Alignment-Free Comparison of Metabolic Pathways

Francesc Rosselló, [1] **Gabriel Valiente** [2]

Keywords: metabolic pathway, similarity measure, hierarchical similarity, information content, gene ontology, clustering, fuzzy clustering, phylogenetic reconstruction, alignment-free comparison

1 Introduction.

Classification of species is a fundamental problem in evolutionary studies. Most contributions to this field have been mainly based on the so-called "molecular approach", that exploits polymorphism information in DNA or proteins to assess the phylogenetic relationship among species [9]. But it is widely agreed that, in this approach, the choice of sequences for comparison greatly affects the final result, since non-local mutations like gene tranfers, gene duplications, gene deletions and functional replacements of genes can prevent the obtention of significant information through the analysis of the evolution of short sequences.

This has led to an increasing interest in phylogenetic reconstruction based on non-genomic information, as for instance information on global biological processes in the organisms. In the last years, the analysis of phylogenetic relationships has been expanded to include the study of the evolution of complete biological processes, as for instance metabolic pathways. A metabolic pathway is a series of individual biochemical reactions, connected via their substrate and product metabolites, that produces a set of metabolites from a set of precursor metabolites and cofactors. Current knowledge about these sets of connected biochemical reactions are organized into databases that contain information about the substrates and products of the reactions, the enzymes that activate them, the substrate specificity of these enzymes, etc. One of the most commonly used database is KEGG [5], where biochemical reactions are organized into a map based on information drawn from the literature. For example, the glycolysis pathway is defined therein as the set of reactions that have been identified experimentally as the most important steps in the conversion of glucose into pyruvate.

As large pieces of the metabolic networks of a wide variety of organisms are being reconstructed and compiled in databases, the use of metabolic pathways is emerging as a new source of phylogenetic information. And with experimental techniques in metabolomics improving constantly, one should expect that most metabolic network data will be soon independent from genomic information, and then metabolic pathways will become a true alternative to biomolecular sequences in phylogeny reconstruction [3]. This has driven research on techniques to measure the similarity of metabolic pathways and to extend these similarity values into phylogenies for different organisms.

[1]Department of Mathematics and Computer Science, Research Institute of Health Science, University of the Balearic Islands, E-07122 Palma de Mallorca, E-mail: cesc.rossello@uib.es

[2]Department of Software, Technical University of Catalonia, E-08034 Barcelona, E-mail: valiente@lsi.upc.edu

2 Results and Discussion.

To assess the performance of the similarity notions and the techniques involved in the reconstruction of phylogenetic relationships from data on structural similarity of metabolic pathways, phylogenetic relationships for model sets of organisms are reconstructed from the similarity measures of the same metabolic pathway for all organisms, and then the phylogenetic trees obtained in this way are compared with the NCBI taxonomy [10], which is based on Ribosomal RNA 16S sequences, using some standard phylogenetic trees comparison tool.

In [2], we have used a new heuristic algorithm for the comparison of metabolic pathways with three enzyme similarity measures (hierarchical, information content, and gene ontology) to assess the structural similarity of metabolic pathways for different organisms. Experimental results on the Glycolysis pathway for a set of 73 organisms representing the three domains of life showed that our method outperforms previous techniques [4, 6, 8].

In this poster, we present two further improvements to these techniques. On the one hand, we have shown in [1] that the resulting phylogenetic trees are usually significantly closer to the NCBI taxonomy than those obtained using average-link hierarchical clustering if a simple fuzzy clustering method is used to compute the phylogenetic relationships from the similarity relation. On the other hand, we show that the resulting phylogenetic trees are also significantly closer to the NCBI taxonomy than those obtained using average-link hierarchical clustering if the measure of structural similarity between metabolic pathways is replaced by an alignment-free similarity measure, known as the universal similarity metric [7] and based on Kolmogorov complexity.

References

[1] Casasnovas, J., Clemente, J.C., Miró-Julià, J., Rosselló, F., Satou, K., Valiente, G.: Fuzzy clustering improves phylogenetic relationships reconstruction from metabolic pathways. In: Proc. 17th Int. Conf. Information Processing and Management of Uncertainty in Knowledge-Based Systems. (2006) Submitted, invited paper.

[2] Clemente, J., Satou, K., Valiente, G.: Reconstruction of phylogenetic relationships from metabolic pathways based on the enzyme hierarchy and the gene ontology. Genome Informatics **16** (2005) 45–55

[3] Forst, C.V., Flamm, C., Hofacker, I.L., Stadler, P.F.: Algebraic comparison of metabolic networks, phylogenetic inference, and metabolic innovation. BMC Bioinformatics (2005) Submitted.

[4] Forst, C.V., Schulten, K.: Phylogenetic analysis of metabolic pathways. J. Mol. Evol. **52** (2001) 471–489

[5] Kanehisa, M., Goto, S., Hattori, M., et al.: From genomics to chemical genomics: New developments in KEGG. Nucleic Acids Res. **34** (2006) D354–D357

[6] Heymans, M., Singh, A.K.: Deriving phylogenetic trees from the similarity analysis of metabolic pathways. Bioinformatics **19** (2003) i138–i146

[7] Li, M., Chen, X., Li, X., Ma, B., Vitányi, P.M.B.: The similarity metric. IEEE T. Inform. Theory **50** (2004) 3250–3264

[8] Liao, L., Kim, S., Tomb, J.F.: Genome comparisons based on profiles of metabolic pathways. In: Proc. 6th Int. Conf. Knowledge-Based Intelligent Information and Engineering Systems. (2002) 469–476

[9] Page, R.D.M., Holmes, E.C.: Molecular Evolution: A Phylogenetic Approach. Blackwell Science (1998)

[10] Wheeler, D.L., Chappey, C., Lash, A.E., Leipe, D.D., Madden, T.L., Schuler, G.D., Tatusova, T.A., Rapp, B.A.: Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. **28** (2000) 10–14