Simulating genes families

Roberto Marangoni

Dipartimento di Informatica, Università di Pisa, Via F. Buonarroti 2
56127 PISA, Italy
marangon@di.unipi.it

Keywords. Genes families, paralogy, duplications.

Introduction

After their complete sequentiation, genomes are clusterized in genes families, the members of which share a significant similarity in their sequences (and often in the structures of their proteic products) but they are often playing different biological roles. When there is such a relationship between two genes, they are called paralogs.
It is of general believe, that paralogs genesis is due to an iterate mechanism of gene duplication with subsequent modification of the copies [1-3]. In a previous work [4] describing a method to reconstruct the history of genes families, a simulator of genes families was introduced in order to bypass the lack of experimental data about genes families history. Working with these simulated data, some interesting features concerning real biological families has been found. Nevertheless, they have not been explored, since they were too far from the main subject of that paper.
In the present work, a simulator similar to that used in the above cited paper has been developed, and many different synthetic data have been generated. The simulation strategy, the biological foundation of it and the comparison between simulated and real sequences are discussed in detail in the poster.

Some Details

1. Simulator structure

The simulator is a home-made software (written in Perl 5.0) that receives as input one gene sequence and generates n simulated paralogs of it, by iterating a duplication-with-modifications process. To maintain the process as close as possible to real biological evolution, the most common variability mechanisms, known from biological studies, have been simulated. In particular, it is known that, by performing a multi-alignment on a set of paralog genes it is possibly to detect [5]: a) common sub-sequences placed in the same order, b) sub-sequences that appear to be inverted and copied to another location (this phenomenon is called "translocation with inversion"), and c)random point mutations and sub-sequences not shared among the genes. Moreover, there are strong indications that copied sequences are often shorter than their templates [5].
The simulation with modification process is depended on some parameters, which allow us to establish how much the copies are similar to their respective templates. More in detail, the generation of simulated paralogs includes the following steps:

1. the simulator marks on the input sequence a number of m different positions, randomly chosen.
2. it extracts, from the previous sequence, $m/2$ sub-sequences included between any interval of the type $[m_{2i}, m_{(2i+1)}]$.
3. an array is built, in which each sub-sequence is originally mapped to the position it has in the input sequence;
4. each array member is, with a probability p (usually low), inverted and translocated to another randomly chosen position, the positions of other array members are shifted accordingly;

5. the paralog sequence is made by assembling each array member in the assigned position, and inserting a new random sub-sequence of length *m*/4 after each array member.

In this way, the generated paralog has a length of about three fourths of its template, and it contains all the feature described in biology. To obtain a family of paralogs it is enough to chose one of the two sequences available (the input and its paralog) and start again at step 1. The main problem is how to chose the sequence to assign as template for the next duplication process. Two possibilities have been explored: a uniform probability on all the available sequences, or a probability that decreases with the "age" of a sequences. In other words, newly generated paralogs have more chances to be chosen as template for another duplication. By setting all the above parameters and the probability to be chosen as template, we obtain simulated paralogs and paralogy tree really different each other.

2. A method to measure the simulated sequences likelihood

If the simulated sequences are similar to real ones, this imply that similarity-based clustering algorithm, like for instance ClustalW, should be unable to distinguish between real and simulated. To test this for each examined family, a mixed set of real and simulated sequences has been built, and given as input to ClustalW, asking for a similarity tree in output. To give a quantitative evaluation of how much real and simulated sequences are mixed in this tree, we defined a μ-index as the ratio between the number of the inner tree nodes that contain both simulated and real in their descending branches, by the total number of nodes in the tree. A, μ values close to 1 identify a good mimetic capability of the simulator with respect of that family, whereas μ values close to 0 suggest that the simulation rules are not so good with respect to the examined family.

3. A brief description of the main results obtained

Simulated data have been generated, with different configurations of the simulator, for about 60 different families in 12 different organisms (both prokaryote and eukaryote). We found that certain simulator configurations are very good with respect to a large number of families (they show $\mu > 0.8$), therefore similarity-based clustering algorithms are unable to recognize which are synthetic sequences and which not. With our surprise, we also found that the μ values for the same family in different organisms are very similar, even if organisms are far on an evolutionary scale. This unexpected result is difficult to understand: in a first approximation it suggests that each family in a genome can have its own evolution rules, not necessarily the same for all the families. The poster will present details about the different configurations used by the simulator and the results obtained for the different families in different organisms.

References

[1] M. Lynch, J.S. Conery, The Evolutionary Fate and Consequences of Duplicate Genes. *Science*, 290:1151-1155, 2000.

[2] H.W. Mewes, K. Albermann, M. Bähr, D. Frishman, A. Gleissner, J. Hani, K. Heumann, K. Kleine, A. Maierl, S.G. Oliver, F. Pfeiffer, A. Zollner, Overview of the yeast genome. *Nature*, 387(6632 Suppl), 7-65 1997.

[3] C. Popovici, M. Leveugle, D. Birnbaum, F. Coulier Coparalogy: Physical and Functional Clusterings in the Human Genome. Biochem. Biophys. Research Comm., 288:362-370, 2001.

[4] N. Pisanti, R. Marangoni, P. Ferragina, A. Frangioni, A. Savona, C. Pisanelli, F. Luccio, PaTre: a method for Paralogy Trees construction. *J. Comp. Biol.*, (in press) 2003.

[5] B. Lewin, Genes (7th. Edition). Oxford Univ. Press, Oxford, 1999.