Molecular Biology, Vol. 39, No. 5, 2005, pp. 741–751. Translated from Molekulyarnaya Biologiya, Vol. 39, No. 5, 2005, pp. 847–858.
Original Russian Text Copyright © 2005 by Gorbunov, Lyubetsky.

BIOINFORMATICS

UDC 577.2.08:681.31:575.582

Identification of Ancestral Genes That Introduce Incongruence between Protein- and Species Trees

K. Yu. Gorbunov and V. A. Lyubetsky

Institute for Information Transmission Problems,
Russian Academy of Sciences, Moscow, 127994 Russia
e-mail: gorbunov@iitp.ru

Received January 25, 2005

Abstract—Two new approaches to detecting potential incongruence between a protein family tree and a species tree are considered. The first approach is based on the substitution of a known mapping of the gene tree *G* into the species tree *S* with a somewhat analogous multivalued *G*-into-*S* mapping. The second approach is based on the elementary concepts of the fuzzy set theory. Two algorithms corresponding to these approaches are described in detail, and their implementation is shown using a simulation example and three protein families from the database of clusters of orthologous protein groups (COGs).

Key words: horizontal gene transfer, search for horizontal transfers, evolutionary event, reconstruction of evolution, phylogenetic species tree, phylogenetic protein tree, tree reconciliation

INTRODUCTION

Rare protein families have an evolutionary history that agrees well with the evolution of the corresponding species. The history of protein families and species is commonly shown by two corresponding trees (or at least two graphs). In this case, the degree of congruence between the evolution of protein family *G* and of the corresponding species *S* is reflected in a match or divergence of the corresponding graphs *G* and *S*. This introduces a constrained but important problem of searching for terminal and inner nodes, i.e., genes of *G* that contribute most to the divergence of *G* and *S*. The problem is constrained since the immediately subsequent problem consists in the interpretation of the event responsible for such an artifactually (irregularly) located gene. The divergence may be due to an artifact in the tree of proteins resulting from significantly different rates of molecular evolution in different lineages, gene duplication in remote ancestors and subsequent loss of many duplicated genes, or horizontal gene transfer. These three cases are at present difficult to distinguish algorithmically. Hence, the problem of interpretation becomes separate from the first problem, searching for artifact genes or proteins. Hereinafter, such genes are referred to as horizontally transferred genes.

Note that a species tree is the result of reconciliation of many gene trees $\{G_i\}$ and may be properly referred to as a *supertree of genes*. Reconciliation implies searching for a tree that is closest to the set of all G_i trees. There are several natural variants that refine the “closeness” analyzed below.

Algorithms for searching *contemporary* (terminal-node) horizontally transferred genes were proposed elsewhere [1, 2]. As an advancement of this work, below we propose two algorithms for searching horizontally transferred *ancestral* genes, i.e., genes corresponding to the internal nodes of a gene tree *G*. Results of the algorithms are illustrated by examples including analysis of some protein families from the database of clusters of orthologous protein groups (COGs; www.ncbi.nlm.nih.gov).

Reconciliation α of a gene tree *G* into a species tree *S* has been defined in [3, 4], which made possible the definition of characteristic *c(G, S)*, called *reconciliation cost*, which shows the difference between trees *G* and *S*.

In contrast to previous algorithms that assume that each edge has unit length, reconciliation cost in [1, 2] is determined taking into account the lengths of tree edges. These publications developed two approaches to searching horizontally transferred genes, based on the following considerations. The first approach is that exclusion of an arbitrary *contemporary* (terminal-node) gene *g* makes it possible to evaluate the degree of the irregularity of its position in a tree *G* from the change in the reconciliation cost. In this case, a numerical description F_g of such change is introduced.

The second approach, proposed in [1, 2] for genes *g*' in the vicinity of gene *g* in tree *G* (within a radius of *r* from *g*), evaluated the remoteness of species *s*', from which gene *g*' was isolated, and species *s*, from which gene *g* was isolated. Accordingly, a numerical characteristic R_g of such remoteness is introduced.

Both approaches select genes in which these characteristics sharply differ (according to the corresponding statistical test) from those of all other genes of a given COG. These approaches were described in detail elsewhere [2].

Both approaches are targeted at searching for recent horizontal transfer events, when both species (the donor and the recipient of the transferred gene) had no time to considerably diverge after the transfer moment. Another approach was proposed [5] to consider relatively old (ancestral) horizontal transfers.

We propose two extensions of the first approach with a broader understanding of the reconciliation α. The first extension develops the idea of comparing a gene tree G and a species tree S on the basis of multivalued mapping (graph) β with edges extending from each node g in G (terminal or internal) to each node in S. The second extension employs a species tree S only and pairwise distances between amino acid sequences, thus replacing the reconciliation α from [3, 4] with fuzzy sets in the tree S. We believe that an advantage of the second extension is that it does not use (at least necessarily) the algorithms for the generation of gene tree G, which are known to present certain problems.

THE ALGORITHMS AND RESULTS OF TESTING ON SYNTHETIC DATA

Let us describe these two extensions hereinafter referred to as the first and second approaches.

The first approach. Genes of a certain COG G are always dealt with. As usual, nodes of a gene tree G and of a species tree S are given some numerical identifiers. In order to simplify explanation, we assume that G and S have the same number of nodes, which are numbered consistently, e.g., by natural numbers, so that node i in the tree G corresponds to a gene isolated from species corresponding to the terminal node i in the tree S. Formally this means that exactly one gene of a certain COG corresponds to each terminal species in S. However, this limitation is irrelevant to our approaches and is not used in our algorithms.

Any *node* g of a tree (of species or genes) corresponds to *set* K of all nodes that are below node g; in this context, node g and *clade* K define each other. For convenience, below this node, g (and the corresponding gene or species) is referred to as an *ancestral* gene or species K similar to its clade for gene tree G and species tree S, respectively. Any set K_0 of terminal nodes corresponds to an *ancestor*—the last node g, the clade of which includes the set K_0. Such node g is an *ancestor* of the set K_0. Clade K of node g is a *minimal clade* including the set K_0. If the set K_0 is a clade, K evidently coincides with K_0.

The first approach developed in [1, 2] relies on a simple idea: the similarity between trees G and S is the greater, the closer to each other are all pairs in the *corresponding clades*; i.e., the closer the similarity between clade g in G and clade $\alpha(g)$ in S. The difference of this similarity from the complete coincidence is considered to be the *reconciliation cost c(G, S)* for trees G and S.

The idea of the aforementioned extension consists in searching for the smallest clade K' "almost containing" clade K (when $K\backslash K'$ is *small* or $K'\backslash K$ is *small*) instead of searching for the smallest clade K' containing clade K. *Small* $K\backslash K'$ corresponds to the case when genes from $K\backslash K' = K\backslash(K \cap K')$ were lost in the present descendants of clade K' but which evolved from the same ancestral gene K in one or more other species. In the case of small $K'\backslash K$, the genes from $K'\backslash K = K'\backslash(K \cap K')$ did not descend from the common ancestral gene K but were transferred to K' or emerged in K' later. In this case, "massive" set $K \cap K'$ plays the role of "normally developing" descendants of the ancestral gene K in descendants of the ancestral species K'. A small number of genes from the set $K\backslash K'$ or genes isolated from species $K'\backslash K$ correspond to genes with irregular position. These two sets $M = K\backslash K'$ and $M' = K'\backslash K$ will be referred to as *components* of clades K and K', respectively.

Graph β includes all clades in G and all clades in S as nodes, and each clade K in G is connected by an edge with each clade K' in S (such edges are referred to as K, K'); there are no other edges in this graph.

The ratio of the power of the component M to the power of clade K that contains it, i.e., the $\frac{|M|}{|K|}$ value, as well as the analogous value $\frac{|M'|}{|K'|}$ for clade K', is calculated for each edge K,K'. Remember that the power $|M|$ of the set M is the number of elements in it. Let us designate the *probability* of component M on edge K, K' as $1 - \frac{|M|}{|K|}$ and the *probability* of component M' as $1 - \frac{|M'|}{|K'|}$. Let us match these two probabilities (which will be referred to as *probabilities of edge K, K'*) with each edge K, K' in graph β. Edge K, K' can be regarded as an analogue of pair $\langle\alpha, g(\alpha)\rangle$ in the algorithms from [1, 2].

The set of genes M can be a clade; however, the last common ancestor M in G can have descendants not included in M. A set of such descendants M^* (which is an empty set in the first case) will be referred to as the *first partner M*. Similarly, the *second partner M** is determined from the ancestor, which is by one edge closer to the root in G than the last common ancestor M. Usually sets M and M^* are not clades.

The first algorithm analyzes the possibility of horizontal transfer between the ancestor of set *M* and the ancestor of its partner *M** in species tree *S* (Fig. 1), where *M* = {*a, b*} and its second partner *M** = {*3, 4*}.

The first algorithm is as follows. A list of all edges *K, K'* of graph β with at least one probability exceeding a certain threshold (e.g., 1/2) is compiled. Both partners *M** are considered for each nonempty component *M* corresponding to such a probability. A pair ⟨*M*, *M**⟩ will be referred to as *candidate* if three conditions are satisfied:

(1) *Proximity in candidate pair* ⟨*M*, *M**⟩, which is determined as the mean distance between elements of these two sets for gene tree *G* (if edge lengths are specified in it) or as pairwise alignments of the corresponding amino acid sequences, must be below a certain threshold.

(2) *Concentration* of the set *M* and concentration of the set *M*' must exceed a certain threshold. Concentration of the set *M* in a species tree *S* is calculated as a proportion of *M* power to the power of the node set in the subtree of the species tree, growing from the last common ancestor for all species from *M*. High concentration values are indicative of horizontal transfer.

(3) *Distance* (determined, e.g., as a number of edges) between the ancestors of all species from *M* and all species from *M** in species tree *S* must exceed a certain significant threshold. (If ancestors of the sets *M* and *M** are close in *S*, the fulfillment of conditions (1) and (2) can be attributed to the relationship between *M* and *M**.) This index is complemented by the threshold condition for the concentration of combined species from *M* and *M**: low values of it are indicative of horizontal transfer.

The greater number of edges of graph β confirms horizontal transfer between the same sets *M* and *M** and that the higher the corresponding probability, the greater weight given to pair ⟨*M*, *M**⟩ as a candidate for the horizontal transfer between the ancestors of *M* and *M**. The *probability* of a horizontal transfer between ancestors of the sets *M* and *M** is described by the value $1 - 2 \times (1 - p_1) \times (1 - p_2) \times \ldots \times (1 - p_n)$, where $p_1, p_2, \ldots, p_n$ are the probabilities of all edges that give rise to the same pair ⟨*M*, *M**⟩ in the manner indicated above.

Simulation example. Ten species *a*, *b*, *c*, *d*, *e*, *1*, *2*, *3*, *4*, and *5* and a species tree ((((*e*, *b*), *a*), ((*d*, *f*), *c*)), ((*1*, *2*), (*3*, *4*))) are given. The tree of a certain COG (gene tree) is as follows: ((((*a*, *b*), (*3*, *4*)), *e*), (((*d*, *f*), *c*), (*1*, *2*))) (Fig. 1). As usual, a node of any tree is referred to as a set of its descendant nodes. In this example, species are divided into two taxa denoted by characters and numbers, respectively. Let us trace the recognition of a horizontal transfer between these taxa (to be precise, between the last ancestors of clades {*3*, *4*} and {*a*, *b*}, i.e., of a transfer with the candidate pair $P = \langle\{3, 4\}, \{a, b\}\rangle$) by the first algorithm (and later by the second algorithm) assuming that gene *e* inherited no properties of the horizontally transferred gene. Let the threshold for the edge probability be 1/2. Any edge connecting clade {*d*, *f*, *c*} with any node as well as edges connecting clades each of which contains two or less elements will be ignored. Such edges are unreliable: in the first case, there is an edge connecting this clade with itself among all possible edges; while in the second case, the denominator in the equation for the edge probabilities is too low.

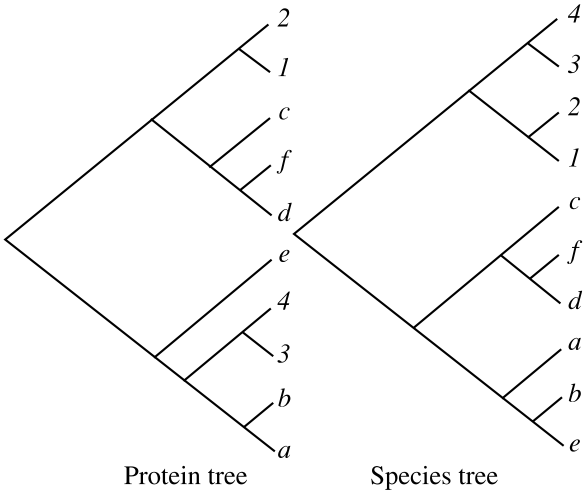

Fig. 1. Simulation of protein and species trees: terminal nodes *a, b, c, d, e, f, 1, 2, 3*, and *4* denote orthologous genes isolated from the corresponding species and these species, respectively.

Let us fix clade *K* = {*ab34*} in the tree *G* and search through all clades *K*' in the tree *S*.

(1) The edge connecting clades {(*ab34*), (*eba*)} with a probability of 1/2 points to component *M* = {*3*, *4*} (which is its second partner *M** = {*a*, *b*}) and with a probability of 2/3 points to component *M* = {*e*}. Partners *M** for this *M* are empty sets or set {*a*, *b*, *3*, *4*}. The second partner is rejected due to the low concentration (while the proximity in the candidate pair exceeds the threshold). The first partner is an empty set and can correspond to a gene transfer with no conservation in the source or to a transfer from a species not represented in the species tree.

(2) The edge connecting clades {(*ab34*), (*ebadfc*)} points again to component *M* = {*3*, *4*} with a probability of 1/2.

(3) The edge connecting clades {(*ab34*), (*34*)} with a probability of 1/2 points to component *M* = {*a*, *b*}, which represents the same candidate pair *P*.

(4) The edge connecting clades {(*ab34*), (*1234*)} points again to component *M* = {*a*, *b*} with a probability of 1/2 and to component *M* = {*1*, *2*} with the same probability. The partners for *M* are empty sets or {*d*, *f*, *c*}

These variants are possible, although the latter does not satisfy the candidate pair proximity requirement.

Let us move to another clade K and search through clades K'.

(5) The edge connecting clades {(*ab34e*), (*eba*)} points to component M = {*3, 4*} with a probability of 3/5.

(6) The edge connecting clades {(*ab34e*), (*ebadfc*)} points to component M = {3, 4} with a probability of 3/5 and to component M = {*d, f, c*} with a probability of 1/2 (the partners for M are empty sets and {*1, 2*}).

(7) The edge connecting clades {(*ab34e*), (*1234*)} points to component M = {*1*, *2*} with a probability of 1/2.

Let us move to another clade K and search through clades K'.

(8) The edge connecting clades {(*dfc12*), (*ebadfc*)} points to component M = {*1*, *2*} with a probability of 3/5 and to component M = {*e*, *b*, *a*} with a probability of 1/2. The first partner for the second component M is {*3, 4*} so that the last variant cannot be excluded (it corresponds to the case when gene *e* descends from a horizontally transferred gene but has lost considerable similarity with it).

(9) The edge connecting clades {(*dfc12*), (*1234*)} points to component M = {*3*, *4*} with a probability of 1/2.

Let us move to another clade K and search through clades K'.

(10) The edge connecting clades {(*12*), (*1234*)} points to component M = {*3*, *4*} with a probability of 1/2. And so forth.

Thus, eight evidence pairs pointed to candidate pair P (six and two of them with a probability of 1/2 and 3/5, respectively), i.e., to an evolutionary event between the ancestors of the sets {*a, b*} and {*3*, *4*} (to be precise, between edges coming to them from the root). The algorithm does not specify which of two directions was realized in this evolutionary event. Pair P is printed, and so forth.

A computer program realizing this approach was tested on biological data (for examples, see below). The testing suggested the following main parameters of the program: edge probability threshold, 0.6; evidence number, around 5. These algorithms were tested on 132 COGs described in detail elsewhere [1, 2]. Analysis of this test is intended for a separate publication.

Second approach. Let us now describe the second approach. For brevity, let us consider the corresponding algorithm for horizontal transfers with the maintenance of the copy in the source. The basic concepts of the fuzzy set theory are used here. This theory provides the rules for the situation when the "certainty $p_g(P)$ of g occurrence in P" is defined for each gene g of a COG and for each set P of genes of the same COG. This certainty is not related to the significance of statistical hypotheses. As usual, the proper certainty (and probability) cannot be formally calculated. Here, we define the certainty of g occurrence in P as the degree of similarity between g and the most similar gene from P. Hereafter, p_g will be used instead of $p_g(P)$.

Let us recall the definition for our situation. A *fuzzy set of genes R* is defined by the specification of a *certainty function*, which gives the certainty p_g of a gene occurrence in R for each gene g of a given COG. A fuzzy set can be conveniently presented as a column (or a row) of numbers from 0 to 100%, the *i*th element of which coincides with the number of gene g (we assume that all genes of the initial COG are numbered or the number of each gene coincides with its identifier g). The fuzzy set theory develops certain rules for operating such columns. Let there be given clade K in a species tree S and a set P of all genes of a given COG isolated from species included in K. We want to generate a *fuzzy set of genes R* from the set of genes P, which requires specification of the aforementioned column p_g. Let us take the simplest variant with the gth member of this column proportional to the similarity between gene g not included in P and the most similar gene g_1 from P. Pairwise distances calculated using the standard method (see Discussion, stage 1), distances in the gene tree G, the proportion of information on gene g presented in set P (the two latter values can be calculated using the Lempel–Ziv algorithm or as indicated in the Discussion of Algorithms, stage 1), etc., can be used as a measure of similarity. In particular, the similarity can be defined using the COG multiple alignment. If gene g is included in P, here we assume the highest possible similarity (100%) for it.

Let us generate a pair of fuzzy sets of genes R and R' corresponding to a pair of nonoverlapping clades K and K' in a species tree S (and the corresponding sets of genes P and P'), i.e., calculate columns p_g and q_g by one of the above-mentioned methods.

Let

$$Q(K, K') = \frac{\sum_g \min(p_g, q_g)}{\sum_g \max(p_g, q_g)}.$$

be the *quality* of their K, K'.

The core M of the initial clades K and K' includes a set of genes g of the initial COG for which $\min(p_g, q_g)$ exceeds a threshold. These genes can be regarded as slightly diverged descendants of a certain, e.g., horizontally transferred, ancestral gene.

The considered algorithm searches for a horizontal transfer through pairs of nonoverlapping clades K and

K' in a species tree S with the quality above a threshold. In addition, two sets of genes (M_1 and M_2) isolated from the core M for these pairs should have a sufficient concentration in tree S, while their combination should be very similar to M and have low concentration in tree S. The algorithm assumes that such pair of clades $\langle K, K'\rangle$ indicates a possible horizontal transfer between the ancestors of sets M_1 and M_2 in the species tree. Horizontal transfer between the ancestors of sets M_1 and M_2 is considered the more probable, the greater the number of clades K and K' of high quality point to the same set pair $\langle M_1, M_2\rangle$.

The probability of a horizontal transfer between the ancestors of species from M_1 and M_2 can be described by the value $1 - 2 \times (1 - p_1) \times (1 - p_2) \times \ldots \times (1 - p_n)$, where $p_1, p_2, \ldots, p_n$ are the qualities of all clade pairs $\langle K, K'\rangle$ that point to these M_1 and M_2 as indicated above.

To illustrate how the second algorithm operates, we assume that the above simulation example has the following matrix of distances between COG genes: $\rho(a, b) = \rho(3, 4) = 1$; $\rho(a, 4) = \rho(a, 3) = \rho(b, 3) = \rho(b, 4) = 1.5$; $\rho(e, a|b|3|4) = 2.5$; $\rho(e, 1|2|c|d|f) = 3.5$; $\rho(a|b|3|4, 1|2|c|d|f) = 4$; $\rho(d, f) = \rho(1, 2) = 1$; $\rho(c, d|f) = 1.5$; $\rho(c|d|f, 1|2) = 2.5$ (vertical line states for "or"). The distance matrix can include any numbers roughly corresponding to the distances in the COG tree shown in Fig. 1. Let $\rho_0 = \rho(g, P)$ be the distance between gene g and the nearest gene g_1 from the set P. Considering the aforementioned proportionality in this example, we assume that the degree of g occurrence in P is calculated from equation $p_g = h(\rho_0)$, where $h(x) = 32 \times (4 - x)$. For instance, the columns p_g and q_g corresponding to sets $P = \{e, b, a\}$ and $P' = \{3, 4\}$ are shown below as two lines:

g	a	B	3	4	e	c	D	f	1	2
$p_g^{\%}$	100	100	80	80	100	16	16	16	16	16
$q_g^{\%}$	80	80	100	100	48	0	0	0	0	0

Simulation example (continued). Let us identify the origin of genes *a*, *b*, *3*, and *4*. Columns p_g and q_g corresponding to these two clades $K = \{e, b, a\}$ and $K' = \{3, 4\}$ in tree species S are calculated. As a result, R "includes" genes *a, e*, and *b* with 100% certainty by definition; genes *3* and *4*, with 80% certainty; and genes *1, 2, c, d*, and *f*, with 16% certainty. R' "includes" genes *3* and *4* with 100% certainty by definition; genes *a* and *b*, with 80% certainty; and gene *e*, with 48% certainty. This is a verbal translation of columns p_g and q_g for R and R' from the previous table considering that $P = K$ and $P' = K'$ in this example.

Let us employ two concepts from the fuzzy set theory. *Fuzzy intersection* of R and R' is described by a column including the minima of p_g and q_g for each COG g; i.e., a column of a new fuzzy set of genes denoted as $R \cap R'$ holds $\min(p_g, q_g)$ in the ith position. In this example, the fuzzy intersection of R and R' "includes" genes *a, b, 3*, and *4* with 80% certainty and gene *e* with 48% certainty. *Fuzzy union* of R and R' is specified by a column including the maxima of p_g and q_g for each COG g; i.e., a column of another fuzzy set of genes denoted as $R \cup R'$ holds $\max(p_g, q_g)$ in the ith position. In this example, the fuzzy intersection of R and R' "includes" genes *a, b, 3, 4*, and *e* with 100% certainty and genes *1, 2, c, d*, and *f* with 16% certainty. In this case, the quality $Q(K, K')$ of the initial clade pair $\langle K, K'\rangle$ is 0.634, which is higher than the adopted threshold of 0.6.

Consideration of another pair of clades $K = \{e, b\}$ and $K' = \{3, 4\}$ by the algorithm yields a slightly higher (better) quality of 0.657, which is due to a smaller fuzzy intersection of R and R'. However, the core $M = \{a, b, 3, 4\}$ is the same. The algorithm finds 14 clade pairs with a quality exceeding the threshold that correspond to the same core M. Thus, our algorithm points to the set $M = \{a, b, 3, 4\}$ in the species tree.

After finding M, the algorithm checks that the set M has *low* concentration in a species tree S but splits into two parts $M_1 = \{a, b\}$ and $M_2 = \{3, 4\}$, each of which has *high* concentration (the algorithm uses two thresholds). It follows that genes *a, b, 3*, and *4* are descendants of the ancestral gene that was horizontally transferred between the ancestors of species sets M_1 and M_2. The algorithm also specifies that gene *e* included in the fuzzy intersection with 48% certainty can be a considerably diverged descendant of a transferred gene. The algorithm makes the decision about this gene according to the threshold values.

Thus, we see that both approaches yield similar results in the simulation example, which counts in their favor. Our testing (and even this example) shows that many pairs of clades K and K' indicate a transfer between ancestors of the same sets M_1 and M_2; i.e., the decision of the algorithm is supported statistically.

TESTING ON BIOLOGICAL DATA

Further tests of the algorithm proceeded in biological COG databases. Below, the results of calculations for three COGs are shown. Initial data included trees and amino acid sequences mentioned in [2] (Initial Data section). The species tree is shown in Fig. 2, where full species names are given.

1. COG 180 (tryptophanyl-tRNA synthetase). The corresponding tree is shown in Fig. 3. The first program proposed sets M_1 = {Bha, Bsu, Sau} and M_2 = {Vch, Ech, Buc, Hin, Pmu} as candidates for horizontal transfer (the second set is a partner of the first one). The transfer is supported by six edges; the concentration values also confirm it.

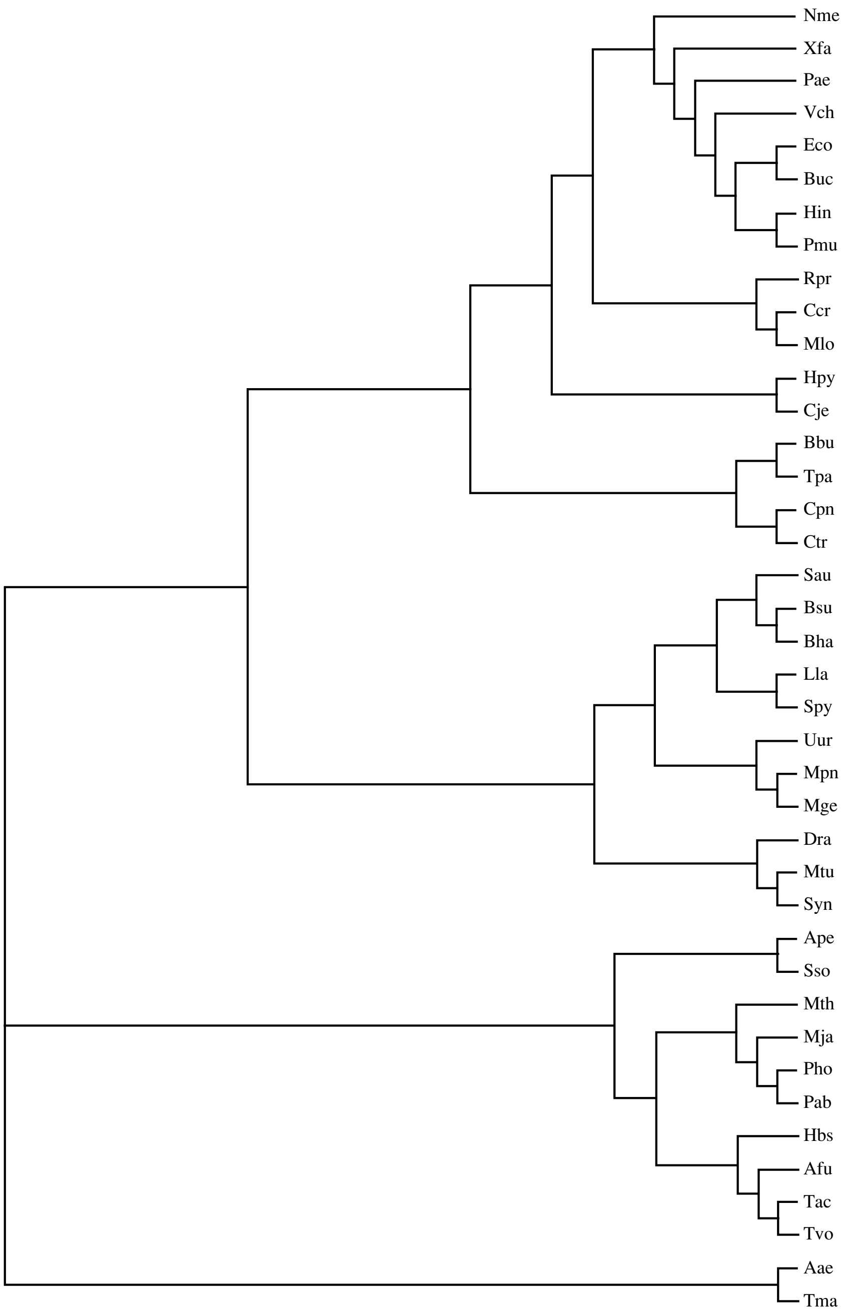

Fig. 2. Species tree of microorganisms. Archaea: Afu, *Archaeoglobus fulgidus*; Hbs, *Halobacterium* sp. NRC-1; Mja, *Methanococcus jannaschii*; Mth, *Methanobacterium thermoautotrophicum*; Tac, *Thermoplasma acidophilum*; Tvo, *Thermoplasma volcanium*; Pho, *Pyrococcus horikoshii*; Pab, *Pyrococcus abyssi*; Ape, *Aeropyrum pernix*; Sso, *Sulfolobus solfataricus*; gram-positive bacteria: Spy, *Streptococcus pyogenes*; Bsu, *Bacillus subtilis*; Bha, *Bacillus halodurans*; Lla, *Lactococcus lactis*; Sau, *Staphylococcus aureus*; Uur, *Ureaplasma urealyticum*; Mpn, *Mycoplasma pneumoniae*; Mge, *Mycoplasma genitalium*; alpha-proteobacteria: Mlo, *Mesorhizobium loti*; Ccr, *Caulobacter crescentus*; Rpr, *Rickettsia prowazekii*; beta-proteobacteria: Nme, *Neisseria meningitidis* MC58; gamma-proteobacteria: Eco, *Escherichia coli*; Buc, *Buchnera* sp.; Pae, *Pseudomonas aeruginosa*; Vch, *Vibrio cholerae*; Hin, *Haemophilus influenzae*; Pmu, *Pasteurella multocida*; Xfa, *Xylella fastidiosa*; epsilon-proteobacteria: Hpy, *Helicobacter pylori*; Cje, *Campylobacter jejuni*; chlamydia: Ctr, *Chlamydia trachomatis*; Cpn, *Chlamydia pneumoniae*; spirochetes: Tpa, *Treponema pallidum*; Bbu, *Borrelia burgdorferi*; DMS group: Dra, *Deinococcus radiodurans*; Mtu, *Mycobacterium tuberculosis*; Syn, *Synechocystis*; Aae, *Aquifex aeolicus*; Tma, *Thermotoga maritima.*

Second program. Six clade pairs with the highest quality in the species tree share the same core M = {Bha, Bsu, Sau, Vch, Eco, Buc, Hin, Pmu, Hpy, Mtu}. During clustering, the algorithm abandoned two latter species and split the remaining set into two clusters M_1 and M_2 identical to those proposed by the first pro-

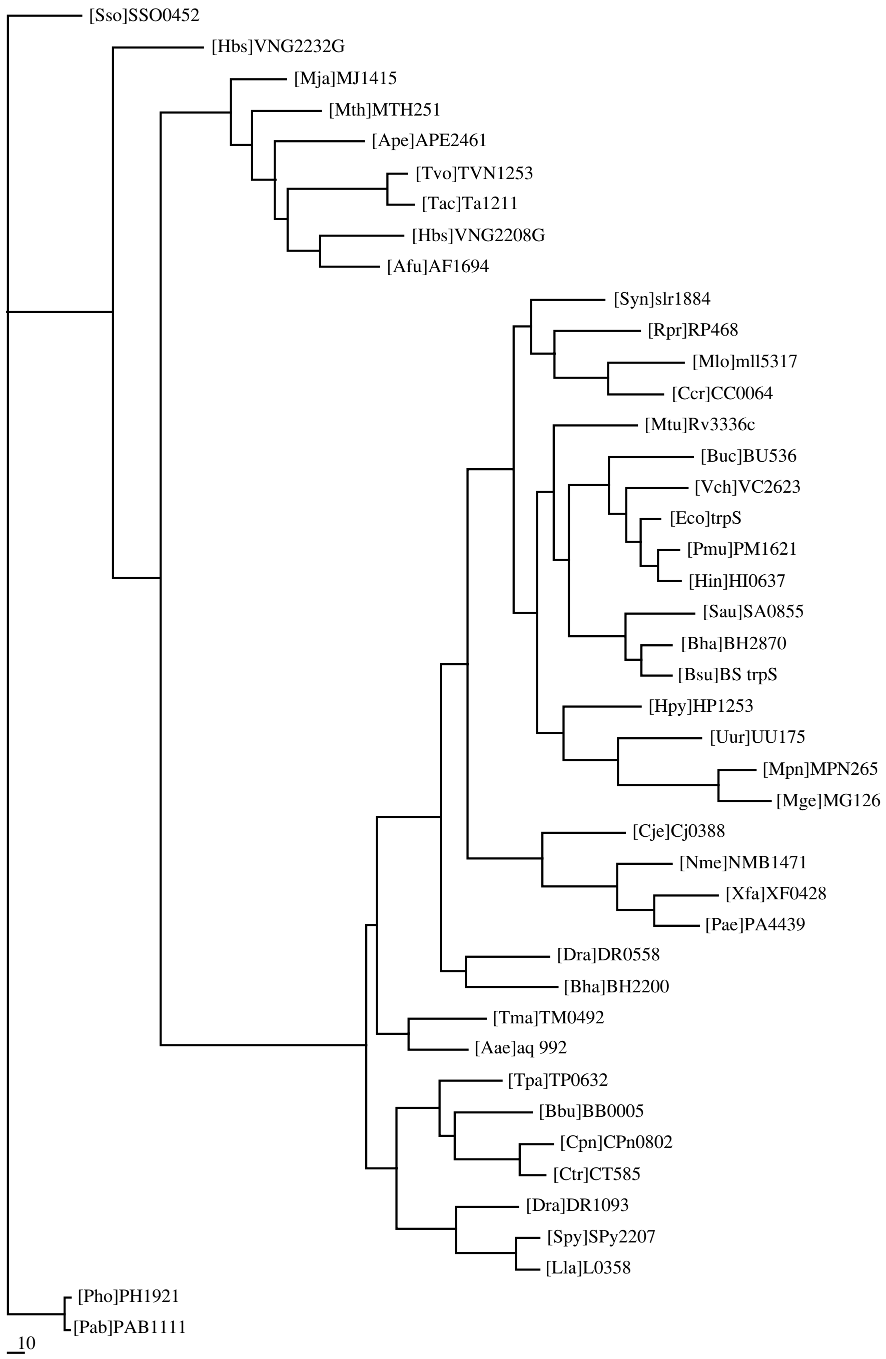

Fig. 3. Tree of COG 180 (tryptophanyl-tRNA synthetase).

gram. The algorithm suggests that the presence of Hpy and Mtu species in the intersection can be due to additional evolutionary events at the molecular level.

2. COG 171 (NAD synthase). The corresponding tree is shown in Fig. 4. The quality of this COG is notably lower as compared to the previous one. Accordingly, the indications of the edges identified by the first program are more diverse. Many of them point to the horizontal transfer between two species groups identified for the previous COG: 16 edges support gene transfer between the ancestors of sets {Sau} and {Buc, Eco}. Two edges point to gene transfer between the ancestors of sets {Sau, Bsu, Bha, Lla, Spy, Dra} and {Pae, Vch, Eco, Buc} (the gene tree does not include species Pmu and Hin). One edge points to gene transfer between the ancestors of sets {Bha, Bsu, Sau, Lla, Spy} and {Vch, Ech, Buc}. Finally, a large group of edges point to gene transfer between the ancestors of {Ccr} and {Mtu}. Such vari-

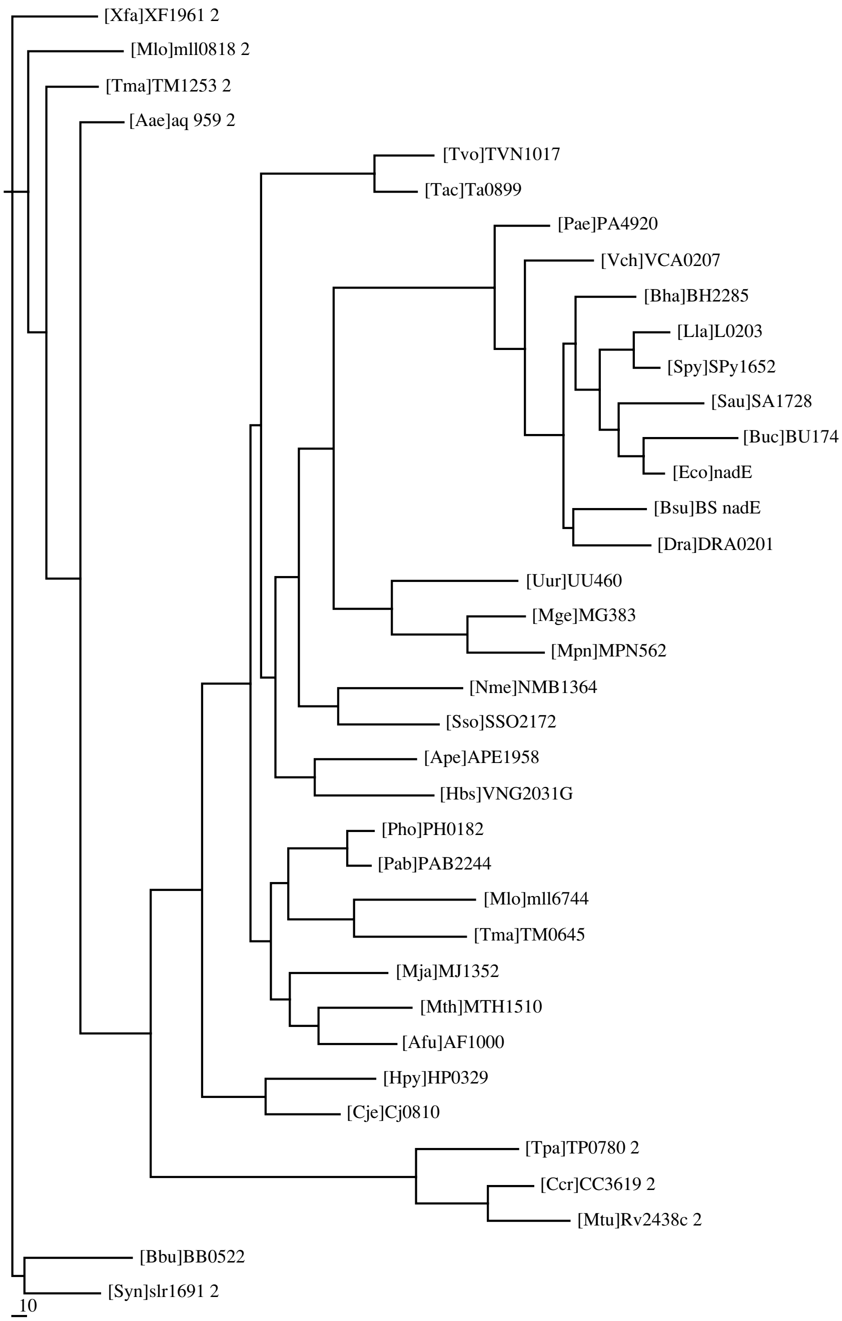

Fig. 4. Tree of COG 171 (NAD synthase).

ation suggests that several evolutionary events could take place.

Second program. The ten best clade pairs share the same core M = {Buc, Eco, Vch, Pae, Dra, Bha, Bsu, Sau, Spy, Lla}. The algorithm splits it into two natural clusters M_1 = {Buc, Eco, Vch, Pae} and M_2 = Bha, Bsu, Sau, Spy, Lla}. The algorithm suggests that horizontal transfer took place between the ancestors of these sets, while a gene transfer from Dra to one of them took place later. Another clade pair shares the core {Ccr, Mtu} and the algorithm indicates possible gene transfer between these species.

3. COG 525 (valyl-tRNA synthetase). A horizontal gene transfer from a group of archaea to Rpr was mentioned in [2], which was confirmed by our programs as well. The first program supported it with more than ten edges. However, the transferred gene has lost high similarity to the archaeal genes (in particular, it follows from the gene tree; Fig. 5). Hence, this transfer can hardly be detected by the second program. At the

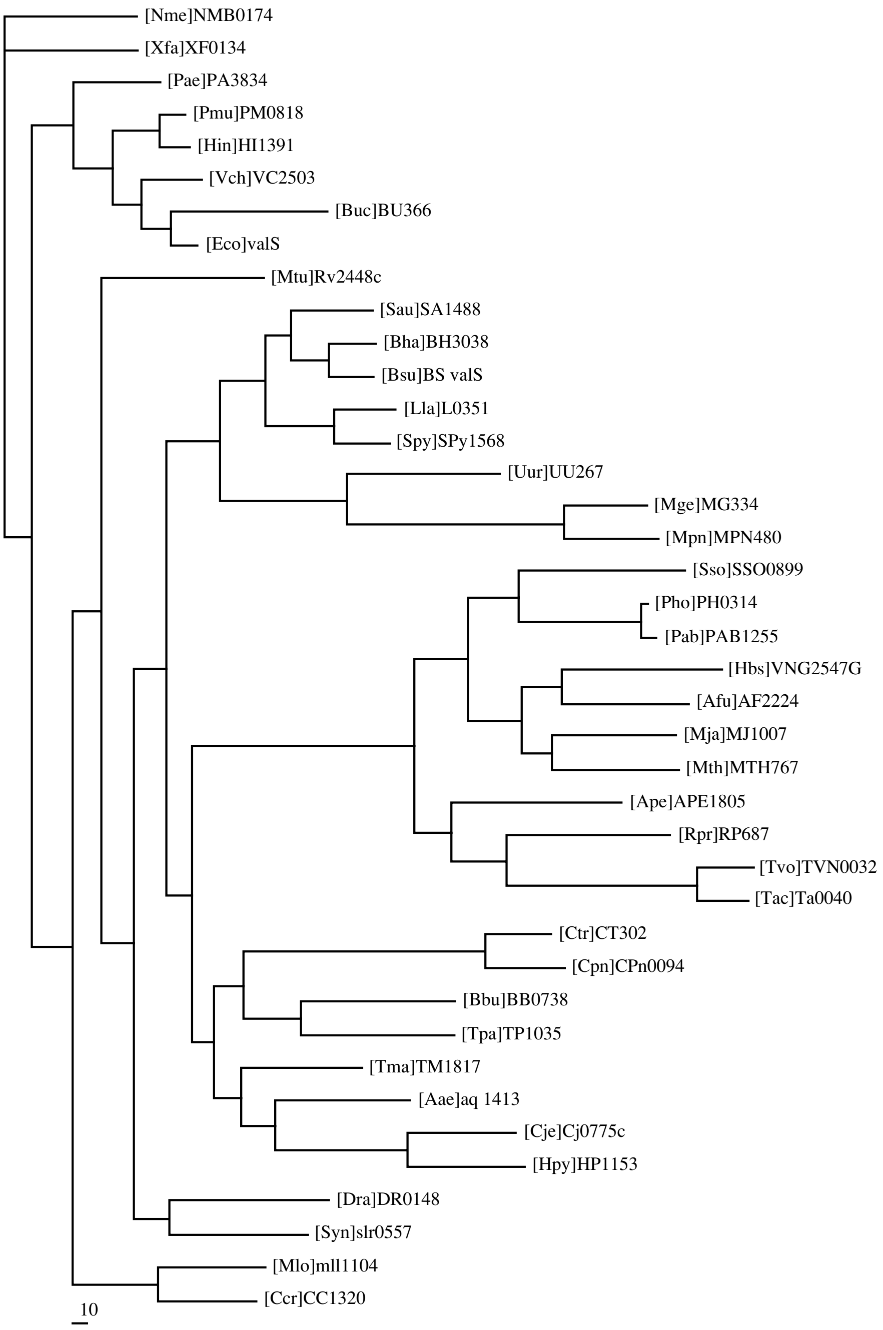

Fig. 5. Tree of COG 525 (valyl-tRNA synthetase).

same time, both archaeal and Rpr genes considerably differ from all other COG genes, which allows the following technique to be applied. The second program is executed with two low threshold values. The first threshold determines the certainty of gene inclusion into a fuzzy clade (which is considered as zero if less than the threshold), while the second threshold determines the certainty of gene inclusion into the intersection of two fuzzy clades. These parameters, *porog_prin* and *porog_per*, are considered in detail in the Discussion (stage 1 and stage 2, respectively). In the case of this COG, they were taken equal to 0.3. More than ten pairs of clades with the best quality support the same large core including the complement {Rpr, archaea} of the set of all COG genes. Processing such large sets offers only potential gene transfers close to the root, which are of no interest, while processing of the {Rpr, archaea} reveals the aforementioned transfer.

This technique also allows gene transfers to be revealed in the absence of the copy in the source.

Indeed, set M of descendants of a transferred gene is unlikely to share significant similarity to the genes isolated from close species in the species tree. Accordingly, at low *porog_prin* and *porog_per* values, it looks like a complement (or "almost" a complement) of a certain large subtree and is recognized by the second program as a core for a certain clade pair.

DISCUSSION OF ALGORITHMS

Let us begin with informal considerations of the details of realization and justification of the first approach. Let node K (or K') be taken as silent if an edge extends from it in graph β with the probabilities $\langle 1, 1\rangle$ and leads to a node corresponding to exactly the same clade. Indications of an edge emanating from a silent node are ignored, since there is a clearly preferable edge linking the silent node to itself in another tree in this case. Such considerations allow us to reduce the number of edges considered by this algorithm.

All descendants of a horizontally transferred gene both in the recipient and donor species have good chances to appear in the same component for a certain edge K, K' in graph β (this *statement* is substantiated in the following paragraph and thereafter). This can be component M of clade K in the gene tree G or an analogous component M' of clade K' in the species tree S; in the second case, these genes were isolated from species in M'. Remember that the designations of the genes and the corresponding species are the same. Then, the first algorithm can reveal a horizontal transfer through analysis of such a component.

Horizontal transfer occurred and the source copy was preserved. In this case, descendants of the source gene and descendants of the transferred gene share common traits. It is clear that some descendant genes could be lost; for instance, some descendants of the donor and recipient species could reject it and maintain the "native" gene. It is also natural to assume that the groups M_1 and M_2 of descendants that conserved the common gene will be represented by a single joint clade $K = M_1 \cup M_2$ in the gene tree. Let us consider the smallest clade K' including the set M_1. Let $K' = M_1 \cup M_1^*$ and M_1^* do not intersect with M_2. Then $K \backslash K' = M_2$, and pair $\langle K, K'\rangle$ is an edge supporting horizontal transfer between the ancestor of the set M_2 and the ancestor of (its partner) set M_1. This horizontal transfer is supported by other edges.

Horizontal transfer occurred and the source copy was lost. In this case, group M_1 or M_2 is empty in the gene tree (e.g., M_1) and the first partner of set $M = M_2$ is empty in the gene tree. However, the algorithm can reveal horizontal transfer in this case as well, provided its source and target are distant in the species tree. Indeed, let us consider the smallest clade $M_2 \cup M_2^*$ in a species tree containing set M_2 (to which a gene was transferred). Let K' be the immediate ancestor (parent) of node $M_2 \cup M_2^*$ and v_1 be another immediate descendant (child) of node K'. Let us consider clade V_1 including the set v_1 and its parent K in a gene tree, where V_2 is a child of node K not including v_1. Genes of the sets M_2 and V_1 have essentially different origins in the species tree, while genes of the sets M_2^* and V_1 have a similar origin. Therefore, V_2 does not intersect with M_2 and K includes M_2^*. Then $K' \backslash K = M_2$ i.e., edge K, K' supports a transfer to the set M_2. Similarly, an edge linking a parent of the M_2 ancestor in G (let V be his other child) to a parent of the V ancestor in S supports such transfer.

Let us consider a possible realization of the *second algorithm* in detail. It includes three stages.

Stage 1. Clade K transformation into fuzzy set R, i.e., calculation of column p_g. Let clade K be fixed in a species tree S and P be a set of genes of the fixed COG G isolated from species in K. The *proportion of information* $d(g, K)$ (in total information contained in g) is calculated for each gene g of COG G not included in P. This proportion is taken as unity for the genes included in P. Each gene g_1 of P is aligned with gene g (e.g., by the Smith–Waterman algorithm using the amino acid substitution matrix and other parameters as in http://www.zbh.uni-hamburg.de/research/BM/torda/sub_mat). For each position i of gene g and each i-containing block b of length *dlina* (a parameter of the algorithm), the *value* $q_g^K(i, g_1, b)$ is calculated. This value is zero if block b was aligned with at least one deletion (in g or g_1), otherwise it equals the sum of pairwise distances between amino acids in aligned blocks b and b_1. The quality $q_g^K(i)$ of position i equals the highest $q_g^K(i, g_1, b)$ for all pairs $\langle g_1, b\rangle$ in which b contains i and g_1 takes on all values from P. If this quality $q_g^K(i)$ is below the threshold *porog_blok* (a parameter of the algorithm), we take $q_g^K(i)$ as zero. Let $S(g, P)$ be the sum of quantities $q_g^K(i)$ for all positions i of gene g. It is divided by the maximum achievable value $S_{\max}(g)$ of the function $S(g, P)$ for any gene g. Naturally, the maximum value is reached for the alignment of identical sequences; i.e., $S_{\max}(g)$ equals the sum of $q_g^{\{g\}}(i)$ for all i. Thus, $p_g = d(g, K) = S(g, K)/S_{\max}(g)$.

If $d(g, P)$ is less than *porog_prin* (a parameter of the algorithm), we take it as zero. For instance, we defined the following parameters: *dlina* = 10, *porog_blok* = 30, and *porog_prin* = 0.5.

In the cases when a good multiple alignment or a tree of the initial COG was available, pairwise alignments induced by this multiple alignment were used or the degree of similarity was calculated from the distances in the gene tree. Qualities of clades K and K' determined on this basis appear to be more biologically adequate.

Stage 2. Calculation of the quality of the initial pair of clades K and K' in a species tree. Here, we derive the same equation for the quality on the basis of the fuzzy set theory and the powers of fuzzy intersection and fuzzy union. For two fuzzy sets R and R' generated for two initial nonoverlapping clades K and K', the quality $Q(K,K')$ is calculated as the proportion of the power of the fuzzy intersection of fuzzy sets R and R' to the power of their fuzzy union. The power of fuzzy intersection is the sum of minima of certainty that elements are included in R and R' ($\min(p_g, q_g)$), while the power of fuzzy union is the sum of the corresponding maxima ($\max(p_g, q_g)$). This quality is combined with a bonus P for the transfer remoteness in time, where P equals the parameter *drevn_priz* multiplied by the difference between the mean power of gene sets P and P', and unity (corresponding to two single-element sets). For instance, *drevn_priz* = 0.01. Transfers remote from the root are forbidden at other steps of the algorithm.

Stage 3. Processing of results. For pairs of initial clades K and K' from S with $Q(K, K')$ no less than *porog_pary* (e.g. 0.7), the genes are recognized in the fuzzy intersection of R and R', for which inclusion into this intersection is no less than *porog_per* (e.g., 0.5). These genes constitute the core M. The clustering program is used to try to recognize subsets M_1 and M_2 of the core M, which demonstrate their high individual concentration (in tree S) and low concentration of their combination (all four tests described for the first approach are applicable if M_1 and M_2 are considered as partners) and for which the union of M_1 and M_2 almost equals M ("almost" is defined by another parameter of the algorithm). If such M_1 and M_2 are found, a message is generated about potential horizontal transfer between the ancestors of sets M_1 and M_2.

CONCLUSIONS

The problem of reconstruction of evolutionary events at the molecular level seems topical. The first step in its solution requires algorithms that automatically search the nodes in a protein family and species trees responsible for significant mismatch between these trees. Two new algorithms were proposed for the automatic search of such *internal* nodes, which correspond to *ancestral* genes and species. In our previous studies [1, 2], a similar problem was solved largely for the *terminal* nodes of these trees, which correspond to *contemporary* genes and species.

The proposed algorithms rely on mathematically different approaches, but their testing yielded consistent results. The algorithms were tested in many ways on the same simulation and natural data: the results were presented for a single simulation tree of protein sequences (Fig. 1) and for trees of three COGs (Figs. 3–5). In the case of simulation, the algorithm recognized all nodes responsible for the tree incongruence, and no other nodes. In the case of COGs, the nodes recognized by the algorithm also agree with the results of biological analysis, e.g., [1, 2].

ACKNOWLEDGMENTS

We are grateful to L.I. Rubanov for his help in this study and valuable discussions, and to the reviewer for beneficial comments on language style.

REFERENCES

1. V'yugin, V.V., Gelfand, M.S., Lyubetsky, V.A. 2002. Tree reconciliation: Reconstructing the evolution of species by phylogenetic gene trees. *Mol. Biol.* **36**, 650–658.
2. V'yugin, V.V., Gelfand, M.S., Lyubetsky, V.A. 2003. Identification of horizontal gene transfer from phylogenetic gene trees. *Mol. Biol.* **37**, 674–687.
3. Eulenstein O., Vingron M. 1995. *On the equivalence of two tree mapping measures. Arbeitspapiere der GMD*, vol. 936. Bonn: Bonn Univ. Publ.
4. Guigo, R., Muchnik, I., Smith, T.F. 1996. Reconstruction of ancient molecular phylogeny. *Mol. Phyl. Evol.* **6**, 189–213.
5. Lyubetsky V.A., V'yugin V.V. 2004. Measuring the dissimilarity between gene and species trees, the quality of a COG. *The Fourth International Conference on Bioinformatics of Genome Regulation and Structure*, Novosibirsk: Inst Tsitol. Genet. Ross. Akad. Nauk, vol. 2, pp. 281–284.