Identifying Communities structure in paralogous genes

ID - 242

Brilli Matteo[1], Fani Renato[1], Lio' Pietro[2]

[1]Dept. of Animal Biology and Genetics, University of Florence, via Romana 17/19, Florence, Italy

[2]Computer laboratory, University of Cambridge

Motivation

Comparative genomics research has speed up in the last few years thanks to the completion of a large number of genomes and the development of software tools helping in managing the huge amount of data produced. Very powerful approaches to trace the evolutionary histories of genes and proteins exist (e.g.: maximum likelihood phylogenetic methods) but they generally are time-consuming, while large scale comparative analyses necessitate of fast (but less precise) methods. The choice is often based on the local alignment tools BLAST and FASTA. These tools find homologies among regions of a query sequence and the sequences stored in a database, listing all those sequences showing homologies to the queries, ranked by the statistical significance of the alignment(s) found. However, with large amounts of sequences the outputs of these tools become intractable; for this reason the usual choice is the conversion of the blast output into an adjacency matrix describing the homology network.

In recent years, several methods have been proposed to reveal the community structure of very heterogeneous network. Even if there is now a wide range of clustering algorithms, only a restricted number can successfully handle networks with thousands of nodes.

Methods

We have implemented the Markov Clustering developed by Van Dongen (2001), which simulates a random walk on the stochastic matrix derived from the homology adjacency matrix and a recent method proposed by Guimera et al (2004) that is instead based on simulated annealing to obtain clustering by direct maximization of the modularity M.

The modularity has been introduced by Newmand and Girvan (2004). It is a measure of the difference between the number of links inside a given module and the expected value for a randomized graph of the same size and degree distribution. For these two algorithms we show the effectiveness of different measures of entropy and centrality.

We incorporate those algorithms into a Blast-family of programs which are extensively used in comparative genomics and data mining to detect homologous sequences with known function/structure.

Results

We applied our package to the analysis of Gamma-proteobacterial plasmids and retrieved several communities composed by proteins involved in antibiotic resistance, conjugal transfer and a huge amount of transposases.

Moreover, we characterize and discuss the community structure associated with proteins involved in Nitrogen Fixation, which comprises proteins involved in aminoacid metabolism, chlorophyll biosynthesis and other central processes.

Email: matteo.brilli@dbag.unifi.it