MEASURING THE DISSIMILARITY BETWEEN GENE AND SPECIES TREES, THE QUALITY OF A COG

Lyubetsky V.A.*, V'yugin V.V.

Institute for Information Transmission Problems, Russian Academy of Sciences, 127994, Moscow, Bolshoi Karetnyi per., 19, Russia
* Corresponding author: e-mail: lyubetsk@iitp.ru

Keywords: *evolution, phylogenetic trees, mathematical model computer analysis*

Summary

Motivation: The availability of genome-scale sequence data from diverse taxa makes it possible to derive new hypotheses about ancient evolutionary events from comparative analysis of large gene sets. Important groundwork of this goal is to find good strategies for comparing COG trees with species trees, to estimate the quality of the COGs and corresponding the trees to compare evolutionary models underlying the reconstructions and, in particular, to integrate approaches allowing inferences about evolutionary scenarios and gene duplication-loss patterns.

Results: In this study we reconstruct selected details of the ancestral history of Archaea and Bacteria within the outlined framework.

Introduction

It is well known that phylogenetic trees derived from different protein families are often incongruent and essentially differ from each other and the species tree. This may be explained by poor choice of the evolutionary model and associated reconstruction biases, as well as by discrepacies between gene and organism evolutionary history due to speciation, gene duplication, gene loss and gene gain (horizontal gene transfer, genesis from non-coding DNA), and so on. In this paper we consider several integral characteristics measuring dissimilarity between gene and species trees as estimates of the quality of the COG or the COG tree. We identify COGs (clusters of orthologous groups of proteins) of different quality. The main purpose of our analysis is reconstruction of selected details of the ancestral history of Archaea and Bacteria. For this analysis, we use the combined gene duplication-loss model extended to incorporate some gene gain events. Any method of tree comparison is based on an evolutionary hypothesis and the corresponding mathematical model of evolution. We compare two such hypotheses.

Models and Algorithms

We employ mapping from a gene tree G into a species tree S introduced in Mirkin *et al*. (Mirkin *et al*., 1995), and extensively used in V'yugin *et al*. (V'yugin *et al*., 2003). Suppose that two binary trees are given, a gene tree G (for a fixed COG) and a species tree S (including all species present in this COG). The unique *tree mapping* α*:* $G \to S$ was defined in (Mirkin *et al*., 1995; V'yugin *et al*., 2003). We consider two methods of gene and species tree comparison. **The first method** is based on comparison of the combinatorial structures of the trees G and S. We use the tree mapping a and compare the neighborhood O_g of the gene g in the gene tree G and its image (under α) in the species tree S. We assume that the gene g is "ambiguous" in position on the species tree if $\alpha(g)$ and $\alpha(O_g')$ are far apart in the species tree S (where O_g' is exactly the neighborhood O_g with the gene g omitted), the numerical characteristic R_g is a measure of this difference. High values of R_g reflect positional ambiguity of the gene g in the species tree (for details refer to (V'yugin *et al*., 2003)). In this paper we define **the integral characteristics of the gene tree** $\langle R_g \rangle = (1/m)\Sigma_g R_g$, i.e. the mathematical expectation of the R_g statistic over the corresponding COG, where m is the

number of genes in the COG. **The second method of comparison** is based on the gene duplication-loss model. A measure of *dissimilarity* $c(G,S)$ of a gene tree G and a species tree S (which is the sum of one–side duplications and intermediate nodes [1, 2]) was introduced. Thus, for any COG tree G we calculate the cost $F=c(G,S)$ (for $\alpha: G \to S$). Further, we reduce the gene tree G_g by gradually removing and replacing genes g from the gene tree G and calculate the cost F_g (for $\alpha: G_g \to S$). The relative change in the costs of two tree mappings α is calculated with the formula $dF_g = (F_g-F)/F$. The corresponding **integral characteristics** of the COG is the expectation $<dF_g>$ of dF_g over all its genes g. When we analyse a set of COGs, we denote by $<<R_g>>$ and $<<dF_g>>$ the expectations of $<R_g>$ and $<dF_g>$, respectively, for all the COGs.

Results and Discussion

A set of maximum-likelihood trees constructed for COGs was analysed in (V'yugin *et al*., 2003). For any statistic $f(g)$ we consider the corresponding p-value $p(g) = \text{card}(\{g': f(g') \geq f(g)\})/m$, where card(.) is the number of elements in the set (.) and m is the same number in the domain of the function f. As the case study for such analysis we selected 13 COGs of the 132 COGs studied in (V'yugin *et al*., 2003), which possessed extreme values of $<R_g>$ (and for which $p(g) < 0.1$). A fragment of this set is givenin the upper part of Table 1. In an analogous manner we analyzed the rest of COGs.

Table 1. Fragments of the COG list sorted by the $<R_g>$ value, where $<<R_g>> = 1.4623$; $<<dF_g>> = 0.6985$

COG	$<R_g>$	p-value for $<R_g>$	$<dF_g>$ in %	p-value for $<dF_g>$
COG0351	2.57	0.0076	-1.6399	0.023
COG0171	2.26	0.015	-0.62327	0.16
COG0547	2.19	0.023	0.10913	0.39
COG0169	2.14	0.015	0.66859	0.58
COG0573	2.11	0.038	0.50343	0.52
COG0135	2.1	0.045	-1.2455	0.076
COG0581	2.03	0.053	-0.42877	0.23
COG0221	1.95	0.061	-0.52252	0.2
COG0159	1.93	0.068	-1.2074	0.083
COG0597	1.92	0.076	1.1511	0.69
COG0340	1.9	0.083	-0.47401	0.21
COG0105	1.89	0.091	0.20002	0.4
COG1488	1.89	0.098	0.92679	0.61
.....	.....	.....	.....	.....
COG0060	1.5	0.3	1.1761	0.7
COG0012	1.47	0.39	0.40489	0.48
COG0016	1.38	0.58	1.363	0.73
COG0049	1.29	0.77	0.099124	0.37
COG0048	1.25	0.83	-0.039805	0.34
COG0051	1.25	0.84	0.20717	0.41
COG0052	1.22	0.86	0.75545	0.58
COG0013	1.19	0.92	0.98059	0.64

A computer program selects 247 genes g from the remaining 109 COGs, for which $p(g) < 0.1$ for p-values defined for any of the two above defined statistics. We refer to these genes as *gained genes*. The gained genes are considered candidates for horizontal transfers and other gene gain events. We further consider the following two options. First, we assume that there are no gain

events and calculate numbers of gene duplications, losses and gains (separately for each COG). In Table 2 we give the total (over 109 COGs) of these numbers (duplication, loss and gain separately). Secondly, all the genes identified as gain events by our approach were excluded from the domain of the corresponding tree mapping α and the same total estimates were calculated (the first case is called **non-GAIN scenario**, the second is **GAIN scenario**).

Table 2. Total number of duplications in groups of species

Group of species	non-GAIN scenario	GAIN scenario
Archaea	149	143
Gram-positive bacteria	54	55
Alpha-proteobacteria	7	7
Gamma&Beta-proteobacteria	207	202
Epsilon-proteobacteria	0	0
Clamydia&Spirochaetes	2	2
DMS	5	4
Thermotoga&Aquifex	0	0

Massive gene duplication attributed to the root of a phylogenetic group could be interpreted as a result of possible "*genome duplication*". Such is the set of 92 gene duplications assigned to the root of the subtree of Archaea. Another large group of 83 gene duplications was found in the gamma-proteobacteria group and assigned to the root of the species subtree (((*Pmu,Hin*),(*Eco,Buc*)),*Vch*). It also might result from ancient genome duplication (see Fig.). Massive gene duplication in the ancestor of *Vibrio cholerae* was independently postulated in (Heidelberg *et al.*, 2000).

Conclusion

A mathematical model of gene duplication and loss was applied to compute numerical characteristics measuring discrepancy between the gene trees and the species tree. Using integral characteristics of COG quality, we excluded a set of gene trees from the analysis. We also conclude that the total number of gene duplications assigned to internal nodes of phylogenetic groups are almost independent of the scenario chosen, GAIN r non-GAIN.

References

Mirkin B.G., Muchnik I., Smith T. A biologically consistent model for comparing molecular phylogenies // J. Comput. Biol. 1995. V. 2. P. 493–507.

V'yugin V.V., Gelfand M.S., Lyubetsky V.A. Identification of horizontal gene transfer from phylogenetic gene trees // Mol. Biol. 2003. V. 37(4). P. 571–584. (In Russian).

Heidelberg J.F. *et al*. DNA sequence of both chromosomes of the cholera pathogen Vibrio cholerae // Nature. 2000. V. 406. P. 477–483.

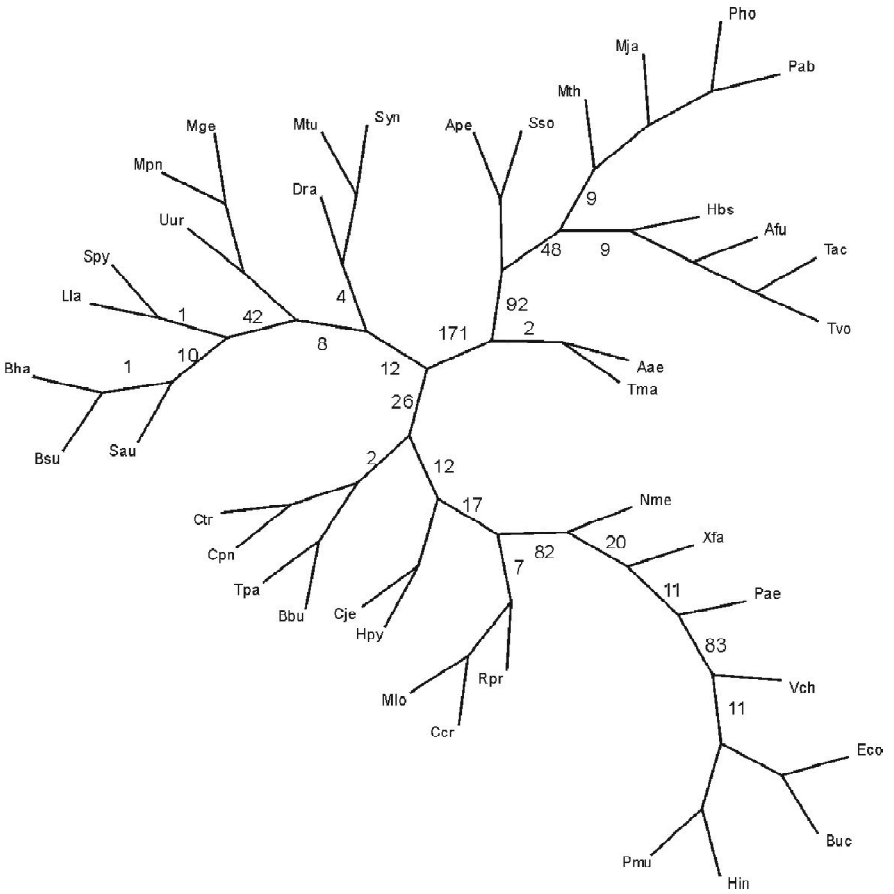

Fig. Total number of duplications assigned to groups of species (for the non-GAIN scenario).