Krajowa Konferencja Zastosowań Matematyki w Biologii i Medycynie

Zakopane - Kościelisko, 1st–5th September 2011

APPLICATION OF SELF-ORGANIZING MAPS TO DESCRIPTION OF RELATIONSHIP BETWEEN AMINO ACID COMPOSITION OF PROTEINS AND ECOLOGICAL PROPERTIES OF MICROORGANISMS

Maciej Sobczyński[1], Paweł Mackiewicz[2]

[1,2]Department of Genomics, Faculty of Biotechnology, University of Wrocław
ul. Przybyszewskiego 63/77, 51-148 Wrocław
[1]macsebsob@poczta.onet.pl,[2]pamac@smorfland.uni.wroc.pl

ABSTRACT

We have tested usefulness of Self-Organizing Maps (SOM) in classification of proteins coming from different prokaryotic taxa. The final neural network was carefully selected based on three criteria: Bayesian Information Criterion, topological error, and spatial autocorrelation. The carried out analyses showed a clear relationship between amino acid composition of proteins and environment in which live studied species. Interesting differences were observed in the composition of domain and non-domain regions as well as proteins classified to various functional groups. The studies indicate that SOM can be successfully applied in huge data sets such as whole proteome studies delivering significant biological information.

INTRODUCTION

Rapid increase of numerous completely sequenced genomes delivers huge data that require special large-scale analyses carried out by sufficient statistical and data mining methods. From biological point of view, a very interesting subject is the study of amino acid composition of proteins. This feature may reflect mutational and selectional constraints on the coded proteins, as well as may be related with the taxonomic affiliation of organisms and their environmental conditions. The most characteristic amino acid usage was observed in thermophiles [1]-[3] and halophiles [4], [5]. The former approaches used standard multidimensional analyses for example Principal Component Analysis or Correspondence Analysis. Here we applied Self-Organizing Maps (SOM) atypically to analyze differences in amino acid composition of prokaryotic proteins. Thanks to this approach we were able to present sets of proteins (i.e. proteomes) in multidimensional space and calculate distances between them. These distances express differences and specific amino acid compositions of the analyzed proteomes. The presented method enabled to identify relationships between the amino acid composition and environmental factors that influence the analyzed organisms. The SOM turned out to be a very sensitive method because it was able to distinguish even very similar proteomes.

MATERIAL AND METHODS

The applied neural network was chosen based on three parameters. First parameter was Bayesian Information Criterion (BIC) defined as:

$$BIC = n \cdot \ln\left(\frac{RSS}{n}\right) + k \cdot \ln(n) \tag{1}$$

where:
n is number of teaching vectors (proteins),
k is number of neurons.

$$RSS = \sum_{i=1}^{n} e_i \tag{2}$$

is sum of quantization errors that are defined as:

$$e_i = d(x_i, m_{x_i}) \tag{3}$$

where:

x_i is teaching vector *i*,

m_{x_i} is centroid of x_i ,

d is distance function (Euclidean in this case).

The teaching vector x_i describes the percentage amino acid composition of protein *i*. The more neurons are in the network, the lower RSS and the lower BIC are because teaching vectors are closer to their centroids. However, this decrease in BIC becomes smaller when the number of neurones *k* increases because the second component of Eq. (1) is larger. Then for big *k*, the BIC may again receive high values. Therefore it is important to make a trade-off between the goodness of fit and the number of neurons. The neural network is the better if it is characterized by the smaller BIC. We tested all 190 rectangular topologies whose dimension ranged from 2×2 to 20×20 neurones.

The second parameter describing the goodness of fit was the topological error *te* defined as:

$$te = \frac{1}{n}\sum_{i=1}^{n} u(x_i) \tag{4}$$

where:
the function $u(x_i)$ gets 0 when the first two Best Match Units, BMU1x_i and BMU2x_i, are neighbors, and 1 when they are not neighbors. The network with smaller *te* is better.

The third criterion of goodness of fit was the spatial autocorrelation. Quantization errors at different neurons may not be independent. For example, measurements made at neighboring units may be closer in their values than measurements made at distant locations of the same network. The spatial autocorrelation measures the correlation of quantization error with itself through the space.

A distance between two sets of analyzed proteomes, $\mathbf{B}_1$ and $\mathbf{B}_2$, was measured by:

$$d(\mathbf{B}_1, \mathbf{B}_2) = \frac{1}{2}\sum_{k=1}^{K}\left|\frac{n_{k\mathbf{B}_1}}{m} - \frac{n_{k\mathbf{B}_2}}{N-m}\right| \tag{5}$$

where:

$n_{k\mathbf{B}_1}$ and $n_{k\mathbf{B}_2}$ are numbers of proteins classified to neuron *k* and coming from the set $\mathbf{B}_1$ and $\mathbf{B}_2$, respectively,

m and *N-m* are total numbers of proteins in the set $\mathbf{B}_1$ i $\mathbf{B}_2$, respectively.

The distance $d(\mathbf{B}_1,\mathbf{B}_2)$ asymptotes 0 and 1.

The analyzed data set contained 434 000 teaching vectors (proteins) belonging to 194 archaeal and bacterial species. Their proteomes were grouped according to environmental factors that are optimal for growth of these organisms. Moreover, proteins were divided into three groups according to their function: information storage and processing (Isp), cellular processes and signaling (Cps), and metabolism (Metab). The biological functions of proteins were identified based on their classification to Clusters of Orthologous Groups (COG). Additionally, we analyzed separately the amino acid composition of domain and non-domain regions in proteins. COG classification and domain searches in Conserved Domain Database (CDD) were made by rpsblast software. Numerical analyses were carried out in R package (`www.r-project.org`).

RESULTS AND DISCUSSION

Fig. 1 presents relationship between BIC values and dimension of tested neural networks. The best networks with the lowest BIC value consisted of 10×10 neurons and 5×19 neurons. Finaly, 10×10 map was chosen to further analyses.

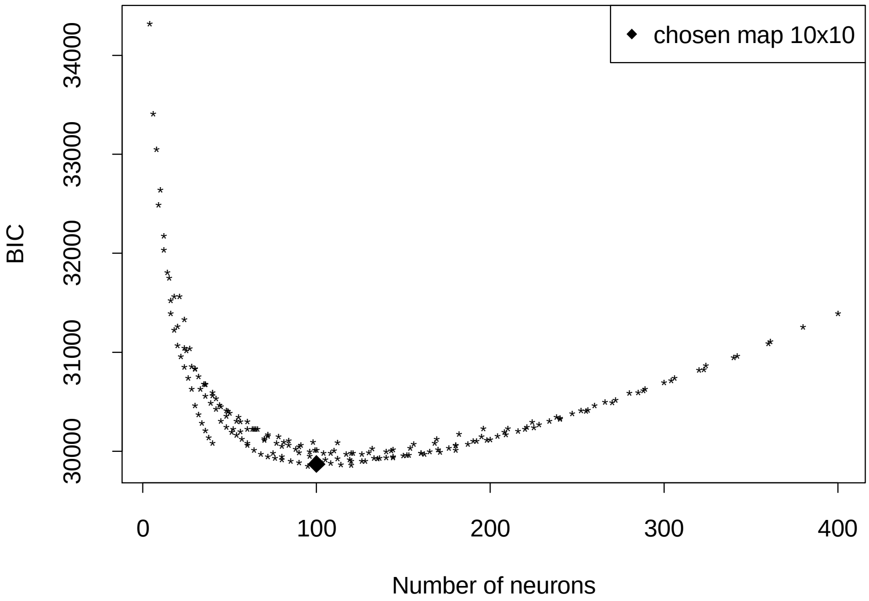

Figure 1. Relationship between Bayes Information Criterion (BIC) and size of map.

Distances $d(\mathbf{B}_1,\mathbf{B}_2)$ between sets of proteins derived from different microbes classified according to their characteristic environmental factors are shown in Fig. 2 and Table 1. The results indicate that temperature has the strongest relation to the amino acid composition of proteins. The distance value for hyperthermophilic microbes is $d(\mathbf{B}_1,\mathbf{B}_2) = 0.4461$ which means that hyperthermophiles are more similar to themselves up to 44.6% than to the other species. Relation to host cell is the second factor which strongly influences the amino acid composition. Intracellular species (i.e. parasites or endosymbionts) show very specific amino acid usage and the distance between their proteins and the rest is $d(\mathbf{B}_1,\mathbf{B}_2) = 0.3705$.

Fig. 3 shows the relationship between different environmental factors and amino acid composition in three sets of proteins divided according to their functional grouping. It was observed that temperature has the strongest influence on the amino acid composition, but it depends on functional classification of proteins. In the case of hyperthermophilic species, the largest distance $d(\mathbf{B}_1,\mathbf{B}_2) = 0.5665$ is for proteins responsible for information storage and processing (Isp), which indicates that these set of proteins has the most different amino acid usage in comparison to non-hyperthermophiles. This difference is much bigger than in the case of proteins responsible for cellular processes and signaling (Cps), $d(\mathbf{B}_1,\mathbf{B}_2) = 0.3655$ and metabolism

(Metab), $d(\mathbf{B}_1,\mathbf{B}_2) = 0.2951$. On the other hand, proteins responsible for metabolism have the most disparate amino acid usage in intracellular microbes in comparison to the rest species.

The most unique amino acid composition of hyperthermophilic proteins results likely from constraints imposed on thermostability of their structure [1]–[3] whereas the composition of proteins from intracellular species is usually modeled by the higher rate of mutations accumulation (see [6] and references therein).

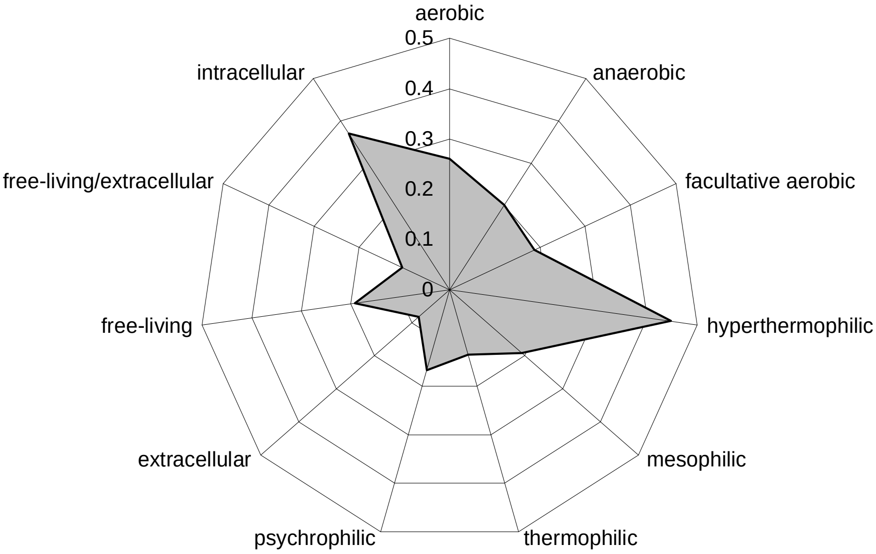

Figure 2. Relationship between environmental factors and amino acid composition measured by distances $d(\mathbf{B}_1,\mathbf{B}_2)$ between sets of all proteins.

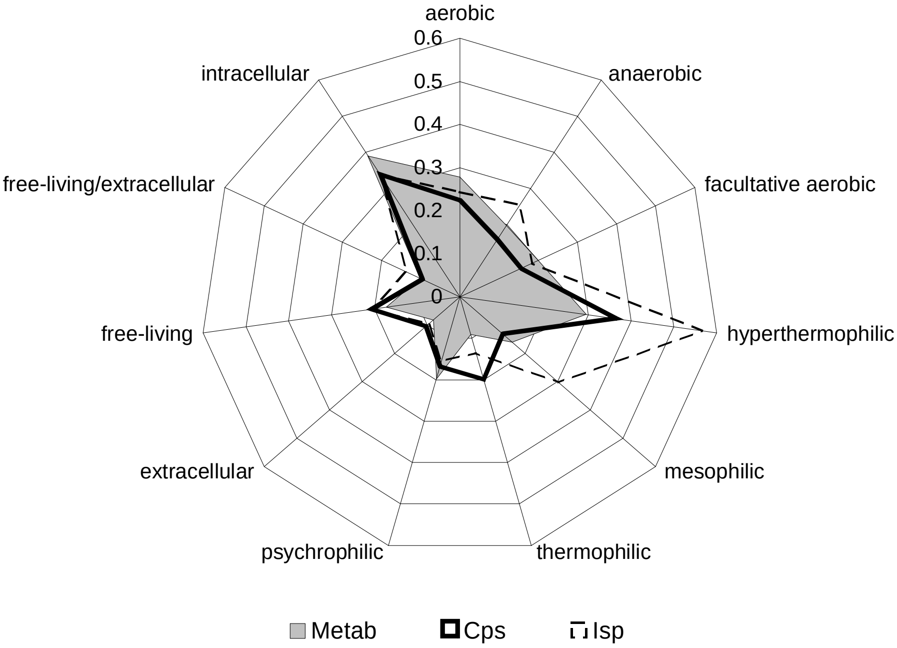

Figure 3. Relationship between environmental factors and amino acid composition measured by distances $d(\mathbf{B}_1,\mathbf{B}_2)$ between sets of proteins divided into three functional groups.

Comparison of amino acid composition of domain and non-domain protein regions according to environmental factors is presented in Fig. 4 and Table 1. The composition of protein domains is generally much weaker connected with environmental factors than the composition of non-domain regions. For example, the amino acid composition of domain regions in aerobic species (i.e. living oxygenic conditions) are much more similar to other species (i.e. anaerobic and facultative aerobic) than the composition of non-domain regions. In the first case the distance between them is $d(\mathbf{B}_1,\mathbf{B}_2) = 0.2467$ whereas in the second case is $d(\mathbf{B}_1,\mathbf{B}_2) = 0.383$. Remarkable is fact, that differences in amino acid composition of hypethermophiles' proteins to others are very close for domain and non-domain regions in proteins. In both cases the distances are large but very similar, $d(\mathbf{B}_1,\mathbf{B}_2) = 0.4091$ and $d(\mathbf{B}_1,\mathbf{B}_2) = 0.4523$, respectively. Analogous results are in the case of intracellular microbes. They are characterized by different amino acid usage in comparison to other species (i.e. free living/extracellular, free-living, and extracellular), but the differenses are very similar for domain, $d(\mathbf{B}_1,\mathbf{B}_2) = 0.3665$, and non-domain, $d(\mathbf{B}_1,\mathbf{B}_2) = 0.3817$, regions in proteins.

It is usually accepted that non-domain regions are less functionally and structurally constrained, therefore they should accumulte more substitutions than domain regions. Assuming that and the observed stronger relationship between amino acid composition and environmental factors for non-domain regions than for domains, one may deduce that different environment conditions cause variuos mutational patterns in genomes and encoded proteins. However, it cannot be excluded that some of these substitution has a structural significance, for example reinforcing stability of protein loop usually formed by non-domain regions.

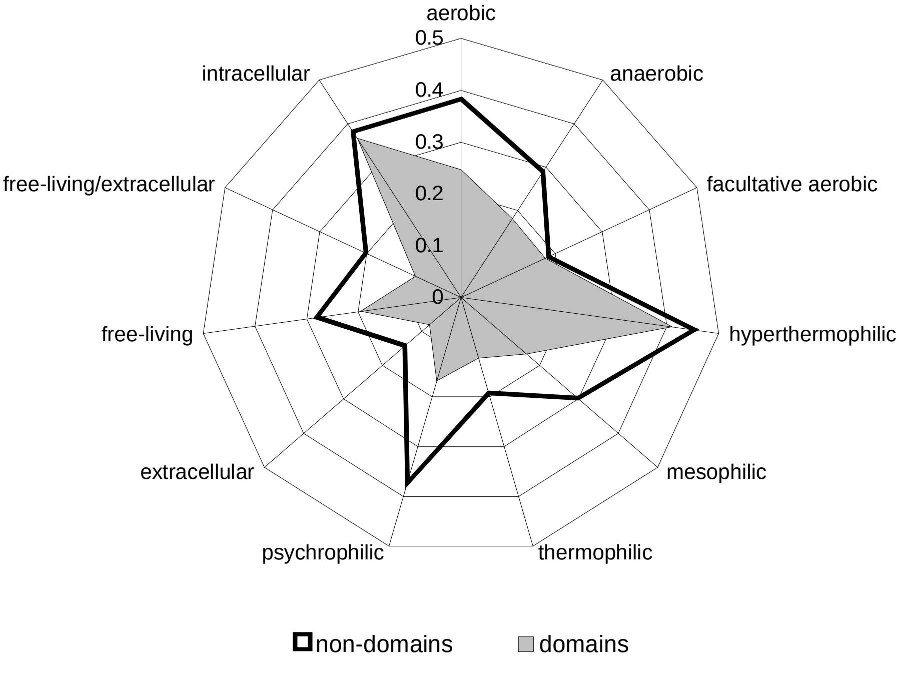

Figure 4. Relationship between environmental factors and amino acid composition measured by distances $d(\mathbf{B}_1,\mathbf{B}_2)$ for domain and non-domain regions in proteins.

Table 1. Distances $d(\mathbf{B}_1,\mathbf{B}_2)$ for different sets of proteins as a measure of relationship with environmental factors.

Environmental factors		**Non-domain regions**	**Domain regions**	**All proteins**	**Metab**	**Cps**	**Isp**
Oxygen	aerobic	0.383	0.2467	0.2599	0.2779	0.2244	0.2427
	anaerobic	0.2888	0.1806	0.201	0.1974	0.1577	0.2543
	facultative aerobic	0.1854	0.1789	0.1877	0.2002	0.1557	0.1844
Temperature	hyperthermophilic	0.4523	0.4091	0.4461	0.2951	0.3655	0.5665
	mesophilic	0.2976	0.1671	0.1914	0.1611	0.1316	0.3032
	thermophilic	0.1932	0.1222	0.1337	0.0914	0.1988	0.1348
	psychrophilic	0.3742	0.1676	0.1654	0.1974	0.1674	0.154
Relation to host cell	extracellular	0.1432	0.0801	0.0812	0.082	0.1045	0.0961
	free-living	0.2805	0.1946	0.1915	0.1716	0.2069	0.2076
	free-living/extracellular	0.2017	0.0962	0.1049	0.1014	0.0961	0.1359
	intracellular	0.3817	0.3655	0.3705	0.39	0.3379	0.3365

Functional classification of proteins: Isp - information storage and processing, Cps - cellular processes and signaling, Metab - metabolism.

REFERENCES

[1] T. Kawashima, N. Amano, H. Koike, S. Makino, S. Higuchi, Y. Kawashima-Ohya, K. Watanabe, M. Yamazaki, K. Kanehori, T. Kawamoto, T. Nunoshiba, Y. Yamamoto, H. Aramaki, K. Makino, and M. Suzuki: *Archaeal adaptation to higher temperatures revealed by genomic sequence of Thermoplasma volcanium*, Proc. Natl. Acad. Sci. USA **97** (2000), 14257-14262.

[2] D. P. Kreil and Ch. A. Ouzounis: *Identification of thermophilic species by the amino acid compositions deduced from their genomes*, Nucleic Acids Res. **29** (2001), 1608-1615.

[3] A. Pasamontes and S. Garcia – Vallve: *Use of multi-way method to analyze the amino acid composition of a conserved group of orthologous proteins in prokaryotes*, BMC Bioinformatics **7** (2006), 257.

[4] S. P. Kennedy, W. V. Ng, S. L. Salzberg, L. Hood, and S. DasSarma: *Understanding the adaptation of Halobacterium species NRC-1 to its extreme environment through computational analysis of its genome sequence*, Genome Res. **11** (2001), 1641-1650.

[5] A. Oren and L. Mana: *Amino acid composition of bulk protein and salt relationships of selected enzymes of Salinibacter ruber, an extremely halophilic Bacterium*, Extremophiles **6** (2002), 217-223.

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