ISSN 0026-8933, Molecular Biology, 2009, Vol. 43, No. 5, pp. 881–893. © Pleiades Publishing, Inc., 2009.
Original Russian Text © K.Yu. Gorbunov, V.A. Lyubetsky, 2009, published in Molekulyarnaya Biologiya, 2009, Vol. 43, No. 5, pp. 946–958.

UDC 575.852

Reconstructing the Evolution of Genes along the Species Tree

K. Yu. Gorbunov and V. A. Lyubetsky

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

Received December 9, 2008
Accepted for publication January 20, 2009

Abstract—A model and algorithm are proposed to infer the evolution of a gene family described by the corresponding gene tree, with respect to the species evolution described by the corresponding species tree. The model describes the evolution using the new concept of a nested tree. The algorithm performance is illustrated by the example of several orthologous protein groups. The considered evolutionary events are speciation, gene duplication and loss, and horizontal gene transfer retaining the original gene copy. The transfer event with the loss of the original gene copy is considered as a combination of gene transfer and loss. The model maps each evolutionary event onto the species phylogeny.

DOI: 10.1134/S0026893309050197

Key words: evolution along species tree, gene tree nesting into species tree, gene duplication, gene loss, horizontal gene transfer

INTRODUCTION

The issue of molecular evolution of species and taxa as well as of genes and proteins has a long-term history reflected in numerous publications and internet sites, for example, [1–3]. Voluminous literature describes the construction of phylogenetic trees for protein families (see [4–8] and references herein). The species tree is constructed by one of the two methods—either based on the sequences of many genes concatenated into one alignment or as a supertree. The latter means that a generalized tree is searched for over a specified set of gene trees so that it would, in a certain sense, best fit each of these trees, and the found tree is taken as the species tree. For example, the species tree is searched for as the tree minimizing the number of gene duplications for a certain correspondence of each gene tree to the sought species tree or as the tree minimizing the total number of gene duplications and losses [9]. The tree can be searched for based on the minimization of losses only [10] or as the tree "containing," in a way, all the given gene trees [11]. Usually, the concept that one tree is contained in another or corresponds to another means that the first tree is *nested* into the second one: the first is the gene tree *G* and the second is the species tree *S*. The classical definition of a nesting of tree *G* into tree *S*, traditionally designated as α [12], gives the information about the numbers and sites of duplications, gaps, and losses and contains the definition of the *cost* for the nesting α. The correlation between the numbers of unilateral duplications, gaps, and losses is determined with the help of α nesting [13].

The α nesting is also used for finding the horizontal gene transfers [3]. The horizontal gene transfers, both ancient and modern, being among the most popular research issues, are searched for using statistical criteria [3], differential stochastic equations [14], and fuzzy sets [15]. The determination of genes or signal profiles in the internal nodes of a species tree gives the information about its evolutionarily significant branches [16]. The expansion of the problem of searching for the genes horizontally transferred into the species tree leaves by considering and searching for the genes horizontally transferred into ancestral nodes of the species tree is considered.

Note that the horizontal gene transfers were earlier taken into account at the biological level without any model [17]. The idea of nesting a gene tree into a species tree was proposed later [18], as well as the problem of taking into account the horizontal gene transfers; however, neither definitions nor algorithms for the construction of such nesting was proposed, to say nothing about the accounting of gene transfers.

Another popular scope of problems is the reconstruction of evolutionary events in a nucleotide sequence: substitutions of letters, insertions, and deletions [19]. A separate group of problems is connected with the reconstruction of evolutionary events for an overall gene family (more precisely, clusters of orthologous protein groups, COGs), namely, gene duplications, losses and gains, and horizontal gene transfers [3, 15], and, finally, the reconstruction of ancestral regulatory sites or their characteristics from a set of

the extant regulatory sites or their characteristics along a given evolutionary tree [16, 21].

Statistical characteristics, such as the numbers of duplications, losses, or horizontal gene transfers, in one genome or one subtree inherited from a node of the evolutionary tree, are also of interest [3].

Another topic is the comparison of various scenarios, for example, permission of horizontal gene transfers versus their prohibition (then duplications and losses play their role). It is possible, thus, to determine the mean number of duplications and losses per one horizontal transfer [3]. Other problem statements connected with the comparison of scenarios are also studied [20].

Molecular evolution of the regulatory systems themselves has recently excited great interest; here, various types of regulation are considered: the regulations based on DNA–protein interactions [21, 22], secondary mRNA structure [19, 23–25], competition of RNA polymerases transcribing the common locus from complementary DNA strands [26, 27, unpublished data of the authors], and various posttranscriptional and posttranslational regulations [23, 28].

We consider below a new (at least, as far as we know) statement of the problem where *the gene tree G and species tree S are given* and the correspondence between the events that took place in the gene family from *G* and the evolutionary events represented in *S* is searched for. These events are matched with the help of an *inner* tree *G′*, defined below.

PROBLEM STATEMENT

Evolution consists in speciation as well as a duplication of genes, sometimes numerous, in the species genome, thereby producing a large number of paralogs, which, then, independently evolve. This evolution leads to divergence of gene nucleotide sequences from both one another and the sequence of their common ancestor; changes in the position of gene in the genome, its specificity, and regulatory mechanism; nonorthologous gene substitution; gene loss; and horizontal transfer. Here we concentrate on taking into account the horizontal gene transfers. The applied algorithm belongs to the class of methods of weighted maximum parsimony. Find below the precise definitions.

We want to describe a model and an algorithm that would allow us to study some aspects of development of a gene family described by the tree *G* within a family of species describe by the tree *S*. This model will be based on considering an auxiliary tree, *G′*, which we name the *inner* tree.

All the considered trees are *rooted* and *growing downwards*; in addition, an edge coming upwards is attached to the trees *G* and *S*. This edge is named *root edge* and the new highest node, the *superroot* (hereinafter, we use only the superroot of *G*). The set of edges in the species tree *S* is partitioned into the fragments named *time layers*. It is assumed that a gene can be horizontally transferred from the edge, *a*, of tree *S* to the *incomparable* edge, *b*, of this tree. Naming it incomparable, we mean that both edges, *a* and *b*, are located in one time layer and it is impossible to get from *a* to *b* along the tree *S* moving only in one direction—from the root or to the root. If the layers are not prespecified, the model and algorithm proposed below are also applicable and work in the assumption that there is only one time layer covering the overall tree. The section on the algorithms, paragraph (*c*), describes the algorithm for partitioning the edges into time layers; this is of a preliminary character, because it is based on the assumptions with a vague biological status. The question on specifying time layers in the tree *S* from a biological standpoint is beyond the goals of this work.

Let each leaf of the tree *G* be marked with the name of a gene and each leaf of the tree *S* be named with the name of a species and a set of gene names taken from this species; this set also can be empty. The trees *G* and *S* are binary; the auxiliary tree *G*' is also binary and has *crosses* at some of its leaves, which means a loss of the gene ascribed to this leaf.

Now we define this tree *G*' (named *inner* tree), which, in the case when the horizontal gene transfers are not considered, describes, in another manner, the above mentioned nesting α. Then, the transfers will be also considered by complicating the definition of inner tree *G′*, and, in this case, it will not reduce to nesting α (the nesting α will be defined below; it is equivalent to the original nesting, which we do not recall here). Examples of artificial inner trees (illustrating the definition) without and with horizontal gene transfers are shown in Fig. 1.

Imagine the edges of tree *S* as hollow ducts and name them *ducts*, considering that the duct does not contain *its beginning and end.* Name the nodes of tree *S* the beginning (or end) of the corresponding ducts. Consider a certain tree *G*' located within these ducts as follows. The tree *G*' has a superroot in the root duct and the edges coming downwards completely inside ducts. Several edges can come inside one duct; an edge can branch at the point where the corresponding duct branches (which means that the gene doubled in connection with speciation) and can also branch within the duct (which means gene duplication); a cross mark can be ascribed to a leaf of the tree (gene loss); all branching in the tree *G*' not connected with speciation are regarded as duplications. Only the trees *G*' with the minimal sum of the numbers of duplications and losses each taken with a certain positive weight are considered. The properties of *G*' and the below described algorithms are independent of the

values of these weights (certainly, they should be reasonably selected for a biologically meaningful computation). The problem is to find the inner tree *G*' in the given species tree *S*.

Such *G*', after removing from it the leaves marked with crosses, is isomorphic to the initial gene tree *G* (note that the gene names indicated in leaves must be preserved). Removal of the leaves marked with a cross means also the removal of their nearest common ancestor, including the reconstruction of the tree *G*' necessary to preserve its binary structure. Such tree *G*' is unique in the sense that two inner trees are isomorphic to each other (and without crosses, also to *G*), and the two isomorphic nodes are localized to the same duct or node of the tree *S*. When speaking about isomorphism *G*' and *G*, it is always supposed that all the crosses from *G*' are removed, as indicated above. Thus, the tree *G*' exists and is unique; its existence follows from an evident correctness of the algorithm used for its construction, which is described in the section on algorithms, paragraph (*a*).

The inner tree *G*' determines the mentioned nesting α in the following manner: the node *x* from *G* is related to the node *y* corresponding to it in isomorphism, and the node *y* is related to the node $\alpha(x)$ in *S*, which is the end of the duct containing *y*. If *y* coincides with this end, then $\alpha(x) = y$. This stipulation is connected with the fact that the duct is considered without its beginning and end.

Now consider the possible horizontal transfers. For this purpose, in the binary *inner tree G'*, it is allowed to draw an edge ("arrow") from one of its nodes *a*, located inside a duct, to another node *b*, located in an incomparable duct or at its end; such edge means a *horizontal gene transfer*. The *gene transfer* (more precisely, the gene transfer *retaining* the original gene copy) is the transfer when the duct containing node *a* retains the gene copy from this node, while the other copy is transferred along the arrow to the incomparable duct. In some cases, the event of *gene transfer with the loss of original gene copy* is considered. This is the sequence of two events—gene transfer with retaining the original copy and the subsequent loss of this copy. Figures 1b and 2 demonstrate artificial cases of gene transfer without and with retaining of its original copy.

We still consider only the trees *G*' with the minimal value of the function amounting to the sum of the numbers of duplications, losses, and horizontal gene transfers, with the retaining of the original gene copy, with certain weights ascribed to these events. This tree is still named *inner* tree. The weights are named the *costs* for each event. In the computation results, shown in the section describing the testing results, they take the following values: the cost of each lost is set 2, that of each duplication is 3, and that of each transfer with the retaining of the original gene copy is 11 (the cost

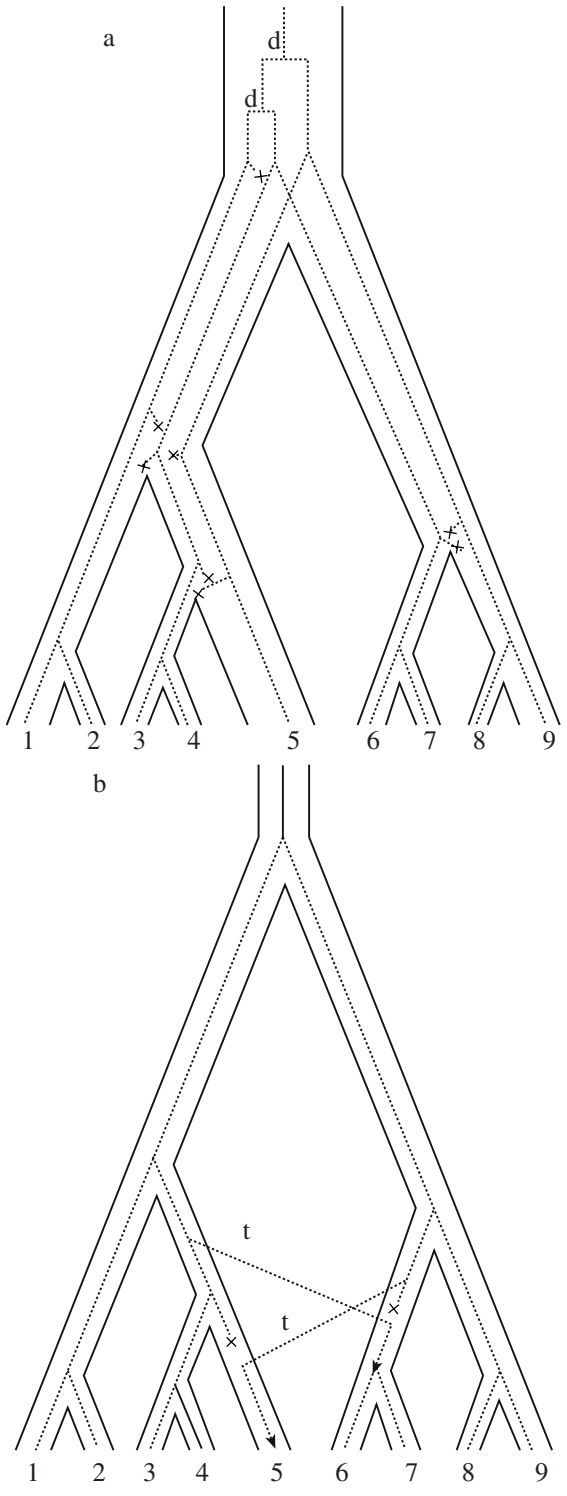

Fig. 1. The inner tree (a) without horizontal transfers and (b) with horizontal transfers. The inner tree is shown with a dotted line in the species tree. Its nodes corresponding to gene duplications are denoted with d, the nodes corresponding to speciation are without designations, and those corresponding to losses are marked with cross. Numbers indicate the leaves of the inner tree and concurrently the leaves of the species tree. The edges (as broken lines) show the horizontal transfers with and without retaining the original gene copy; they are additionally denoted with t.

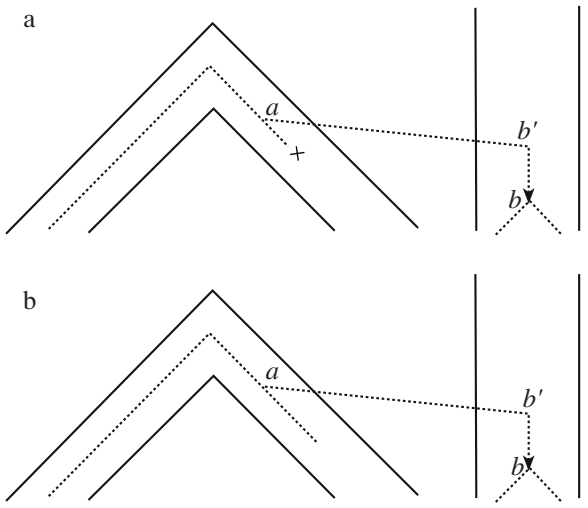

Fig. 2. Horizontal gene transfers (a) without and (b) with retaining the original gene copy. The inner tree is shown with dotted line. Each edge corresponding to the transfer is composed of two segments.

for transfer without retention, 13). We selected these numerical values based on observing the frequencies of the corresponding events ([29] and our unpublished data), although the question on selection of the cost values is far from being grounded.

Such tree G' exists but now is not unique. Its existence follows from the evident correctness of the algorithm used for its construction, described in the section Algorithms, paragraph (*b*).

ALGORITHMS FOR CONSTRUCTING INNER TREE AND TIME LAYERS

(a) The Case When Horizontal Gene Transfers Are Not Taken into Account

The inner tree G' is constructed as a set of values of isomorphism from G onto G'. By analogy with α, this isomorphism is named *nested.* Staring from the root $r(G)$ of tree G, we move along the tree G from its root to the leaves. It is evident from this construction that there exists only one G' isomorphic to G.

Let r be the node of the inner tree G' that corresponds to the root $r(G)$ of tree G. The following three positions are possible for r: (1) r coincides with $r(S)$, the root of tree S; (2) r is located inside the root duct; or (3) r is located below the end of the root duct (with the indication of one of the two directions). The following three relations between two pairs (A,B) and (C,D) are possible, where A and B are the sets of genes ascribed to the leaves of two subtrees in G, with the roots being daughter roots of $r(G)$, and C and D are the sets of genes ascribed to the leaves of S, with roots being daughter roots of $r(S)$. Then

(1) If $(A = C, B = D)$ or $(A = D, B = C)$, then assume (1);

(2) If $((A \neq C \,\&\, A \neq D)$ or $(B \neq C \,\&\, B \neq D))$ and $(C \neq \varnothing \,\&\, D \neq \varnothing)$, then assume (2); and

(3) If $((A \neq C \,\&\, A \neq D)$ or $(B \neq C \,\&\, B \neq D))$ and $(C = \varnothing$ or $D = \varnothing)$, then assume (3) and continue the root edge of the inner tree G' into the nonempty part and end the edge coming into the empty part with a cross (loss).

It is evident that the choice of some other variant than (1)–(3) in each of the three cases will generate excess duplications and losses and prevent the construction of the nesting.

Then the isomorphism and inner tree G' are constructed by induction: consider the edge e' of the tree G', which is isomorphic to the edge e of tree G from its construction and is located inside the duct d of tree S. Let r be the node of the tree G' that corresponds to the end a of the edge e. The following three positions are possible for r: (1) the node r coincides with a (speciation); (2) the node r is located within the duct d (duplication); or (3) the node r is located below the duct d (loss). Reasoning as above for the initial step of induction with the substitution of a for $r(G)$, d for $r(S)$, $C \cap (A \cup B)$ for C, and $D \cap (A \cup B)$ for D. Using the inductive hypothesis $(A \cup B) \subseteq (C \cup D)$, continue the construction until the leaves in G are reached.

(b) The Case When Possible Horizontal Gene Transfers Are Taken into Account

For the edge e in G, designate as $T(e)$ the subtree in G that starts from e so that e is the root edge for $T(e)$. The nestings of tree $T(e)$ into S is constructed for all the pairs ⟨edge e from G, duct d from S⟩ are listed; for each pair, the nesting is searched for among the nestings of the tree $T(e)$ in S for which the beginning e is in d. The desired nesting corresponds to the pair ⟨root

edge in G, root duct in $S\rangle$. When making the list, the edges e located farther from the root are considered first, and at a fixed e, the ducts d located at later time layers are considered first. Designate as x the end of the edge e in G and as y, its image in the nesting of G into S.

The initial step in induction is the case when e and d lead to the leaves $l_e = x$ and l_d, respectively. Then the cost for nesting is 0, if gene l_d is present among the genes in l_e; otherwise the cost of nesting is equal to the cost of horizontal transfer without retaining the original gene copy. Indeed, the only way to stretch a branch to the gene l_e in S is to draw it into the duct leading to the leaf containing l_e. In any case, y coincides with l_d.

The induction step for the next pair $\langle e, d\rangle$ is as follows. Designate as e_1 and e_2 the edges formed by the branching of the edge e at the node x (if it is not a leaf) and, correspondingly, as d_1 and d_2 the two ducts coming from d. Select one of the following possibilities:

(1) The edge e does not end with a leaf. The node x corresponds to duplication within d. According to the inductive hypothesis, the nestings corresponding to the pairs $\langle e_1, d\rangle$ and $\langle e_2, d\rangle$ are already constructed. Their combination gives the nesting for $\langle$e,d$\rangle$, the cost of which is the sum of the costs for $\langle e_1, d\rangle$, $\langle e_2, d\rangle$, and duplication. Then the node y is located within the duct d.

(2) The edges e and d do not end with leaves. The node x corresponds to the branching of the duct d. According to the inductive hypothesis, the nestings corresponding to the pairs $\langle e_1, d_1\rangle$ and $\langle e_2, d_2\rangle$ and the pairs $\langle e_1, d_2\rangle$ and $\langle e_2, d_1\rangle$ are already constructed. Their combination gives the nesting for <e,d>, the cost of which is the sum of the costs for $\langle e, d\rangle$ and $\langle e_1, d_1\rangle$, and $\langle e_2, d_2\rangle$ and $\langle e_1, d_2\rangle$, $\langle e_2, d_1\rangle$. Then the node y coincides with the end of the duct d.

(3) The duct d does not end with a leaf. The node x does not correspond to an event within the duct d or its branching, and the nearest event in the edge e is a loss. In this case, the edge d passes through the branching of the duct S in S, turning into one of the ducts, d_1 or d_2 (both variants are considered; however, for definiteness, let the edge e continue in the duct d_1) and the loss takes place in the other duct (in G', a sequence of several successive edges corresponds to the edge e). According to the inductive hypothesis, the nesting corresponding to the pair $\langle e, d_1\rangle$ is already constructed. The cost for the corresponding nesting for $\langle e, d\rangle$ is the sum of the costs for $\langle e, d_1\rangle >$ and the loss. The position of the edge y is specified by the inductive hypothesis.

(4) The edge e does not end with a leaf. The node x corresponds to the horizontal gene transfer with retaining the initial gene copy in d. Two variants are considered: the transfer is specified by the edge e_1 or the edge e_2 (for definiteness, let it be the edge e_1). Enumerate all the ducts d_1 in S whereto the transfer from the duct d is possible. According to the inductive hypothesis, the nestings corresponding to the pairs $\langle e_1, d_1\rangle$ and $\langle e_2, d\rangle$ are already constructed. Their combination gives the nesting for $\langle e, d\rangle$, the cost of which is the sum of the costs for $\langle e_1, d_1\rangle$ and $\langle e_2, d\rangle$ and the cost for transfer. Then the node y is located within the duct d.

(5) The node x corresponds to neither any event in the duct d nor its branching, and the nearest event in the edge e is a horizontal transfer. In this case, the gene is horizontally transferred without retaining the original copy from the duct d. Enumerate all the ducts d_1 in S whereto the transfer from the duct d is possible. Note that it is impossible to transfer the edge e without retaining the original copy from the duct d_1 once again to a certain duct d_2, because it is "cheaper" to perform a direct transfer to the duct d_2 (the case when d_2 is comparable with d is now prohibited; see the Notes below). Thus, only variants (1)–(4) are possible for e in the duct d_1. Variants (1), (2), and (4) are connected with the branching of e into e_1 and e_2, which already allows the inductive hypothesis to be applied and the position of node y within d_1 (variants (1) and (4)) or at the end of d_1 (variant (2)) to be determined. Note that variant (4) can be omitted from consideration, because if the transfer from d_1 to d_2 with retaining the original copy has occurred, then the equivalent course of events—when the first the copy was transferred to e in d_1 with retaining and then the retained copy was transferred from e to d_2 without retention—will have the same cost. In variant (3), there are two logical possibilities: (i) if e passed through the branching of the duct d_1 and turned to the duct d_3 localized to the same time layer as d_1 and incomparable with d, then it is cheaper to directly transfer it to d_3 (thereby, this variant is prohibited) and (ii) if the duct d_3 is located in a later time layer (or below d), then we apply the inductive hypothesis (also relative to the position of the node y).

It is evident that the algorithm constructs the inner tree G' and the isomorphism of G and G' over time no longer than the cube of the number of genes. This isomorphism is named the *nesting*.

Figure 3 shows the scheme of the algorithm described in paragraph (*b*) for variant (4). The algorithm from paragraph (*a*) can be considered as a particular case of the algorithm from paragraph (*b*): if we assume the cost for horizontal transfer high enough, the algorithm from paragraph (*b*) will output a (unique) solution without any transfers.

Notes. When determining the cost for horizontal transfer, we assume that the partition of ducts into time layers is such that the total cost of moving from duct a into its daughter duct b, which comprises the transfer from a to an incomparable duct c followed by the transfer from c to b, is large as compared with the total cost for moving from the duct a to its daughter duct b downward the tree S together with the cost for

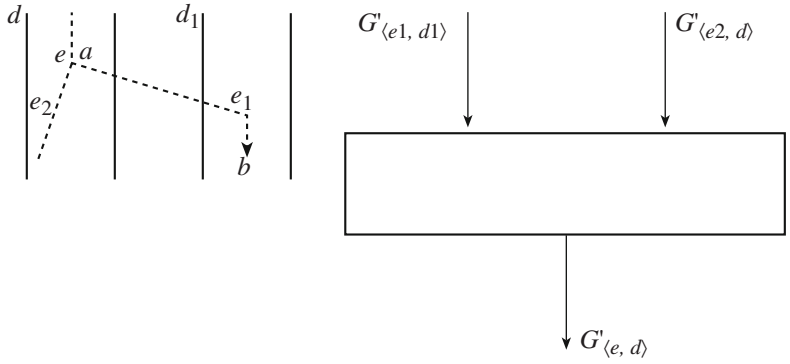

Fig. 3. Scheme of the algorithm for constructing the inner tree G' illustrating variant (4) in paragraph (b) from the section Algorithms. The rectangle (one step of the algorithm) comprises the following actions: the edge e is drawn upwards from the branching of the edges e_1 and e_2; two trees are connected; their nodes become near-root; and a new node appears to form the root of new tree. The algorithm applies induction to the pair $\langle e,d\rangle$: Let the trees $G'_{\langle e_1, d_1\rangle}$ and $G'_{\langle e_2, d\rangle}$ be already constructed by the algorithm. The tree $G'_{\langle e, d\rangle}$ is searched for. The case of a horizontal gene transfer with the retaining of the original gene copy from duct d to duct d_1 along the arrow $e_1 = \overrightarrow{a, b}$ (variant (4) in the same section).

all losses during this moving. This is fulfilled if the partition into time layers does not contain two comparable ducts in one layer. In particular, the below described algorithm for the separation of time layers possesses this feature. If this condition is not satisfied, our algorithm needs supplementary items to the above description.

In certain cases, it is necessary to permit the horizontal transfers from outside the considered cluster of orthologous groups (COG). For this purpose, it is convenient to supplement the tree with an outgroup, i.e., the duct coming from the root to leaf with an empty set of genes. Then the transfers from this duct will be interpreted as the outside transfers. Our algorithm can be generalized for this case.

Frequently the initial gene tree G can be rootless. In this case, the algorithm exhausts all its edges, roots each edge, and constructs the described nesting. According to the quality of nesting position, the algorithm determines the possible position of the root in the initially rootless tree.

When partitioning the ducts into time layers, it is sometimes natural to partition a long duct into several short ducts; this is why the *edges having only one daughter* sometimes appear in the species tree. The above described model and algorithm are easily generalizable for his case.

(c) Computing Time Layers if They Are Not Specified

Moving from leaves to root, calculate (in arbitrary units) the time $r(v)$ from each node v in the species tree to the corresponding leaf. Then ascribe to one time layer all the ducts localized to the same time interval. In this process, we can meet so long ducts that they go beyond one time interval; then these ducts are partitioned into new regions, ducts, each falling into its own interval; in this process, the edges in S with one daughter are formed. According to the above Notes, all this is applicable to such tree S.

Thus, we need to construct the function $r(v)$. If v is a leaf, assume $r(v) = 0$. According to inductive hypothesis, calculate the distances $r(v_1)$ and $r(v_2)$ for two daughters v_1 and v_2 of a certain node v. Possibly, they are calculated in different scales: the distance for v_1 is calculated in the units t_1 and is, correspondingly, a_1t_1 and for v_2 it is calculated in units t_2 and is a_2t_2.

We need to calculate $r(v)$. Let the length of the edge e_1 from v to v_1 be xt_1 and the length of the edge e_2 from v to v_2 be yt_2, where x and y are the desired values. Here we have an evident equality

$$(a_1 + x)t_1 = (a_2 + y)t_2, \quad (1)$$

and will proceed from the following principle: the ratio of lengths of the edges e_1 and e_2 is *inversely proportional* to the ratio of the mean numbers f_1 and f_2 of the branchings encountered on the way from v along e_1 to a certain leaf (the averaging is performed over leaves) and analogously, along e_2. Thus,

$$xt_1f_1 = yt_2f_2. \quad (2)$$

Consider the following variants:

(1) Both edges v_1 and v_2 are leaves, then $r(v) = 1$.

(2) The node v_1 is leaf and v_2 is not leaf; then $a_1 = 0$ and $a_2 > 0$. Designate $k = f_2/f_1$. In this case, Eqs. (1) and (2) provide for directly expressing the length z of the edge (v,v_1) in units t_2: if $xt_1 = zt_2$, then $zt_2 = a_2t_2 + yt_2$, $z = a_2 + y$, $ky = a_2 + y$, and $y = a_2/(k - 1)$.

(3) The node v_2 is leaf and v_1 is not leaf; then $a_1 > 0$ and $a_2 = 0$. This variant is symmetrical to variant (2).

(4) Both nodes v_1 and v_2 are not leaves. In this case, it follows from Eqs. (1) and (2) that $(a_1 + x)yf_2 = (a_2 + y)xf_1$. This equation with the variables x and y determines a curve (a hyperbola or a straight line) on the plane. To find the solution of it, find first the point (x^*,y^*) outside the curve and then project it onto the curve. For this purpose, apply the above formulated principle of inverse proportionality to find the edge length ratios e_1 to e_{11} and e_1 to e_{12}, where e_{11} and e_{12} are the edges connecting v_1 with its daughters v_{11} and v_{12}. Let f_{11} and f_{12} be the mean numbers of branchings on the way from the node v to leaf along the edges e_1, e_{11} and e_1, and e_{12}, respectively. Then two approximate estimates of x are k_1l_1 and k_2l_2, where $k_1 = f_{11}/f_1$, $k_2 = f_{12}/f_1$, $l_1 = a_1 - r(v_{11})$, $l_2 = a_1 - r(v_{12})$. Taking a geometric value of these two estimates, we get x^* and, analogously, y^*:

$$x^* = \frac{\sqrt{f_{11}f_{12}(a_1 - r(v_{11}))(a_1 - r(v_{12}))}}{f_1},$$

$$y^* = \frac{\sqrt{f_{21}f_{22}(a_2 - r(v_{21}))(a_2 - r(v_{22}))}}{f_2}.$$

where the designations in equation for y^* are analogous to those for x^*.

As the solution (x, y), take the projection of the point (x^*, y^*) on the mentioned curve, which is found by a standard algorithm.

Using Eq. (1), find the ratio $t_1/t_2 = (a_2 + y)/(a_1 + x)$. Then in one of the subtrees, for example, the subtree with the root v_1, turn from the scale utilizing the measurement units t_1 to the scale with the units t_2 and assume $r(v) = a_2 + y$. Analogously, in the subtree with the root v_2, it is possible to turn from the scale with units t_2 to the scale with units t_1 and assume $r(v) = a_1 + x$ (the algorithm each time selects the variant that provides for avoiding too large and too small of values).

ALGORITHM TESTING RESULTS AND DISCUSSION

Here we demonstrate the results of testing the algorithm according to our model and compare them with the results obtained by *completely different algorithms*: the algorithm finding the horizontal gene transfers into leaves [30] and the algorithm for determining the transfers to ancestral nodes of the species tree [15]. Note that the former method [30] is based on two criteria for detection of the transferred gene, a leaf in the gene tree G: (1) removal of the leaf leads to the maximal decrease in the cost for nesting and (2) when mapping α, the neighborhood of this leaf in the species tree S is located far from the leaf itself. The results of the method proposed here and the mentioned algorithms [15, 30], in some cases, confirm to one another and, in other cases, are different; the latter results are discussed below.

The following species were used:

Archaea*: Archaeoglobus fulgidus* (Afu), *Halobacterium* sp. *NRC-1* (Hbs), *Methanococcus jannaschii* (Mja), *Methanobacterium thermoautotrophicum* (Mth), *Thermoplasma acidophilum* (Tac), *T. volcanium* (Tvo), *Pyrococcus horikoshii* (Pho), *P. abyssi* (Pab), *Aeropyrum pernix* (Ape), and *Sulfolobus solfataricus* (Sso). **Gram-positive bacteria:** *Streptococcus pyogenes* (Spy), *Bacillus subtilis* (Bsu), *B. halodurans* (Bha), *Lactococcus lastis* (Lla), *Staphylococcus aureus* (Sau), *Ureaplasma urealyticum (*Uur), *Mycoplasma pneumoniae* (Mpn), *M. genitalium* (Mge). **α-Proteobacteria:** *Mesorhizobium loti* (Mlo), *Caulobacter crescentus* (Ccr), and *Rickettsia prowazekii* (Rpr). **β-Proteobacteria:** *Neisseria meningitidis MC58* (Nme). **γ-Proteobacteria:** *Escherichia coli K12* (Eco), *Buchnera* sp. *APS* (Buc), *Pseudomonas aeruginosa* (Pae), *Vibrio cholerae* (Vch), *Haemophilus influenzae* (Hin), *Pasteurella multocida* (Pmu), and *Xylella fastidiosa* (Xfa). **ε-Proteobacteria:** *Helicobacter pylori* (Hpy) and *Campylobacter jejuni* (Cje). **Chlamidia:** *Chlamydia trachomatis* (Ctr) and *C. pneumoniae* (Cpn). **Spirochetes:** *Treponema pallidum* (Tpa) and *Borrelia burgdorferi* (Bbu). **Other:** *Deinococcus radiodurans* (Dra), *Mycobacterium tuberculosis* (Mtu), *Synechocystis* (Syn), *Aquifex aeolicus* (Aae), and *Thermotoga maritime* (Tma).

The species tree shown in Fig. 4, was constructed with the help of the TIQMAX algorithm [31]. Figures 5a and 6a show the trees for three COGs; the species name (in square brackets) and the abbreviated name of the gene taken from the corresponding species are indicated for each leaf; in these figures, the tree nodes are numbered, and the root is denoted with zero. The trees were constructed using the standard PhyloBayes and PhyML algorithms [2].

Figures 5b and 6b show the fragments of the inner tree G' for these two COGs found by the proposed algorithm; the nodes of the inner tree are also numbered (with a smaller type size), and the leaves of the outer tree have abbreviated species names. When describing the inner tree in the text, we put the duct number in front of parentheses (which coincide with the number of its end by definition) and parenthesized the nodes of the inner tree corresponding to the genes duplicated or lost in the corresponding duct (the number of primes indicate the number of the initial gene copy that was lost).

(1) COG0012 (hypothetical GTPase). The COG tree is shown in Fig. 5a, and a fragment of the inner tree is shown in Fig. 5b. The algorithm taken from [30] predicts that Chlamydia are a hypothetical source for a horizontal transfer of the gene ***bu191*** into the bacterium *Buchnera aphidicola* (from γ-Proteobacte-

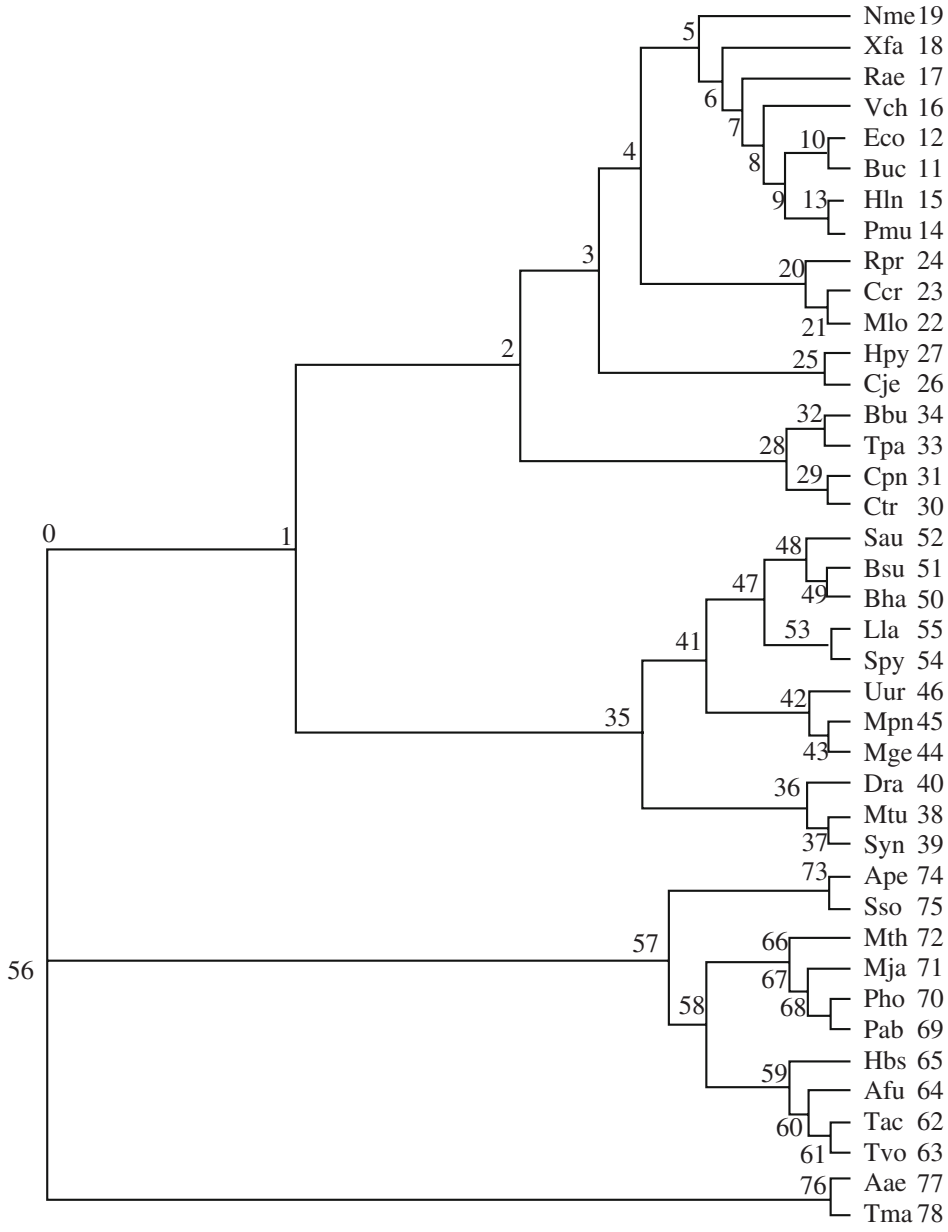

Fig. 4. The species tree. The species names and their abbreviations are listed at the beginning of the section on testing algorithms; tree edges are numbered, and the root is denoted with zero.

ria). With several restrictions, this algorithm also suggested that the gene ***sll0245*** was horizontally transferred from spirochetes into the genome of *Synechocystis* sp. Several values characterized the significance of the latter assumption; however, they were considerably smaller that those for the gene ***bu191***

The algorithm proposed here has predicted that the evolution of this COG contained

Nine duplications: 0(0,20); 1(24,25,58); 2(26); 8(42); 20(53); and 58(1);

38 losses: 1(1',21'); 2(59'); 3(76'); 4(33'); 5(53'); 10(46'); 11(43'); 13(43''); 16(43'''); 20(36'); 22(56'); 23(54'); 24(54''); 25(36'',53''); 28(36''',53'''); 29(76''); 34(29'); 35(26'); 36(59''); 38(75'); 39(52'); 40(75''); 41(50',75'''); 57(21'',23'); 60(6'); 61(16'); 64(2'); 65(2''); 66(2'''); 73(1''); 76(1'''); 77(21'''); and 78(23'');

Two horizontal transfers with the retaining of the initial copy of gene 15 from duct 67 into duct 60 and gene 17 from duct 60 to duct 73; and

One horizontal gene transfer without the retaining of the initial copy of gene 29 from duct 33

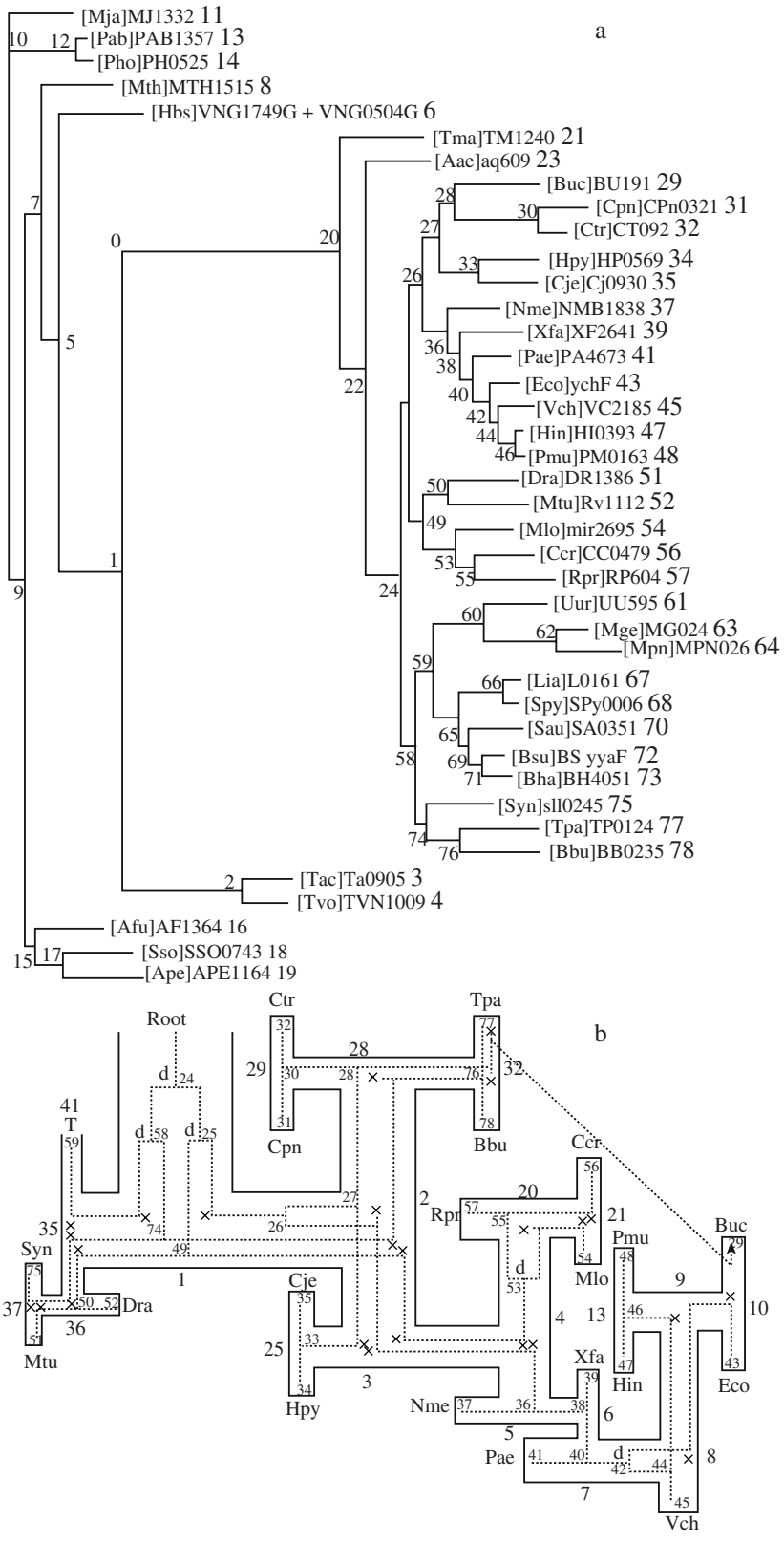

Fig. 5. (a) The tree of COG0012 (hypothetical GTPase) and (b) a fragment of its nesting into the species tree. (a) Tree edges are numbered; the root is denoted with zero; and the square brackets contain the species name to which the corresponding gene belongs. (b) The designations are as in Fig. 1., the nodes of species tree are indicated outside ducts, inner nodes are numbered, leaves are provided with the abbreviated species names, the numbers of inner tree nodes are given inside ducts with smaller type size, and the tree growing from the node marked with T is not shown.

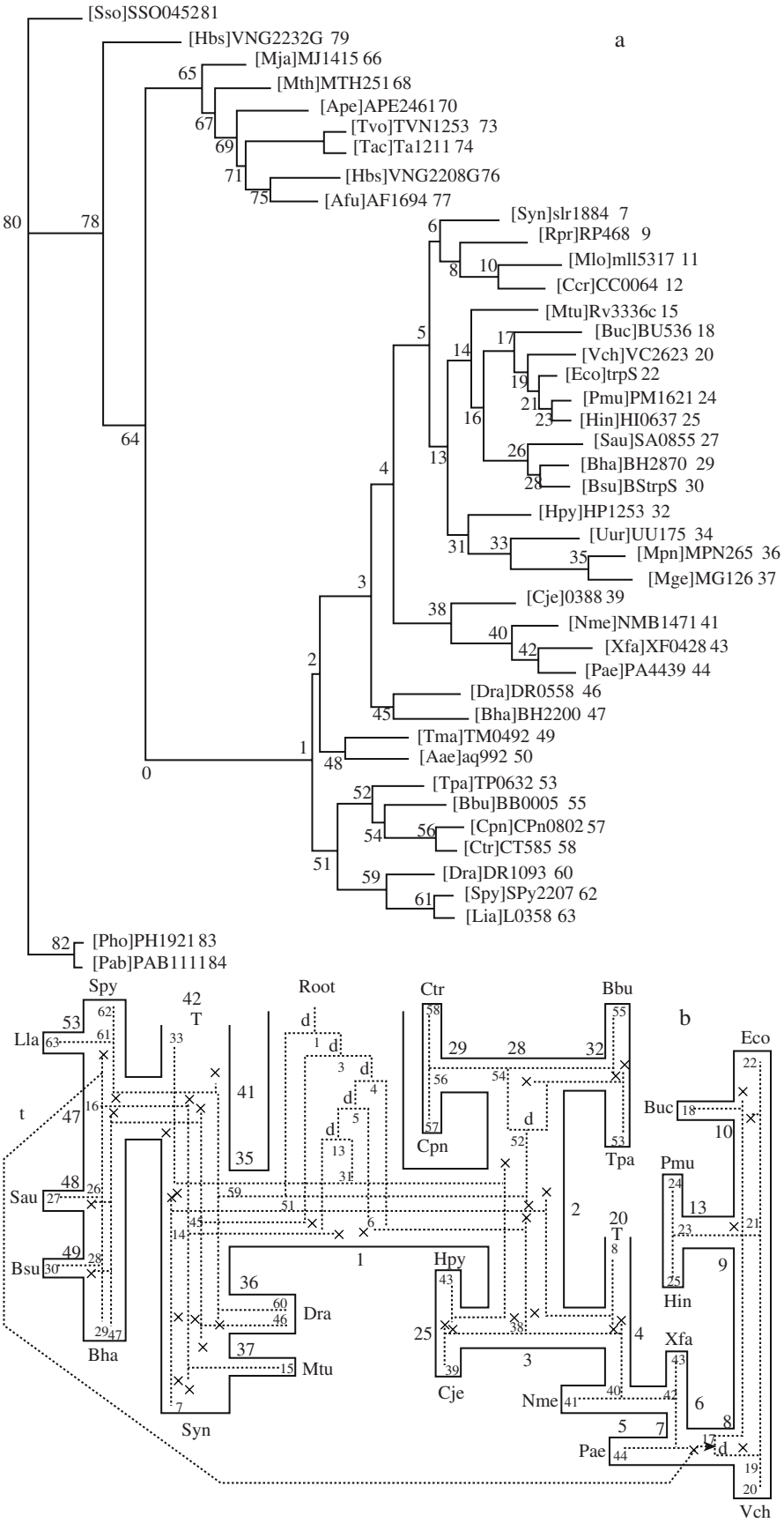

Fig. 6. (a) The tree of COG0180 (tryptophanyl-tRNA synthetase) and (b) a fragment of its nesting into the species tree. (a) Designations are as in Fig. 5a and (b) designations are as in Fig. 5b.

(or in the case of another inner tree, from contiguous duct 34) to duct 11.

We see that the last transfer corresponds to the earlier predicted transfer of the gene ***bu191*** On the contrary, the earlier predicted transfer of the gene ***sll0245*** is not confirmed: under the proposed model, we infer that this "transfer" results from three ancient duplications (the leaf Syn in Fig. 5b). In [30], this prediction had considerably lower significance characteristics; and it was impossible in that work to compare the transfer with consequences of duplications.

The algorithm proposed here also has found two new potential transfers with retaining the initial gene copy, which occurred within the Archaea, namely, the transfer of a common ancestor of the genes *AF1364*, *SSO0743*, and *APE1164* from a common ancestor of the species Mja, Pho, and Pab to a common ancestor of the species Afu, Tac, and Tvo, followed (after duplication within the same duct) by the transfer of a common ancestor of the two last genes to a common ancestor of the species Sso and Ape.

Figure 5b shows a fragment of the constructed nesting bounded by the subtree of the species tree with the root in node 1 and, correspondingly, the COG tree with the root in node 24 (thus, the separate group of Archaea and the ancient species Tma and Aae, whose position in the species tree is not completely unambiguous, are not shown). The nesting of the subtree of COG tree with root 59 into the subtrees of species tree with root 41 appeared identical (i.e., it contains no nontrivial events) and is not shown (the nesting of the root of this subtree is denoted with T), d indicates duplications, the only horizontal transfer is shown by the arrow and t, and losses are denoted with crosses.

Here we have two inner trees G' with the same minimal value of the cost for nesting; they differ in the duct, Tpa or Bbu, from which the gene was transferred (Fig. 5b shows the variant with Tpa).

(2) COG0180 (tryptophanyl-tRNA synthetase). The COG tree is shown in Fig. 6a. The algorithm described in [3, 15] suggested a possible horizontal transfer between the ancestors of {Bha, Bsu, Sau} and {Vch, Eco, Buc, Hin, Pmu}. Our algorithm predicts that the evolution of this COG contained

Ten duplications: 0(0,1); 1(3); 8(17); 28(52); 57(64,65,67,78); and 59(71);

46 losses: 1(64'); 2(45'); 3(52'); 8(44'); 11(22'); 12(18'); 13(18''); 16(18'''); 20(40'); 26(32'); 27(39'); 28(38'); 29(53'); 33(55'); 34(53''); 37(46',60'); 38(7'); 42(47',61'); 48(61''); 51(47''); 52(47'''); 53(47''''); 56(51'); 57(48'); 59(66',68',82'); 60(79'); 61(77'); 64(72'); 65(72''); 66(71',79''); 67(68''); 68(66''); 71(82''); 72(66''',82'''); 73(66'''',68''',79'''); 74(81'); 75(70'); and 76(64'');

Two horizontal gene transfers with the retaining of the initial copy of gene 17 from duct 48 to duct 8 and gene 32 from duct 42 to duct 25; and

Two horizontal gene transfers without the retaining of the initial copy of gene 8 from duct 40 to duct 20 and gene 15 from duct 53 to duct 38.

To refine the obtained scenario and detect the most reliable transfers, we increased the weight for transfer by unity, i.e., assumed the weight for transfer with the retaining of the initial copy to be 12 and, consequently, that without retaining to be 14. The modified scenario contained

13 duplications: 0(0,1); 1(3,4,5,13); 8(17); 28(52); 57(64,65,67,78); and 59(71);

60 losses: 1(64'); 2(14',45'); 3(52'); 4(32'); 5(8'); 8(44'); 11(22'); 12(18'); 13(18''); 16(18'''); 20(40'); 25(8''); 26(32''); 27(39'); 28(8''',32''',38'); 29(53'); 33(55'); 34(53''); 35(38''); 36(33'); 37(46',60'); 38(7'); 39(15'); 40(7'',15''); 41(7'''); 42(16',47',61'); 47(33''); 48(61''); 51(47''); 52(47'''); 53(47''''); 56(51'); 57(48'); 59(66',68',82'); 60(79'); 61(77'); 64(72'); 65(72''); 66(71',79''); 67(68''); 68(66''); 71(82''); 72(66''',82'''); 73(66'''',68''',79'''); 74(81'); 75(70'); and 76(64''); and

One horizontal gene transfer without the retaining of the initial copy of gene 17 from duct 53 to duct 8 (or for another inner tree, from duct 48). This transfer corresponds to the first transfer from the previous scenario (ducts 48 and 53 are contiguous) and is retained with further increase in the weight for transfer through the score of 18 in the case of retaining the initial copy (through 20 without retaining). This ancestral transfer corresponds to the transfer between the ancestors of {Bha, Bsu, Sau} and {Vch, Eco, Buc, Hin, Pmu}, found in [15].

Figure 6b shows a fragment of the second of these inner trees (without Archaea, Tma, and Aae). The designations are the same as in Fig. 5b.

In this case, there are two trees G' with the same value of the minimal cost for nesting: in one case (as is shown in Fig. 6b), the gene is transferred from duct 53 without retaining the initial copy; in the other, the gene is lost in duct 53 immediately after speciation, while the transfer with retention takes place from contiguous duct 48.

CONCLUSIONS

So far, the gene tree G and species tree S were correlated using the nesting α of the former into the latter. This nesting maps the edges in G (events with genes) onto the nodes in S (events with species). Consequently, this correlates the duplication and missing (and, indirectly, losses) of a gene and speciation events. However, it is clear that many other important events occur with genes during speciation. In this work, we propose a model that correlates the duplica-

tions and losses with speciation events in a more explicit manner and, what is most important, correlates the horizontal gene transfers with speciation. This matching for given G and S is provided for by the inner tree G'. Its role consists in locating G "inside" S in a certain way: imagine the edges of tree S as hollow ducts; then G' is in a proper sense located inside S. This gives the opportunity to detail the corresponding events, for example, to distinguish between the horizontal gene transfers with retaining the initial gene copy and with loss of the gene.

Testing of the algorithm and discussion of the corresponding results demonstrate that the horizontal transfers predicted by this model match the predictions of other works, which are characterized as reliable. In the case of several less reliable published predictions, our model replaces the horizontal transfers with several ancient duplications. This provides at least the opportunity under a definitely formulated model to "exchange" horizontal transfers for a certain number of duplications, the possibility to localize an event to the end or inner part of a duct, and so on. We propose to expand the model by including into it other molecular events occurring with nucleotide sequences in an analogous manner.

ACKNOWLEDGMENTS

We are grateful to V.V. Aleshin for helpful discussion of this work.

REFERENCES

1. *Mathematics of Evolution and Phylogeny.* 2005. Ed. Gascuel O. Oxford, MA: Oxford Univ. Press.
2. http://evolution.genetics.washington.edu/phylip/software.serv.html.
3. Lyubetsky V.A., Gorbunov K.Yu., Rusin L.Y., V'yugin V.V. 2006. Algorithms to reconstruct evolutionary events at molecular level and infer species phylogeny. In: *Bioinformatics of Genome Regulation and Structure II*. Springer Sci. & Business Media, Inc., pp. 189–204.
4. Nei M., Kumar S. 2000. *Molecular Evolution and Phylogenetics*. Oxford, MA: Oxford Univ. Press.
5. Gascuel O., Steel M. 2007. *Reconstructing Evolution: New Mathematical and Computational Advances*. Oxford, MA: Oxford Univ. Press.
6. Page R.D.M., Holmes E.C. 1998. *Molecular Evolution: A Phylogenetic Approach.* Oxford: Blackwell.
7. Wolf Y., Rogozin I., Grishin N., Tatusov R., Koonin E. 2001. Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol. Biol.* **1**, 1–22.
8. Durand D., Haldorsson B.V., Vernot B. 2006. A hybrid micro–macroevolutionary approach to gene tree reconstruction. *J. Comput. Biol.* **13**, 320–335.
9. Hallett M.T., Lagergren J. 2000. New algorithms for the duplication-loss model. *Proc. Fourth Annu. Internat. Conf. Comput. Mol. Biol. RECOMB 2000* ACM, pp. 138–146.
10. Chauve C., Doyon J.-P., El-Mabrouk N. 2007. Inferring a duplication, speciation and loss history from a gene tree (extended abstract). In: *Comparative Genomics, RECOMB 2007 International Workshop*. Eds Tesler G., Durand D. Springer, 4751 of LNCS, pp. 45–57.
11. Willson S. 2004. Constructing rooted supertrees using distances. *Bull. Math. Biol.* **66**, 1755–1783.
12. Guigo R., Muchnik I., Smith T.F. 1996. Reconstruction of ancient molecular phylogeny. *Mol. Phylogenet. Evol.* **6**, 189–213.
13. Eulenstein O., Mirkin B., Vingron M. 1998. Duplication-based measures of difference between gene and species trees. *J. Comput. Biol.* **5**, 135–148.
14. Novozhilov A.S., Karev G.P., Koonin E.V. 2005. Mathematical modeling of evolution of horizontally transferred genes. *Mol. Biol. Evol.* **22**, 1721–1732.
15. Gorbunov K.Yu., Lyubetsky V.A. 2005. Identification of ancestral genes that introduce incongruence between protein- and species trees. *Mol. Biol.* **39**, 847–858.
16. Gorbunov K.Yu., Lyubetsky V.A. 2007. Reconstruction of ancestral regulatory signals along a transcription factor tree. *Mol. Biol.* **41**, 918–925.
17. Smith M.W., Feng D.F., Doolitte R.F. 1992. Evolution by acquisition: The case for horizontal gene transfers. *Trends Biochem. Sci.* **17**, 489–493.
18. Page R.D.M., Charleston M.A. 1997. Reconciled trees and incongruent gene and species trees. In: *Mathematical Hierarchies in Biology*, vol. 37. Eds. Mirkin B., McMorris F.R., Roberts F.S. Rzhetsky A. Am. Math. Soc., pp. 1–14.
19. Lyubetsky V.A., Zhizhina E.A., Rubanov L.I. 2008. The Gobbsean approach to the problem of evolution of gene expression regulatory signal. *Probl. Peredachi Inform.* **44**, 52–71.
20. Mirkin B.G., Fenner T.I., Galperin M.Y., Koonin E.V. 2003. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol. Biol.* **3**, 1–34.
21. Johnston A.W., Todd J.D., Curson A.R., Lei S., Nikolaidou-Katsaridou N., Gelfand M.S., Rodionov D.A. 2007. Living without Fur: The subtlety and complexity of iron-responsive gene regulation in the symbiotic bacterium *Rhizobium and other alpha-proteobacteria. Biometals*. **20**, 501–511.
22. Gerasimova A.V., Gelfand M.S. 2005. Evolution of the NadR regulon in Enterobacteriaceae. *J. Bioinform. Comput. Biol.* **3**, 1007–1019.
23. Seliverstov A.V., Putzer H., Gelfand M.S., Lyubetsky V.A. 2005. Comparative analysis of RNA regulatory elements of amino acid metabolism genes in Actinobacteria. *BMC Microbiol.* **5**, 1–14.
24. Vitreschak A.G., Mironov A.A., Lyubetsky V.A., Gelfand M.S. 2008. Functional and evolutionary analysis of the T-box regulon in bacteria. *RNA*. **14**, 717–735.
25. Gorbunov K.Yu., Lyubetskaya E.V., Asarin E.A., Lyubetsky V.A. 2009. Modeling evolution of the bacterial regulatory signals involving secondary structure. *Mol. Biol.* **43**, 527–541.

26. Zghidi W., Merendino L., Cottet A., Mache R., Lerbs-Mache S. 2007. Nucleus-encoded plastid sigma factor SIG3 transcribes specifically the *psbN* gene in plastids. *Nucleic Acids Res*. **35**, 455–464.

27. Favory J.-J., Kobayshi M., Tanaka K., Peltier G., Kreis M., Valay J.-G., Lerbs-Mache S. 2005. Specific function of a plastid sigma factor for *ndhF* gene transcription. *Nucleic Acids Res*. **33**, 5991–5999.

28. Seliverstov A.V., Lyubetsky V.A. 2006. Translation regulation of intron containing genes in chloroplasts. *J. Bioinform. Comput. Biol*. **4**, 783–793.

29. Lyubetsky V.A., V'yugin V.V. 2003. Methods of horizontal gene transfer determination using phylogenetic data. *In Silico Biol*. **3**, 17–31.

30. V'yugin V.V., Gelfand M.S., Lyubetsky V.A. 2003. Identification of horizontal gene transfer from phylogenetic gene trees. *Mol. Biol*. **37**, 673–687.

31. V'yugin V.V., Gelfand M.S., Lyubetsky V.A. 2002. Tree reconciliation: Reconstruction of species phylogeny by phylogenetic gene trees. *Mol. Biol*. **36**, 807–816.