METHOD OF HORIZONTAL GENE TRANSFER DETERMINATION USING PHYLOGENETIC DATA

Lyubetsky V.A. [*], ***V'yugin V.V.***

Institute for Information Transmission Problems RAS, Moscow, Russia, e-mail: lyubetsk@iitp.ru

***Key words**: evolution, phylogenetic methods, trees consensus, horizontal gene transfer, mathematical models of evolution*

Resume

Motivation: An algorithm for comparative analysis of multiple trees reconstructed for representative protein families are discussed. This algorithm is based on the hypotheses of gene loss and horizontal gene transfers and uses stochastic methods and optimization. Some practical results are discussed. We describe a species tree comprising 40 prokaryotic organisms constructed by our algorithm on the basis of 132 individual groups of orthologous proteins (COGs) from GenBank of the National Center for Biotechnology Information (USA). We also describe a method for determination of horizontally transferred genes and its practical applications.
Results: An algorithm for horizontal gene transfer determination is developed. Several horizontally transferred genes were detected using this algorithm.
Availability: The software is available on request from the authors.

Introduction

The availability of numerous complete genome sequences from diverse taxa induces the development of new phylogenetic approaches, which incorporate information derived from comparative analysis of large gene sets. We consider two closely related approaches to reconstruction of phylogenetic relationships between species and to determination the horizontal gene transfer—the events occurring on molecular level-using these phylogenetic data.
An algorithm (V'yugin, Lyubetsky, 2001; V'yugin et al., 2002) for reconstructing phylogenetic species trees is described. This algorithm uses a set of possibly contradictory gene (protein) trees as initial data and produces a species tree as a census for these gene trees. The algorithm uses a model of gene duplications and losses explaining and measuring the dissimilarity between single gene tree and species tree (V'yugin, Lyubetsky, 2001). We also apply this model to determination of molecular events of horizontal gene transfer. Several computer experiments were realized using this algorithm. One of them is based on the data obtained from the National Center for Biotechnology Information, USA (Wolf et al., 2001). The data includes 132 maximum-likelihood trees constructed on the basis of 132 clusters of orthologous groups of proteins (COGs). The genomic sequences extracted belong to 40 living organisms from 13 groups: Archae (10), gamma-proteobacteria (7 organisms), gram-positive bacteria (8), alpha-proteobacteria (3), epsilon-proteobacteria (2), *Chlamydia* (2), spirochetes (2), beta-proteobacteria (1), cyanobacteria (1), *Deinococcus radiodurans* (1), *Aquifex aeolicus* (1), *Thermotoga maritima* (1), and *Mycobacterium tuberculosis* (1). The corresponding 132 maximum-likelihood protein trees were used as initial data in our method for constructing a census species tree. Species tree S is constructed as a tree closest to the 132 gene trees G_n. The measure of similarity is represented by a cost functional F. Definition of this functional is based on a model of evolution (V'yugin, Lyubetsky, 2001).

Methods and Algorithms

A stochastic algorithm of species tree reconstruction. The algorithm is based on a hypothesis of that the dissimilarity between gene trees constructed for different protein families results from a lineage-specific gene losses and duplications and horizontal gene transfer. We define a natural homomorphism (embedding) of gene tree into species tree and compare the gene and species trees by cost of this embedding. The value of functional F represents this cost (V'yugin, Lyubetsky, 2001). Following the principle of Occam's razor, we find a species tree minimizing this cost F of embedding (i.e. we try to find a species tree minimizing the total cost of molecular events of gene duplications and losses during the evolution of species). The search algorithm runs on a set of randomly generated (for example, 1000) initial species trees S_0. Any such tree S_0 is transformed using the method of nearest neighbor interchange to obtain a local minimum of the functional F. To specify this local minimum, we use an *a priory* probability distribution in the set of initial species trees. This probability distribution is defined as follows using 132 gene trees. Recall first that a distance between two leaves a and b is defined as a number of edges in the path between them. For any species a, an empirical probability distribution $p(b|a)$ that species b defines with a an elementary two-elements tree (i.e. they are located at a distance 2). We could define more detailed conditional distributions (for small trees of species located at a distance 3, 4, and so on), but this requires larger sets of initial data. The needed empirical probability distribution is defined using statistics of distribution of genes in 132 COG

[*] Corresponding author

trees. Let N_a be the number of COGs containing species *a*, and let $N_{a,b}$ be the number of COGs containing *a* and *b* located at a distance 2. We define $p(b|a) = N_{a,b}/N_a$. Then $1 - S_b\, p(b|\mathrm{a})$ is a probability that *a* forms a one-element elementary tree (i.e. no species locates at a distance 2 from *a*). The initial species *a* is defined using a uniform pseudorandom number generator, its neighbor *b* is defined using a generator of conditional probability distribution $p(.\ |\ a)$. We repeat this procedure to generate the next pair and so on. A random binary tree S_0 is generated on the basis of these elementary trees. This tree S_0 serves as the initial tree for the algorithm searching for optimal tree that gives a local minimum to the cost functional *F.* A variety of initial trees generate a variety of resulting trees produced by the search algorithm. As a final result, we output a consensus tree computed on the basis of a subset of these trees with sufficiently small values of the functional *F.* The numbers reflecting the reliability of corresponding clusters are assigned to the edges of this consensus tree. A resulting species tree explaining the evolutionary phylogenetic relationships of 40 living organisms (listed above) was constructed. This tree is close to a tree obtained in (Wolf et al., 2001) by comparative analysis of multiple trees reconstructed for representative protein families (approach (v) Wolf et al., 2001). Our species tree has very good suggestion for pairs of leaves and sufficiently good suggestion for the main 11 groups of organisms listed above. The difference between our tree and the best species tree from (Wolf et al., 2001) is only in the relative position of epsilon-proteobacteria group and the (Aae, Tma)-pair.

A method for detection of horizontally transferred genes. Horizontal gene transfer is a transfer of genes between organisms without reproduction. There are several hypotheses concerning the mechanisms of this transfer, for example, DNA can be transferred by infected bacteriophages or via mating mediated by plasmides (Lorencz, Wackernagel, 1994). In this section, we describe a method for detecting the genes suspected of being horizontally transferred. This method is based on a hypothesis implying that an event of horizontal gene transfer involves essential dissimilarity between the gene and species trees. We use two methods for estimating this dissimilarity. First is based on comparison of neighborhoods of a gene in the gene tree and its image in the species tree. If two genes are located at a small distance in the gene tree but their images are dispersed in the species tree, this indicates a possible horizontal transfer of one of them. We measure dispersion of a gene neighborhood under the homomorphism of gene tree into species tree. Let v1*,* $v_2, \ldots, v_n$ be all the genes located in the neighborhood of gene *v* of a radius *r,* and let $s_1, s_2, \ldots, s_n$ and *s* be the corresponding species (their images in species tree). The distances $r(v,v_i)$ in the gene tree and distances $r(s,s_i)$ in the species tree are calculated, where, *i* =1, …, n. We also calculate the average values as

$$r(v) = (1/n)\ S_i\, r(v,v_i) \text{ and } r(s) = (1/n)\square\ S_i\, r(s,s_i).$$

The ratio $p = r(s)/r(v)$ reflects the degree of average dissipation of the gene *v* vicinity in the species tree. Large values of this ratio can be interpreted as reflecting pathology in the location of gene *v* in the species tree. The computer program outputs the list of the genes suspected of being horizontally transferred. Each gene in this list is supplied with certain confidence information. In our analysis, we also take into account the diversity of COG (gene) trees. A high cost of COGs embedding into species tree reduced the confidence level of the results concerning genes from this COG.

The second approach uses a hypothesis that the temporary deletion of a transferred gene from the gene tree and updating of its image in the species tree after this deletion implies essential decrease in the value of cost functional *F* (for example, the cost can be decreased by many sigma from a mean). To apply this method, a normalization of the gene trees is needed. The edges of maximum likelihood gene trees are supplied with the numbers (lengths) reflecting the time of evolution of corresponding genes. The longest lengths have the biggest impact on the total value of functional *F.* This, wrong locations of these longest edges (which can result from inaccuracy of the maximum likelihood method) bring about the biggest error to the value of the functional. To eliminate this effect, we normalize excessively long leaf edges in the gene tree. For any COG tree, the cost *F* of embedding of the corresponding tree in the species tree is calculated. We remove temporarily each gene *g* from gene tree *G* to obtain a reduced gene tree G_g and compute the cost F_g of embedding the gene tree G_g in the species tree. The relative change in the cost of embedding is calculated as $dF_g = (F_g - F)/F.$ We sort all the genes from the COG by absolute values of dF_g. We suppose that the order numbers of genes suspected of being horizontally transferred has a high correlation with large absolute values of dF_g. More correctly, the mean and variance of dF_g were used to compute the confidence information for each gene from any COG. Additional confidence information was computed as follows. For each gene *g*, a value

$$dF_{cp}{}^{r}(g) = (1/n(r))\ S_{g:r(s,g)<r,\ s\neq g}\, dF_s$$

was calculated, where $n(r)$ is the number of leaves in the neighborhood of *g* of a radius *r.* In our experiments, we used *r* = 1, 2, 3, 4, and 5. A large value of

$$k_g{}^{r} = (dF(g))/(\square\ dF_{cp}{}^{r}(g))$$

suggests that gene *g* in the species tree displays a pathological location.

This approach can be applied more efficiently to a case of horizontal transfer of groups of genes. Combined lists of genes suspected of being horizontally transferred were formed on the basis of both methods. Each gene in the list is supplied with confidence information. This information can serve as a tool for an expert analyzing the molecular events in the process of evolution.

Implementation and Results

In this section, we present some results of detecting genes suspected of being horizontally transferred in the descending order of their reliability levels.

1) *yicF* is a gene in the COG0272 extracted from the genome of *E. coli* (gamma-proteobacteria group). The gene was selected as (a) a temporary deletion of it from the gene tree gives a large deviation of the value dF_g from the mean value of this variable (18 sigma) and (b) the nearest neighbors of this gene in the gene tree embed into group of gram-positive bacteria located at a long distance from gamma-proteobacteria in the species tree (the ratio $p = r(s)/r(v)$ defined above is equal to 8). We suppose that the group of gram-positive bacteria is the source of this horizontal transfer.
2) *aq946* is a gene in the COG0571 extracted from the genome of *Aquifex aeolicus;* it also displays a large deviation of the value dF_g from the mean value of this variable (15 sigma), and the nearest neighbors of this gene in the gene tree embed into group of alpha-proteobacteria. This group is supposed to be the source of this horizontal transfer.
3) *VNG1097G* is a gene in the COG0215 extracted from genome of *Halobacterium sp.* NRC-1 (the family Archae). A temporary deletion of this gene from the gene tree gives a large deviation of the value dF_g from the mean value (10 sigma), and the nearest neighbors of this gene in gene tree embed near *Deinococcus*, which is located in the species tree at a long distance from Archae.
4) *RP687* is a gene in the COG0525 extracted from the genome of *Rickettsia prowazekii* (the group of alpha-proteobacteria). The nearest neighbors of this gene in gene tree embed near Archae, which is located in the species tree at a long distance from alpha-proteobacteria group. A deletion of this gene also changes essentially the value of the functional.
5) *VNG2507G* is a gene in COG0167 extracted from the genome of *Halobacterium sp.* NRC-1 (the group of Archae). The nearest neighbors of this gene in gene tree embed near the group of *Mycobacterium tuberculosis, Synechocystis,* and *Deinococcus radiodurans,* which is located in the species tree at a long distance from Archaea. A deletion of this gene also changes essentially the value of the functional.

References

1. Lorencz M.G., Wackernagel W. (1994). Bacterial gene transfer by natural genetic transformation in the environment. Microbial Reviews. 58:563-602.

2. V'yugin V.V., Gelfand M.S., Lyubetsky V.A. (2002). Trees reconciliation: species trees reconstruction by phylogenetic gene trees. Mol. Biol. (accepted for publication).

3. V'yugin V.V., Lyubetsky V.A. (2001). On algorithm of horizontal gene transfer searching based on phylogenetic protein trees. Informatsionnye Protsessy. 1(2):167-177.

4. Wolf Y., Rogozin I., Grishin N., Tatusov R., Koonin E. (2001). Genome trees constructed using five different approaches suggest new major bacterial clades. BMC Evol. Biol. 1:8.