neuron k and belonging to the proteome set $\mathbf{B}_1$ and $\mathbf{B}_2$, respectively, m and N-m are total numbers of proteins in the $\mathbf{B}_1$ and $\mathbf{B}_2$, respectively. The distance $d(\mathbf{B}_1, \mathbf{B}_2)$ ranges from 0 to 1. The teaching vector x_i represented the percentage composition of 20 amino acids in protein i belonging to one of two compared proteomes. We tested all possible 190 rectangular topologies of SOM with dimension ranged from 2 × 2 to 20 × 20 neurons. Finally, we selected the topology that minimized the three criteria: Bayesian Information Criterion, topological error and spatial autocorrelation.

Next, using SOM as a classifier, the evolutionary algorithm was applied to identify subsets of the most distinctive amino acids that discriminated maximally two sets of proteins, i.e. maximized the distance (fitness function) between the proteomes. The population of chromosomes (potential solutions) were subjected to mutations with the probability *pmut* = 0.02 and recombinations with the probability *pcross* = 0.75 in subsequent generations. Every chromosome was the binary vector of the length 20, in which particular elements v_{nj} corresponded to an appropriate amino acid:

$$\mathbf{ch}_n = [v_{n1}, v_{n2}, \ldots, v_{n19}, v_{n20}],$$

where $v_{nj} = 1$ when j*th* amino acid was included in the calculation of the distance between proteomes and $v_{nj} = 0$, when it was not considered. We used deterministic tournament selection with two chromosomes per tournament to create parental population. In the applied elitist strategy the best fitted chromosome passed to the next generation.

GECCO'13 Companion, July 6–10, 2013, Amsterdam, the Netherlands.
ACM 978-1-4503-1964-5/13/07.

3. RESULTS AND DISCUSSION

Tab. 1 presents original and optimized distances calculated between sets of proteins coming from different ecological groups of microorganisms. The largest distances, i.e. the most distinctive proteomes according the amino acid composition had hyperthermophilies and microorganisms living inside their host cells. The application of evolutionary algorithm enlarged significantly differences calculated on SOM between the real proteomes. The maximized distance increased significantly 2.3 times on average in proteomes of psychrophiles, 2.0 times in proteomes of anaerobes, 1.6 times in proteomes of hyperthermophiles, 1.5 times in proteomes of aerobes and 1.4 times in proteomes of intracellular microorganisms. In contrast to the results obtained for the real proteomes, we did not observe such significant increase in maximized distances calculated for artificial proteomes. The distance calculated for these proteomes grew on average only 1.22 times in the best cases, i.e. for proteomes of anaerobes.

Table 1. Original (orig) and mean maximized (max) distances on SOM calculated in 300 independent runs for real and artificial microbial proteomes of given ecological groups.

Ecological group	Real		Artificial	
	orig	max	orig	max
hyperthermophiles	0.446	0.723	0.256	0.290
psychrophiles	0.165	0.387	0.434	0.448
aerobes	0.260	0.390	0.349	0.413
anaerobes	0.201	0.411	0.239	0.292
intracellular	0.371	0.510	0.448	0.518

Analyzing elements of chromosomes evolving in the applied algorithm we selected amino acids that maximized differences between proteomes calculated on SOM. As an example, Fig. 1 shows average frequencies of amino acids selected by the best chromosomes for proteomes of hyperthermophiles. The uneven distribution of these frequencies for real proteomes is clearly visible in contrast to the artificial ones, for which the frequencies oscillate close to 0.5 (dashed line) when all amino acids have the same discriminative power. The most distinctive amino acid selected by EA in comparisons of hypertherthmophilic proteomes vs. others was glutamine (Q), which appeared in all solutions. Valine (V), tyrosine (Y) and asparagine (N) were present very often, with frequency above 0.8. On the other side, leucine (L) appeared only in 9% of the best results. In solutions for psychrophilic proteomes dominated arginine (R) with 0.94 frequency whereas with frequency higher than 0.70 occurred amino acids with amide in their side-chain: glutamine (Q) and asparagine (N). Proteomes of aerobes were the best differentiated from others by basic amino acids: arginine (R) and lysine (K), selected in almost 80% chromosomes. Quite often (> 70%) appeared also alanine (A) and glutamic acid (E). In the case of anaerobic proteomes, only glutamine (Q) appeared diagnostic, with very high frequency 0.97. Two amino acids, serine (S) with frequency 0.90 and phenylalanine (F) with frequency 0.71 differentiated intracellular proteomes.

The obtained results indicate that, in contrast to real proteomes, amino acids in artificial sequences have the same power to discriminate proteomes consisted of such sequences although the absolute distances on SOM for artificial proteomes were sometimes even larger than for real proteomes. It suggests that differences between real proteomes can to some extent result from different nucleotide biases characteristic of particular genomes [2]. However, the identification of amino acids by the evolutionary algorithm, significantly differentiating the real proteomes suggests that the unique amino acid usage can also be modeled by selection, e.g. on higher stability and functional effectiveness of proteins in specific environmental conditions. The combination of SOM and AE methods seems to be successful approach in analyses of huge biological data and selection of the most discriminating variables from a noise.

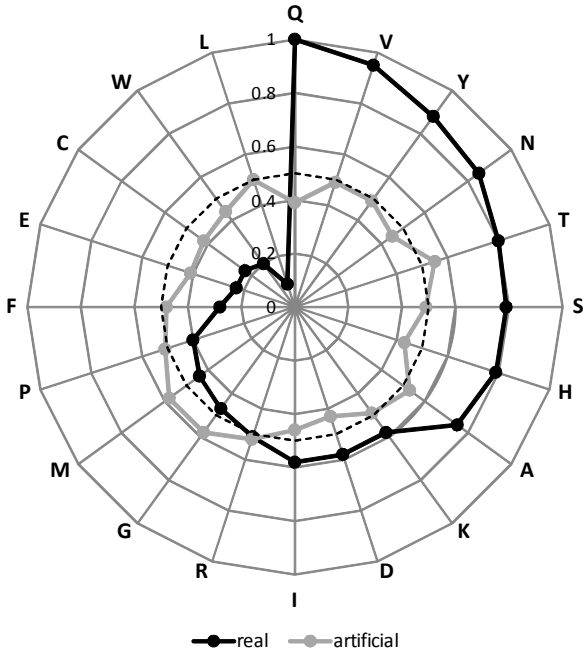

Figure 1. Average frequencies of amino acids calculated in 300 independent runs for the best chromosomes discriminating real and artificial proteomes of hyperthermophiles.

4. REFERENCES

[1] Eiben, A.E., and Smith, J.E. 2003. *Introduction to Evolutionary Computing.* Springer.

[2] Kiraga, J., Mackiewicz, P., Mackiewicz, D., Kowalczuk, M., Biecek, P., Polak, N., Smolarczyk, K., Dudek, M. R., and Cebrat, S. 2007. The relationships between the isoelectric point and: length of proteins, taxonomy and ecology of organisms. *BMC Genomics* 8, 163.

[3] Kohonen, T. 2001. *Self-Organizing Maps.* Springer Verlag, Berlin.

[4] Sobczyński, M., and Mackiewicz, P. 2011. Application of self-organizing maps to description of relationship between amino acid composition of proteins and ecological properties of microorganisms. *Proceedings of the Seventeenth National Conference on Applications of Mathematics in Biology and Medicine*. Zakopane-Kościelisko, 1-6.09.2011, 17, 96-101.

[5] Tekaia, F., Yeramian, E., and Dujon, B. 2002. Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis. *Gene* 297, 1-2, 51-60.