The use of Self-Organizing Maps and genetic algorithms to investigate a proteomes

Zakład Genomiki, Wydział Biotechnologii, Uniwersytet Wrocławski

Autor rozprawy: mgr Maciej Sobczyński

Promotor: Dr hab. Paweł Mackiewicz

Summary

Amino acid composition of proteins can be modelled by both selectional and mutational pressures. The influence of selection is related with functional and structural requirements on proteins, whereas mutations modify directly nucleotide composition of protein coding sequences. To study the influence of these pressures on amino acid composition and its dependence from evolutionary relationships and ecological properties of almost 200 prokaryotic microorganisms, Kohonen neural networks (Self-Organizing Maps, SOM) were applied. This approach was chosen in order to compare amino acid composition of proteins contextually and find compositional patterns/prototypes of particular proteomes, which were encoded in centroids of learned neurons.

The size of nets for a particular data set were thoroughly selected using Bayesian Information Criterion, topological error and spatial autocorrelation. Proteins coming from proteomes of organisms showing particular environmental properties were classified on nets based on their amino acid composition. Distances between groups of proteomes were based on fractions of their proteins assigned to given neurons. The applied method appeared a good approach in discrimination of proteomes described by many parameters.

The carried out analyses showed clear relationship between amino acid composition and environmental conditions of studied organisms. Organisms living in extremely high temperatures (hypertermophiles) and inside cells of other organisms (endparasites and

endosymbionts) were characterized by the most different composition from other ecological groups. The relationship of amino acid composition with environmental features was most distinct for sequences of regions located between protein domains than sequences of these domains. It may suggest some influence of mutational pressure on their composition. However, the changes in the composition can be also an adaptation of proteins to specific environmental conditions. It was also shown that pairs of species with the smaller number of common ecological features differed more in amino acid composition than it could be expected from their evolutionary relationships (measured by the number of substitutions in rRNA sequences).

On the other hand, pairs of species with the larger number of common ecological features differed less in their amino acid composition than it could be expected from their relationships.

In this dissertation, evolutionary algorithm (EA) was also applied to find subsets of amino acids that discriminate best the compared proteomes. To eliminate the potential influence of selection pressure on proteome composition and to study the effect of mutation pressure, artificial proteins were generated based on the global nucleotide composition of appropriate genomes. Then they compared with the real proteins. The analysis showed that, in contrast to the real sets, it was not possible to find amino acids that discriminate the compared sets of proteins because each amino acid occurred in the solution of the algorithm with the same frequency 0.5. As a result of this it was not possible to increase differences between the analysed sets. Moreover, the found amino acids showed no correlation in their co-occurrence.

Obtained results allow to state that it is very unlikely that amino acid composition of protein results only from mutational pressure. Selection pressure plays here a crucial role. In agreement with that, the elimination of selection effect, i.e. an important force influencing proteome composition, gives results unobserved in nature.