Modelling co-evolution of regulatory systems, genes and species with supercomputers

K.Yu. Gorbunov, V.A. Lyubetsky

Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute), 19 Bolshoy Karetny, 127994 Moscow, Russia

gorbunov@iitp.ru

1 Abstract

An original model is developed to describe the co-evolution of regulatory systems, genes and species (ref. to sections 2-5). The model was extensively tested on simulated and biological data. We will present the underlying supercomputer techniques and the obtained results.

2 Reconciliation of a set of gene trees

A long recognized problem is inference of a tree S that reconciles a set of input trees G_i, with leaves in each G_i being assigned homologous sequences from an i-family (homologous genes or regulatory regions with or without the regulated gene, etc.). Usually the tree S is a tree of species or other taxonomic units. Assume leaves in S are labelled with species names, leaves in G_i – with pairs of gene-species names; paralogs are allowed. We further develop a traditional approach to find the tree S such that it minimizes the total cost of mappings of individual trees into S, [6]. Let us call S a supertree, each G_i – a gene tree, each sequence – a gene. A mapping of G_i into S implies a fixed set of evolutionary events. The *standard set* contains only gene duplications and losses. The *extended set* additionally contains horizontal gene transfers, gains, etc. The *total cost* is the sum of "individual" *costs* for all G_i mappings into the tree S, and is similar to the cost from [6] in case of the standard set. Is computing total/individual costs possible in small polynomial time? This question is tackled below.

Under a traditional approach, the supertree building problem is NP-hard, i.e., any algorithm to solve it correctly must possess exponential complexity. Numerous heuristics exist, but they generally do not find the global minimum of the total mapping cost. We proposed a reformulation of this problem that allows a computationally effective deterministic algorithm and meets many biological prerequisites. Namely, the supertree S is sought for such that it contains the majority of clades from input trees G_i. With the standard event set, the algorithm is mathematically correct and possesses the running time of $O(n^3Чm^3)$, where n is the number of gene trees, and m is the total number of species [2]. For simplicity, here we assume that the average number of leaves in input trees is a multiple of m. With an extended event set, the algorithm is heuristically correct and cubic in complexity. The authors are unaware of analogous approaches in published literature. A relevant biological discussion of our approach is provided in [3].

The algorithm realization is available at http://lab6.iitp.ru/en/super3gl/.

3 Reconciliation of gene and species trees

Edges in S may be broken by inserting additional nodes, thus formally producing another tree S_0, with nodes producing two descendant edges or only one edge. It imposes time slices such that horizontal transfers are allowed only within one slice, see [4]; in particular cases S=S_0. With the extended event set, we developed an algorithm that reconciles any gene tree G and S_0, i.e., correctly computes the mapping of G into S_0 and its cost in time $O(|G|Ч|S_0|)$, which gives $O(|S|^3)$. Here | | is the number of nodes in a tree. A mathematical proof is given in [4, 5] (refer also to a later study [1], which has used [4, 5]). Refresh that a phylogenetic net of genes or species is an acyclic directed graph with one vertex (the "root") that can be connected by a path with any other node, and terminal vertices (the "leaves"); the leaves are labelled with species names or species-gene name pairs. An important special case is a binary net, where for each node, except for the root, one of the following is true: the node possesses only one incoming edge and no outcoming edges, or two outcoming and one incoming edge, or two incoming and one outcoming edge. In a species net, introducing time slices ordinarily generates vertices with one incoming and one outcoming edge. The definitions and costs of mapping of a gene net into a species net and mapping of a gene tree into a species tree are identical.

The algorithm realization is available at http://lab6.iitp.ru/en/embed3gl/.

4 Mapping a gene tree to a species tree and its cost

Thus, Problems form the sections 2-3 are reduced to the extended event set case. We proposed a possible definition in [4, 5] and formulate the idea below; informal motivation can be found in [4].

Below e runs over all edges of G, and d – over all edges (also referred to as *tubes*) of S_0. The root is pictured at the top in all. A formal edge (the so-called *root edge*) enters the root from above. By definition,

$e'<e$ if an edge e' is strictly below e, and $e'\leq e$ if $e'<e$ or $e'=e$. One of at most 15 evolutionary events can occur on fixed edge e in tube d; such events are marked by index i. Two tubes d and d' belong to the same (temporal) *slice*, by definition, if the lengths of path from the root tube to d and d' are equal, correspondingly. Assume d is a terminal tube, and e is a terminal edge. If gene e belongs to species d, then the mapping $f_{<e,d>}(e) = <d,fin>$ and the cost $c(<e,d,fin>) = 0$. Otherwise, if e belongs to species d', then the mapping $f_{<e,d>}(e,fin) = <d,tr(d,d')>$ and the cost $c(<e,d,fin>) = 13$, where 13 is the cost of gene transfer without retention (all costs are conditional). Here $tr(d,d')$ indicates that gene e is transferred from tube d to tube d' such that no copy of e remains in d, and 0 is the index of such event. Thus, mapping $f_{<e,d>}(e,i)$ is defined by induction. Now we exemplify further induction steps $f_{<e,d>}(e') = <d', mark>$ where $e' < e$ and d' is in a slice later or equal then the slice d.

Assume d is a tube with single descendent tube d_1 and $pass = \arg_i \min c(e,d,i)$. Then $f_{<e,d>}(e) = <d,pass>$, $f_{<e,d>}(e') = f\ _{<e,d1>}(e')$, and $c(<e,d,pass>) = \min_i c(<e,d_1,i>)$. Here, the "*pass*" indicates the survival of gene e down to the next tube.

Assume d is a tube with two descendent tubes d_1 and d_2 and $passl = \arg_i \min c(e,d,i)$. Then $f_{<e,d>}(e) = <d,passl>$, $f_{<e,d>}(e') = f_{<e,d1>}(e')$, and $c(<e,d,passl>) = \min_i c(<e,d_1,i>)+2$, where 2 is the cost of loss. Here, the "*passl*" indicates the survival of gene e into tube d_1 and loss of its copy in tube d_2. Symmetric cases are not discussed everywhere.

Denote e_1 and e_2 two descendent edges of non-terminal edge e and $passlr = \arg_i \min c(e,d,i)$. Then $f_{<e,d>}(e) = <d,forklr>$, $f_{<e,d>}(e') = f_{<e1,d1>}(e')$ if $e' \leq e_1$, and $f_{<e,d>}(e') = f_{<e2,d2>}(e')$ if $e'\leq e_2$; $c(<e,d,passlr>) = \min_i c(<e_1,d_1,i>)+\min_i c(<e_2,d_2,i>)$. Here, the "*forklr*" indicates divergence of gene e into e_1 in tube d_1 and e_2 in tube d_2.

For non-terminal edge e and $dupl = \arg_i \min c(e,d,i)$, $f_{<e,d>}(e) = <d,dupl>$, $f_{<e,d>}(e') = g_{<e1,d>}(e')$ if $e' \leq e_1$, and $f_{<e,d>}(e') = f_{<e2,d>}(e')$ if $e'\leq e_2$; $c(<e,d,dupl>) = \min_i c(<e_1,d,i>)+\min_i c(<e_2,d,i>)+3$, where 3 is the cost of duplication. Here, the "*dupl*" indicates the duplication of gene e within tube d into e_1 and e_2.

For non-terminal edge e assume $\exists d'\ [tr1(d,d') = \arg_i \min c(e,d,i)]$, where d' be a tube in the same slice with d that differs from d and minimizes the value $\min_i c(<e_1,d',i>)$ over d'. Then $f_{<e,d>}(e) = <d,tr1(d,d')>$, $f_{<e,d>}(e') = f_{<e1,d'>}(e')$ if $e' \leq e_1$, and $f_{<e,d>}(e') = f_{<e2,d>}(e')$ if $e' \leq e_2$; $c(<e,d,tr1(d,d')>) = \min_i c(<e_1,d',i>) + \min_i c(<e_2,d,i>)+11$, where 11 is the cost of transfer with retention. Here, the "$tr1(d,d')$" indicates that gene copy e_1 is transferred from tube d into tube d', and gene copy e_2 remains in d.

5 Gene tree reconstruction

The authors propose the problem restatement and a polynomial algorithm that correctly infers the gene tree in special cases. In simulations of a general case the algorithm was shown to be very fast and in about 75% cases reconstruct a tree very close to that produced by PhyloBayes v.3.3 in much longer time. The tree is sought for among trees consisting of clades from a prebuilt set P that possesses the following property: each set in P can be split into two subsets also from P, and so on until singlet sets are obtained that correspond to alignment rows (ref. to further as rows). The solution found by the algorithm depends heavily on the "correctness" of set P.

In the algorithm, selected "non-informative" columns of the initial alignment $\boldsymbol{X}$ are omitted as in [7], dynamic programming is then applied to the refined alignment for a maximum likelihood inference of one tree that contains all clades from P per alignment column. The obtained set of trees is reconciled using a supertree building algorithm described in section 2 and in [2, 3]. Set P is constructed with induction. All singlet sets (i.e. consisting of a single row) are included in P; at each induction step, if sets X_1 and X_2 from P do not intersect, and for their union X the conditions $S(X) \geq l(|X|)$, $m^2(X) + \beta > M^{\,2}(X)$ are true then X is included in P. Define $I(X)$ the sum of self-similarities of each row from X divided by the total length of all rows from X. Define $S(X)$ the sum of similarities of all row pairs from X divided by the product of total pairs length and $I(X)$. Define $m(X)$ the minimal value of S over all pairs from X; $M(X)$ its maximal value over all pairs that contain one row from X and another – from its complement. Let $l(m)$ be a linear function, where m decreases from $|\boldsymbol{X}|$ to 2, $l(|\boldsymbol{X}|) = S(\boldsymbol{X})$, and the derivative equals α, the algorithm parameters are α and β.

The algorithm realization is available at http://lab6.iitp.ru/en/gene3gl/.

References

[1] J.-P. Doyon et al. An Efficient Algorithm for Gene/Species Trees Parsimonious Reconciliation with Losses, Duplications and Transfers. *Lecture Notes in Bioinformatics*, 2010, **6398**, 93-108.

[2] K. Gorbunov, V. Lyubetsky. The tree nearest on average to a given set of trees. *Problems of Information Transmission*, 2011, **47**(3), 274-288.

[3] K. Gorbunov and V. Lyubetsky. Fast Algorithm to Reconstruct a Species Supertree from a Set of Protein Trees. *Molecular Biology* (*Mosk*), 2012, **46**(1), 161-167.

[4] K. Gorbunov and V. Lyubetsky. Reconstructing the evolution of genes along the species tree. *Molecular Biology* (*Mosk*), 2009, **43**(5), 881-893.

[5] K. Gorbunov and V. Lyubetsky. An algorithm of reconciliation. *Information Processes*, 2010, **10**(2), 140-144 (in Russian).

[6] R. Guigo et al. Reconstruction of ancient molecular phylogeny. *Molecular Phylogenetics and Evolution*, 1996, **6**, 189-213.

[7] V. Lyubetsky et al. Removing noise in a multiple protein alignment. *Information processes*, 2005, **5**(5), 380-391 (in Russian).