Molecular Biology, Vol. 36, No. 5, 2002, pp. 650–658. Translated from Molekulyarnaya Biologiya, Vol. 36, No. 5, 2002, pp. 807–816.

BIOINFORMATICS

UDC [577.2.08:681.3]+575.852

Tree Reconciliation: Reconstruction of Species Phylogeny by Phylogenetic Gene Trees

V. V. V'yugin[1], M. S. Gelfand[2], and V. A. Lyubetsky[1]

[1] *Institute for Problems of Information Transmission, Russian Academy of Sciences, Moscow, 101447 Russia; E-mail: lyubetsk@iitp.ru*

[2] *State Scientific Center GosNIIGenetika, Moscow, 113545 Russia*

Received March 14, 2001; in final form, March 19, 2002

Abstract—It is well known that phylogenetic trees derived from different protein families are often incongruent. This is explained by mapping errors and by the essential processes of gene duplication, loss, and horizontal transfer. Therefore, the problem is to derive a "consensus" tree best fitting the given set of gene trees. This work presents a new method of deriving this tree. The method is different from the existing ones, since it considers not only the topology of the initial gene trees, but also the reliability of their branches. Thereby one can explicitly take into account the possible errors in the gene trees caused by the absence of reliable models of sequence evolution, by uneven evolution of different gene families and taxonomic groups, etc.

Key words: phylogenetic species tree, phylogenetic protein tree, tree mapping, tree reconciliation, mitochondrial genomes, eukaryotes

INTRODUCTION

Reconstruction of species philogeny is one of the main problems in evolutionary biology. Paleontological records are often incomplete and contradictory; therefore, this reconstruction is impossible without using the methods of molecular biology. This implies using amino acid sequences of modern related biopolymers to derive a gene tree within the framework of one or another model of molecular evolution. More or less realistic models of evolution entail unsolvable calculation problems; therefore, methods of approximation must be used for deriving the gene tree (see reviews [1, 2]).

A phylogenetic tree may be derived using various genes and gene-encoded proteins. These trees are referred to as "gene trees." Experience shows that gene trees often have variable structure depending on the selected gene family (the structure reflects completeness of the initial data and precision of the applied algorithm). As a rule, a phylogenetic gene tree differs from the species tree even if there were no errors related with selection of an inadequate model of molecular evolution, incompleteness of the initial data, or drawbacks of either algorithm or computer program. Therefore, the task can be formulated as follows: to find a "consensus" tree (of species) best reconciling the given family of gene (protein) trees. We consider a biological model which suggests that the differences originate from gene duplication and gene loss as well as from the errors in gene tree formation induced by, e.g., unequal rate of evolution. One more essential reason, the lateral gene transfer, will be discussed elsewhere.

If gene duplication occurs in an ancestor species, the two copies start changing independently of each other, both being inherited by subsequent generations.

Duplication of a gene in the ancestor species may result in two offspring species carrying genes a_1 and a_2 diverged earlier than the two species. In this case the difference between the genes a_1 and a_2 within a genome may be higher than that of orthologous genes, e.g., a_1, from two distinct genomes. If one of the two original gene copies, say a_1, is lost in one of the offspring species, comparison of the remaining gene a_2 with the sample of a_1 from other species would result in deviation in the species tree topology. Goodman and coworkers [3] appeared to be the first who considered the problem of phylogeny comparison for the genes and for the species [3]. The duplication and loss model was used to produce the species tree in some works [4–6].

For this purpose the procedure usually starts with comparative analysis of the tree structure for gene tree and species tree. The minimal number (cost) of elementary operations (these operations should have reasonable biological interpretation) needed to fit the gene tree into the species tree is calculated. This approach is based on the assumption that the gene tree most similar to the species tree, with a corresponding set of biologically reasonable transforming operations, most faithfully reflects gene evolution within the involved species. Therefore, preference is given to

0026-8933/02/3605-0650$27.00

the phylogenetic tree of species with minimal total cost of transformation operations for all gene trees. This work introduces a new more complex function of the tree transformation cost. Following our procedure, only reliable nodes are included in the "consensus" tree, while the nodes that are insignificant or reflect species phylogeny in a wrong way should be rejected. We also suggest an algorithm to produce a species tree most similar to the given set of gene trees.

We applied the new algorithm to the data of two types. The first suggested gene duplications and loss in the process of species evolution; the data were taken from the studies of higher eukaryotic proteins. The second considers tree deviations induced only by incompleteness of the data and errors of the tree-producing methods. As an example, we applied this approach to study mitochondrial genomes.

Phylogenetic trees were derived for eukaryotic families using the local minimum of the cost function. Our calculations have shown that the use of weighted values for gene trees provides considerable stability of the results upon varying the initial tree within our algorithm of the local minimum search.

TAXONOMY TREES AND THEIR HOMOMORPHISMS

A classification task starts with a set of initial operational taxonomic units. In our case, these units are species or groups of species. Mathematically this can be presented as a numeric set $I = \{1, 2, \ldots, N\}$. Let us consider two binary trees: gene tree G and species tree S. Let each of these trees have N leaves corresponding to one-member subsets $\{1\}, \{2\},\ldots, \{N\}$. The internal nodes of the trees are designated by the sets formed by the elements of I as follows. If direct offspring of a node are subsets A and B, than the node is the union of these two $A \cup B$. In this case the root of the tree corresponds to the set I. A node is identified with the corresponding set. It is evident that for any two node sets, either one is a subset of the other (if one node is an offspring of the other) or their intersection is empty (if not). Therefore, each phylogenetic tree is composed of clusters which are subsets of the set I.

The gene tree G and the species tree S produce a unique map designated α

$$\alpha: G \longrightarrow S$$

as follows: for each $g \in G$ the value of $\alpha(g)$ is determined as minimal for set-theoretic inclusion $s \in S$ with $g \subseteq s$. This mapping should obviously be considered as tree *homomorphism*, since if $g \subseteq g'$, then $\alpha(g) \subseteq \alpha(g')$.

Let us consider an internal node or root of the g tree, designating cg and Cg its left direct offspring and its right direct offspring, respectively. If g is not a root, then pg designates the direct ancestor of g.

If the gene tree and the species tree are of similar structure, then the mapping would be isomorphic since any node s from S would be a reflection of some node g from G (i.e. $s = \alpha(g)$) and $g \subseteq g'$ would be equivalent to $\alpha(g) \subseteq \alpha(g')$ for all g and g' nodes of the gene tree.

Consider the main features which show the difference between tree homomorphism α from tree isomorphism. These are *duplications* in the range of definition, i.e. two nodes g and g' where g' is a direct offspring of g and $\alpha(g) = \alpha(g')$, and also *intermediate gaps* in the range of values, i.e. node s in species tree S fits $\alpha(g) \subseteq s \subseteq \alpha(pg)$ (it means that node s is located strictly between nodes $\alpha(g)$ and $\alpha(pg)$). This node $s \in S$ is named g-gap (or g-intermediate node). Consider the set of all g-intermediate nodes I_g. Obviously, the set of all gaps coincides with

$$M(S,G) = \bigcup_{g \in G} I_g.$$

The (g, s) pair, where $s = \alpha(g)$ is named *one-side duplication* if one of the following is true: either $\alpha(g) = \alpha(cg)$ or $\alpha(g) = \alpha(Cg)$. If both these equations are true, than the (g, s) pair is named *two-side duplication*. Consider the set of one-side duplications $O(G, S)$.

Eulenstein and Vingron [7] introduce a measure of dissimilarity for gene tree G and species tree S as the *function of comparison cost*

$$c(G, S) = |M(G, S)| + |O(G, S)|.$$

This function also shows the difference of homomorphism α from isomorphism.

Let us consider a biological interpretation of these events taking into account the cost of the tree comparison. The following model of evolutionary history of the genes in the process of species development is assumed. The gene shows divergence at species divergence, i.e. when an ancestor species divides into two offspring species. Each of the offspring genes is transferred to one of the two offspring species. Further evolution of these genes is mutually independent, and this gene pair is named *orthologous*. Another possible event is deviation of one gene into two identical copies, say a_1 and a_2, within one and the same species with subsequent independent development of the two genes within this species and its offspring. In this case the gene pair is named *paralogous*. A difference between gene tree G and species tree S may be explained by duplication of some gene a at the root into two copies a_1 and a_2; in the process of further evolution the offspring of both copies are diverged. It may happen that no offspring of copy a_2 is detected in an offspring species. To explain this event one should consider *gene loss*, which happened at the node of the respective tree edge first showing no offspring of the given gene. Examples of duplications, losses and α-mapping are shown in Figs. 1 and 2. The errors of

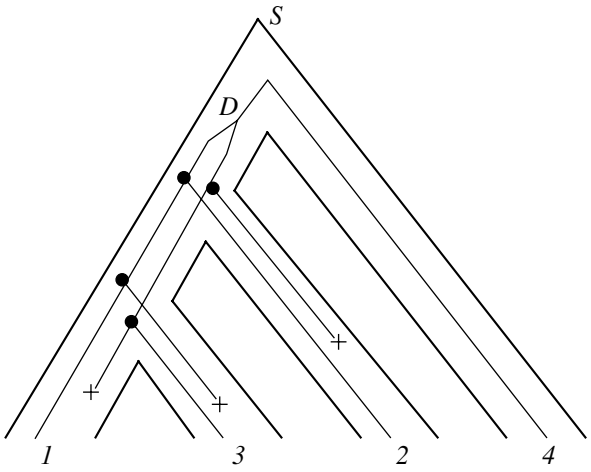

Fig. 1. Gene evolution in the process of species origin (one duplication *D* and three losses *S* are shown) on tree *S* of four species.

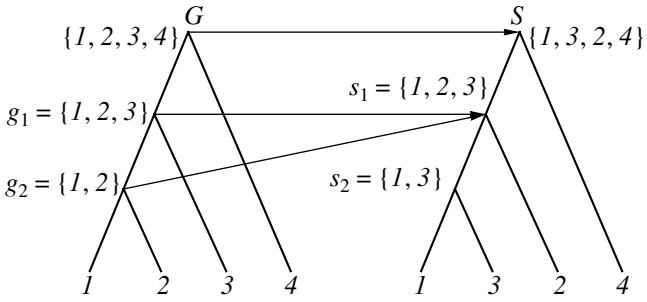

Fig. 2. Homorphism α transformation of gene tree *G* into species tree *S* (shown are a one-side duplication and two gene loss events.

the tree mapping can be one more source of dissimilarity between the gene tree and the species tree (see below).

One-side duplications and intermediate nodes are selected to estimate the losses, since each of these events can be related with single event of gene loss (see [8]). Therefore, the measure of dissimilarity for the two trees is determined as the total number of the loss events required to explain this distinction.

Some algorithms of the tree mapping from the sets of informative macromolecular sequences include calculation of numeric values showing the path between these sequences. These paths are considered as the lengths $c(a, b)$ assigned to the edges of the phylogenetic tree, and the lengths are in turn related with time (assuming equal rates of evolution for the considered genes).

In this work we estimate reliability of a node (i.e., of the length $c(a, b)$ of the respective edge) for a phylogenetic tree using the index of the node support for bootstrap pseudoreplica.

Given the lengths $c(a, b)$ of the tree edges introduce the differential cost function for the trees *G* and *S* as

$$L(G, S) = \sum_{(g, \alpha(g)) \in O(G, S)} c(g, pg) + \sum_{(g, \alpha(g)) \in M(G, S)} c(g, pg)|I_g|,$$

where the first term shows losses from duplications and the second shows losses related with the missing nodes. Consequently, the function of the differential cost for the trees is determined by total number of gene loss events taken with different weights. The weights may be estimated as either transformed index of sequence similarity assigned to the nodes of the given edge, or as bootstrap support of the respective cluster, i.e., part of the tree underlying the given edge. In this work we use the latter.

ALGORITHM AND CALCULATION TECHNIQUE

It is easy to develop an algorithm allowing calculation of the tree differential cost function for average time $O(N\log N)$, where *N* is the number of species. This is the case for α transformation from *G* into *S*. Calculation is based on the following principles: for leaf *g* of the gene tree $\alpha(g)$ is equal to that of the respective leaf of the species tree; for an internal node $\alpha(g)$ is equal to the lowest (considering natural partial ordering of the nodes of the species tree) common ancestor of $\alpha(cg)$ and $\alpha(Cg)$, calculated for average time $O(\log N)$. Improved algorithms of the lowest common ancestor search in binary tree allow one to reduce the required time for these calculations and for the whole algorithm to $O(N)$ [9].

The number of one-side duplications and gaps for the nodes of the species tree *S* is calculated in parallel with mapping α. For this purpose, a counter of duplication cost $d(s)$ is introduced for each node *s* of the tree *S* (with initial value $d(s) = 0$); a counter of intermediate node cost $i(s)$ (initial value $i(s) = 0$) is also introduced. The counters work in parallel with α-mapping as follows (remember that $\alpha(g)$ is mapped from the leaves to the root of the tree *G*): if $\alpha(g) = \alpha(cg) = s$ and this is not true for another direct offspring *g'*, than take $d(s) = d(s) + c(g, pg)$; if $\alpha(g)$ is not a parent of $\alpha(g)$, than for any *s* when $\alpha(g) \subset s \subset \alpha(pg)$ take $i(s) = i(s) + c(g, pg)$. The average working time for this algorithm is $O(N\log N)$, the longest time is $O(N^2)$.

Given the genes trees $G_1, G_2, \ldots, G_n$, let us consider the problem of mapping the species tree S when the value of

$$c(S) = c(G_1, S) + c(G_2, S) + \ldots + c(G_n, S) \qquad (1)$$

reaches its minimum. In general, the time required for this would grow exponentially with increasing set of the initial data.

The suggested algorithm implies gradual rearrangement of the current species tree S around each of its nodes aiming to produce tree S with local minimum of $c(S)$ as calculated from (1). The initial species tree S_0 can be produced e.g. by random number generator. First the depth of rearrangement h is assumed. For each node p of the current tree S all nodes located at depth h under the given node are considered and rearranged in various ways together with respective subtrees. The values of $c(S)$ are calculated at the same time. Then the nearest neighbor, i.e., the new current tree S with minimal $c(S)$ is selected, and the procedure is repeated. Various methods of node sequence search can be used. Calculations are run till cost $c(S)$ stops to decrease. The time of one scanning of all nodes is estimated as $O(nT(h))$, where n is a number of the nodes of the tree and $T(h)$ is the number of all binary trees of the depth h; obviously, this algorithm is rather fast. The mathematical aspect of this algorithm was discussed by us earlier [16].

A more complex stochastic algorithm can be used to scan the nodes of the species tree (to search for the minimal value of the cost function), implying given and constantly recalculated probability distribution to select a subsequent node for local rearrangement. It is also possible to always select an offspring providing stronger decrease of $c(S)$.

We considered one more algorithm where the final tree is produced by iteration of the procedure: first the species tree is produced for the first i species; a current tree with i leaves is obtained, which reconciles only i-species-related parts of the gene trees, then the tree is extended to a similar tree of $i + 1$ (or $i + p$, where p is fixed) leaves. This algorithm obviously depends on ordering of the initial species set; rather convenient are the cases allowing "natural" ordering of the considered species. Combination of this algorithm with that described above allows one to obtain a functional of somewhat improved quality; however, with our data (probably not specific) the resulting species trees are almost the same.

The biological aspect of this type of a problem suggests selection of units (species or higher-order taxa) for taxonomy analysis for various aims of classification: from estimating similarity of given taxa to certain known groups to producing complete species tree for the studied units.

The data files of the selected species were downloaded from genetic databases of the GenBank type (www.ncbi.nlm.nih.gov). Then the obtained files were tested for completeness of information: they should contain a rather large group of protein families or genes common for all selected organisms. The gene trees were produced for each family, and the species tree was derived from these gene trees. We used the CLUSTAL software package for multiple alignment of the sequences related to the protein families (biomaster.uio.no/clustalw.html). Gene trees were produced using the PHYLIP package [10].

Then we applied software TIQMAX implementing the proposed algorithm to produce a species tree from the set of gene trees at minimal cost. The program searches only for the local minimum of S depending on the initial species tree S_0, therefore we generated random initial species trees S_0 and used TIQMAX to find local minimum S for each of these. As a result, we obtained the sequence of locally optimized species trees, ordered for increasing value of the cost comparison functional. The frequency of occurrence was calculated for each of these trees (in the process of optimization using random initial trees).

SOME EXAMPLES OF CALCULATION

Species Tree Derived from Nine Protein Families

We applied the described procedure to classify 14 groups of animals: bone fishes (salmon, trout, catfish), lizards and snakes (iguana, anoles, gecko, cobra), crocodiles (crocodile, alligator), ducks (duck, goose), hens (hen, pheasant, turkey), hares (rabbit), predators (dog, fox, cat), artiodactyls (bull, ram), perissodactyls (pig), rodents (rat, mouse), and primates (marmoset, monkey, human).

Nine protein families were analyzed: aldolases, α-fetoproteins, lactate dehydrogenases, prolactins, rhodopsins, trypsinogens, tyrosinases, vasopressins, and Wnt-7 (see [11]). Nine gene trees were derived from these data.

In most cases only the data for one or two proteins were available for analysis within species (except for well-studied mammalian and avian groups). We were mainly interested in analysis of the higher-order taxa, therefore sometimes we used the data for different species within one higher-order group.

Calculations were run for the two types of data. In the first case we calculated the percent weight of the tree edges using the bootstrap method and reflecting reliability of the respective clusters. In the second case the weights were considered equal to one unit, i.e. only tree topology was analyzed. We performed 1000 calculations for each of the two cases to produce a locally optimized species tree derived from an initial species tree obtained using random number generator.

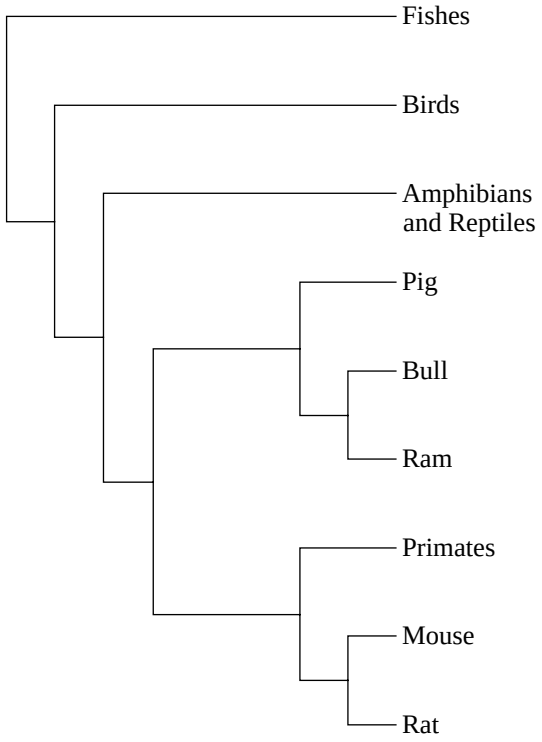

Fig. 3 Vertebrate tree produced from nine protein families using program TIQMAX with bootstrap edge weights.

The species tree in Fig. 3 was obtained in all 1000 variants (with bootstrap weights) and in 747 variants (with one-unit weights); in the latter case the comparison cost was higher for all other trees.

Tree Mapping for Mitochondrial Genes

At this step we tested the suggested algorithm producing a consensus species tree using various trees of mitochondrial proteins. Mitochondrial genomes from the following organisms were selected for analysis:

Rhodophyta (red alga): *Porphyra*, *Cyanidoschyzon*; Bryophyta: *Marchantia* (liverwort); Oomycetes: *Phytophthora*; Chryzophyceae: *Chrysodidymus*; Bicosoecida: *Cafeteria*; Fungi: *Allomyces*; Mastigophora (flagellates): *Reclinomonas*; Metazoa: *Metridium* (actinia), *Platyneris*, *Lumbricus* (worms), *Katharina* (mollusk), *Drosophila* (fly), *Anopheles* (mosquito), *Locusta* (locust), *Ixodes* (tick), *Artemia* (crustacean), *Penaeus* (shrimp), Echinodermata: *Asterina* (starfish), *Florometra*; vertebrates: *Branchiostoma* (lancelet), *Petromyzon* (lamprey), *Squalus* (ray), *Salmo* (salmon), *Xenopus* (frog), *Alligator* (alligator), *Gallus* (hen), *Didelphis* (opossum), and *Homo* (human).

Figure 4 shows the phylogentic tree of these species according to taxonomy of the GenBank (http://www.ncbi.nlm.nih.gov/Taxonomy).

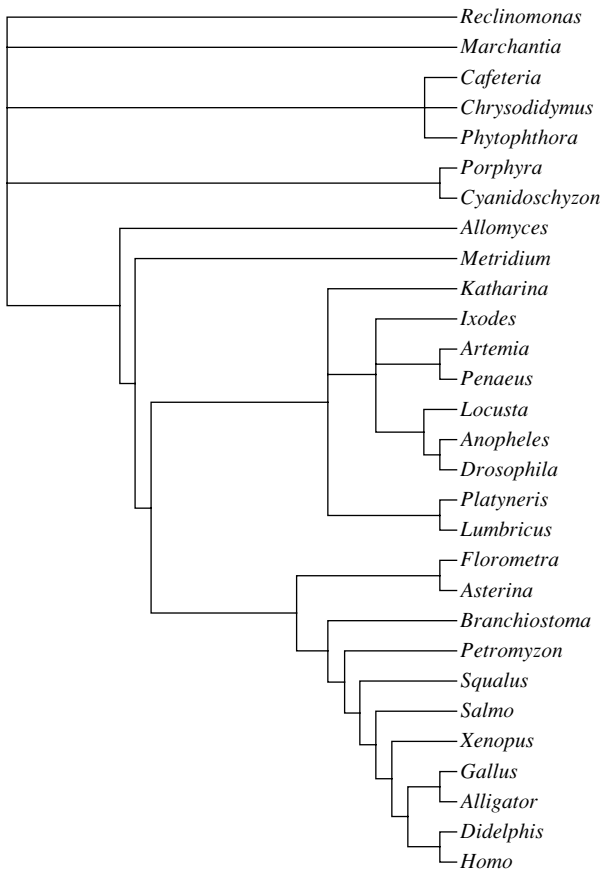

Fig. 4. Taxonomy of eukaryotes according to http://www.ncbi.nlm.nih.gov/ Taxonomy (*Reclinomonas* in not classified).

The mitochondrial genome data were downloaded from the NCBI database of GeneBank. Then we selected a 13-protein cluster corresponding to the genes found in mitochondria of all the organisms under study: NAD1, NAD2, NAD3, NAD4, NAD4L, NAD5, NAD6, COX1, COX2, COX3, ATP6, ATP8, CYTB.

The tree was obtained for each protein as follows: The sequences were aligned, then variants of the tree were obtained using the minimal cost method (programs SEQBOOT and PROTPARS of the PHYLIP package); the consensus gene tree was produced from these variants using program CONSENSE. Each internal node of the resulting tree had a numeric value reflecting its reliability (assigned to the respective cluster by the bootstrap method). These numerical values were used as edge lengths to obtain a species tree using TIQMAX.

We run 1000 calculations of the locally optimized tree depending on the initial species tree obtained using the program generating random marked trees The tree shown in Fig. 5a was obtained in 971 cases, other 29 trees had higher comparison cost.

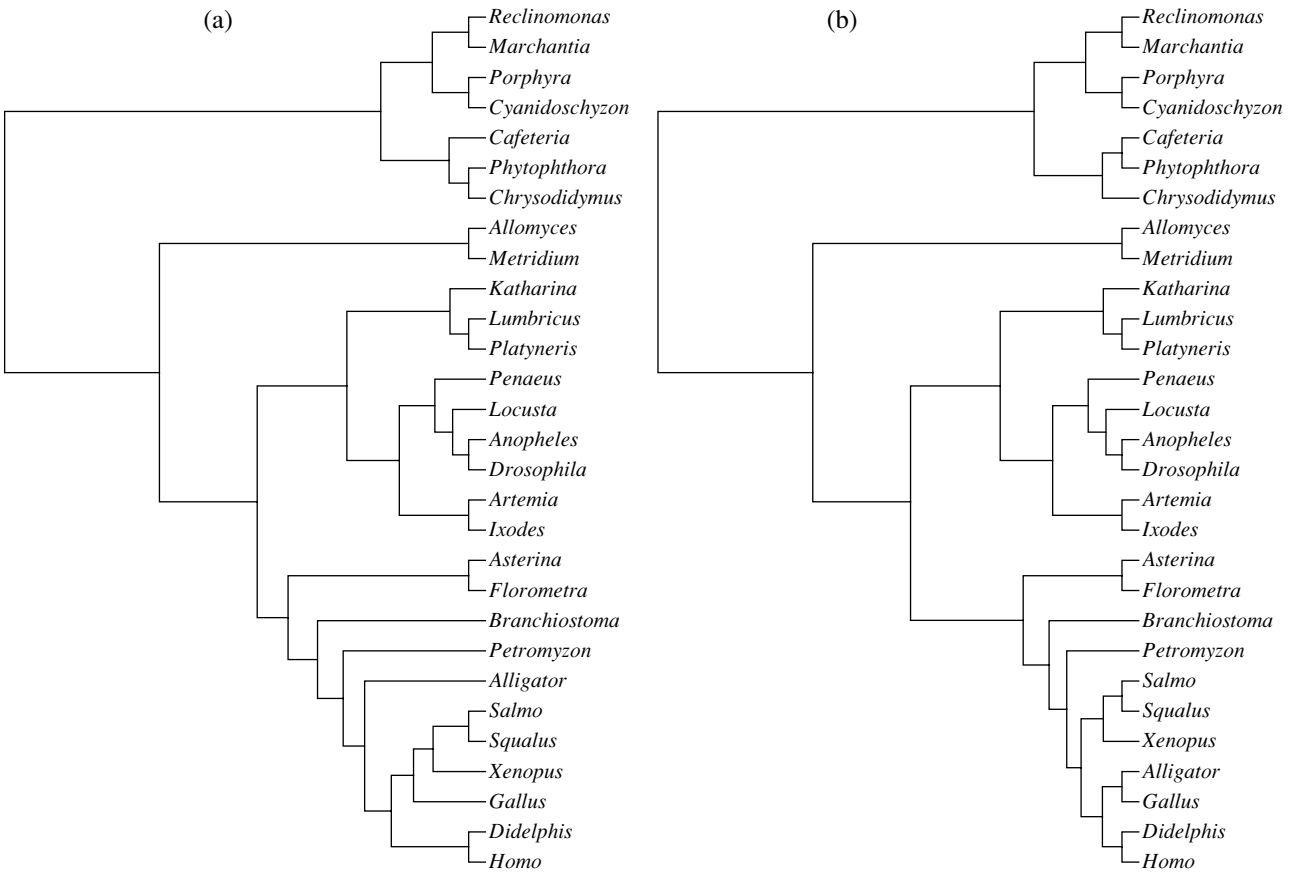

Fig. 5. (a) Phylogeny of eukaryotes derived from the set of 13 mitochondrial genes using tree reconciliation programs TIQMAX with bootstrap edge weights. The tree root was always selected aiming to maximal accordance with traditional classification. (b) Phylogeny of eukaryotes derived from the set of 13 mitochondrial genes using tree reconciliation programs TIQMAX with one-unit weights of the tree edges.

For the same set of the gene trees with one-unit weights, i.e. when only topological structure of a tree was considered, the 1000 random initial trees produced 355 trees with the minimal cost. In total we obtained six slightly distinct species trees occurring in 83, 74, 72, 47, 46, and 33 cases). One of these is shown in Fig. 5b; this tree is most distinct from the tree shown in Fig. 5a. All trees of higher cost had more deviation from the basic tree shown in Fig. 4.

In order to test the efficiency of our algorithm as a tool to produce the species tree basing on mitochondrial genes, we tried two more ways to produce the consensus species tree from gene trees. The first implied joining (concatenation) of all aligned sequences into one long sequence and subsequent using of PROTPARS to produce a tree considered to be a species tree. (Fig. 6). One more way was to produce a consensus tree using program CONSENSE from gene trees obtained earlier (Fig. 7). All three methods had similar results.

DISCUSSION

The complete set of species following earlier work [11] produced a large variety of trees with distinct topology but with similar weights. The reason probably lies in scarce data: many taxa were represented only with one or two protein families. Similar results were obtained earlier [12]: 12 species trees equally represented the set of 53 gene trees. Approximation by merging related poorly represented taxa and by removal of isolated low-numbered taxa produces rather acceptable and stable (considering a variety of initial species trees) though uninteresting tree (Fig. 3). Advancing in this direction is considerably hindered by the absence of available protein sets represented in a variety of taxonomic groups.

On the other hand, the cost of comparison can be considered as a measure of tree reconciliation even in the absence of duplications. With this procedure the trees can be produced from the sets of orthologous proteins. Contrary to the algorithms using only topology of the gene trees, we count for the lengths of those

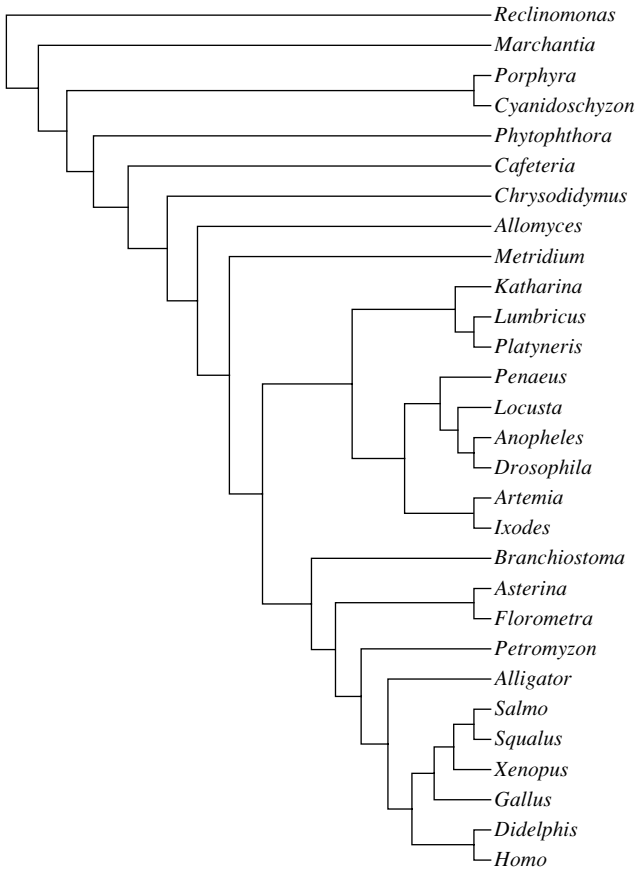

Fig. 6. Phylogeny of eukaryotes derived from the set of 13 mitochondrial genes using tree reconciliation programs PROTPARS (PHYLIP) and concatenated sequences.

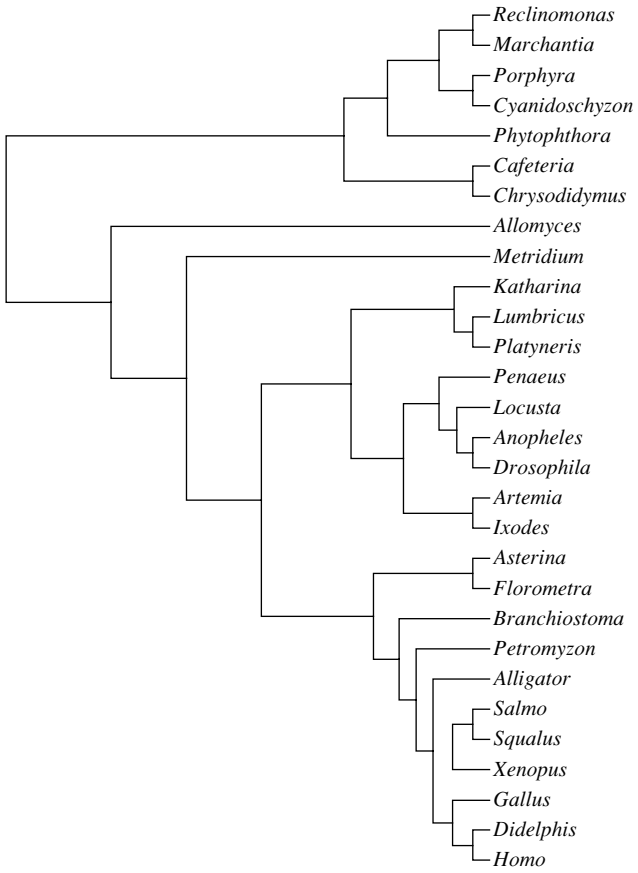

Fig. 7. Phylogeny of eukaryotes derived from the set of 13 mitochondrial genes using tree reconciliation programs PROTPARS and CONSENSE (PHYLIP).

gene tree edges which are inconsistent with the species tree: short (i.e. weakly supported) edges have lower cost than longer ones.

As an example, we analyzed complete mitochondrial genomes of various eukaryotes. The tree of the analyzed species, names following the GenBank taxonomy [13], is shown in Fig. 4.

The reconciled trees were produced using three distinct procedures. Figure 5a shows the trees obtained using the program TIQMAX described in this work (taking the bootstrap weights into account) and Fig. 5b shows the trees considering only topology of the gene trees. Figure 6 shows the result obtained using the program PROTPARS to analyze the amino acid sequence concatenated from all studied proteins for each of the species. Finally, Fig. 7 shows the results obtained using the program CONSENSE of software PHYLIP to analyze gene trees.

The resulting trees are slightly distinct. General taxonomy of Metazoa is reconstructed similarly, except for joining *Metridium* and *Allomyces* only in the TIQMAX tree (Fig. 5a,b). Bilateria are subdivided into Protostomia and Deuterostomia, the former include Arthropoda and Mollusca/Annelida (always forming one taxon), and the latter include Echinodermata and Chordata. Deuterostomia within the PROTPARS tree (Fig. 6) have "altered" edges of Echinodermata and *Branchiostoma*.

At the same time, in all three produced trees (Figs. 5a, 6, 7) alligator (*Alligator*) rather than ray (*Squalus*) is an outspecies for the other Gnathostomata, ray (*Squalus*) and salmon (*Salmo*) form a taxon which includes frog (*Xenopus*) as the next member, while hen (*Gallus*) either falls into this taxon (TIQMAX and PROTPARS), or forms a cluster with mammals corresponding to the taxon of warm-blooded animals (CONSENSE). No clusters were related to canonical taxa Tetrapoda and Archozauria.

One may note some peculiarities also in the Arthropoda taxon: in all three trees (Fig. 5a, 6, 7) it is subdivided into Ixodes/*Artemia* and *Penaeus*/Hexapoda, while in traditional taxonomy (Fig. 4) it includes ticks (Ixodes), insects (Hexapoda), and crustaceans (*Artemia* and *Penaeus*); these are taxa of same rank within the subtype Arthropoda.

Cladogram of gene trees produced using the maximal parsimony method (outsider taxa are shown in bold)

	Vertebrata	Arthropoda	Bryophyta, Rhodophyta, Stramenopiles
COX1	(*Petromyzon*, ((*Squalus*, *Salmo*), ((*Alligator*, *Xenopus*), (*Gallus*, (*Didelphis*, *Homo*)))))	(Ixodes, (*Penaeus*, (*Artemia*, (*Locusta*, (*Drosophila*, *Anopheles*)))))	((*Marchantia*, (*Reclinomonas*, (*Cyanidoschizon*, *Porphyra*))), (*Cafeteria*, (*Chrysodidymus*, *Phytophthora*)))
COX2	(*Alligator*, (*Petromyzon*, (((*Squalus*, *Salmo*), *Xenopus*), (*Gallus*, (*Didelphis*, *Homo*))))	(*Artemia*, (((*Locusta*, Ixodes), (*Penaeus*, (*Drosophila*, *Anopheles*))), **Mollusca**/**Annelida**))	(*Cyanidoschizon*, ((*Chrysodidymus*, *Reclinomonas*), ((*Porphyra*, *Marchantia*), (*Cafeteria*, *Phytophthora*))))
COX3	(*Petromyzon*, (((*Squalus*, *Salmo*), (*Alligator*, *Gallus*)), (*Xenopus*, (*Didelphis*, *Homo*))))	((***Marchantia***, Ixodes), (**Mollusca**/**Annelida**, (*Artemia*, (*Penaeus*, (*Locusta*, (*Drosophila*, *Anopheles*))))))	(((*Cyanidoschizon*, (*Allomyces*, (*Phytophthora*, *Porphyra*))), (*Chrysodidymus*, *Reclinomonas*)), *Cafeteria*); *Marchantia*
ATP6	(*Petromyzon*, (((*Squalus*, (*Salmo*, *Xenopus*)), (*Alligator*, *Gallus*)), (*Didelphis*, *Homo*)))	(Ixodes, (*Artemia*, (*Locusta*, (*Penaeus*, (*Drosophila*, *Anopheles*)))))	((*Porphyra*, *Cyanidoschizon*), ((*Reclinomonas*, *Marchantia*), (*Phytophthora*, (*Cafeteria*, *Chrysodidymus*))))
ATP8	No cluster formed	No cluster formed	No cluster formed
CYTB	(((*Squalus*, *Salmo*), *Xenopus*), (*Gallus*, ((*Alligator*, *Didelphis*), *Homo*))); *Petromyzon*	(Ixodes, (*Artemia*, (*Penaeus*, (*Anopheles*, (*Locusta*, *Drosophila*)))))	(((*Reclinomonas*, *Marchantia*), (*Cyanidoschizon*, *Porphyra*)), (*Cafeteria*, (*Chrysodidymus*, *Phytophthora*)))
ND1	(*Petromyzon*, (((*Squalus*, *Salmo*), *Xenopus*), ((*Alligator*, *Gallus*), (*Didelphis*, *Homo*))))	((Ixodes, *Artemia*), (*Penaeus*, (*Locusta*, (*Drosophila*, *Anopheles*))))	(*Cafeteria*, (*Chrysodidymus*, (*Phytophthora*, (*Cyanidoschizon*, (*Porphyra*, (*Reclinomonas*, *Marchantia*))))))
ND2	((*Petromyzon*, *Alligator*), (((*Squalus*, *Salmo*), *Xenopus*), (*Didelphis*, *Homo*)))	(((Ixodes, *Artemia*), **Annelida**), (*Penaeus*, (*Locusta*, (*Drosophila*, *Anopheles*))))	(*Chrysodidymus*, (*Cafeteria*, (*Phytophthora*, ((*Reclinomonas*, *Marchantia*), (*Porphyra*, *Cyanidoschizon*)))))
ND4	((*Petromyzon*, ***Asterina***), (*Xenopus*, ((*Alligator*, (*Squalus*, (*Salmo*, *Gallus*))), (*Didelphis*, *Homo*))))	No cluster formed	(((((*Reclinomonas*, *Marchantia*), (*Chrysodidymus*, *Metridium*)), (*Porphyra*, *Cyanidoschizon*)), *Phytophthora*), *Cafeteria*)
ND5	(*Petromyzon*, (*Alligator*, ((((*Squalus*, *Salmo*), *Xenopus*), *Gallus*), (*Didelphis*, *Homo*))))	(*Artemia*, (Ixodes, (**Mollusca**/**Annelida**, (*Penaeus*, (*Locusta*, (*Drosophila*, *Anopheles*))))))	((((*Reclinomonas*, *Marchantia*), (*Porphyra*, *Cyanidoschizon*)), *Phytophthora*), (*Chrysodidymus*, *Cafeteria*))
ND6	(*Alligator*, (*Petromyzon*, (*Gallus*, (((*Squalus*, *Salmo*), *Xenopus*), (*Didelphis*, *Homo*)))))	((**Mollusca**/**Annelida**, (Ixodes, *Artemia*)), (*Penaeus*, (*Anopheles*, (*Locusta*, *Drosophila*))))	(*Chrysodidymus*, (*Metridium*, (*Marchantia*, (*Reclinomonas*, (*Phytophthora*, (*Cyanidoschizon*, (*Porphyra*, *Cafeteria*)))))))

It should be noted that polyphyletic evolution of crustaceans was recently shown [17]. A tree was produced were *Penaeus* (shrimp) formed one taxon with insects, while *Daphnia* and *Artemia* were located separately.

In all three trees *Reclinomonas* forms one taxon with *Marchantia*, this taxon then forms a cluster with red algae (*Porphyra* and *Cyanidoschizon*). At the same time, taxon Stramenopiles is formed only within the TIQMAX tree (Fig. 5); it includes Bacillariophyceae (diatom algae), Bicosoecida (including *Cafeteria*), Chrysophyceae (golden algae, including *Chrysodidymus*), Oomycetes (including *Phytophthora*), Phaeophyceae (brown algae), Xanthophyceae (yellow-green algae), and a range of other taxa.

Traditional systems place the recently described flagellate *Reclinomonas* [14] close to *Cafeteria* [15], while the relationships of the taxa represented by *Cafeteria*, *Chrysodidymus* (golden alga), *Phytophthora* (oomycete), red algae, and liverworts remain under discussion (e.g., the group Stramenopiles is not always recognized [15]).

The table lists the results obtained for the trees mapped using the maximal parsimony method from amino acid sequences of individual mitochondrial proteins. These results show close relationship of fishes and amphibians: eight trees of eleven have a cluster (ray, salmon), and six have a cluster ((ray, salmon), frog). Lamprey is an outside member of this cluster in five trees. Mammals almost always form a cluster (opossum, human), while the relationship of

alligator and hen is quite uncertain, and the Archozauria taxon was formed only in three cases.

In eight trees, arthropods show clustering of *Penaeus* with insects. The relationship between the groups Bryophyta, Rhodophyta, and Stramenopiles is rather complicated: these three groups usually form one cluster, however, this cluster has unstable branching. Nevertheless, cluster (*Reclinomonas*, *Marchantia*) is formed in six trees and cluster (*Reclinomonas*, *Marchantia*, Rhodophyta) is formed in five trees.

We conclude that the analysis of gene trees shows discrepancies between reconciled species trees and the existing taxonomy of vertebrates and arthropods. These discrepancies do not result from errors in the reconciling algorithm, but rather reflect the properties of gene trees produced by standard methods. These discrepancies may be explained upon more detailed analysis of mitochondrial protein evolution in various taxonomic groups, this being beyond the scope of the present work.

At the same time, our results show the applicability of the method in case when the initial set contains the analyzed gene families in all taxa. Moreover, our results suggest relationship of *Reclinomonas*, an unclassified eukaryote, with green plants (Bryophyta).

ACKNOWLEDGMENTS

Software TIQMAX was developed by S. Shelayev and O. Zverkov.

The authors are very grateful to V.V. Aleshin for his advice concerning data processing technology and valuable scientific discussion of the obtained results.

This work was supported by the Russian Foundation for Basic Research (00-15-99362), INTAS (99-1476), LLCR (CRDF, RB0-1268), and HHMI (55000309).

REFERENCES

1. Weir, B., *Analysis of Genetic Data*. Translated under the title *Analiz geneticheskikh dannykh*, Moscow: Mir, 1995.
2. Waterman, M.S., *Introduction to Computational Biology*, Chapman and Hall, 1995.
3. Goodman, M., Czelusniak, J., Moore, G.W., Romero-Herrera, A.E., and Matsuda, G., *Syst. Zool.*, 1979, vol. 28, pp. 132–163.
4. Guigo, R., Muchnik, I., and Smith, T., *Mol. Phyl. Evol.*, 1996, vol. 6, pp. 189–213.
5. Page, R.D.M. and Charlstone, M.A., *Mol. Phyl. Evol.*, 1997, vol. 7, pp. 231–240.
6. Page, R.D.M., *Bioinformatics Application Notes*, 1998, vol. 14, pp. 819–820.
7. Eulenstein, O. and Vingron, M., *Arbeitspapiere der GMD*, Bonn, Germany, 1995, vol. 936.
8. Eulenstein, O., Mirkin, B., and Vingron, M., *J. Comput. Biol.*, 1998, vol. 5. 135–148.
9. Schieber, B. and Vishkin, U., *SIAM J. Comput.*, 1988, vol. 17, pp. 1253–1262.
10. Felsenstein, J., *Cladistics*, 1989, vol. 5, pp. 164–166 [http://evolution.genetics.washington. edu/phylip.html].
11. Page, R.D.M., *Mol. Phyl. Evol.*, 2000, vol. 14, pp. 89–106.
12. Page, R.D.M. and Charleston, M.A., *Mathematical Hierarchies in Biology*, DIMACS, 1997, vol. 37.
13. Wheeler, D.L., Chappey, C., Lash, A.E., Leipe, D.D., Madden, T.L., Schuler, G.D., Tatusova, T.A., and Rapp, B.A., *Nucleic Acids Res.*, 2000, vol. 28, pp. 10–14.
14. Flavin, M. and Nerad, T.A., *J. Eukar. Microbiol.*, 1993, vol. 40, pp. 172–179.
15. Kusakin, O.G. and Drozdov, A.L., *Filema organicheskogo mira* (Phylema of Organic World), St. Petersburg: Nauka, part 2, 1998.
16. V'yugin, V.V., Gorbunov, K.Yu., and Lyubetsky, V.A., *Problems of Control and Modeling in Complex Systems, Proc. 2nd Int. Conf.*, Samara, 2000, pp. 130–137.
17. Hwang, V.W., Friedrich, M., Tautz, D., Park, C.J., Kim, W., *Nature*, 2000, vol. 413, pp. 154–157.