DIMACS Technical Report 95-28
August 1995

A Biologically Meaningful Model for Comparing Molecular Phylogenies [1]

by

Boris Mirkin,[2]
DIMACS
Rutgers University
P.O.Box 1179 Piscataway NJ 08855-1179
and CEMI of Russian Academy of the Sciences, Moscow, Russia

Ilya Muchnik[3]
RUTCOR, Rutgers University,
New-Brunswick, NJ 08903

Temple F. Smith
Biomolecular Engineering Research Center,
College of Engineering, Boston University, Boston MA 02215

[1]The authors acknowledge the hospitality of the Center for Discrete Mathematics & Theoretical Computer Science (DIMACS), and also thank the organizers of the DIMACS Special Year in Mathematical Support for Molecular Biology, 1994-95 (in particular, they are pleased to acknowledge the support of NSF grants STC 91-19999 and BIR-94-12594 to Rutgers University).
[2]Research Associate
[3]Research Associate

DIMACS is a cooperative project of Rutgers University, Princeton University, AT&T Bell Laboratories and Bellcore.

DIMACS is an NSF Science and Technology Center, funded under contract STC–91–19999; and also receives support from the New Jersey Commission on Science and Technology.

ABSTRACT

In the framework of the problem of combining different gene trees into a unique species phylogeny, a model for duplication/speciation/loss events along the evolutionary tree is introduced. The model is employed for embedding a phylogeny tree into another one via so called Duplication/Speciation principle requiring that the gene duplicated evolves in such a way that any of the contemporary species involved bears only one of the gene copies diverged. The number of biologically meaningful elements in the embedding result (duplications, losses, information gaps) is considered a (asymmetric) dissimilarity measure between the trees. The model duplication concept is compared with that one defined previously in terms of a mapping procedure for the trees. A graph-theoretic reformulation of the measure is derived.

Contents

1 Introduction

The discipline of the phylogeny reconstruction, as based on the data on homologous biomolecules in a set of currently living species, started in sixties (see Fitch and Margoliash (1967)) and became a working tool for evolutionary analysis, both for theoretical and practical needs (see, for example, Nei (1987), Beanland and Howe (1992)).

The current strategy for the phylogeny reconstruction is based on the separate consideration of different gene families which are represented by so called homologous sequences. The homologous sequences are supposed to be descendants of a common ancestor in such a way that the more similar sequences correspond to the more recent common predecessors. An evolutionary tree constructed for a particular family or super-family serves as a structural support for analysis of the history of corresponding functions. This can be considered as the first step in the following two-step strategy for reconstructing evolutionary tree for species.

1. Constructing a phylogenetic tree for each of the gene families independently.
2. Combining particular gene trees into a unique tree representation.

This work focuses on Step 2: combining different, often contradictory, gene trees into a unique evolutionary tree called "species tree" (Fitch (1970), Goodman et al. (1979), Nei (1987)). The problem is considered in the literature as the problem of reconciling different phylogenetic trees. There are several ideas which have been suggested for the last two decades (see, for a representative sample, works by Robinson (1971), Waterman and Smith (1978), Smith and Waterman (1980), Robinson and Foulds (1981), Margush and McMorris (1981), Hendy, Little and Penny (1984), Mirkin and Rodin (1984), Day (1985), Day and McMorris (1985), Leclerc (1985), Adams (1986), Barthélemy, Leclerc and Monjardet (1986), Farach and Thorup (1993)). A common feature of all these methods is that they consider phylogenetic trees as formal mathematical objects and try to find as much alike in the structures as possible. The shortcoming of this approach, in the present authors' opinion, is that the concepts of "similar structures" involved are purely formal, having no independent biological meaning. For instance, there is no formal difference between inconsistencies that arise from different techniques applied to the same data or the difference in the data themselves.

To get a closer look at the matter, let us consider an example involving the following six taxa: 1. Arthropoda 2. Mammalia. 3. Echinodermata. 4. Protozoa. 5. Embryophita. 6. Clorophycea.

In Fig. 1(a) a phylogenetic tree is presented for these taxa, as reconstructed from the data on a protein family (tubulin α_1 chain) in Guigó, Muchnik and Smith (1994). Fig. 1(b) presents another plausible evolutionary history for the same taxa. What can be recommended for combining these trees via the mathematical approach discussed?

There are two major algorithmic approaches: (a) consensus, and (b) approximation. A consensus tree is a tree reflecting the features common to both of the original trees. Such a consensus tree usually is defined in terms of the clusters forming the original trees. The tree in Fig. 1(a) has 12, 123, 46, and 456 as its non-trivial clusters, while those for the tree in Fig. 1(b) are 23, 23456, 456, and 56. The only common cluster is 456; the corresponding consensus tree is presented in Fig. 1(c); we can see this as an example of

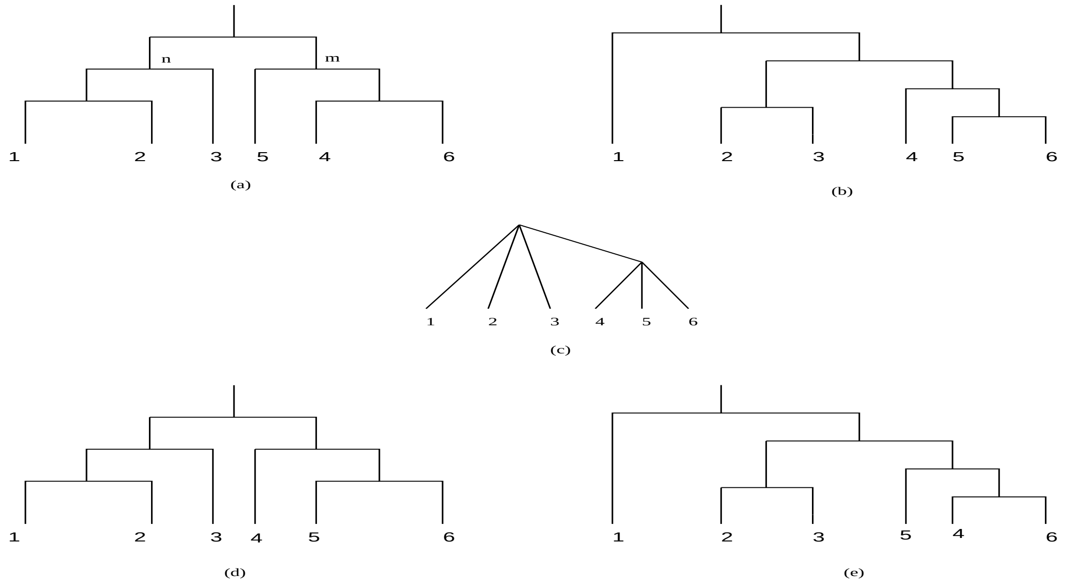

Figure 1: Results of the pure mathematical approach to combining trees (a) and (b): consensus tree (c) and approximation trees (d) and (e).

a tree having very little biological relevance, since no information is presented beyond 123 to 456 opposition. When there are more initial trees, there are fewer common clusters. We can modify the concept of “ a common cluster” to include those clusters which occur in, say, in 80 per cent of the initial trees. But this does not help too much: the pattern of the consensus tree remains very ambiguous in most cases. Another problem cannot be treated with this approach: different gene trees involve different subsets of the set of taxa analyzed. Restricting ourselves to the common subset of the taxa means eliminating all the other taxa, which drastically reduces the information available. Another possibility, an independent analysis of different subsets, may lead to contradictory “common” clusters (which can be considered the basis for so called “consensus impossibility theorems” [see, for example, Barthélemy, Leclerc and Monjardet (1986) and Margush and McMorris (1981)]).

The other, approximation approach, is supposed to overcome the ambiguity of the structure of the combined consensus tree. Based on a measure of similarity/dissimilarity between trees, the combined tree is sought as a tree of a specified structure maximizing its total resemblance (minimizing its total dissimilarity) to the original trees. Still, the measures employed have no explicit biological meaning. Let us consider, for instance, a popular dissimilarity measure, the set-theoretic distance, which is equal to the number of clusters occurring in exactly one of the two trees considered (see, for example, Robinson and Foulds (1981), Day (1985)). The distance between trees in Fig. 1 (a) and (b) is equal to 6 since each has three “unique” clusters. It can be proved rather easily that, in this case, either of the trees in Fig. 1 (a), (b) minimizes the total distance (equal to 0+6) from two of these trees considered as the input. Moreover, some other binary trees also are approximation trees minimizing the total distance (two of them are presented in Fig. 1, (d) and (e) as having their distances from the trees in Fig. 1(a), (b), equal to 2 and 4, respectively). We can see how the ambiguity in the structure of the consensus tree has been transformed into ambiguity of the solutions when we imposed the binary divergence structure. So these purely mathematical approaches cannot avoid great ambiguity in the results.

Are there any biologically meaningful dissimilarity measures possible? To address the issue, let us analyze biological meaning of the contradictions among the gene trees. Obviously, all the contradictions are in the pattern of the divergence. It is the phenomenon of divergence which should be presented and explained in the combined species tree (Fitch (1970), Goodman et al. (1979), Nei (1987)). Let us recall that the gene divergence can be the result either of speciation (gene divergence between species) or of duplication (gene divergence within species) (Ohno (1970)). If gene divergence occurs only with speciation, the gene and species trees are identical: the phylogeny of the gene follows the speciation history. However, if a gene is duplicated, gene divergence may occur without subsequent speciation and the gene phylogeny does not necessarily follow the pattern of the species phylogeny. This happens with so called *paralogous* evolution events representing gene lineages that arose from gene duplication prior to species separation (Fitch (1970)). Taking into account the paralogous evolution allows us fitting any gene lineages within their species lineages as explained with corresponding gene duplications hypothetically making the difference between gene and species topologies. This idea was elaborated by Goodman et al. (1979) and put in a wider biological context by Guigó, Muchnik and Smith (1994) who made the first attempt to identify and precisely locate the gene duplications in eukariotic history. These authors have developed some formalisms for annotating the evolutionary tree with duplication events by means of so called mapping of the trees (this concept was employed also in Page (1994)). The number of duplications and losses posited is a dissimilarity measure to be minimized by the reconstructed evolutionary tree. (The dissimilarity measure considered by Goodman et al. (1979) includes also the number of the nucleotide replacements needed).

Such a development of the approximation approach has two important advantages in applications. First, it can be applied when the data on different single genes involve different subsets of species; this is a typical situation which can hardly be handled by the mathematical approaches cited above. Second, the reconstructed tree is annotated with information specifying mutation events along the tree, thus giving biological information.

However, the tree mapping concept itself (implicitly considered in Goodman et al. (1979), and explicitly in Page (1994) and Guigó, Muchnik and Smith (1994)) turns out to be too simplistic and formal (as demonstrated below in Section 4). Moreover, the descriptions of mutation events in Goodman et al. (1979) and Guigó, Muchnik and Smith (1994) are rather technical: the definitions of the duplication and subsequent events (speciation/losses in both papers and some more complicated concepts in Goodman et al. (1979)) are not substantiated with any biologically meaningful model, which makes all their constructions fairly controversial.

The primary goal of this work is to propose a biologically meaningful model for duplication/speciation/loss events for explaining and counting the inconsistencies between a single gene and species trees.

The remainder of the report consists of four Sections. Section 2 introduces the main concepts of gene duplication, loss, and information gap in graph-theoretic terms. Those concepts are a biological context for presented method for embedding the gene phylogenies in the species tree discussed in Section 3. The procedure presented involves comparison of every subtree of a particular gene tree with corresponding subtree in the species tree. Every time when the comparison implies an inconsistency, a duplication of the gene is postulated to explain the inconsistency. This duplication event then is reflected in the species tree with the event's history leading to the current situation in currently living species; the history along the tree involves certain gene copy losses and information gaps which accompany the duplication event during the evolutionary history. The total number of duplication/loss/gap events required to explain all the inconsistencies is considered the inconsistency measure between the gene and species tree. In Section 4, equivalence between this inconsistency measure and that of Guigó, Muchnik and Smith (1994) is discussed (the measure suggested by Goodman et al. (1979) seems more complicated). It turns out that the concept of mapping between the trees used in the latter papers, does not allow, in general, properly classifying corresponding loss/gap events, although it takes into account the duplication events appropriately. A conjecture is made that the two indices coincide, which has been strongly supported with numerous testing computations. In Section 5, a set-theoretic reformulation of the inconsistency measure suggested is given, which connects it with so called incompatible pairs of the nodes. This reformulation simplifies the measure since it allows to catch all the duplication/loss/gap events in one run over the trees without that multiple history annotating process along all the inconsistent subtrees which is defined in the original model in Section 3. On the other hand, the reformulation has a biological meaning since the concept of the incompatible pair of the nodes corresponds to those of the ancestor species who keep both of the gene copies diverged. In Section 6, the results are briefly discussed in the context of the problem of reconstructing the molecular phylogeny.

2 Model for Duplications in Phylogenies

The model for evolutionary history will be rooted binary tree with the leaf set labeled by biological taxa. In this text, we will identify the leaves with their labels thus considering the N-element set of the labels I as the set of the taxa. Any interior node n is considered a subset (cluster) of its subordinate leaves, interpreted as the descendants of the common ancestor of the leaves represented by node n. For instance, in tree (a), Fig. 1, node n represents leaf set consisting of $1, 2$, and 3, which will be denoted $n = 123$ (thus relaxing, for the sake of simplicity, both curled brackets and commas in the traditional set-theoretic denotation, $n = \{1, 2, 3\}$, in this case). Similarly, node m represents cluster $m = 456$. This allows representing the tree as the set of its leaf clusters n. Tree (a), Fig. 1, is thus represented by its clusters $12, 123, 46, 456$ corresponding to the interior nodes, and $I = 123456$ corresponding to the root. The leaves themselves are considered trivial singleton clusters. This allows us using denotation $n \in T$ for a vertex n of tree T since T is just a set of the clusters corresponding to the nodes in a drawn representation of the tree. The equivalence of the nodes (vertices) and corresponding clusters will be exploited through all the text below. The major evolutionary relation "m is a descendant of n" is expressed, in set-theoretic setting, just as "$m \subset n$"(which indicates strict inclusion, in contrast to denotation $m \subseteq n$ allowing equality of m and n).

The two immediate descendants (children) of n are denoted $a(n)$ and $b(n)$ (also considered as the clusters of the subordinate leaves). In tree (a), Fig. 1, clusters 123 and 456 are the children of the root cluster I while the latter child itself has its children 46 and 5. Assigning the symbols $a(n)$ and $b(n)$ to the siblings is quite arbitrary.

If $n_1, n_2, ..., n_l$ $(l \geq 3)$ is a path (sequence of different nodes) connecting node n_1 with its descendant n_l, the node clusters in a path are obviously nested: $n_l \subset n_{l-1} \subset ... \subset n_1$ and node n belongs to the path between n_1 and n_l if and only if $n_l \subseteq n \subseteq n_1$; n is called *intermediate* if $n_l \subset n \subset n_1$. The siblings of the vertices $n_2, ..., n_l$ in a path $n_1, ..., n_l$ will be referred to as *collateral* ones.

Before introducing mathematical formalisms, let us consider how a duplication history can be annotated in phylogenetic tree. Let us consider a five-taxa phylogeny presented in Fig.2 (a). The black and white circles show two copies of a gene A emerged sometime before the root ancestor appeared. Then taxon 1 diverged with only one, black circle, copy left while the other, white circle, was lost (the loss is shown with an x-mark). The others, cluster 2345, kept both of the gene copies. Then taxa 2 and 345 diverged, again speciated, that is, keeping one of the copies evolved each while the other (black circle in 2 and white circle in 345) was lost in the contemporary taxa. The descent of the speciated ancestral taxon 345 conserved the black circle copy of gene A (perhaps along with less significant changes). Fig. 2 (b) presents a bit more complicated story when we have a partial information on gene B copies in contemporary taxa: 1 and 2 have black circle copies, 3 has white circle copy, while nothing is known on the B copy contents in the other taxa, 4, 5, and 6, the more so about their recent ancestors represented by clusters 56 and 456. Considering that in the evolutionary framework, we can say that the gene B duplicated transcended to 23456 ancestor diverged then in two siblings, one, 23, still keeping both of the copies, and the other, 456, lacking any information about its gene B stock, which is reflected in the question mark in Fig. 2 (b). This kind of information lacking is a frequent situation which should be reflected in mathematical models.

The three kinds of duplication history elements presented in Fig. 2, (a) and (b), – duplication itself, loss, information gap – will be considered the only information needed to explain the difference between gene trees.

Still, the duplication histories shown in Fig. 2, (a) and (b), with their dotted lines and circles and crosses and question marks, seem too picturesque to be represented mathematically as they are. More formal pictures representing the same histories are shown in Fig. 2, (c) and (d), where the circles are substituted by signs, + (black circle) and - (white circle), the loss and question marks, by boxes in corresponding vertices, and the dotted lines are just omitted (since they do not bear any supplementary information). Actually, the box marks

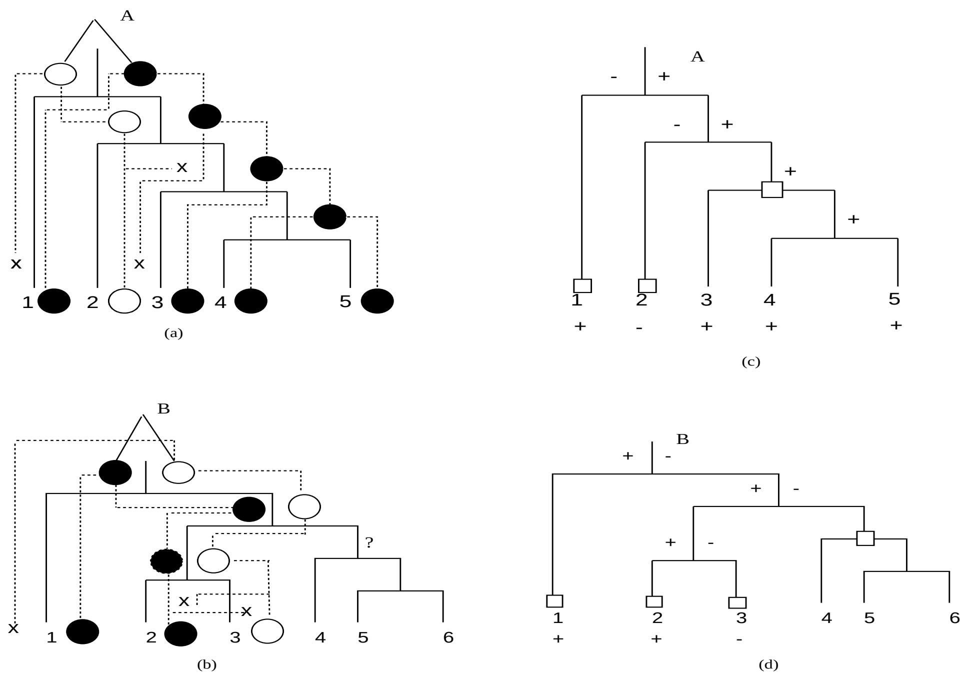

Figure 2: The evolutionary history of duplications A (a) and B (b) represented with sign labeled trees (c) and (d), respectively.

also could be omitted since the sign label pattern indicates their location quite clearly: they correspond to maximal speciated (loss) or unlabeled (information gap) clusters in the tree. Thus, it is only the pattern of sign labeling must be presented in duplication modeling to have all the duplication history represented in the tree. This leads us to the following definition involving the denotations: a binary rooted tree T with leaf set I representing a species phylogeny; set $T(n)$ of all its vertices (leaves and n included) being subordinates of node n; that is, in set-theoretic setting, $T(n) = \{m \in T : m \subseteq n\}$ is the subtree of T having n as its root; and set $L = \{\emptyset, \{+\}, \{-\}, \{+,-\}\}$ of all the subsets of the sign set $\{+,-\}$.

Definition 1 *A mapping $\delta : T(n) \to L$ will be referred to as a gene duplication in node $n \in T$ iff (1) it is monotone, that is, $\delta(m) \subseteq \delta(n)$ when $m \subseteq n$, and, (2) it is saturated, that is, $\delta(n) = \{+,-\}$. Vertex $m \in T(n)$ is called mixed if $\delta(m) = \{+,-\}$, speciated if $\delta(m) = \{+\}$ or $\delta(m) = \{-\}$, and gapped if $\delta(m) = \emptyset$.*

The definition means that the duplication δ emerges in node n (condition 2) and then evolves through all its descent by only losing the copies sometimes, one (speciation) or both (lack of information) of them, as represented by the signs "+" and "-".

As we have seen in Figure 2 (see also Fig.3 and 4 where the duplications occur in interior vertices, not in the root only), the maximal mixed, speciated and gapped vertices have particular evolutionary meaning: the maximal mixed vertex n represents the duplication event itself; the maximal speciated vertices correspond to losses; and the maximal gapped vertices correspond to the ancestors with completely unknown duplicated gene stock. The total number of these maximal vertices will be called complexity $c(\delta)$ of duplication δ; it counts for the total number of the evolutionary distinct events connected with the duplication (one for the duplication plus the number of losses plus the number of information gaps). In Figure 2, the complexity of duplication A (a) is 4 (1 duplication + 3 losses) and of B (b) is 5 (1 duplication + 3 losses + 1 gap).

When a species tree T is supplied with a set of duplications $\Delta = \{\delta_1, ..., \delta_p\}$ (related perhaps to different vertices in T), their total complexity is expressed as just the sum of the complexities of the individual duplications, $c(\Delta) = \sum_{k=1}^{p} c(\delta_k)$, since there is no evolutionary mechanism is presumed here for interaction among the duplications.

We consider that there is no information on the duplications available beyond the considerations related to inconsistencies among the different gene trees. This makes us narrowing the class of duplications considered to catch the features of the losses and gaps obtained due to the inconsistencies between the species tree and single gene trees. This leads us to introduce the following definition.

Definition 2 *A mapping $\delta : T(n) \to L$ will be referred to as an operational duplication in node $n \in T$ iff*

1. δ is a duplication,

2. any singleton leaf cluster is either speciated or gapped,

3. for any gapped singleton leaf cluster $i \in n$, $i \in m \subset n$ implies m is either gapped or mixed.

In other words, a gene duplication is an operational duplication if is satisfies the two supplementary conditions: there is no contemporary taxon keeping both of the duplicated gene copies (2), and the information gap may not involve any preliminary speciation (3). These conditions, actually, assume a kind of Occam's Razor for the hypothetical duplications explaining the inconsistencies observed and will become quite clear in the context of the tree comparing procedure presented in the next Section. At the first glance, item (2) imposes certain biological restrictions implying that no species in a particular gene tree may relate to two or more copies of the same gene. However, nothing prevents us from considering any of those copies as specific different genes leaving the restriction only for the hypothetical ones to be introduced to explain the inconsistencies.

Yet a relaxed version of the duplication concept is needed to be applied when a gene involves just a part of taxa in the species tree. A duplication δ will be referred to as a *partial duplication* if it is not defined for all the nodes of subtrees of form $T(l)$ for some $l \in T(n)$.

3 Method for Comparing a Gene Tree with a Species Tree

3.1 Root Inconsistency and Duplication

Let T be a species tree or one of its subtrees and G be a single gene tree such that the leaf set J of G is included in the leaf set I of T: $J \subseteq I$. These trees will be referred to as *root-consistent* if each of the sets of the leaf descendants $a(g), b(g)$ of the children of the root g in G is contained in a child-set $a(t)$ or $b(t)$ of the root t of T, and as *root-inconsistent*, otherwise. The root-consistency means that the "root branches" of the tree G do not contradict the "root branches" of T; an event causing the divergence of the root descendants occurring in the species tree is reflected in the divergence of the specific gene family which corresponds to gene tree G.

If the trees G and T are root-inconsistent, a duplication event in the root of T can be defined to explain the inconsistency. To cope with the fact that duplication does not necessarily alters the topology, we need a postulate as follows.

Duplication/Speciation Principle. *Root-inconsistency of the trees G and T means that a duplication event in the gene corresponding to tree G occurred at the root of the species tree T and evolved in T in such a way that contemporary organisms corresponding to the leaves in $a(g)$ have one of the divergent gene copies and the leaves in $b(g)$ the other.*

The principle gives unambiguous recommendations for assigning the diverged copies to the elements of J while leaving $I-J$ subject to an arbitrary decision. Let us define a simplest mapping $\delta_g : T \to L$ satisfying the principle:

1. for any $i \in I$, $\delta_g(i)$ is defined according to the following rule

$$\delta_g(i) = \begin{cases} + & if \quad i \in a(g) \\ - & if \quad i \in b(g) \\ \emptyset & if \quad i \in I - J; \end{cases}$$

2. $\delta_g(n) = \{+\}$ if $n \cap a(g) \neq \emptyset$ and $n \cap b(g) = \emptyset$;

3. $\delta_g(n) = \{-\}$ if $n \cap b(g) \neq \emptyset$ and $n \cap a(g) = \emptyset$;

4. $\delta_g(n) = \{+, -\}$ if $n \cap a(g) \neq \emptyset$ and $n \cap b(g) \neq \emptyset$;

5. $\delta_g(n) = \emptyset$ if $n \subseteq I - J$.

These values $\delta_g(n)$ $(n \in T)$ can be constructively defined just in one bottom up run through tree T.

For a set Δ of mappings $\delta_g : T \to L$, let us recall that $\delta \in \Delta$ is minimum in Δ if, for any other $\gamma \in \Delta$, $\delta(n) \subseteq \gamma(n)$ for any $n \in T$. As it is well known, the minimum is unique if exists.

Theorem 1 *Mapping δ_g is the minimum duplication in the set of all operational duplications satisfying the Duplication/Speciation principle.*

Proof: The root of T is mixed since it corresponds to set I which includes $J = a(g) \cup b(g)$ and, therefore, contains elements of both of the clusters, $a(g)$ and $b(g)$. Let us prove that if a node is speciated or gapped, then its subordinate nodes (leaves included) have the same label pattern. Indeed, for any subordinate n' of non-mixed n, the inclusion $n' \subset n$ holds, which proves the property. This proves that the mapping δ_g is a duplication. The item (3) in Definition 2 follows from item (1) in definition of δ_g, while item (4) in Definition 2 follows from the fact stated above that the speciated clusters have all their descent equally speciated. Thus, mapping δ_g is an operational duplication.

To prove minimality of δ_g, let us consider any node $n \in T$ which satisfy either $n \cap J \neq \emptyset$ or $n \in I - J$. If $n \in I - J$, then $\delta_g(n) = \emptyset$ by definition, which guarantees that $\delta_g(n)$ is minimum. Otherwise, n is speciated, thus, being minimum by the Duplication/Speciation principle, or mixed one containing both kinds of leaves, from $a(g)$ and $b(g)$, which is necessarily follows from monotonicity of δ and the Duplication/Speciation principle. The theorem is proved. $\square$.

Thus, δ_g is the only duplication explaining the root-inconsistency to satisfy the Duplication/Speciation principle in the most parsimonious way.

Complexity $c(\delta_g)$ shows the extent of the biologically meaningful difference (that is, the duplication, losses and gaps required to) between G and T, due to their root-inconsistency, and will be denoted as $c(g, T) = c(\delta_g)$.

3.2 Embedding a Gene Tree into Species Tree

To compare the entire gene tree G with species tree T, we need comparing all the subtrees of G with T, which will be done, initially, for the case when $J = I$ (the other case, $J \subset I$, is treated in the next subsection).

The comparing procedure consists of sequential comparisons of the gene tree subtrees $G(n)$ (by all nonsingleton $n \in G$) with those subtrees of T_G, the leaf sets of which include

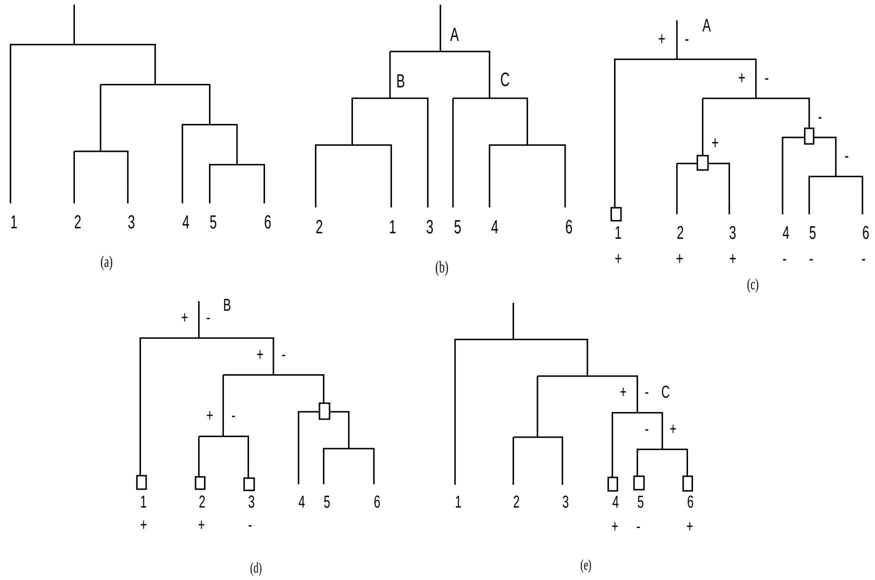

Figure 3: Embedding gene tree (b) onto tree (a) requires three duplications corresponding to nodes A, B, and C in (b); the minimum duplications δ_A, δ_B, and δ_C are shown in (c), (d), and (e), respectively; the losses and information gaps are represented with square boxes.

n. If $G(n)$ is root-consistent with a subtree $T(m)$ of T, we proceed to subtrees of $T(m)$. When $G(n)$ is root-inconsistent with a subtree $T(m)$, the minimum duplication mapping $\delta_{nm} : T(m) \to L$ is defined as the operational duplication of gene G in vertex $m \in T$ to explain the inconsistency.

Let us point out, that the resulting set of duplications δ_{nm} does not depend on the order in which the subtrees $G(n)$ are considered, which makes the procedure just a multiple defining the minimum duplication mappings requiring at most $O(N^3)$ steps (of comparing two species) where $N = |I|$: any of $O(N)$ subtrees $G(n)$ is compared with at most $O(N)$ subtrees $T(m)$, each comparison involves checking for the root-consistency only requiring thus $O(N)$ steps at most; defining the duplications takes at most N^2 steps which are added.

The total inconsistency $C(T/G)$ between T and G can be defined quantitatively through all the duplication events, losses, and information gaps counted across all the tree T, which means that $C(T/G)$ is the sum of the complexities of all the duplications δ_{nm} emerged, $C(T/G) = \sum_{n \in G} C(n,T)$. Obviously, values $C(T/G)$ and $C(G/T)$ can be different (which is illustrated in Fig. 3 and 4 below).

In such a way, an evolution history of gene G is displayed on the evolution tree T with a set of duplications along with corresponding losses and gaps as presented in Fig.3. In the picture, a real-life pattern of a gene tree (b) is shown as compared to the corresponding part of the species tree (a) (see Fig. 1). There are three subtrees, in Fig.3 (b), shown by their roots A, B, and C, which are root-inconsistent with the corresponding subtrees in the species

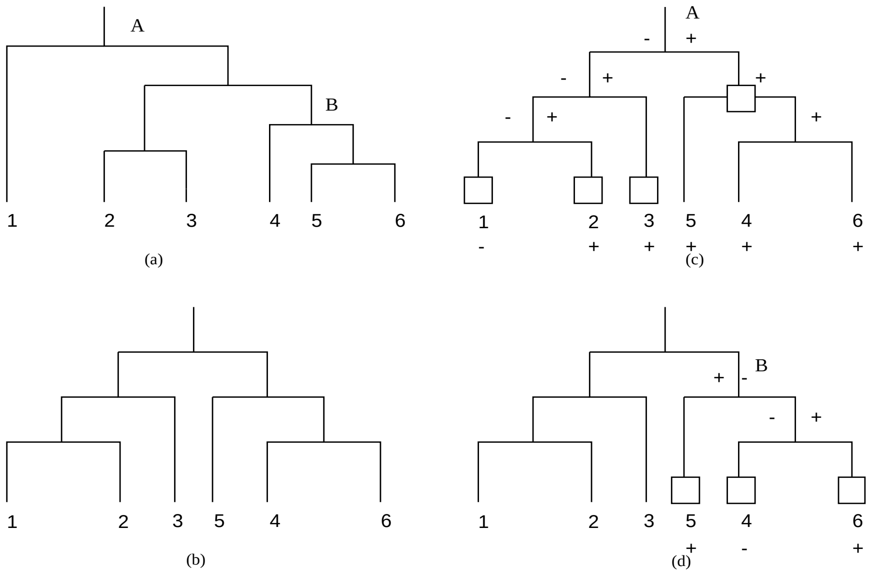

Figure 4: Embedding tree (a) in tree (b) requires two duplications corresponding to nodes A and B in (a); the duplications are presented in (c) and (d), respectively, with square boxes for the losses.

tree implying corresponding duplications, as shown in Fig.3, (c), (d), and (e), respectively. The maximal speciated nodes for each of the duplications are shown in Fig.3 (c, d, e) with square boxes; the node 456 in Fig.3 (d) is gapped since the gene subtree of duplication B involves only three taxa while it is root-inconsistent with the entire species tree, and it is also square-boxed. Thus, the total number of the duplications is equal to 3, the number of the losses to 9, and the number of the gaps to 1, which gives $C(a/b) = 13$ as the measure of inconsistency between the trees in Fig.3 (a) and (b). In Fig.4, embedding tree (a) in tree (b) is shown; it requires only 2 duplications and 7 losses; no information gap is involved, which makes the overall inconsistency measure $C(b/a) = 9$.

3.3 Embedding a Smaller Gene Tree

Embedding gene tree G in species tree T with $J \subset I$ may involve an ambiguity if J is significantly smaller than I. The ambiguity arises when the gene leaves are insufficient to distinguish between different clusters in the species tree: vertices $m, n \in T$ may be different while $m \cap J = n \cap J$. Let us develop a corresponding formalism.

Let us denote $n_J = n \cap J$, for any $n \in T$, and call the vertices $m, n \in T$ J-equivalent if $m_J = n_J$. Since the relation of J-equivalence is obviously an equivalence relation, it partitions I in nonoverlapping classes of J-equivalent vertices. The class of the vertices $n \in T$ having $n \cap J = n_J$ will be denoted $< n_J >$. Let us refer to the class $< \emptyset >$ consisting of vertices n with $n \cap J = \emptyset$ as the *vanishing* class.

Theorem 2 *Any not vanishing J-equivalent class $< n_J >$ is represented with a path (in tree T), all the collateral siblings of which belong to the vanishing class $< \emptyset >$.*

Proof: Indeed, $m, n \in < n_J >$ implies that $m \cap n \neq \emptyset$ and, thus, $m \subset n$ or $n \subset m$. It follows from that that all the intermediate vertices also belong to $< n_J >$. The collateral siblings are collateral children of all the nodes of the path, except for the minimum one, none of whose children belongs to the path. Let, say, $n \in < n_J >$ has $a(n) \in n_J$, that is, $n_J = a(n)_J$. This implies that the sibling, $b(n)$, has no elements from J at all, that is, $b(n)_J = \emptyset$. The theorem is proved. □.

Such a simple structure of J-equivalent classes allows easily "cutting out " all the leaves $I - J$ from T with simultaneous contracting the classes in just nodes n_J; all the nodes from the vanishing class are cut out, too. The tree contracted, actually, consists of the nonempty node clusters n_J representing not vanishing J-equivalent classes $< n_J >$. This contracted species tree will be denoted T_G. The contracting procedure takes $O(N^2)$ steps at most, and does not change the order of complexity of the whole procedure (moreover, the subtree comparing procedure can be organized without computing T_G itself: just, in any comparing step, belonging of the entities to J is to be checked preliminarily).

Since the leaf sets of G and T_G coincide, the embedding procedure above can be performed for gene tree G and species tree T_G along with all the duplications defined for vertices $n_J \in T_G$.

After that any duplication in T_G must be unfolded, that is, expanded onto tree T. To do that, we need some more symbols. For a duplication δ_k in node $k \in T_G$, let $T_k = \cup_{n_J \in T_G(k)} < n_J >$; so, T_k consists of those nodes of T which are J-equivalent to some $n_J \in T_G(k)$.

Definition 3 *A partial duplication $\delta_m : T(m) \rightarrow L$ (in T) is referred to as unfoldment of a duplication $\delta_k : T_G(k) \rightarrow L$ (in T_G) if (1) $m \in T_k$, (2) $\delta_m(n)$ is not defined when $n \in < \emptyset >$ where $< \emptyset >$ is the vanishing class, (3) for every $n_J \in T_G(k)$, $\delta_m(n) = \delta_k(n_J)$ for some $n \in < n_J >$.*

Condition (3) (along with the monotonicity) guarantees that the evolutionary history of δ_m in T is consistent with that of δ_k in T_G; condition (1) says that the history begins in a proper place; and condition (2) provides no history for those species and their ancestors that are not presented in the gene tree G.

It appears, no duplication unfolding can change its complexity.

Theorem 3 *The complexity of a duplication δ_k in T_G coincides with that of any of its unfoldments δ_m in T: $c(\delta_k) = c(\delta_m)$.*

Proof: Let us consider, in T, J-equivalent classes $< n_J >$ and $< p(n_J) >$ for a node n_J and its parent $p(n_J)$ from T_G. As it follows from Theorem 2, they consist of all the nodes $n_1, ..., n_l$ of a path in T, which implies that $\delta_m(n_l) \subseteq ... \subseteq \delta_m(n_1)$ since $n_l \subset ... \subset n_1$ and δ_m is monotone. If n_J is a maximal speciated node (under δ_k, say, $\delta_k(n_J) = \{+\}$), then

$\delta_k(p(n_J)) = \{+,-\}$. Thus, $\delta_m(n') = \{+\}$ and $\delta_m(n'') = \{+,-\}$ for some n', n'' in the path, which implies that there is one and only one maximal speciated (by +) node in the path consisting of $< n_J > \cup < p(n_J) >$. The proof that maximal gapped node n_J corresponds to one and only one maximal gapped node $n \in < n_J > \cup < p(n_J) >$ is quite similar, which proves the theorem. □.

The proof of the Theorem, actually, shows the structure of the unfoldments of arbitrary duplications. When δ_k in T_G is an operational duplication, its operational unfoldments have even simpler structures.

Theorem 4 *For any operational unfoldment δ_m of an operational duplication δ_k in $k_J \in T_G$, $\delta_m(n) = \delta_k(n_J)$, for any $n \in < n_J >$, where n_J is any node $n_J \in T_G(k_J)$ such that $n_J \neq k_J$.*

Proof: If $n_J \in T_G(k_J)$, $n_J \neq k_J$, is mixed, then, for its children, $\delta_k(a(n_J)) \neq \delta_k(b(n_J))$. This implies that no node $n \in < n_J >$ may be not mixed: if n is gapped or speciated, all its descent must have the same pattern of δ_m, which contradicts the assumption. If $n_J \in T_G(k_J)$ is speciated or gapped, both of its children must have the same pattern of δ_k, which implies that no $n \in < n_J >$ may have a different pattern under δ_m: the mixed pattern is impossible by the monotonicity, and different pattern of speciation (or information gap) contradicts to the assumed pattern of the children $a(n_J)$ and $b(n_J)$. This proves the statement.□.

The Theorem shows that operational unfolding an operational duplication δ_k in T_G may have an ambiguity only in location of the duplication event itself, and justifies the following procedure.

Unfolding Procedure

1. For the class $< k_J >$ of the nodes of T that are J-equivalent to the duplication-event node k_J itself, pick up any node $m \in < k_J >$ and define it as the node where the duplication emerged, thus putting $\delta_m(m) = \{+,-\}$. Take all the descendants $n \in < k_J > \cap T(m)$ of m in $< k_J >$, and define $\delta_m(n) = \{+,-\}$ for all of them.

2. For any other class $< n_J >$, $n_J \in T_G(k_J)$, define $\delta_m(n) = \delta_k(n_J)$, for any $n \in < n_J >$.

3. If $n \in < \emptyset >$ (that is, $n \subseteq I - J$), leave $\delta_m(n)$ not defined.

Thus, there is only one kind of the ambiguities: location of the duplication mutation event since any $m \in < k_J >$ may be used as that. The other important observation is that the complexity of the unfolded duplication does not change, $C(T/G) = C(T_G/G)$, which allows us using the contracted tree T_G for counting the inconsistency measure $C(T/G)$.

4 The Duplications Defined in Terms of Mapping of the Trees

Let us consider two phylogenetic trees, G and T, having coinciding leaf sets, which was proved above sufficient for measuring the inconsistency between them. For any node $g \in G$, let us denote as $M(g)$ the node of T being its least common ancestor, that is, the smallest cluster satisfying inclusion $g \subseteq M(g)$. This map M, considered first by Goodman et al. (1979), is referred to as *mapping of G into T* in Page (1994) and Guigó, Muchnik and Smith (1994). Evidently, if $g' \subseteq g$ then $M(g') \subseteq M(g)$.

In Goodman et al. (1979) and Guigó, Muchnik and Smith (1994), a duplication concept is introduced in terms of mapping M. To distinguish it from that defined above, we refer to it as to M-duplication.

Definition 4 *An M-duplication is an interior node $g \in G$, such that M(g)=M(a(g)) or/and M(g)=M(b(g)).*

This concept exactly corresponds to the duplication event defined when corresponding subtrees of G and T are root-inconsistent.

Theorem 5 *Node $g \in G$ is an M-duplication if and only if subtrees $G(g)$ and $T(M(g))$ are root-inconsistent.*

Proof. Let $t = M(g) = M(a(g))$. It means that $g \subseteq t$ but neither $a(t)$, nor $b(t)$ contains $a(g)$; that is, $a(g) \cap a(t) \neq \emptyset$ and $a(g) \cap b(t) \neq \emptyset$. Therefore, the subtrees $G(g)$ and $T(M(g))$ are root-inconsistent. Reversely, let subtrees $G(g)$ and $T(M(g))$ are root-inconsistent. This means that at least one of the children of g, say, $a(g)$, overlaps both of the children of $M(g)$, say, a and b. This implies that $M(a(g)) = M(g)$ since $M(a(g))$ can be neither a, nor b. The proof is over.$\Box$.

Let us define the mutation cost function associated to any non-terminal node of G, as it was introduced (under the name of "loss function") by Guigó, Muchnik and Smith (1994), although in slightly reformulated terms.

Definition 5 *The cost L(g) associated with $g \in G$ is defined as: (a) the total number of the intermediate nodes between $M(a(g))$ and $M(g)$ and between $M(b(g))$ and $M(g)$ if both $M(a(g))$ and $M(b(g))$ are proper subsets of $M(g)$; (b) the number of the intermediate nodes between $M(a(g))$ (or, respectively, $M(b(g))$) and $M(g)$ plus one if only $M(a(g))$ (or, respectively, $M(b(g))$) is a proper subset of $M(g)$ while the other child is mapped at $M(g)$; (c) zero, if both of the children of g are mapped at $M(g)$, that is, $M(a(g)) = M(b(g)) = M(g)$.*

The total cost function $c(G,T)$ associated with the mapping of G into T was defined in Guigó, Muchnik and Smith (1994)) as the sum of all penalties $L(g)$ by non-terminal $g \in G$ and of the number of all M-duplications. In numerous testing examples, without any exception, this cost function values have appeared coinciding with the values of the inconsistency function as defined above. This (along with certain theoretical considerations) leads us to believe that the following statement is true.

Conjecture. For any phylogeny trees G and T defined with the same leaf set, $c(G,T) = C(T/G)$.

If the conjecture is proved, use of the cost function $c(G,T)$ in the general evolutionary tree reconstruction will be justified theoretically.

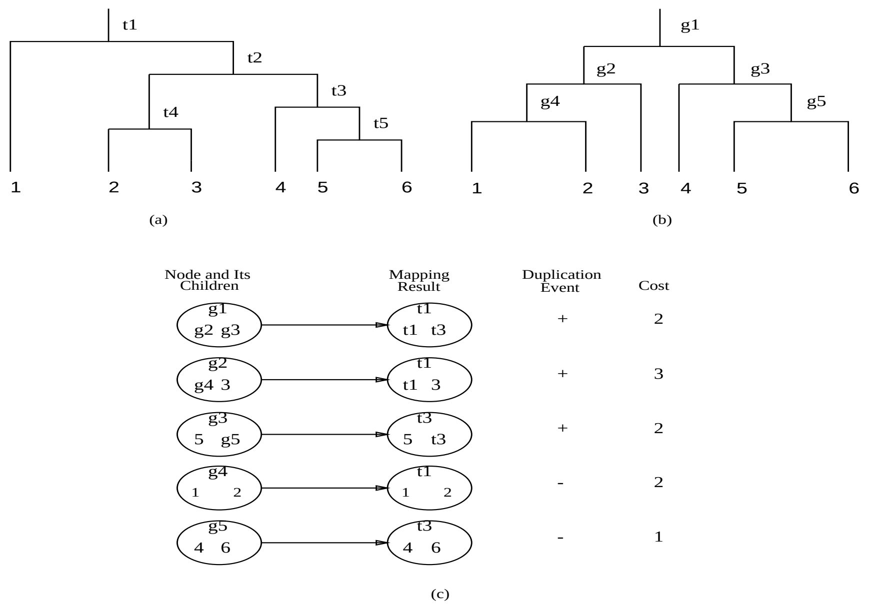

Figure 5: Mapping gene tree (b) into tree (a) as presented in (c) along with duplications and costs counted to give the total cost value 3+10 = 13.

Although the number and the location of the duplication events are captured by the mapping concept properly (see Theorem 5 above), the losses, as based on the model developed in this work, generally speaking, cannot be recognized with the cost function, which can be demonstrated with the example of trees in Fig.4, (a) and (b). In Fig.5 (c) all the results of mapping of the tree in Fig.5 (b) into tree in Fig.5 (a), are shown. We can see that while mapping node g4 (along with its children, 2 and 1) into t1 (along with the same children, 2 and 1), two intermediate nodes, t2 and t4, are counted as the cost associated with g4. Obviously, those nodes, t2 and t4, may not be assigned to the same duplication event since if it could have been done, it would have given values 2+2=4 to the duplication event A, or 3+2=5 to B, or 2+2=4 to C in Fig.2, none of which is true.

The information gap concept is missed completely in the theories based on the mapping of the trees, although the fact that a gap (or some gaps) must be present in the minimum duplication can be reformulated in the mapping terms, as follows.

Theorem 6 *An M-duplication $g \in G$ corresponds to the minimum duplication containing no gaps if and only if $g = M(g)$.*

Proof: Indeed, $g = M(g)$ means that there is no leaves, in subtree $T(M(g))$, beyond $g = a(g) \cup b(g)$, which implies that all the leaves in $T(M(g))$ are speciated in duplication δ_g. On the other hand, if no leaf $i \in T(M(g))$ has $\delta_g(i) = \emptyset$, then $T(M(g)) = a(g) \cup b(g)$, which proves the statement.□.

So, if, for an M-duplication g, $g \subset M(g)$, all the leaves from $M(g) - g$ will be gapped under the corresponding duplication, although the mapping itself cannot provide us with the information how many maximum gapped vertices are due to δ_g.

5 A Set-Theoretic Reformulation of the Inconsistency

For any two phylogenetic trees G and T (having their leaf sets coinciding) and for any two interior nodes $g \in G$ and $t \in T$, let us refer to those nodes as *incompatible* ones, if the statement in (1) holds:

$$g \not\subseteq t \text{ and } a(g) \cap t \neq \emptyset \text{ and } b(g) \cap t \neq \emptyset \tag{1}$$

Obviously, if g and t are incompatible and $p(t)$ is the parent of $t \in T$, then g and $p(t)$ will be also incompatible unless $g \subseteq p(t)$. An incompatible pair (g, t) will be referred to as *maximal incompatible pair* if it satisfies that last property, that is, $g \subseteq p(t)$. Obviously, for any incompatible pair (g, t), there exists a maximal incompatible pair (g, t'), where $t \subseteq t'$, which is defined in a unique way.

The importance of those maximal pairs in the present context follows from the following observation.

Theorem 7 *For any trees G and T, if a pair (g, t) is maximal incompatible then trees $G(g)$ and $T(p(t))$ are root-inconsistent. Reversely, if those trees are root-inconsistent while*

$g \subseteq p(t)$, then either (g,t) or (g,t') is maximal incompatible, where t' is the other child of $p(t)$.

Proof. Let pair (g,t) be maximal incompatible. Thus $g \subseteq p(t)$ but $g \not\subseteq t$. Since both children $a(g)$ and $b(g)$ overlap t, they cannot be included in the other child of $p(t)$. On the other hand, if both of them are included in t, then g, as their union, also is included in t which contradicts the definition of incompatibility and implies that $G(g)$ and $T(t)$ are not root consistent.

Conversely, let trees $G(g)$ and $T(p(t))$ be root-inconsistent, and $g \subseteq p(t)$. That means that at least one of the clusters $a(g), b(g)$ intersects both children t, t' of $p(t)$ and the other one intersects at least one of those children, say t. Then both $a(g)$ and $b(g)$ intersect t, but their union g is not included in t because at least one of $a(g), b(g)$ overlaps t'. This implies that (g,t) is incompatible, which completes the proof. □.

Now we can characterize nonmaximal incompatible pairs in terms of the duplications related to corresponding maximal incompatible pairs.

Theorem 8 *For any maximal incompatible pair (g,t), pair (g,n) is incompatible if and only if n is a mixed node due to the corresponding duplication in tree $T(p(t))$ after embedding tree $G(g)$ into $T(p(t))$.*

Proof. In the embedding process, leaves of $a(g)$ get one sign label and of $b(g)$ the other. Any mixed node n in $T(p(t))$ must overlap both $a(g)$ and $b(g)$ which is equivalent to (1) since $g \not\subseteq n$. Conversely, if (g,n) is incompatible, than $n \in T(p(t))$ because n overlaps g and thus, $p(t)$; on the other hand, n cannot be ancestor of $p(t)$ since n does not include g, which proves $n \subseteq p(t)$. Since n overlaps both $a(g)$ and $b(g)$, it is mixed after embedding is done. The proof is finished.□.

Let us denote the number of all the incompatible pairs as $I(G,T)$ and the number of the maximal incompatible pairs as $M(G,T)$.

Theorem 9 *The numbers of the incompatible pairs are related to the inconsistency measure in the following way: $C(T/G) = I(G,T) + 3M(G,T)$.*

Proof. To prove the equality, let us define a correspondence between the incompatible pairs and the items (duplication event, loss, gap) counted in $C(G/T)$. The set of all incompatible pairs can be partitioned onto subsets I_g containing incompatible pairs (g,t) with g fixed. Let us consider such a g fixed, and take corresponding maximal incompatible pair $(g,t) \in I_g$. Theorem 7 implies that pair $(g,p(t))$ corresponds to a duplication event counted in $C(g,T)$, although this pair is not counted in $I(G,T)$.

Let us now show that the number of elements in I_g plus 2 is equal to the number of losses/gaps due to the duplication $(g,p(t))$ associated with pair (g,t). Along with the uncounted duplication event $(g,p(t))$ corresponding to (g,t) as a maximal incompatible pair, that will lead to $|I_g| + 3 = C(g,T)$. Summing up these equalities, by g contained in the incompatible pairs, will give the statement of the theorem.

Let us consider tree $T(p(t))$ along with corresponding duplication δ_g, due to $G(g)$. Evidently, maximal speciated nodes in that tree form a partition of the set of its leaves. Let us get off all those nodes (along with their subordinates) from $T(p(t))$. That means, that we leave in the obtained cut tree U, only the mixed nodes. Thus, due to Theorem 8, the whole number of nodes in U is equal to the number of elements in I_g. Every terminal node in U, obviously, gets both of its children (in $T(p(t))$) speciated and thus is related to two inconsistencies (loss or information gap) exactly, though it is counted in $I(G, T)$ only once. The internal nodes in U can be, obviously, of degree 2 or 3 only. Every node of degree 2 corresponds to exactly one speciation event in $T(p(t))$ (except for the root which is excluded since it is not incompatible); every node of degree 3 is connected to mixed nodes only and, thus, gives no speciation events at all. Thus, in the tree U, any node of degree 1 (terminal node) contributes to $C(g, T)$ with two counts, any node of degree 2 with 1 count, and nodes of degree 3 with no counts. But in a binary rooted tree, the number of the nodes of degree 3 is equal to the number of the terminal nodes minus 2. Thus, the number of the elements in I_g is equal to the total number of losses and gaps due to mutation $(g, p(t))$ minus 2, which proves the theorem. $\square$.

Curiously, the theorem proven shows that the duplication inconsistency measure is associated with the set-theoretic incompatibility of the trees considered in such a way that any maximal incompatible pair gets its weight four (counted four times) as compared to ordinary incompatible pairs (counted once only). In particular, the theorem implies that $C(T/G) = 4I(G, T)$ when all the incompatible pairs are maximal. Since the maximal incompatible pairs correspond to the duplication mutation events while nonmaximal ones just to mixed nodes, this can be put also as follows: the inconsistency measure $C(T/G)$ is equal to four times the number of the duplications needed to explain all the inconsistencies plus number of the other mixed vertices (each counted as many times as it occurs under a duplication).

The theorem reveals also the graph-theoretic nature of the asymmetry of $C(T/G)$: the measure involves different objects in G and T, comparing the triples parent-children $(g, a(g), b(g))$ from G with just the clusters $t \in T$.

6 Discussion

A model for duplication mutations has been formalized to explain the inconsistencies between binary phylogenetic trees derived from different gene families. The model involves a gene duplication history in the species phylogeny described with corresponding speciation/loss events and information gaps occurred due to some particular kinds of the inconsistencies. To make the model work, so called Duplication/Speciation principle is introduced to require that every hypothetical gene duplication evolves in such a manner that in any contemporary species (participating in the inconsistency occurred) only one copy is observable. These explicitly defined concepts distinguish our analysis from earlier models (Goodman et al. (1979), Guigó, I. Muchnik, and T.F. Smith (1994)) where the formalisms introduced have not been supported with any explicit biological interpretation. Yet, it appears, the technical concept of duplication defined in the papers cited is equivalent to the duplication mutation event according to the model presented. In this aspect, the model gives a substantiation to the phylogeny reconstruction methods developed in Goodman et al. (1979) and Guigó, I. Muchnik, and T.F. Smith (1994). However, in the present model, losses can be assigned to particular duplication line descent, which is not the case in that previous papers, so more to say on the information gaps which are completely missed in them.

A set-theoretic reformulation of the tree inconsistency measure derived from the model is presented. Although the reformulation does not increase the computational efficiency of the method (which requires at most $O(N^3)$ computational steps), it allows computing the inconsistency measure without annotating the phylogeny with all the duplication/loss/gap histories thus saving the memory. Amazingly, the reformulated measure involves the unspeciated ancestors only, assigning weight four to the duplication mutation events themselves.

Using the formalism developed, multiplicity of location of some of the duplication events (Section 3.3), in particular, it could be expected that some strategies could be developed for constructing a "consensus" species tree directly from the multiple tree data rather than through evaluation the embedding of those data in a predefined particular species tree as it has been done in Goodman et al. (1979) and Guigó, I. Muchnik, and T.F. Smith (1994).

Some other directions for future developments: relating the model to the original biomolecular data, from which one would like to directly calculate the overall "species" tree (which could, eventually, lead to reconsideration of the sequence alignment procedures); extending the model onto different kinds of biomolecular inconsistencies; including more detailed descriptions of duplication mutation events needed to explain inconsistencies (for instance, in the direction of the "regulation gene expression" considered by Goodman et al. (1979)); considering different kinds of mutation events.

References

[1] Adams, E.N. III. 1986. N-Trees as Nestings: Complexity, Similarity and Consensus, *Journal of Classification*, 3, 299-317.

[2] Barthélemy, J.P., Leclerc, B., and Monjardet, B. 1986. On the Use of Ordered Sets in Problems of Comparison and Consensus of Classifications, *Journal of Classification*, 3, 187-224.

[3] Baum, B.R. 1992. Combining Trees as a Way of Combining Data Sets for Phylogenetic Inference, and the Desirability of Combining Gene Trees, *Taxon*, 41, 3-10.

[4] Day, W.H.E. 1985. Optimal Algorithms for Comparing Trees with Labeled Leaves, *Journal of Classification*, 2, 7-28.

[5] Day, W.H.E., and McMorris, F.R. 1985. A Formulation of Consensus Index Methods, *Bulletin of Mathematical Biology*, 47, 215-229.

[6] Farach, M. and Thorup, M. 1993. Fast Comparison of Evolutionary Trees, Technical Report 93-46, DIMACS, Rutgers University, Piscataway, NJ.

[7] Fitch, W.M. 1970. Distinguishing homologous and analogous proteins, *Syst. Zool.*, 19, 99-113.

[8] Fitch, W.M., and Margoliash, E. 1967. Construction of Phylogenetic Trees, *Science*, 155, 279-284.

[9] Goodman, M., Czelusniak, J., Moore, G.W., Romero-Herrera, A.E.,and Matsuda, G. 1979. Fitting the Gene Lineage Into Its Species Lineage. A Parsimony Strategy Illustrated by Cladograms Constructed From Globin Sequences, *Syst. Zool.*, 28, 132-163.

[10] Guigó, R., Muchnik, I., and Smith, T.F. 1994. Reconstruction of Ancient Molecular Phylogeny, *Molecular Phylogenetics and Evolution, in press.*

[11] Hendy, M.D., Little, C.H.C., and Penny, D. 1984. Comparing Trees with Pendant Vertices Labelled, *SIAM J. Appl. Math.*, 44, 1054-1067.

[12] Leclerc, B. 1985. La Comparaison des Hiérarchies: Indices et Métriques, *Mathematiques et Sciences Humaines*, 92, 5-40.

[13] Margush, T., and McMorris, F.R. 1981. Consensus n-Trees, *Bulletin of Mathematical Biology*, 43, 239-244.

[14] Nei, M. 1987. *Molecular Evolutionary Genetics.* Columbia University Press, New York.

[15] Page, R.D.M. 1994. Maps Between Trees and Cladistic Analysis of Historical Associations Among Genes, Organisms, and Areas, *Syst. Biol.*, 43, 58-77.

[16] Robinson, D.F. 1971. Comparison of Labeled Trees with Valency Three, *Journal of Combinatorial Theory*, 11, 105-119.

[17] Robinson, D.F., and Foulds, L.R. 1981. Comparison of Phylogenetic Trees, *Mathematical Biosciences*, 53, 131-144.

[18] Smith, T.F., and Waterman, M.S. 1980. How Alike are Two Trees, *American Math. Monthly*, 87, 552-553.

[19] Waterman, M.S., and Smith, T.F. 1978. On the Similarity of Dendrograms, *Journal of Theor. Biology*, 73, 789-800.