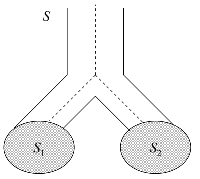
THE PROBLEMS OF RECONCILING GENE AND SPECIES TREES, MAPPING A GENE TREE INTO A SPECIES TREE, AND GENE TREE INFERENCE

Konstantin Gorbunov and Vassily Lyubetsky

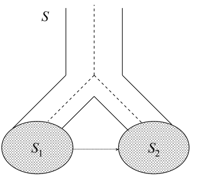
Institute for Information Transmission Problems of the Russian Academy of Sciences;
19 Bolshoy Karetny per., Moscow, 127994, Russia; gorbunov@iitp.ru; lyubetsk@iitp.ru;

Problem 1. A long recognized problem is inference of a tree *S* that amalgamates a set of input gene trees. We further developed a traditional approach to find the tree *S* such that it minimizes the total cost (gene duplications and losses) of mappings of individual gene trees into *S* [1,2]. An algorithm is novel mathematically correct and possesses the cubic running time in *n* and in *m*, where *n* is the number of gene trees, and *m* is the total number of species. Is a correct inference of the tree *S* possible in polynomial time with the horizontal gene transfer events? Is a correct inference of a phylogenetic net *S* (instead of a species tree *S*) possible in polynomial time at least with gene duplication and loss events?

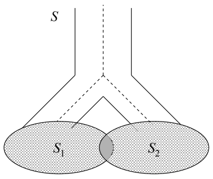
Our condition: the supertree *S* is sought for such that it contains the majority of clades from input trees G_i .
Our method: inductive joining trees S_1 and S_2 and rooting it at the joint node.



But: if horizontal gene transfers are allowed, the descendants of genes entering to S_1 can transfer to S_2, which makes the precise optimization difficult.

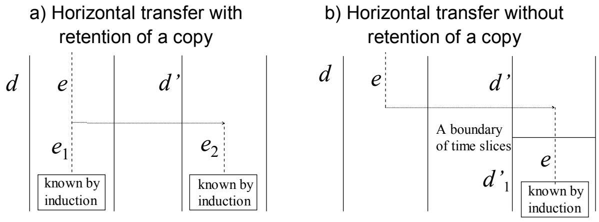


And: if instead of trees there are phylogenetic nets, then species sets in S_1 in S_2 can intersect each other, which makes the precise optimization difficult.



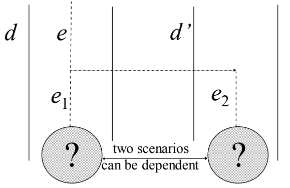
Problem 2. We suggested a novel mathematically correct algorithm to map (reconcile) a gene tree *G* into *S* (with time slices) that possesses the cubic running time in |*S*|, [3-6]. Could one does the same for phylogenetic nets?

Our condition: a horizontal transfer is possible only between branches of the species tree *S* that lie in the same time slice.
Our method: inductive bottom-up construction of embeddings of subtrees of *G* into subtrees of *S*.

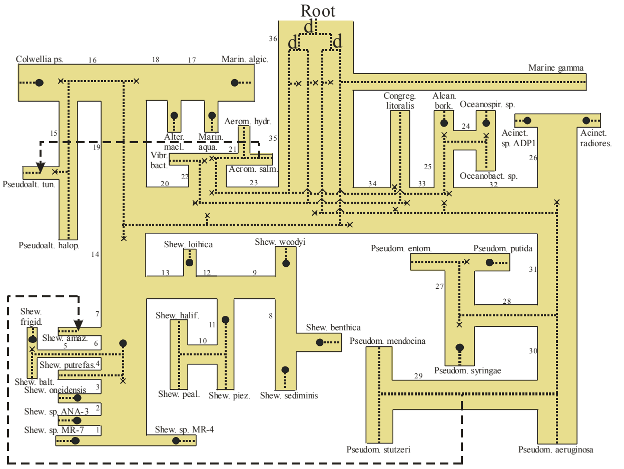

a) Horizontal transfer with retention of a copy

b) Horizontal transfer without retention of a copy

But: if instead of trees there are phylogenetic nets, then the two scenarios in the point a) are not independent which makes the precise optimization difficult.



Problem 3. We suggested a novel definition of an evolutionary scenario that is an embedding of a gene tree (or another evolutionary tree) into a species tree for the joint case of duplications, losses, horizontal gene transfers, gains, etc [4-6]. On the figure in the right hand side we show an example of an optimal evolutionary scenario of proline synthesis genes regulation binding sites that our algorithm has constructed. Edges of a species tree are shown as black-out tubes, evolution of sites is shown as dotted lines inside of the tubes. Duplications are shown by symbols “d”, losses – as short branches with daggers, horizontal transfers – as edges with arrows. Emergencies of a new sites were modeled as transfers from a special tube (outgroup, nor shown in the figure), which leads from the root to a leaf (without any specie) and are shown in the figure as dark circles.
A problem: how to define the embedding taking into account dynamics of molecular sequences and chromosome structures?



Problem 4. We suggested a novel heuristic algorithm to reconstruct a gene tree on the base of an multiple alignment. The tree is sought for among trees consisting of clades from a prebuilt set *P*. We have developed a technique of construction of reasonable *P*. After the construction of *P* for each column *i* of the alignment we construct by dynamic programming a marked tree T_i , such that all its clades lie in *P* and the sum of similarities of symbols on edge ends is maximal. To ensure grouping of equal symbols in clades we calculated similarity of a symbol *b* with its son symbol b_1 with account of a simplest estimate on length of the edge (bb_1). We define that the length is equal to the difference of heights of the vertex *b* and the vertex b_1 where the height of a vertex is the maximal number of edges on a path from the vertex to a leaf. If *b* is not equal to b_1 then the usual similarity of *b* with b_1 was multiplied by $n/(n–1+d)$ where *n* is the number of alignment rows and *d* is the length of the edge (bb_1). If *b* is equal to b_1 then the similarity was not modified. Such modification makes advantageous placing pairs of unequal symbols at ends of the longest edges. Thus we ensure grouping of leaves with equal symbols in one clade. At the final stage for each set *M* from *P* we consider all possible partitions of *M* on two parts M_1 and M_2 and chose the best partition to join the already constructed trees $T(M_1)$ and $T(M_2)$ under the common root. To ensure dependence of *T*(*M*) not only on *M* but on topology of a whole tree, we used the following technique, which will be demonstrated by an artificial example. Let the alignment in the alphabet {A,B,C} be shown in the frame. Below in trees numbers 1, 2, 3, 4 denote the rows.

AAAA
BAAA
CAAA
CBBB

Let the similarities be: *s*(A,A)=*s*(B,B)=*s*(C,C)=3, *s*(A,C)=–1, *s*(B,C)=–0.95, *s*(A,B)=–0.9. For simplicity we assume that *P* consists of all 15 possible sets. For all *i,k* we construct the marked tree T_{ik} which is the tree T_i marked according the *k*-th column of the alignment. For example, the tree T_{11} is (A_1,(B_2,(C_3,C_4)C)C)C, while the tree T_{21} is any binarization of the tree ((A_1,B_2,C_3)C,C_4)C)C. We see that in this tree the root of the clade {1,2,3} is marked by the symbol C. Then the tree *T*({1,2,3}) is constructed **under the condition** that its root is marked by C in the first column. We have: *T*({1,2,3})=(1,(2,3)), and the final tree is *T*({1,2,3,4})=((1,(2,3)),4). It is easy to verify that this tree is optimal. Note that without the used condition the final tree would be *T*({1,2,3,4})= (((1,2),3),4) which is not optimal and that the known method “Neighbor joining” gives the same non-optimal tree.

A problem: Is there such mathematically correct algorithm?

REFERENCES

[1] K.Yu. Gorbunov, V.A. Lyubetsky. The tree nearest on average to a given set of trees. *Problems of Information Transmission*, 2011, Vol. 47, No. 3, P. 274–288.

[2] K.Yu. Gorbunov, V.A. Lyubetsky. Fast Algorithm to Reconstruct a Species Supertree from a Set of Protein Trees. *Molecular Biology* (*Moscow*), 2012, Vol. 46, No. 1, P. 161–167 .

[3] K.Yu. Gorbunov, V.A. Lyubetsky. Reconstructing the evolution of genes along the species tree. *Molecular Biology* (*Moscow*), 2009, Vol. 43, No. 5, P. 881–893 (received December 9, 2008; accepted for publication January 20, 2009).

[4] K.Yu. Gorbunov, V.A. Lyubetsky. An algorithm of reconciliation of gene and species trees and inferring gene duplications, losses and horizontal transfers. *Information Processes*, 2010, Vol. 10, No. 2, P. 140–144 (in Russian).

[5] J.-Ph. Doyon, C. Scornavacca, K.Yu. Gorbunov, G.J. Szollosi, V. Ranwez, V. Berry. An Efficient Algorithm for Gene/Species Trees Parsimonious Reconciliation with Losses, Duplications and Transfers. Article in the book: Comparative Genomics, Lecture Notes in Computer Science, Springer-Verlag Berlin Heidelberg, 2010, Vol. 6398, P. 93–108.

[6] K.V. Lopatovskaya, K.Yu. Gorbunov, L.Yu. Rusin, A.V. Seliverstov, V.A. Lyubetsky. The evolution of proline synthesis transcriptional regulation in gammaproteobacteria. *Moscow University Biological Sciences Bulletin*, 2010, Vol. 65, No. 4, P. 211–212 .