Nucleotide, gene and genome evolution : a score to bind them all

Wandrille Duchemin, Vincent Daubin, Eric Tannier

Nucleotide, gene and genome evolution : a score to bind them all

Wandrille Duchemin [*†1], Éric Tannier [*1,2], Vincent Daubin [*1]

Session Phylogénie 1
mardi 28 15h20
Amphi Mérieux

[1] Laboratoire de Biométrie et Biologie Évolutive (LBBE) – CNRS : UMR5558, Université Claude Bernard - Lyon I, INRIA – 43 boulevard du 11 Novembre 1918, F-69 622 Villeurbanne Cedex, France
[2] BEAGLE (Insa Lyon / INRIA Grenoble Rhône-Alpes / UCBL) – INRIA, Institut National des Sciences Appliquées [INSA] - Lyon, Université Claude Bernard - Lyon I (UCBL) – Antenne INRIA Lyon la Doua, Bâtiment CEI-1, 66 Boulevard Niels Bohr, F-69 603 Villeurbanne, France

The history of genomes is usually studied through a pipeline of independent, successive steps, each answering a specific question by optimizing a specific metric. For instance sequences homology must first be assessed in order to delineate gene families. The sequences of each gene from a family are then aligned together. This alignement is used to generate an history of the gene family, for instance a unique phylogenetic tree (a "gene tree") or a distribution of phylogenetic trees if one wants to account for the uncertainties of the reconstruction. Using a species tree, it is then possible to associate each node of a gene tree with a species (ie. a node in the species tree) and one or several evolutionary events (for instance: speciation, duplication or horizontal gene transfer). This process is called a reconciliation. It is also possible through this method to detect the loss of a gene copy in a lineage of the species tree. The result of such an association is termed a reconciled gene tree. The reconciled tree of different gene families are then combined with relationships information such as gene-order, interactions, co-regulation or co-expression that links some extant gene copies together. Finally, ancestral relationships can be inferred along with some information about the history of these relationships.

Each of these steps represents a complex task and each has a consequent body of literature attached to it. As each step is performed independently, it is possible that the optimization of a previous step sets the following ones in a sub-optimal space of solutions. For instance, the gene tree with the maximal likelihood might require numerous evolutionary events of transfer and loss to be reconciled with the species tree, while a gene tree only slightly less likely with respect to the alignement could allow some simpler reconciliation scenarios. Furthermore, as this pipeline is done for each gene family separately, choices are made that might downplay the amount of coevolution that can be expected from genes evolving in the same species and with sometimes related function or neighbouring positions on a chromosome [Liang2010]. Some methods jointly infer the gene tree topology and the gene tree reconciliation, yielding a more comprehensive view of the gene history [Szöllősi2013b, Scornavacca2014], but they still consider each gene family independently from each other. For instance, two genes both undergoing a duplication event in the same species are usually seen as two separate duplication events (and thus costs twice the cost of a duplication in a parsimonious framework). Yet, if this two genes happen to be neighbours in that species, then it is likely that only one duplication event occurred which encompassed both genes.

Here we will focus on integrating three levels of inference in order to allow for the co-evolution of genes: gene tree reconstruction, gene tree reconciliation and adjacency tree building. In this work, we define a gene as a block of nucleotides that cannot be broken into sub-genes or undergo internal rearrangement. In practice, it can be a protein domain, a entire coding sequence, or any segment of a chromosome. We define adjacencies as binary relationships between genes. Adjacencies might represent diverse notions of relatedness such as co-function, co-expression

*. Intervenant
†. Corresponding author : wandrille.duchemin@univ-lyon1.fr

or co-occurrence on a chromosome. The history of a group of homologous adjacencies can be summed up in an adjacency tree [Bérard2012].

We propose a score that integrates gene tree reconstruction, reconciliation and adjacency history building into a parsimonious framework, but also incorporates a notion of coevents: events regrouping several gene copies. By introducing this notion, we account for the coevolution between neighbouring genes. We are able to propose solutions that may not be optimal when gene families are considered as independently evolving but yield more coherent global scenarios of evolution when all gene families are allowed to co-evolve.

Methodologically, several algorithms are assembled to provide the different component of the proposed score. By using conditional clades probability [Höhna2012], we can estimate the score of a single gene tree from a posterior distribution of trees. We can also compute the score of a gene tree reconciliations, including events of duplication, loss and lateral gene transfer possibly from extinct or unsampled lineages of the species tree [Szöllősi2013a]. This is typically done using the parsimonious reconciliation algorithm of [Doyon2010] as implemented in TERA [Scornavacca2014]. The score of gene adjacency histories are computed using the algorithm of DeCoLT [Patterson2013], which computes the parsimonious adjacency histories given reconciled gene tree and extant adjacencies. Coevents are computed from the results of the DeCoLT algorithm and used to correct a weighted sum of the scores yielded by the different algorithms in order to obtain the global score. Optimizing such a score through an exhaustive exploration of the space of solution would be intractable, as the spaces of all gene trees topologies, all gene trees reconciliation, and adjacency history of all gene families would have to be combined.

We propose an exploration strategy based on heuristic and local moves. The idea is that the initial solution has been obtained by optimizing each element (gene family, for instance) without taking any notion of coevent into account. Thus, to get a better global score, coevents whose global score correction compensate the loss of local optimality must be proposed.

This approach can be used at a variety of scale, e.g. by considering protein domains as unit to reconstruct the history of modular proteins, or by considering genes to reconstruct the history of chromosomes and metabolic networks. Other than providing more coherent evolutionary scenarios, coevents can prove a useful tool to study the dynamics of genomic events of duplication, loss and transfer. They could be used to study their size (not being limited at the size of a gene) or give a better estimate of the frequency of these events (because they can effectively detect several neighbouring gene undergoing the same event as one event rather than several).

References

[Doyon2010] J.-P. Doyon, C. Scornavacca, V. Ranwez, V. Berry (2010). An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers. In: *Proceedings of the 2010 International Conference on Comparative Genomics, RECOMB-CG'10*, Springer-Verlag, Berlin,Heidelberg, 93-108

[Liang2010] Z. Liang, M. Xu, M. Teng, et al. (2010) Coevolution is a short-distance force at the protein interaction level and correlates with the modular organization of protein networks. *FEBS Letters* 19:4237-4240

[Bérard2012] S. Bérard, C. Gallien, B. Boussau, et al. (2012) Evolution of gene neighborhoods within reconciled phylogenies. *Bioinformatics* 28(18):i382-i388

[Höhna2012] S. Höhna, A. Drummond (2012) Guided tree topology proposals for Bayesian phylogenetic inference. *Systematic Biology* 61:1–11.

[Patterson2013] M. Patterson, G. J. Szöllősi, V. Daubin, É. Tannier (2013) Lateral gene transfer, rearrangement, reconciliation. *BMC bioinformatics* 14(Suppl 15):S4

[Szöllősi2013a] G. J. Szöllősi, É. Tannier, N. Lartillot, V. Daubin (2013) Lateral Gene Transfer from the Dead. *Syst Biol* 62:386-397

[Szöllősi2013b] G. J. Szöllősi, W. Rosikiewicz, B. Boussau, et al. (2013) Efficient exploration of the space of reconciled gene trees. *Systematic Biology* 6:901-912

[Scornavacca2014] C. Scornavacca, E. Jacox, G. J. Szöllősi (2014) Joint amalgamation of most parsimonious reconciled gene trees. *Bioinformatics* 6:841-848

Mots clefs : phylogeny, reconciliation, adjacency, coevent