Laboratory of mathematical methods and models in bioinformatics

Institute for Information Transmission Problems, Russian Academy of Sciences

lyubetsk@iitp.ru,
http://lab6.iitp.ru/en/

The directions below may be of mutual interest:

I) Inferring scenarios of **co-evolution** of multiple genes and regulations across the species tree.
E.g., **co-evolution** of a **species**, **regulon**, regulation **factor** and its binding **site**.

II) Inferring clusters of evolutionary events across the species tree.
E.g., several genes involved in one metabolic pathway often undergo evolutionary perturbations at the same areas of the species tree

To develop with tasks I-II:

1) a concept of the scenario, where each event is assigned a **particular type and the area in the species tree**;

2) an algorithm of embedding of a gene tree onto the species tree: an algorithm of **constructing the scenario**;

3) obtain a confident and representative set of gene trees and a corresponding species tree;

4) an accurate and fast algorithm of **supertree** construction (also as an independent research direction)

III) **Reconstructing** evolution of a regulatory region (= a certain **regulation**) or a **gene** along a gene tree with or without defining the tree topology and branch lengths; and **inferring time slices**.

An original approach is proposed in Lyubetsky, Zhizina, Rubanov, 2008;

it can be further elaborated together

There are two fundamentally different approaches: “inferring the **events**” and “inferring the **sequence – structure evolution**”.

I’ll speak on the “**events**” (here) and then on the “**sequence – structure evolution**” (the latter is quite broad, see file **Directions 2-4**).

We tried to merge them but this is a separate task

I) Co-evolution of

species, genes and regulatory elements

II) The evolution of gene

(defined by gene tree *G*)

along species tree *S* and

clustering of evolutionary events.

The separate problem of time slices here!

Example result of task I: co-evolution of species, genes and regulatory elements

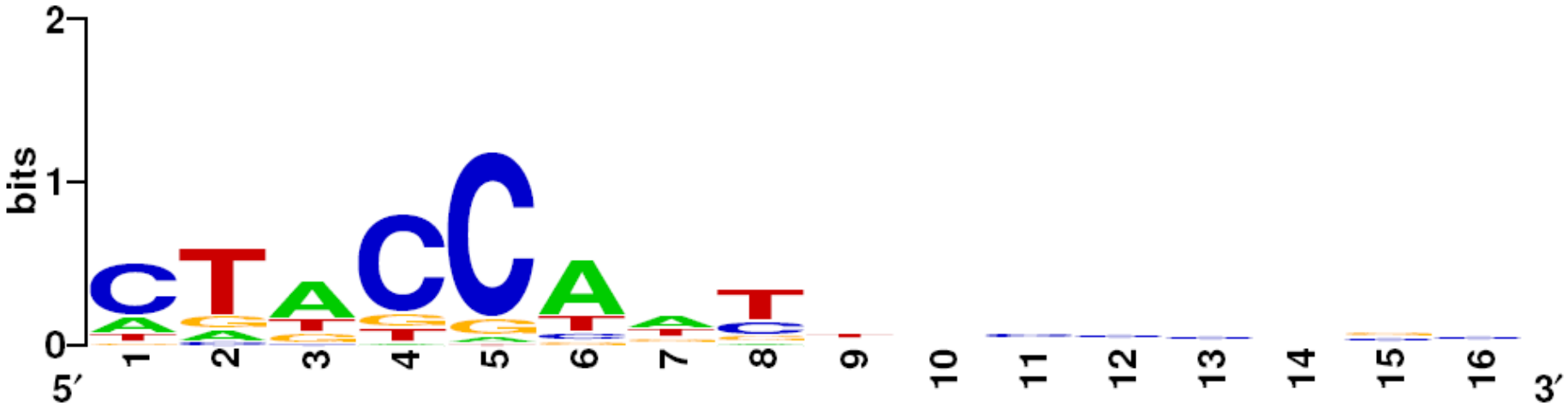
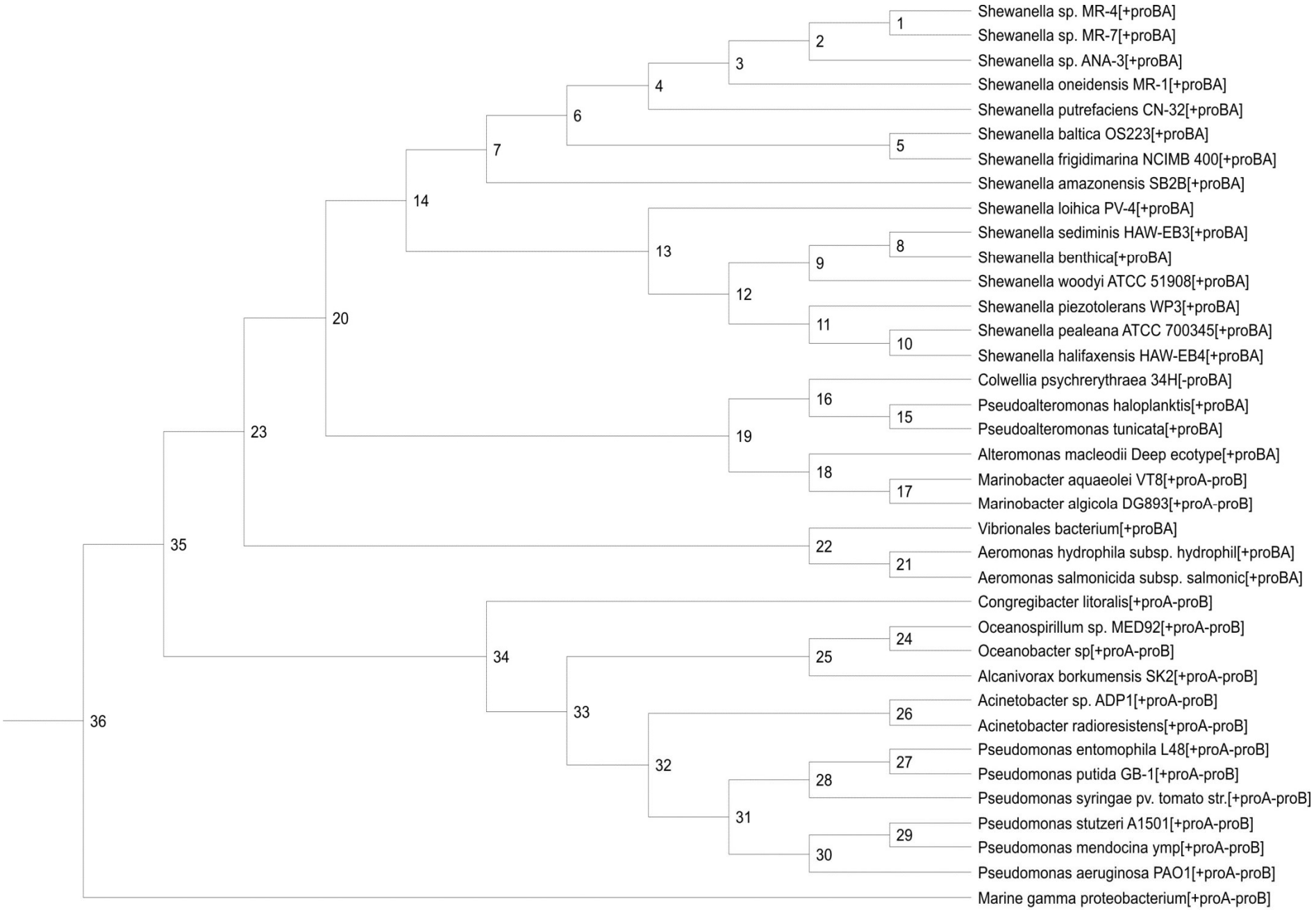
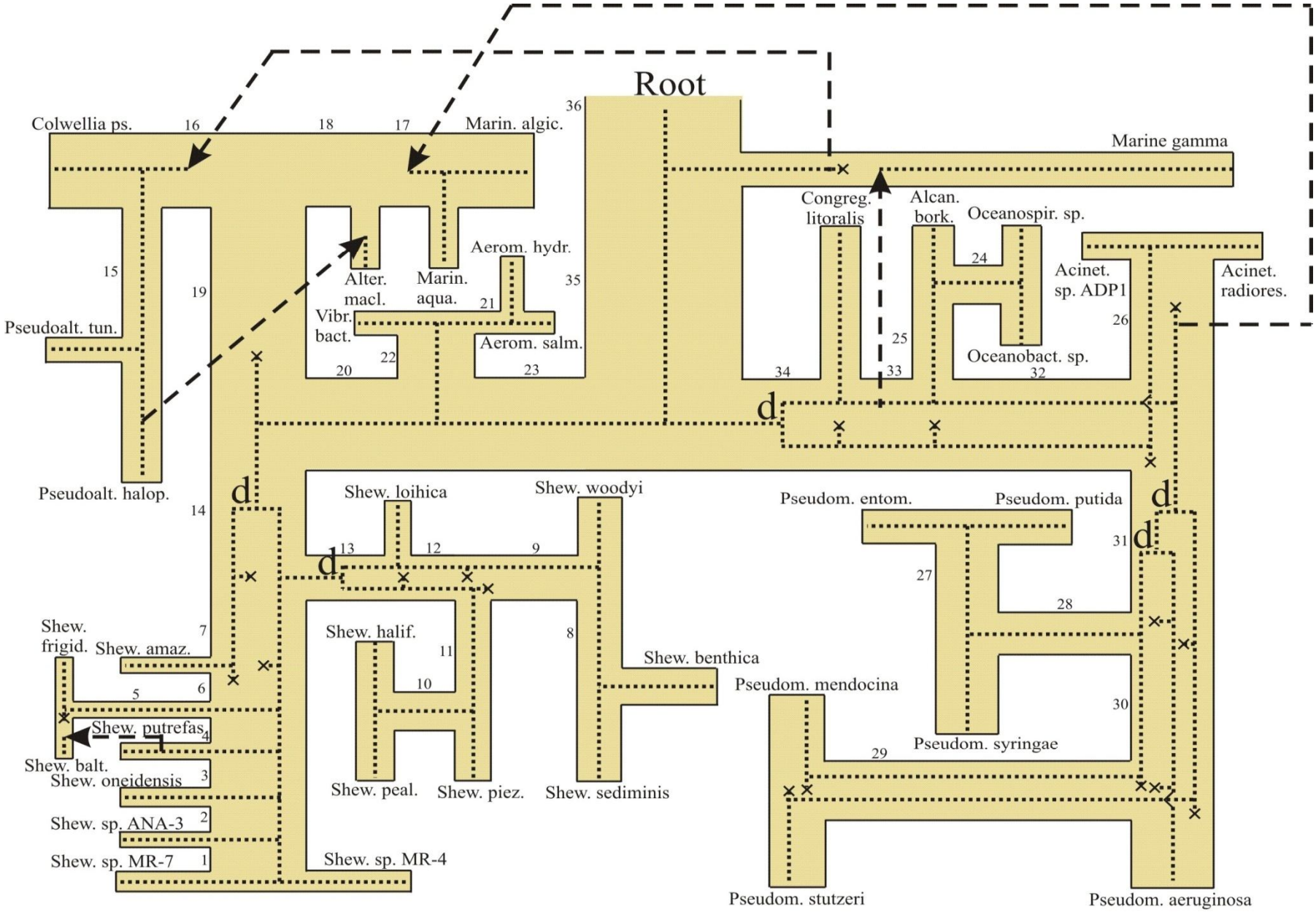


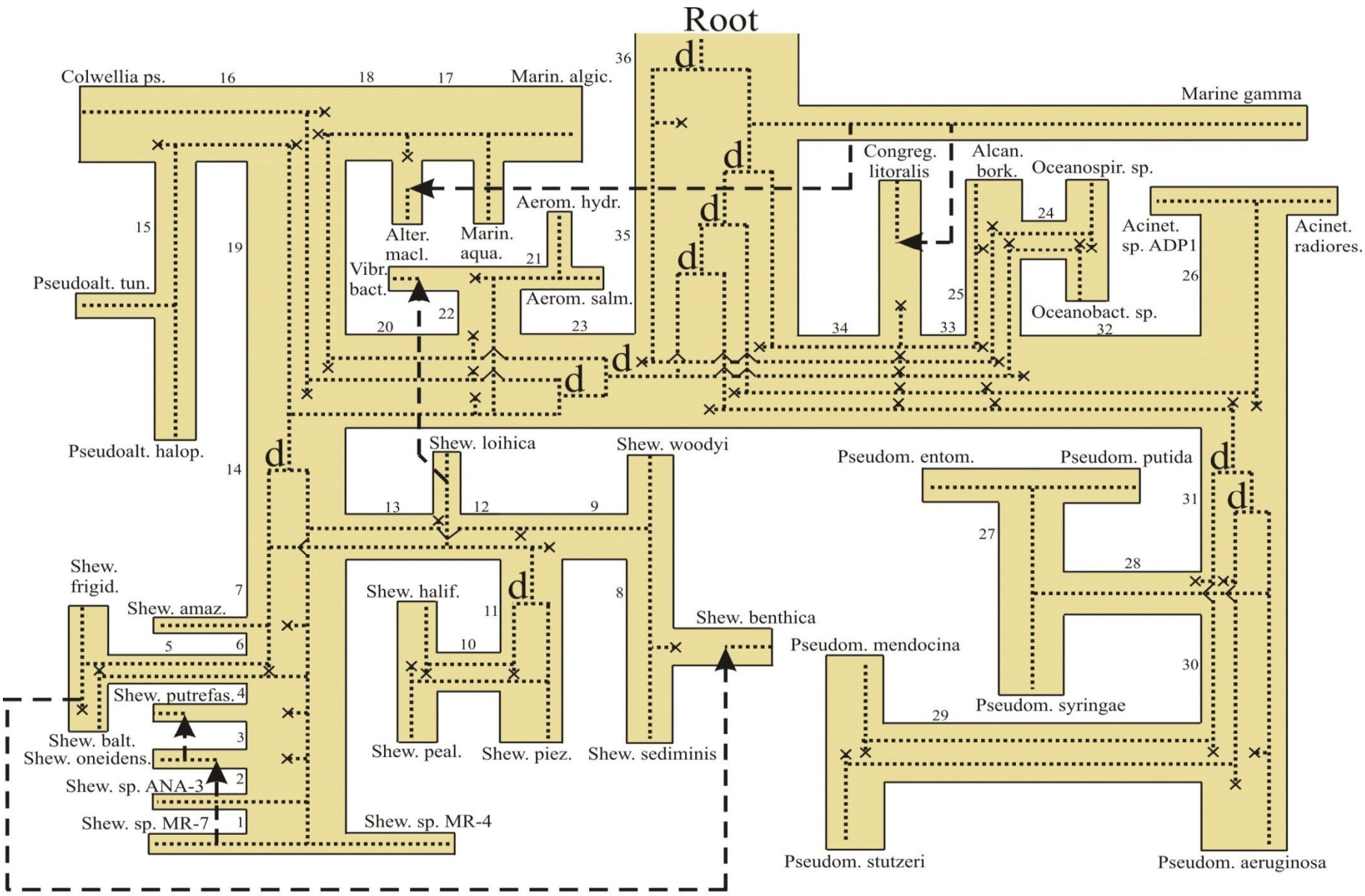
Fig. 1. Frequency profile of the 8bp long binding **site** and its weakly conserved 3'-end upstream **genes** *proA* and *proB* widely represented among γ-proteobacteria. The genes often form the *proBA* operon. We identified a TetR family protein, an ortholog of the NP_249058 protein from *P. aeruginosa* PAO1, as a transcription **factor**. **Sites, genes, factors and species evolve together. The question is how?**



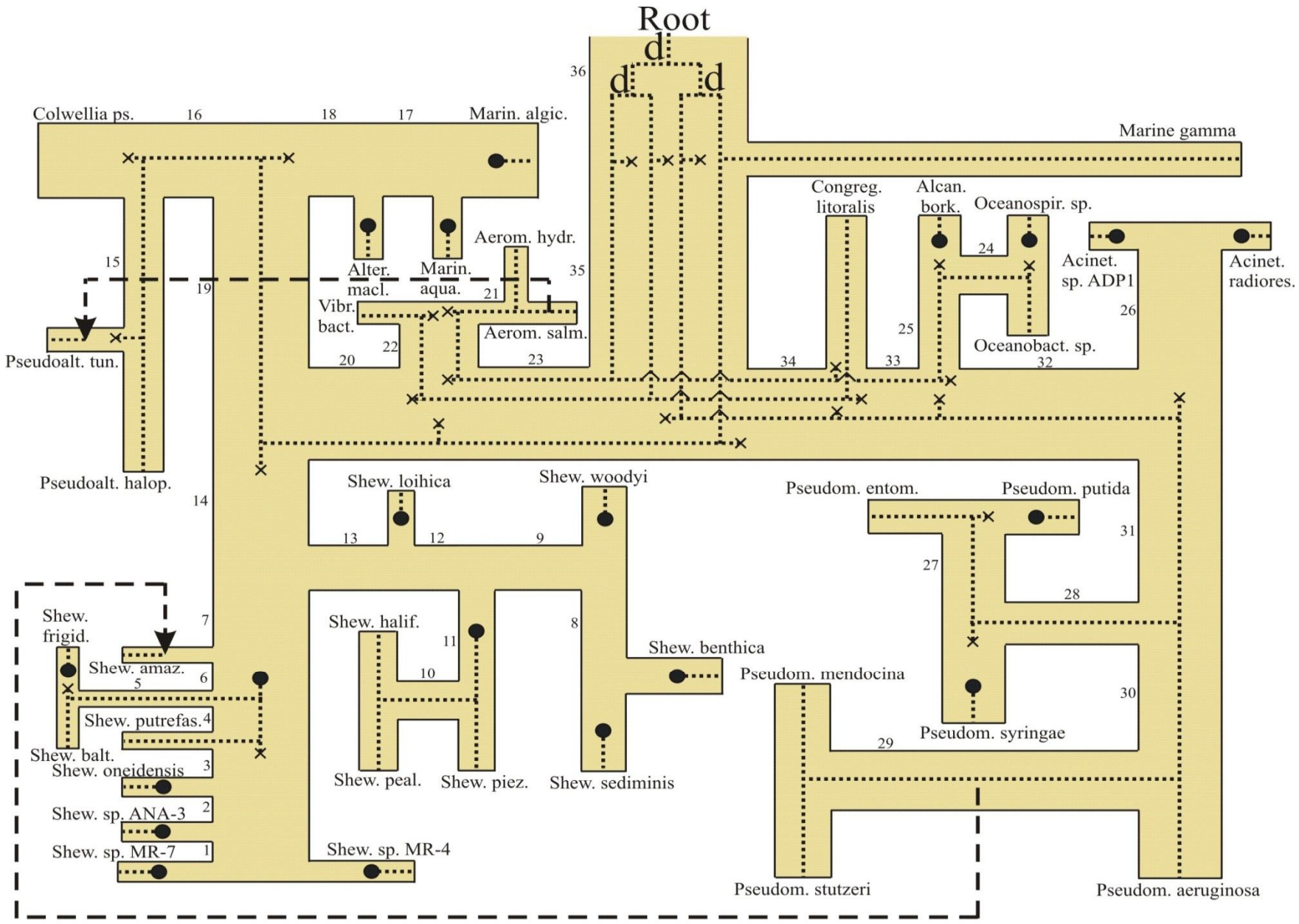
Supertree *S* = species tree *S*



The supertree *S* with *proBA* evolutionary scenario:
S in beige color, shown inside the tubes of *S*



The supertree *S* with the factor evolutionary scenario:
S in beige color, shown inside the tubes of *S*



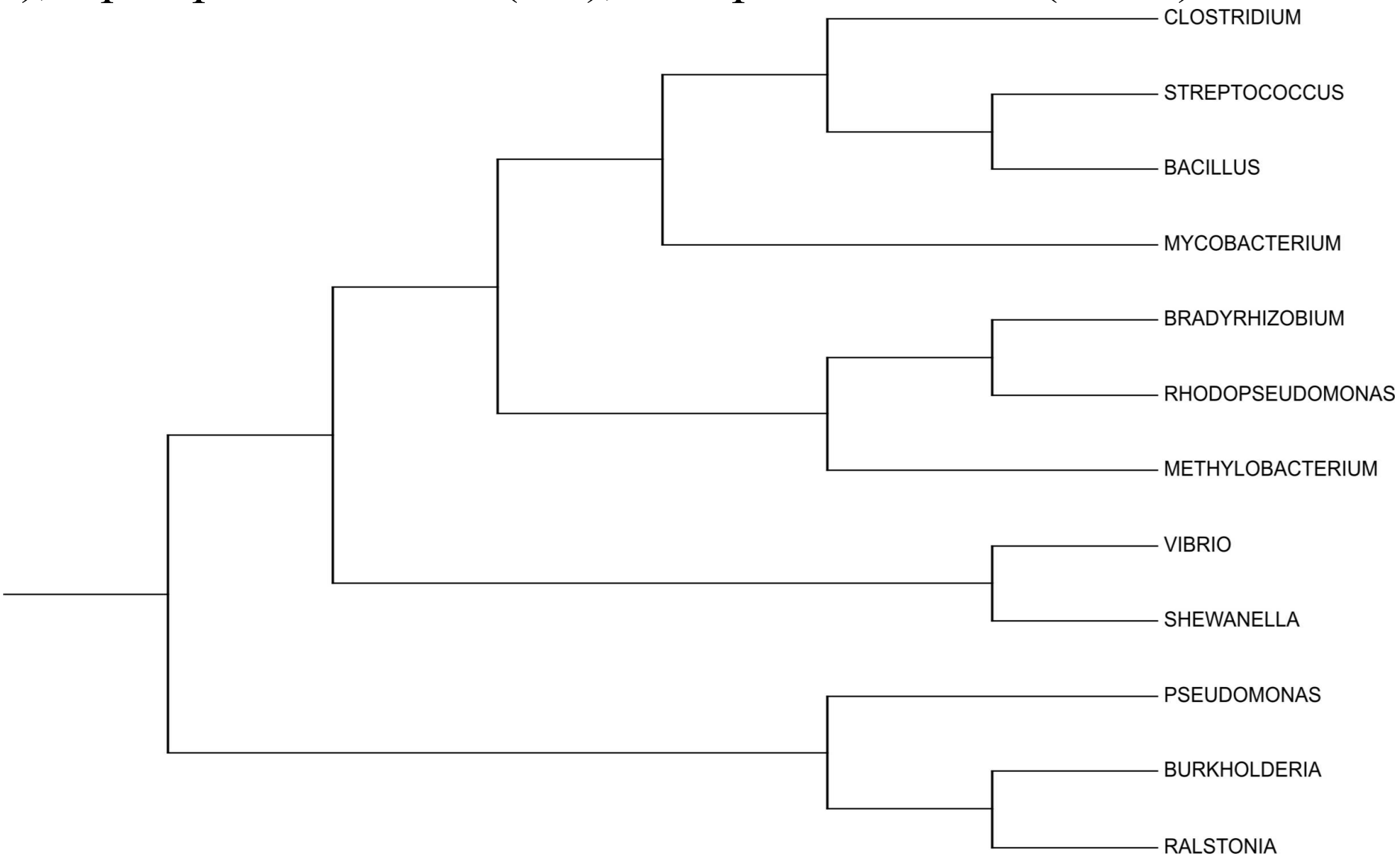
The supertree *S* with the site evolutionary scenario:
S in beige color, shown inside the tubes of *S*

Example results of task II:

In analyses of 1500 genes, 138 HGTs are found to occur in the genus **PSEUDOMONAS** being a donor of 25 and acceptor of 29 HGTs. Other HGTs were distributed more of less equally among other genera.

Genera: firmicutes (3), actinobacteria (1), alpha-proteo (3), beta-proteo (2)

The upper 4 genera are Gram-positive, the rest are Gram-negative. We see the clades of firmicutes (1-3 from the top), acinobacteria (4), alpha-proteobacteria (5-7), beta-proteobacteria (11-12).



How to accomplish tasks I-II?

What is needed (= a working plan):

1) Definition of the evolutionary scenario (= embedding *f* of *G* into *S*)?

2) Fast algorithm of constructing supertree *S from set $\{G_i\}$ and embedding *f* of *G* into *S*.**

3) Single evolutionary event costs validation.

4) Gene tree data mining and rooting (!).

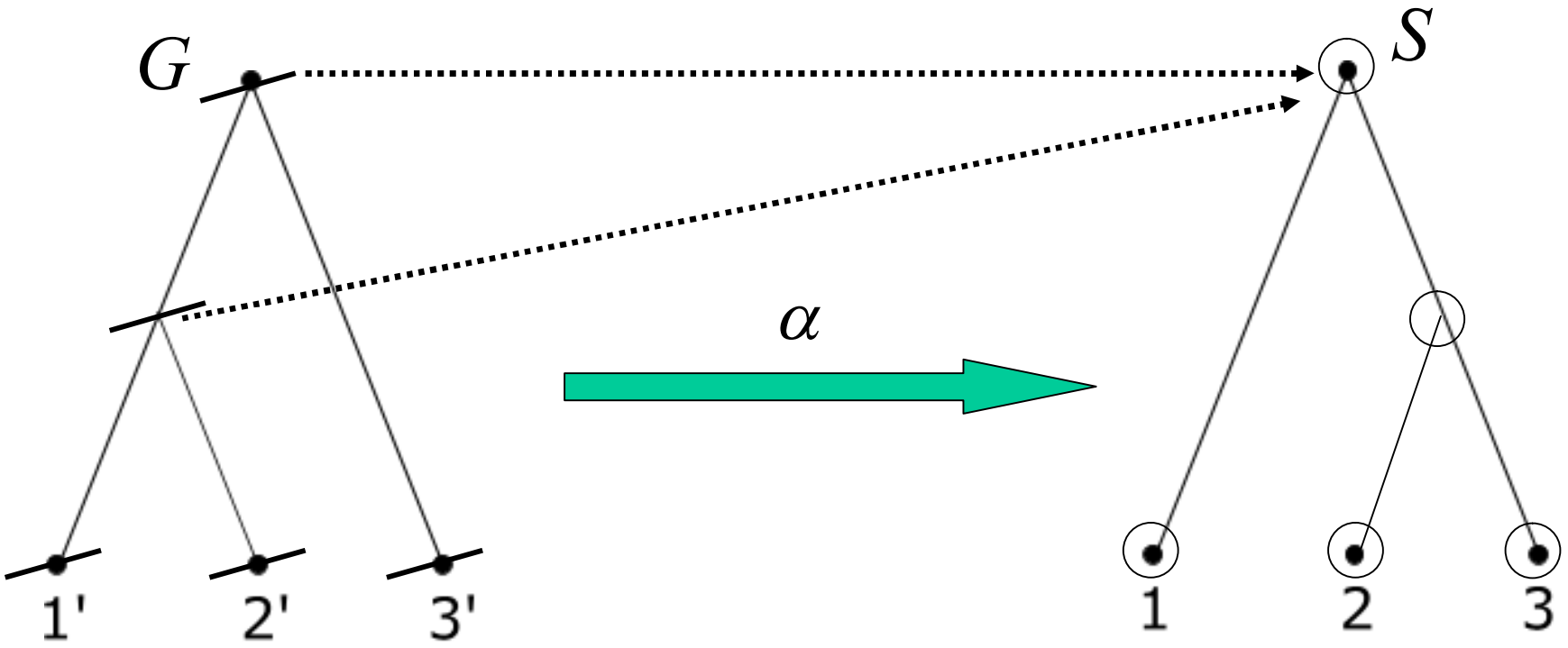
5) Clustering of evolutionary events and its robustness against costs etc.; biological interpretation of clusters

"Valid" definition and cs formalization of an embedding is a **fundamental task**.

One **well known solution does not account for gene losses** (at least as events) **and transfers** (at all)

(Guigo R., Muchnik I., Smith T.F. 1996 Reconstruction of ancient molecular phylogeny. *Mol. Phyl. Evol.* 6; Mirkin and et al, …):

A known approach to reconcile the evolution of gene and species is embedding α and its cost $c(G,S)$

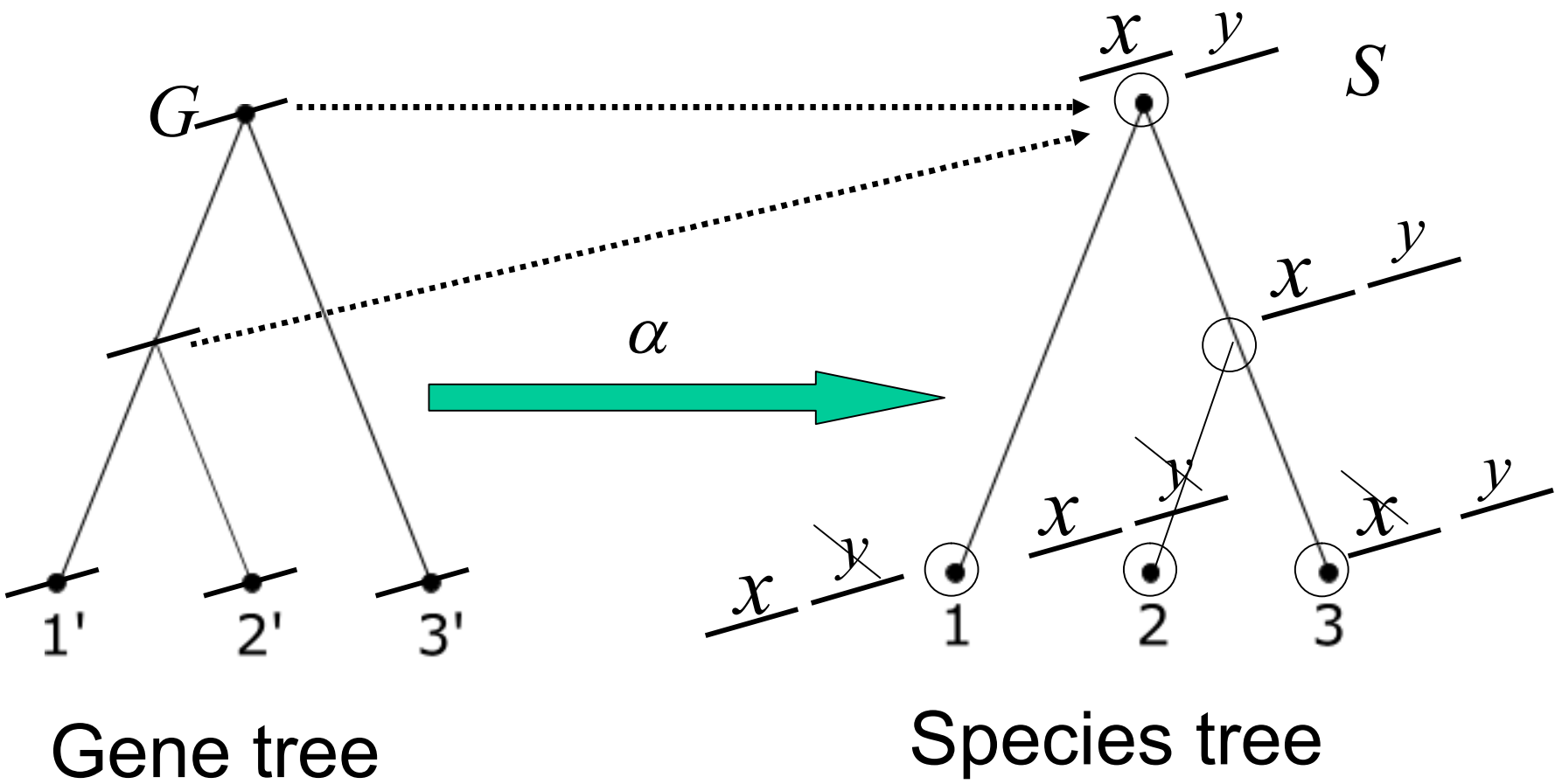


Gene tree

Species tree

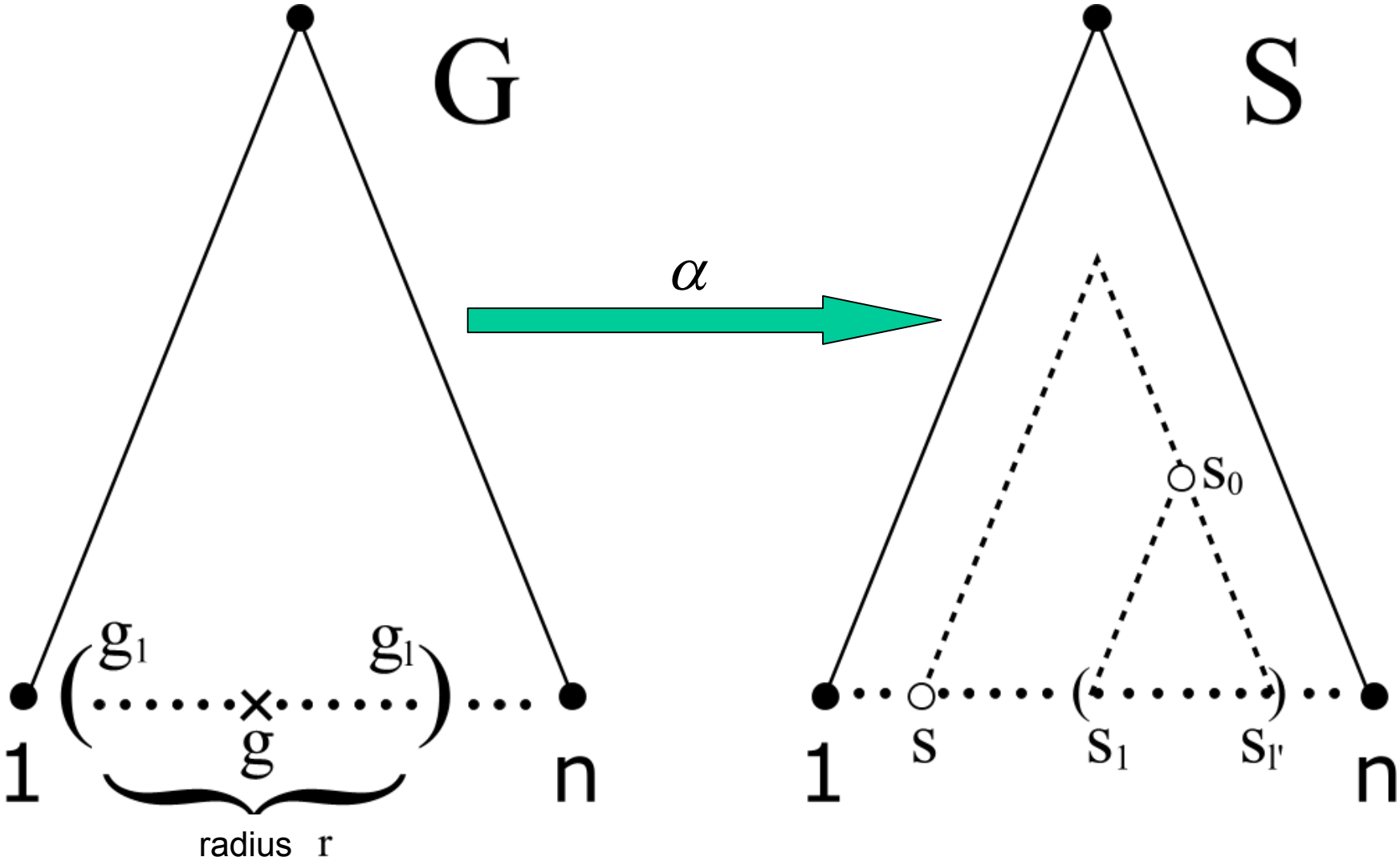
$c(G,S) = 4$

Inferring gene losses from embedding α based on a "theorem":



The number of **losses** is a sum of one-way duplications and gaps. Here the sum is 3 (HGTs not concerned at all)

In the context of this approach “alpha” we introduced a **TEST for putative recent HGTs**: gene ***g* is embedded into *s* but its neighborhood embeds far from *s***



1) Under this definition we revealed a long biologically reasonable list of HGTs and drawn some general conclusions, e.g.:

on average, one putative recent HGT decreases the number of losses by **4.4**:

$$lost_{new} = lost_{old} - 4.4 \cdot t \ .$$

But duplication counts drop only **slightly**.

2) We developed algorithms of finding **RECENT** and **ANCIENT** HGTs

In our **novel approach**:

("tube" is simply an edge in *S*)

we study embedding f of *G* into *S*, such

$f(g)$ is tube *d* or vertex *s* in *S*,

but informally $f(g)$ **is a tag of the particular event type in *d* or *s***.

We developed effective approach

to deal with such embeddings f

Scenario β is minimal embedding f

according to functional

$$c(f,G,S) = c_l \cdot l(f,G_{f^*},S) + c_d \cdot d(f,G,S) +$$

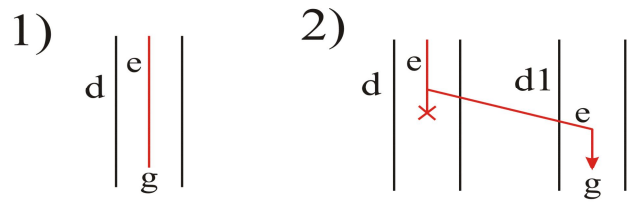
$$+ c_{t^+} \cdot t^+(f,G,S) + c_{t^-} \cdot t^-(f,G,S)$$

1) So, scenario β is a system of complex notations of evolutionary events in the tubes of tree *S*.

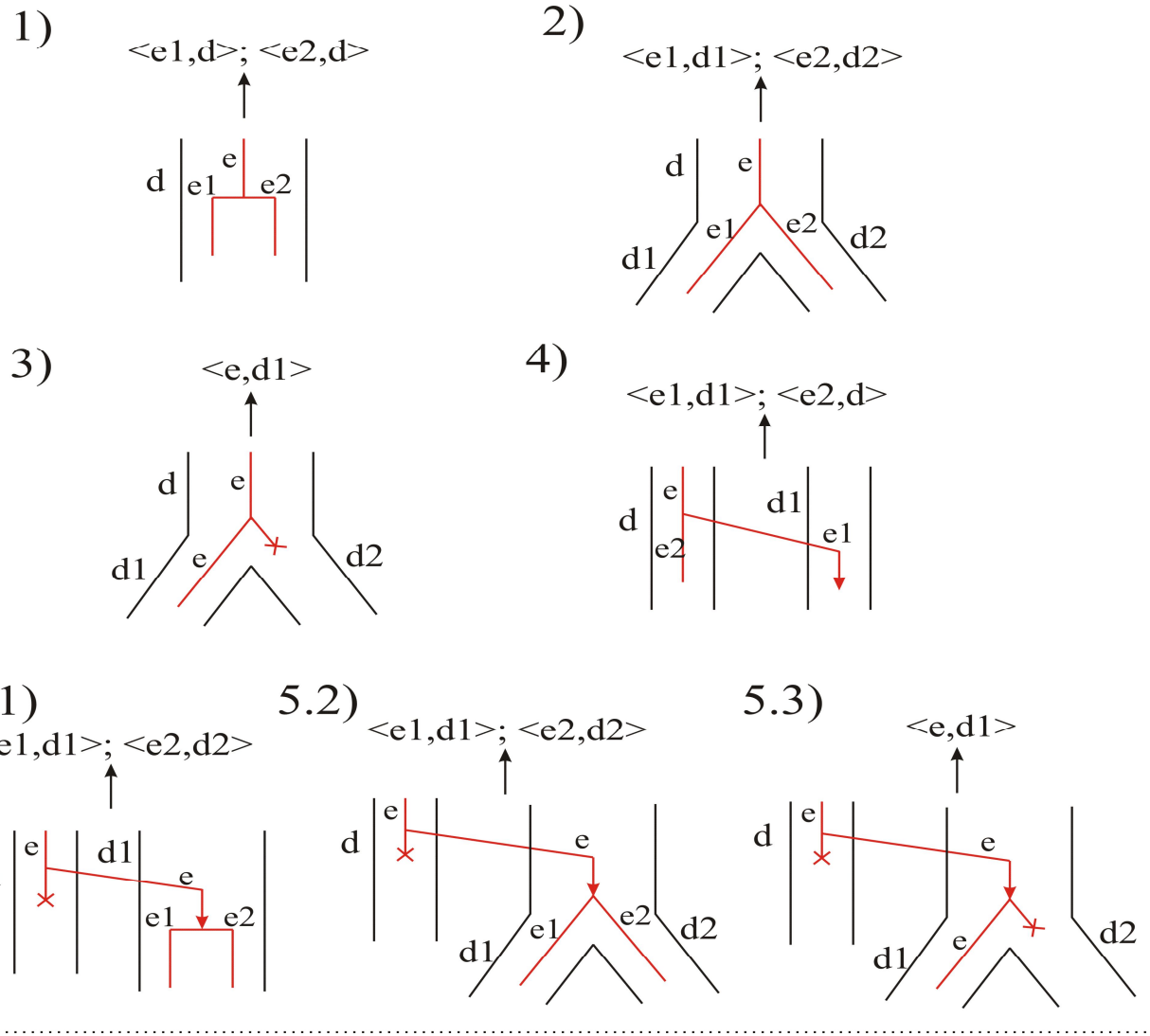
Our algorithm constructs scenarios and has a cubic complexity.

2) We introduce a **concept of time slices in species tree** *S* and developed an *ad hoc* algorithm to compute time slices

Initial step of building f^*



Inductive steps of building f^*



A *scenario* (without HGTs) of gene evolution (defined by gene tree G) along species tree S is minimal mapping f of all vertices $V(G)$ in tree G into vertices $V(S)$ and tubes $E(S)$ in tree S, when the following is true: 1) the super-root in G is mapped into the root tube in S; each leaf g in G maps in S into leaf s, the source of g; 2) if g_1 descends from g and $f(g)$ is a vertex, then $f(g_1) < f(g)$, and if $f(g)$ is a tube, then $f(g_1) \leq f(g)$; 3) let g_1 and g_2 be descendants of g: if $f(g)$ is a vertex, then the shortest path from $f(g_1)$ into $f(g_2)$ in S includes $f(g)$

A *scenario* (with HGTs) of gene evolution (defined by gene tree G) along species tree S is minimal mapping f of all vertices $V(G')$ in a subdivision G' of G into vertices $V(S)$ and tubes $E(S)$ in S, when the following is true: 1) the super-root in G' maps into the root tube in S; each leaf g in G' maps into leaf s in S, the source of g. Let g, g_1, g_2 be vertices in G'; 2) let g_1 descend from g: if $f(g)$ is a vertex, then $f(g_1) < f(g)$, and if $f(g)$ is a tube, then two cases apply. If g_2 is another descendant of g, then for both descendants $f(g_i) \leq f(g)$ or: for one descandant $f(g_i) \leq f(g)$ and for the other $f(g) \neq f(g_j) \sim f(g)$; here $f(g_i)$ is a vertex of a tube and $f(g_j)$ is a tube, i,j=1,2. If g with its parent g' produce a single descendant g_1, then $f(g_1) \leq f(g) \sim f(g') \neq f(g)$ or $f(g) \neq f(g_1) \sim f(g)$; here in the first inequality $f(g_1)$ is a vertex or a tube, $f(g')$ is a tube, and in the second inequality $f(g_1)$ is a tube; 3) let g_1 and g_2 descend from g: if $f(g)$ is a vertex, then the shortest path from $f(g_1)$ to $f(g_2)$ in S includes $f(g)$; if g produces a single descendant, then $f(g)$ is a tube.

Gene duplication is vertex g in G' with two descendants g_1 and g_2, for which $f(g)$ is a tube in S and for both descendants $f(g_i) \leq f(g)$, i=1,2.

Gene loss is pair $\langle e,s \rangle$, where e is an edge in G', s – a vertex in S with two descendants, and $f(e^+) < s < f(e^-)$.

Speciation event (with respect to a given gene) is vertex g in G', for which $f(g)$ is a vertex in S, and each of vertices g and $f(g)$ produces two descendants.

Horizontal transfer with retention is vertex g in G' with two descendants g_1 and g_2, for which $f(g)$ is a tube in S, and one of descendants g_i has $f(g) \neq f(g_i) \sim f(g)$.

Horizontal transfer without retention is vertex g in G' with single descendant g_1, for which $f(g) \neq f(g_1) \sim f(g)$.

Constructing supertree S^*

The problem of building a species supertree given a set of gene trees $\{G_i\}$ is of great applied value. This problem is NP-hard and finding effective solutions requires its biologically valid reformulation.

We proposed such a reformulation and a fast algorithmic solution to build a supertree.

Simulations and a mathematic proof demonstrate that the algorithm is both fast and accurate

In our original approach, binary supertree S^* is sought among such trees, that have all clades contained in **fixed predefined set *P* of possible clades**. In the simplest case *P* consists of **all clades of given gene trees** $\{G_i\}$.

Define *V* from *P* as **basic** if it can be split in two sets from *P*, which can also be split in two, and so on until singlet leaves are obtained

Supertree construction

Given set $\{G_i\}$ of gene trees,

tree $S(V)$ is built inductively:

if V is split into V_1 and V_2 and $S(V_1)$ and $S(V_2)$ are already built, then they are merged on a minimal partition V_1^* and V_2^* according to functional

$$\sum_i [c_l \cdot l(V, V_1, V_2, G_i) + c_d \cdot d(V, V_1, V_2, G_i)] +$$

$$+ c(\{G_i\}, S(V_1)) + c(\{G_i\}, S(V_2))$$

The resulting tree is **minimal** according to functional

$$c(f,\{G_i\},S)=\sum_i\left(\begin{array}{l}c_l\cdot l(l,G_i,S)+c_d\cdot d(l,G_i,S)+\\+c_{t^+}\cdot t^+(f,G_i,S)+c_{t^-}\cdot t^-(f,G_i,S)\end{array}\right)$$

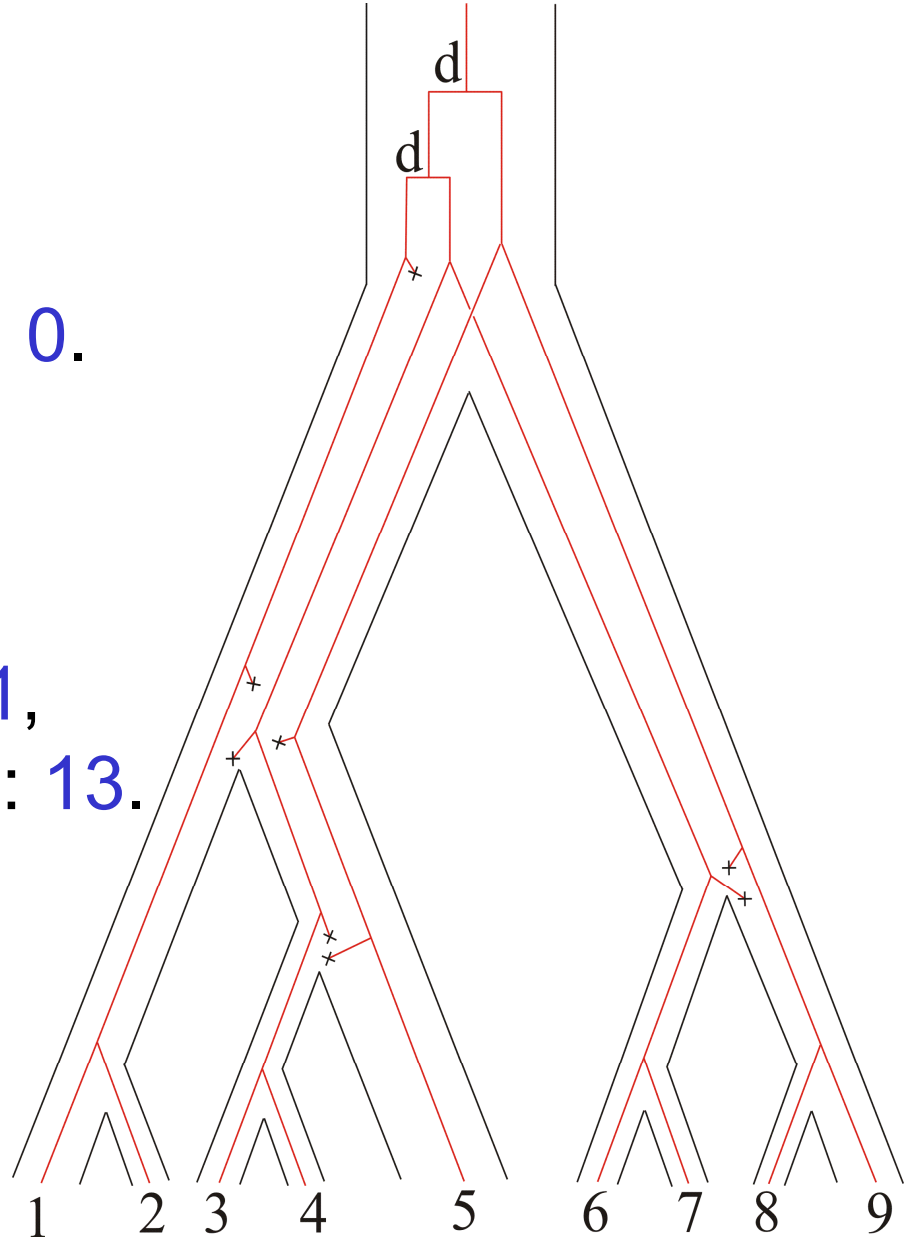
Setting the costs in a scenario without transfers:

single duplication cost: 3,
single loss cost: 2,
single «speciation» cost in *G*: 0.

Later we define:

cost of HGT with retention: 11,
cost of HGT without retention: 13.

Our clustering is robust
against the cost values!



Allowing for horizontal transfers usually simplifies scenarios. Here is a scenario of the same *G* and *S* as before but with HGTs:

Here an edge in the gene tree can transfer from one tube into another within the same slice (the problem of slices!). Under such scenario, no duplications are inferred but only 1 loss and 2 HGTs (one with and one without retention of the donor copy)