ALGORITHMS TO RECONSTRUCT EVOLUTIONARY EVENTS AT MOLECULAR LEVEL AND INFER SPECIES PHYLOGENY

V. Lyubetsky, K. Gorbunov, L. Rusin, V. V'yugin*
Institute for Information Transmission Problems, Russian Academy of Sciences, Bolshoy Karetnyi per. 19, Moscow, Russia, e-mail: vyugin@iitp.ru
* *Corresponding author*

Abstract: Mathematical methods and models for comparative analysis of large sets of protein phylogenies are described. The processes modeled are gene duplication, loss, gain, and horizontal transfer. Initially, a species tree is constructed as a consensus of the corresponding gene trees using probabilistic distribution on source data. Algorithms are further implemented to identify vertices accounting for topological disparities between the gene and species trees, with possibility to infer underlying evolutionary events. The analysis is illustrated on case studies of a prokaryotic protein family and a set of protein phylogenies deduced from families from the COGs database (NCBI). The potential of the described methods to infer phylogeny and gene evolution events is discussed.

Key words: evolution; phylogenetic tree; consensus tree; gene duplication; gene loss; horizontal gene transfer; mathematic models of evolution; stochastic optimization

1. INTRODUCTION

Methods and algorithms described here are aimed at implementing two tasks: reconstruction of prokaryotic species trees and analyzing hypotheses about gene evolution. The main emphasis is placed on original algorithms and their performance, although, due to space limits, only general descriptions are provided along with the necessary references.

Events in gene evolution are usually viewed as gene divergence during species differentiation, gene duplication, gene gain, loss, and horizontal gene

transfer (HGT). Molecular data is protein sequences grouped according to their amino acid and functional similarity into clusters of orthologous groups of proteins (COGs; Tatusov et al., 2001).

The general approach to reconstruct gene evolution events has long been defined (Goodman et al., 1979; Eulenstein et al., 1998). A protein gene family is selected, usually from among COGs, with subsequent assembling of multiple sequence alignment and reconstruction of the gene tree G (also referred to as a protein tree or COG tree). Further analyzed are topological similarity and disparity between the gene trees from the set $\{G_i\}$ in order to reconstruct the species tree and infer gene evolution events, respectively. Topological differences are reconciled to produce the species tree S. Alternatively, when inferring gene evolution events, considerable topological differences between a particular gene tree G (often pertaining to the family $\{G_i\}$) and the species tree S are the basis of the analysis.

Mathematic models of gene evolution are formulated to accommodate the observed differences, and optimization of model parameters is used as a tool to reconstruct evolutionary history of a microbial gene family. The evolutionary model is defined as a procedure of comparing the gene and the species trees, while its parameters are defined as sets of tree vertices with assigned evolutionary events. An optimized model has parameters corresponding to the extremes of the relevant evolutionary characteristics.

2. METHODS AND ALGORITHMS

2.1 Reconstructing the gene tree

For a given protein family (usually, for a COG), a multiple sequence alignment is assembled (routinely we use the program PROBCONS v. 1.09). Sequences with a low level of the overall detectable homology with respect to the other family members are identified using the CORE index (Notremade et al., 2003), which scores each residue for the amount of positional consistency it contains with respect to the other residues occurring in the same column (index computed with the program T-COFFEE v. 2.11). Sequences with the CORE value below the recommended threshold are removed from the alignment.

At the next step, a list of reliable phylogenetic clades is defined. For this purpose, a standard bootstrap analysis is applied. Sufficiently large numbers of bootstrapped replicates are generated for primary data using the program *seqboot* from the PHYLIP v. 3.63 package and further used to estimate the ML distance matrices under selected evolutionary model using the program PUZZLEBOOT. Neighbor joining is used to construct the trees that are

further reconciled to produce a 70 % consensus (facilitated by the programs *neighbor* and *consense*, respectively, from PHYLIP v. 3.63). The groups retained in the consensus comprise the list of *reliable clades*. An ML model to be used whenever else needed is selected from more that 50 empirical models of protein evolution on the basis of significant improvement in the data likelihood according to the likelihood ratio test (Akaike, 1974; Goldman, 1993) and the Bayesian (Schwarz, 1978) information criteria. Model selection is implemented with the program ModelGenerator.

High evolutionary rates often lead to mutational saturation and loss of phylogenetic signal in highly variable regions of the protein molecule. We introduce several functions of conditional entropy in order to range the columns of the *initial alignment* according to the amount of consistency they possess with respect to the list of reliable clades and subsequently screen out for the non-informative ones.

In order to detect the amount of columns needed to be removed from the initial alignment to achieve maximum performance of phylogenetic inference, we implement a criterion based on two statistics. After eliminating a subsequent portion of highest entropy positions, we compute for the resulting alignment (1) the percentage of unresolved quartets of taxa and (2) g_1-statistic. Procedures of estimating the statistics were modified as follows.

(1) Maximum-likelihood mapping. The quartet analysis was conducted so that the phylogenetic signal related to robust clades does not contribute to the percentage of unresolved quartets. Namely, sequences corresponding to the taxa in a reliable clade from the list were substituted with an ancestral sequence reconstructed with ML at the root of the clade, thus defining a *reduced* alignment. Maximum likelihood mapping (Strimmer and Haeseler, 1997) was performed with the program TreePuzzle v. 5.02 and ancestral sequence reconstruction, with the PAML v. 3.14 package.

(2) g_1-statistic. Under the maximum parsimony, the tree length is defined as a minimum number of the changes required to explain its topology. If aligned data are phylogenetically structured, the percentage of shorter trees among a large set of the randomly generated ones will skew the tree length distribution to the left (Hillis and von Huelsenbeck, 1992). The distribution skewness is measured with the g_1 statistic. To preclude the phylogenetic signal related to well-resolved groups from contributing to the distribution skewness, we constrained analysis by generating random topologies in the areas remaining unresolved in the 70 % consensus.

Alignment columns are removed until both statistics reach extreme values. In the cases when the statistics diverge in detecting the optimal alignment, phylogenies are estimated with both alignments and further reconciled in a strict consensus. In the resulting tree, the evolutionary

distances are computed as branch lengths with ML according to the selected evolutionary model.

The entire procedure is iterated until an optimal alignment is found. Phylogenetic trees inferred with the described approach always possess a higher likelihood with respect to the primary data than do the trees estimated with the initial alignment and often do not constitute a confidence set with them.

2.2 Constructing the bacterial species tree

A species tree is produced by reconciling a set of the gene trees $\{G_i\}$. It is defined as such S from the space of all *suitable* species trees that maximizes a certain parameter, e.g., the similarity between S and all G_i. There exist several natural definitions of this 'similarity' (examples are provided below), while *a priory* suitability requirements to be imposed on species trees are not biologically straightforward.

Mapping of and the cost for dissimilarity of trees were introduced by Goodman et al. (1979), Guigo et al. (1996), and Page and Charlstone (1997). This cost definition was modified by V'yugin and Lyubetsky (2002) by substituting the number of edges with the sum of the corresponding edge lengths, introducing edge length normalization, parameter γ, and probabilistic distribution over the primary sequence data (details are discussed below). The interpretation of the edge length in the tree G depends on the tree inference method. It can either be an estimate of the edge robustness or evolutionary distance between vertices.

Let α denote the conventional mapping of the gene tree G into the species tree S and let $c(G, S)$ denote the cost of such mapping, a measure of dissimilarity between α and identical mapping of trees, i.e., the extent to which the tree G is not identical to the tree S. Remember that a duplication event can be thought of as a pair (g, s), where g is a vertex in the gene tree G and s is a vertex in the species tree S satisfying the condition $\alpha(g) = \alpha(g')$ for one or both immediate descendants g' of the gene g (one to the left is designated as cg; one to the right, as Cg). The vertex s of the species tree S is g intermediate, if it is situated exactly between the vertices $\alpha(g)$ and $\alpha(pg)$, where pg stands for the last common ancestor of the vertex g. Let us denote $M(G, S)$ as a set of all g intermediate vertices for all gs from G. A member of the set $M(G, S)$ is also called a gap. The gap corresponds to the edge (g, pg) with the length $l_{(g, pg)}$ in the gene tree G. Guigo et al. (1996) proved a theorem stating that the total number of gene losses equals to the total number of one-side duplications and gaps.

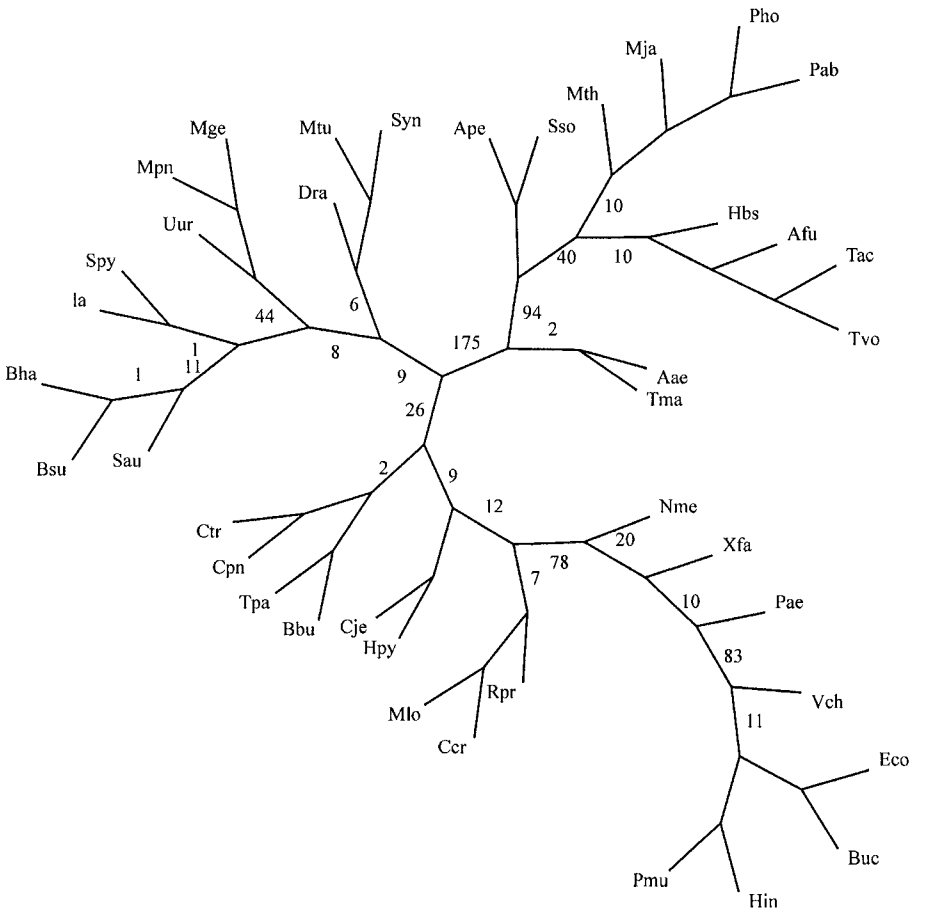

Figure -1. Evolutionary tree of 40 microorganisms from the following groups: Archaea—(Afu) *Archaeoglobus fulgidus*, (Hbs) *Halobacterium* sp. NRC-1, (Mja) *Methanococcus jannaschii*, (Mth) *Methanobacterium thermoautotrophicum*, (Tac) *Thermoplasma acidophilum*, (Tvo) *Thermoplasma volcanium*, (Pho) *Pyrococcus horikoshii*, (Pab) *Pyrococcus abyssi*, (Ape) *Aeropyrum pernix*, and (Sso) *Sulfolobus solfataricus*; Gram-positive bacteria—(Spy) *Streptococcus pyogenes*, (Bsu) *Bacillus subtilis*, (Bha) *Bacillus halodurans*, (Lla) *Lactococcus lastis*, (Sau) *Staphylococcus aureus*, (Uur) *Ureaplasma urealyticum*, (Mpn) *Mycoplasma pneumoniae*, and (Mge) *Mycoplasma genitalium*; Alpha-proteobacteria—(Mlo) *Mesorhizobium loti*, (Ccr) *Caulobacter crescentus*, and (Rpr) *Rickettsia prowazekii*; Beta-proteobacteria—(Nme) *Neisseria meningitidis* MC58; Gamma-proteobacteria—(Eco) *Escherichia coli* K12, (Buc) *Buchnera* sp. APS, (Pae) *Pseudomonas aeruginosa*, (Vch) *Vibrio cholerae*, (Hin) *Haemophilus influenzae*, (Pmu) *Pasteurella multocida*, and (Xfa) *Xylella fastidiosa*; Epsilon-proteobacteria—(Hpy) *Helicobacter pylori* and (Cje) *Campylobacter jejuni*; Chlamydia—(Ctr) *Chlamydia trachomatis* and (Cpn) *Chlamydia pneumoniae*; Spirochetes—(Tpa) *Treponema pallidum* and (Bbu) *Borrelia burgdorferi*; and DMS—(Dra) *Deinococcus radiodurans*, (Mtu) *Mycobacterium tuberculosis*, (Syn) *Synechocystis*, (Aae) *Aquifex aeolicus*, and (Tma) *Thermotoga maritima*. Vertices are assigned the total number of duplications for 132 protein families. The list of the families is taken from Wolf et al. (2001).

Remember that the duplication (g, s) is one-sided if either of the conditions $\alpha(g) = \alpha(cg)$ or $\alpha(g) = \alpha(Cg)$ is true. A one-side duplication (g, s) corresponds to the edge (g, cg) or (g, Cg) with the length $l_{(g, cg)}$ in the gene tree G. A set of all one-side duplications is designated as $O(G, S)$.

The duplication (g, s) is considered to have occurred in the vertex s. The number of such pairs under fixed s defines the number of duplications in the vertex. The total number of duplications in the genome assigned to the vertex s is the sum of all one-side duplications in the vertex over all gene families from a fixed set of families (Figure 1). The statement 'in the genome' implies that the set is assembled to be maximally representative. For an individual protein family, the total number of duplication in descendants of the vertex s is estimated as a sum of one-side duplications in all vertices of the clade contained in s. A more sophisticated procedure is used to infer the number of gene losses in the vertex. Remember that a gene loss in the vertex s corresponds to the pair (g, s), where g contains a duplication, s descends from $\alpha(g)$, and the clade s does not contain either of genes from the clade g', g' being an immediate descendant of g, while the clade ps does contain genes from both clades g' (Eulenstein et al., 1998). This definition is sometimes made more complex with additional conditions imposed on the pair (g, s). The number of losses in s is defined as the number of all such pairs (g, s) under fixed s. Other types of evolutionary events are treated analogously. The total estimates are considered as important characteristics of vertices of the species tree, protein families (genomes), and phylogenetic clades. HGT is considered as a special case of gene gain when its origin can be traced.

The cost of mapping of the gene tree G into the species tree S is defined as

$$c(G, S) = |O(G, S)| + \gamma \cdot |M(G, S)|,$$

where $|\{\cdot\}|$ stands for the cardinality $\{\cdot\}$, i.e., the number of set members. Otherwise, it can be given by two sums:

$$c(G,S) = \sum_{g \in O(G,S)} l_{(g,cg)} + \gamma \cdot \sum_{g \in M(G,S)} l_{(g,pg)} .$$

By minimizing the value of $c(G, S)$ under $\gamma = 1$, the total number of gene losses is minimized. If $\gamma < 1$, the cost favors duplications over gaps. In some cases, only the number of duplications is minimized (Page and Charlstone, 1997).

Thus, the species tree S is produced by minimizing the value

$$c = c(S) = c(G_1, S) + c(G_2, S) + \ldots + c(G_n, S),$$

where all gene trees G_i are already obtained, and the unknown species tree S is being produced under certain *a priory* imposed conditions. This value will also be referred to as a cost of mapping of the gene tree set $\{G_i\}$. From the mathematical standpoint of computational complexity theory, finding the minimum of $c(S)$ is a highly nontrivial task. Importantly, more robust edges of the trees G_i have more impact on minimization of the function $c(S)$. The edge lengths (robustness or divergence times) of trees G_i can be induced on the resulting species tree S.

V'yugin et al. (2003) proposed a partial solution of the known issue with reconstructing species trees caused by long branch attraction artifact. Namely, the minimization of function $c(S)$ is preceded by normalization of the edge lengths in gene trees from $\{G_i\}$. Edge lengths are re-estimated using the formula

$$l(g) = (l(g) - l_{cp})(1 + m)^{-(l(g)/l_{cp})} + l_{cp},$$

where l_{cp} stands for the mean edge length of gene trees. The normalization procedure reduces the impact of extreme edge lengths in G_i on the resulting tree S.

Let us now concern the selection of value for parameter γ, which determines the ratio between the numbers of duplications and losses. Many of the loss events, especially in vertices close to the root of the species tree S, may represent false predictions incurred from incorrect topologies of the source trees. Apart from that, mapping α does not accurately account for the gene gain events (particularly, HGT). A putative gene gain event can be alternatively explained by the topological disparities between the gene and species trees caused by a small number of gene duplications compared to a magnitude-larger number of gene losses. Therefore, a species tree constructed with optimization of an α-based model may be improved by assigning more weight to duplications, which are predicted more accurately. In our experiments, we generally assumed $\gamma = 0.1$.

Our algorithm constructs an optimal tree S as a *specific* local minimum. Since the algorithm produces a local minimum depending on the initial species tree S_0, we developed an *ad hoc* approach to construct the initial S_0. Namely, a probability distribution in the set of all initial species trees is built; it is defined automatically by the family $\{G_i\}$ of gene trees as follows. For any species a and b, the distribution $p(b|a)$ is defined as a probability for both b and a to form an elementary tree (i.e., to be located at a distance less than or equal to some fixed r, for example, $r = 2$). Let N_a be the number of the gene trees containing species a, and let $N_{a,b}$ be the number of trees containing a and b located at a distance r. Then, $p(b|a) = N_{a,b}/N_a$, and

$1 - \sum_b p(b|a)$ is the probability of the event that there is no occurrence of $b \neq a$ in any elementary tree containing a (i.e., there is no species b located at a distance r from a). We considered small species trees defined with the distances $r = 2, 3, 4$, etc., although larger distances require larger sets of primary data. The random binary tree S_0 is generated with the distribution and is taken as an initial tree in the search algorithm. The final output is a consensus tree computed on a subset of the resulting trees with a sufficiently small value of the function c. Edges of this consensus tree are assigned values of support of the corresponding clusters.

2.3 Identification of vertices introducing incongruence between gene and species trees

Optimizing of parameters of the above-described models requires identification of the tree vertices informative with respect to the inferring events of gene evolution. A substrate for this type of analysis is a topological incongruence between some gene trees G_i and the consensus species tree S, which can be accounted for by actual events in gene evolutionary history or artifacts in reconstruction of the source trees, the latter representing a problem of its own.

Three algorithms for detecting sets of incongruent vertices are described. The first algorithm is based on identification of the subset G' of terminal vertices (leaves) in the gene tree G representing the gene gain events. The evolutionary model is two mappings α with different domains: initially, α is defined on the gene tree G and, subsequently, on its subset of the leaves $G \backslash G'$ obtained by excluding the leaves with putative HGTs. The subset $G \backslash G'$ is transformed into a binary tree using a standard procedure. Liberally speaking, $G \backslash G'$ can be considered as a subtree of G. The set G' is a parameter of the model, and its selection (optimization of parameter G') is carried out by maximizing a set-dependent value. Other evolutionary events are defined via mapping α as described above.

The first algorithm consists of the two segments.

The first segment. The terminal gene g in gene tree G is considered as putative HGT, if all genes $g_1, g_2, \ldots, g_n$ in its proximity except for g itself are mapped with α onto the species $s_1, s_2, \ldots, s_n$ distant from the species $s = \alpha(g)$. It is also prerequisite that the set of species $\{s_1, s_2, \ldots, s_n\}$ is located compact enough in the species tree S, i.e., its ancestor s_0 is close enough to the leaves and the distance between s_0 and s is considerably large in S. The genes $g_1, g_2, \ldots, g_n$ are defined as a set of terminal vertices without g separated in G with a distance less than r, where the distance is the length of the path from g to g_i; it either takes into account the edge lengths or does not

if those are unit lengths. The gene set $\{g_1, g_2, \ldots, g_n\}$ is also called the punctured neighborhood of the gene g with a radius r. Usually, under unit lengths, we assumed $r = 4$. Let us provide some more details.

If two terminal genes g and g_1 are located at a small distance in G but species $\alpha(g)$ and $\alpha(g_1)$ in mapping α are at a great distance in S, it may suggest an abnormal position of one of the genes. Hence, the distance $r(g, g_i)$ in the gene tree and the distance $r(s, s_i)$ in the species tree are calculated, where $i = 1, \ldots, n$, and thus the average values are

$$r(g) = (1/n) \sum_i r(g, g_i) \text{ and } r(s) = (1/n) \sum_i r(s, s_i) .$$

The value of $R_g = r(s)/r(g)$ determines the extent to which the size of the species set $\{s_1, s_2, \ldots, s_n\}$ is larger than that of the gene set $\{g_1, g_2, \ldots, g_n\}$. Large values of R_g can be interpreted as suggesting abnormality in location of the gene g in the species tree. Conventional p-values are calculated for the statistic $p(.)$ using the formula

$$p(g) = \left| \left\{ g' \middle| R_{g'} \geq R_g \right\} \right| / m ,$$

where m is the number of all terminal vertices. The computer program selects all genes g with $p(g) \leq p_0$, where p_0 is a threshold. Such genes are considered as abnormally positioned.

The algorithm also selects all cases when the species $s_1, s_2, \ldots, s_n$ are part of a taxonomic group that does not contain species s and its ancestor s_0 is sufficiently separated from s in S. This suggests that this group is a putative origin of a horizontally transferred gene g.

The second segment. Suppose that each abnormally located gene generates a series of invalid duplications and losses under mapping α, which are required to explain incongruence between the gene G and species S trees in the model. Therefore, temporarily omitting the transferred gene g from G and re-estimating α after the deletion entails an essential reduction in the cost $c(G, S)$ of mapping G into S. Therefore, we calculate $c(G, S)$ and subsequently remove each gene g from G to obtain the reduced gene tree G_g and compute the cost c_g of mapping of the new gene tree G_g into the same species tree S. The relative change in the mapping cost is $F_g = (c_g - c)/c$. As above, we use p-values for the statistic F_g for all genes g from the given COG G. Similarly, the computer program selects all the genes g for which $p(g) \leq p_0$.

The mean and standard deviation of the statistic F_g can be used, if the empirical distribution of the statistic F_g is normal. Interestingly, our studies

reveal a considerably high support for the hypothesis of normality of the empirical distribution of F_g and log-normality of R_g for most COGs.

The genes selected at the second segment of the algorithm are interpreted as gained (not only due to HGT, as its origin is not always determined). The genes selected in both segments are considered as gained during an evolutionary event, probably, a HGT.

The first algorithm is designed to detect the recent HGTs, when the recipient and donor species did not diverge greatly in evolution. Gorbunov and Lyubetsky (2005) proposed two novel algorithms as generalization of the first algorithm to be able to detect deeper ancestral HGTs. In this sense, *ancestral* genes are those existing in an internal vertex of the phylogenetic tree. To stress this discrimination, *extant* genes are sometimes referred to as those existing in terminal vertices.

The second algorithm implements a juxtaposition of the gene tree G with the species tree S using graph β instead of mapping α in the first algorithm. Let us define some terminology.

Each *vertex* g in a tree corresponds to the *set* K of all leaves contained in the vertex g, in which sense the vertex g and *clade* K are mutually deterministic. The graph β contains all clades in G and all clades in S as vertices, with each clade K in G connected via one edge with each clade K' in S, edge K, K'; the graph contains no other edges. Let us define the *components* of the edge K, K' as two sets $M = K \backslash K'$ and $M' = K' \backslash K$.

For each edge K, K', we calculate the ratio of the cardinality of component M to the cardinality of the components containing clade K, the $\frac{|M|}{|K|}$ value, and, analogously, the $\frac{|M'|}{|K'|}$ value for the clade K'. Let us remember that the cardinality $|M|$ of set M is defined as the number of its members. The probability of the component M on edge K, K', we define as $1-\frac{|M|}{|K|}$, and, analogously, the probability of the component M' as $1-\frac{|M'|}{|K'|}$. Each edge K, K' in the graph β is assigned the two probabilities, which we define as *probabilities* of the edge K, K'. The edge K, K' can be viewed as an analogue to the pair $< g, \alpha(g)>$ in the first algorithm.

Let us define *the workmate* M^* of the set M of leaves (terminal genes) as a complement of M to the set of all leaves in a certain subtree of G. Two alternatives of defining the subtree are considered: the subtree is rooted in the last common ancestor of all members of the set M; otherwise, it is rooted in the node parental to this ancestor. Let us call these the *first* and *second* workmates. The Sets M and M^* usually are not clades.

The algorithm described tests the possibility of HGT between the ancestor of set M and the ancestor of its workmate M^* in the species tree S. The algorithm is as follows. A list of all edges K, K' in the graph β is defined, for which at least one of the probabilities is above a certain threshold. For each nonempty component M with such probability, both workmates M^* are analyzed. The pair $< M, M^* >$ is called a *candidate pair*, if three simple conditions are satisfied:

(1) *Similarity of the candidate pair* $< M, M^* >$, measured as a mean distance between the elements of the two sets in the gene tree G (if edge lengths are present) or as a percent identity in pairwise alignments of the corresponding sequences, is under a certain threshold.

(2) *Compactness* of the set M in species tree S, defined as the ratio of the cardinality of M to the cardinality of the leaf set in a subtree of S rooted in the last common ancestor of all leaves from M as well as the analogous compactness of the set M^*, are above certain threshold.

(3) *The distance* between the last common ancestor of the set M and the analogous ancestor of M^* in the species tree S exceeds a certain considerable threshold. (If the ancestors are close in the tree S, conditions (1) and (2) may be true simply due to relatedness of M and M^*). This requirement is supplementary to the requirement that the compactness of the union of all species from M and M^* is below a certain threshold. Low values of this compactness, to the contrary, suggest a HGT event.

The more edges are in the graph β that imply the pair $<M, M^*>$ with higher probability, the higher weight is given by the algorithm to the pair as a candidate HGT between ancestors of M and M^*.

Performance of the second algorithm can be assessed on a case example of two trees, species tree $(((a,(e, b)),(3,(4, 5))),((1, 2),(c, d)))$ and gene tree $((((a, b),(c, d)),e),((3,(4, 5)),(1, 2)))$, with $M = \{a, b\}$ and its second workmate $M^* = \{c, d\}$.

For reasons of conciseness, *the third algorithm* will be described for the case when a gene copy persists in the source lineage after HGT. The algorithm is not sensitive to this constraint. It is based on analysis of fuzzy gene sets from a fixed COG.

The fuzzy gene set R is defined by a credibility function, which estimates the 'credibility of membership' in R of each gene from a fixed COG. Let K be a clade in the species tree and P be the set of all genes from the COG belonging to K. The fuzzy set R is given by P, i.e., given is a string of numbers, credibilities p_g, for all genes g from the COG. In the simplest case, p_g is proportional to similarity of the gene g to its closest match g_1 from P. The similarity can be estimated from COGs multiple alignment, from a path in the COG tree, or, in absence of the two former, simply from a percent identity of pairwise alignments of the corresponding sequences (Gorbunovand Lyubetsky, 2005).

Instead of similarity, one may calculate 'informativity about the gene g contained in g_1' or 'informativity about the gene g contained in set P'. To do so, we applied the Lempel–Ziv algorithm originally modified to use the entry g_1 sequence or entry set P. For the basics, one may consult Otu et al. (2003).

Hence, for an arbitrary pair of clades K and K' in the species tree, one can calculate a pair of the corresponding fuzzy sets R and R'. Let the *quality* $Q(K, K')$ of the clade pair be the ratio of the cardinality of 'fuzzy intersection' of R and R' to the cardinality of 'fuzzy union' of R and R', i.e., by definition, $Q(K,K') = \dfrac{\sum_g \min(p_g, q_g)}{\sum_g \max(p_g, q_g)}$. Let the *kernel* M of two primary clades K and K' be a set of genes g from the COG, for which $\min(p_g, q_g)$ is above a certain threshold. The genes from M may be interpreted as descendants of a horizontally transferred gene. The algorithm searches for HGTs as pairs of disjoint clades in the species tree S, with their kernel M containing two gene sets M_1 and M_2, both having sufficient compactness in S and their union closely coinciding with M, and not having high compactness in S (relevant thresholds implied).

Performance of the algorithm can be illustrated on the same case study as provided above after defining reasonable distances between the genes with respect to the gene tree and assuming a simple transformation of the distance x from gene g to its closest match g_1 from P into credibility p_g as $p_g = 32 \cdot (4-x)$.

3. RESULTS AND DISCUSSION

3.1 Reconstruction of bacterial species phylogeny

Consider a typical output of the algorithm described in section 2.1. It was run to infer the phylogeny of 40 microorganisms with 132 protein families. A detailed description of the primary data is provided in V'yugin et al. (2003; Figure 1). The search algorithm runs on a set of 5000 generated initial species trees S_0 as described above. The minimum value of c was 42 648. Robustness of the algorithm can be judged from the observation that two groups of resulting species trees selected by the algorithm, a set of 48 trees with $42\,648 < c < 42\,991$ and a set of 182 trees with $42\,648 < c < 44\,861$, have identical consensus topologies. The same holds true for a number of subsequently constructed species trees. The incongruence between the species tree thus obtained and the best species tree published in Wolf et al. (2001, approach (v)) is negligible and occurs only with respect to the relative position of groups of epsilon-proteobacteria, Aae, and Tma.

3.2 Deciding between two alternative hypotheses

Consider a typical case when decision is to be made in favor of either a small number of HGTs or a considerable number of gene losses. Application of the first algorithm to COG0272 (NAD-dependent DNA ligase) returns the following result: the initial mapping α onto the species tree detects 5 duplications (with 4 existing in the vertex of species tree) and 17 gaps, thus giving 22 gene losses in total. It identifies the gene *yicF* from *E. coli* as largely accounting for the incongruence between the protein and species trees.

After omitting gene *yicF* from the gene tree, the number of duplications reduces to two and the number of gaps, to five, giving a total of seven gene losses.

Thus, assuming HGT with *yicF* decreases the number of losses by 15 (Figure 2). The first algorithm concludes with a high confidence that the gene *yicF* was horizontally transferred from some spirochaete bacteria.

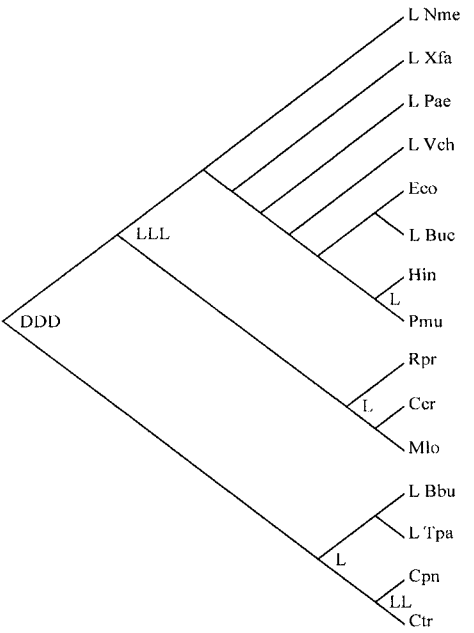

Figure -2. A part of the evolutionary history of COG0272 (NAD-dependent DNA ligase). The hypothesis about the absence of HGT events for gene *yicF* requires assuming 3 additional duplications and 15 additional losses of this and other genes. Duplications are marked with D; losses, with L.

3.3 Reconstruction of ancestral events in gene evolution

We conducted mass analyses of COGs using the tree algorithms (for more detail, refer to V'yugin et al., 2003; Lyubetsky et al., 2003a, b). Consider a typical result obtained for the above-mentioned 132 protein families. Initially, we tested all genes from each COG and selected 365 of those that contribute the most to the incongruence between the gene and species trees. Subsequently, all the 365 genes selected were omitted from their gene trees, and the mapping α of each of the gene trees into the species tree was re-estimated. For both cases, we counted the numbers of gene duplications and losses for each COG. In the first case, called *non-GAIN scenario*, the algorithms detect 1558 gene duplications and 9009 gene losses. The second case is called *GAIN scenario* and produces 1392 gene duplications, 7400 gene losses, and 365 GAIN events. The hypothesis about single GAIN event reduces the number of losses by an average of 4.4 (the difference between 9009 losses in non-GAIN and 7400 losses in GAIN scenario divided by 365 gains). The distribution of total estimated duplications under the GAIN scenario across prokaryotic families is as follows: Archaea, 154 (94 in the root); gram-positive bacteria, 65 (8 in the root); alpha-proteobacteria, 7 (all in the root); beta-proteobacteria, 0; gamma-proteobacteria, 124 (20 in the root); and chlamydias and spirochetes, 2 (both in the root; Figure 1).

Large total numbers of gene duplications (comparable to the number of protein families) assigned to a vertex of the gene tree might suggest *whole genome duplications*. Such are the group of 92 duplications in the root of Archaea and the group of 83 duplications in the root of (((Pmu,Hin),(Eco,Buc)),Vch).

Table -1. Selected number of reconstructed evolutionary events

	non-GAIN scenario			GAIN scenario		
1	2	3	4	5	6	7
COG	Dupl	Loss	Gain	Dupl	Loss	Gain
COG0012	12	63	0	9	48	1
COG0102	13	71	0	11	53	1
COG0143	16	102	0	14	80	2
COG0198	18	99	0	13	67	1
COG0215	16	71	0	11	46	2
COG0272	8	51	0	5	28	1
COG0290	9	41	0	8	28	2
COG0343	14	70	0	13	65	1
COG0544	4	30	0	5	25	1
COG0571	9	59	0	4	22	3
COG0653	8	39	0	7	29	2
COG1160	5	27	0	4	23	1

The computer programs also output mappings of each COG tree into the species tree for purposes of evolutionary history reconstruction under both scenarios for each of the 132 protein families (Lyubetsky et al., 2003b). Selected numbers of evolutionary events thus reconstructed are given in Table 1; columns 2–4 contain inferences under non-GAIN scenario and columns 5–7, those under GAIN scenario.

To continue, let us provide some details on the event reconstructions for selected COGs.

COG0012 (predicted GTPase). *Buchnera aphidicola*, a member of the gamma-proteobacteria group, occurs in the species tree in the same cluster with *E. coli*, but its gene *bu191* is found close to chlamydial genes in the gene tree. We suggest that this group is the source of HGT. Also suggested is that the gene *sll0245* is horizontally transferred to the genome of *Synechocystis* sp. from spirochetes.

COG0215 (aminoacyl-tRNA synthetases and alternative system for amino acid activation). It is suggested that the gene *vng1095G* from *Halobacterium* sp. (halophilic archaebacteria originating from eubacteria) is horizontally transferred from the genome of an organism similar to *Deinococcus radiodurans*. It is likely that the gene *xf0995* from the organism *Xylella fastidiosa*, which occurs in the same cluster, is transferred from some alpha-proteobacteria similar to *Caulobacter crescentus*.

COG0143 (methionyl-tRNA synthetase). The *mlr5926* gene from *Mesorhizobium loti* (alpha-proteobacteria) is a putative HGT from some archaebacteria. Moreover, this event entailed subsequent divergence of paralogous genes in this genome.

COG0102 (ribosomal protein, large subunit) provides an example of a ribosomal gene HGT. The *dr0174* gene from *Deinococcus radiodurans* (L13 protein) is likely transferred from a genome of some gamma-proteobacteria.

COG0198 (ribosomal protein, large subunit). The *bb0489* gene from *Borrelia burgdorferi* (Spirochaeta; encodes L13 protein) is transferred from some gamma- or beta-proteobacteria.

COG0272 (basal replication machinery). The *yicF* gene from *E. coli* (NAD-dependent DNA ligase) is horizontally transferred from spirochaete bacteria. In addition, the *E. coli* genome contains gene *lig*, bearing the same function as *yicF*.

COG0343. The *af1485* gene from *Archaeoglobus fulgidus* (queuine/archaeosine-tRNA ribosyltransferase) is likely to be transferred from eubacteria.

The second and third algorithms converge in inferring the same ancestral HGTs. Thus, for COG0180 (tryptophanyl-tRNA synthetase), the algorithms predicted putative HGTs between the ancestors of groups {Bha, Bsu, Sau} and {Vch, Eco, Buc, Hin, Pmu}. The predictions corresponded to 6 edges in graph β, high densities and 6 pairs of clades in species tree producing the

same kernel M = {Bha, Bsu, Sau, Vch, Eco, Buc, Hin, Pmu, Hpy, Mtu} (refer to descriptions of the second and third algorithms).

Putative HGTs can be alternatively identified with non-phylogenetic approaches based on comparative analyzes of codon usage, frequencies of genomic features, and other contextual characteristics (Garcia-Vallve et al., 2003).