Poster C-33

Evolutionary Trends in Proteomes: exploring the Multidimensional Contributions

Authors:
Tekaia (*Institut Pasteur*)
Yeramian (*Institut pasteur*)

Short Abstract: Various aspects of the complex evolutionary trends that mutually link species can be grasped from the detailed analysis of their proteomes. Using Correspondence Analysis, we derive global pictures for the evolutionary links between species in terms of Genome trees (based on conservation profiles) and of proteomes amino-acid compositions.

Long Abstract:
Evolutionary Trends in Proteomes: Exploring the Multidimensional Contributions

Fredj Tekaia1 and Edouard Yeramian2

1:Unité de Génétique Moléculaire des Levures (URA 2171 CNRS and UFR927 Univ. P.M. Curie), Institut Pasteur, 25, Rue du Dr Roux, F-75724 Paris Cedex 15; e-mail:tekaia@pasteur.fr;
2:Unité de Bio-Informatique Structurale (URA 2185 CNRS), Institut Pasteur, 25, Rue du Dr Roux, F-75724 Paris Cedex 15, France, e-mail:yeramian@pasteur.fr.

Abstract:

Various aspects of the complex evolutionary trends that mutually link species can be grasped from the detailed analysis of their proteomes. However, to reflect properly such complex relationships it is necessary to resort to appropriate multidimensional representations of the mutually dependent histories of proteins in various species (Tekaia et al. 1999). In this direction, using Correspondence Analysis, we derive global pictures for the evolutionary links between species in terms of Genome trees (based on conservation profiles) and of proteomes amino-acid compositions. In both cases, fully taking into account the multidimensional contributions (for n complete genomes), the primary information is represented through n-component vectors, corresponding either to protein conservation profiles or to species-specific amino-acid compositions.

A) Genome tree constructions based on protein conservation profiles in multiple species
The basic idea is that the multi-component “presence-absence” protein conservation profiles permit tracking of common evolutionary histories of genes across multiple genomes. Indeed, the conservation profile of a given protein captures an evolutionary history, expressed as an n-component vector detailing the presence or absence of homologs, in each of n considered species. Through the multidimensional structure of conservation profiles, the evolutionary history of proteins is thus observed jointly across n species: proteins with identical conservation profiles can be associated with identical evolutionary histories. From the

complete set of the considered proteomes, the set of distinct conservation profiles is indicative of the various evolutionary histories.
Genome tree construction can then be based on the core set of shared distinct conservation profiles or on the whole set of distinct conservation profiles resorting to similarity scores between pairs of species (as for example Jaccard score).
The obtained genome trees show a sharp discrimination between the three primary domains of life: Bacteria, Archaea and Eukarya, and display significant correspondences with classically recognized taxonomical groupings, along with a series of departures from such conventional clusterings. These observations can be interpreted in the frame of various evolutionary trends shaping the structure of genes and genomes (Tekaia and Yeramian 2005).

B)Distribution of species following their amino-acid compositions
Based on amino-acid composition data (for 208 completely sequenced organisms with large representation for the 3 phylogenetic domains and various life styles), correspondance analysis reveals a series of striking features in the distribution of species according to amino-acids (Tekaia & Yeramian, work submitted):
1) a sharp discrimination of eukaryotes ;
2) a sharp discrimination for certain lifestyles (hyperthermophiles), with other lifestyles (psychrophiles) being non-discriminated;
These observations confirm previous work (Tekaia et al. 2002), demonstrating discrimination between lifestyles associated with hyperthermophily-thermophily and mesophily.
Overall, distributions of species and amino-acids lead to a coherent picture for some fundamental evolutionary trends, such as the plausible origin of life in hot environments and the chronology of amino-acid recruitment in the genetic code associated with the evolution of species.

Conclusion:
With the rapid expansion of genome sequence data it is expected that multivariate analysis methods will increasingly help for the discovery of important evolutionary trends intrinsically associated with the multidimensional structure of genomic data, as obtained from the comparison of many genomes.

References:

Tekaia F, Lazcano A, Dujon B. (1999). The genomic tree as revealed from whole proteome comparisons. Genome Res. 9:550-7.

Tekaia F, Yeramian E. (2005). Genome Trees from Conservation Profiles.
PLoS Comput Biol.1(7):e75

Tekaia F, Yeramian E, Dujon B. (2002).
Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis. Gene. 297:51-60.