Models, algorithms, and programs for phylogeny reconciliation

Jean-Philippe Doyon, Vincent Ranwez, Vincent Daubin, Vincent Berry

Models, algorithms, and programs for phylogeny reconciliation

Keywords: phylogeny; gene duplication, loss, horizontal gene transfer; parsimony; probability; reconciliation.

1. Institut des Sciences de l'Evolution

Our research focuses on the patterns and mechanisms of evolution in the living world. It aims to better understand the processes involved in the origins and dynamics of biodiversity.

The specific approach taken by scientists at the ISEM is to work on both extant and fossil material and to study plants, animals, and micro-organisms.

2. Jean-Philippe Doyon

jean–philippe.doyon@univ–montp2.fr

Université Montpellier II, Institut des Sciences de l'Evolution
Montpellier, France

Jean–Philippe Doyon received the PhD degree in computer science from the University of Montréal, Canada, in June 2010. His research focuses on comparative genomics, especially combinatorial and probabilistic aspects of reconciliations between gene and species trees, and the development of efficient algorithms in these area.

3. Vincent Ranwez

vincent.ranwez@univ-montp2.fr

Université Montpellier II, Institut des Sciences de l'Evolution
Montpellier, France

4. Vincent Daubin

vincent.daubin@univ-lyon1.fr

Université Lyon 1, Laboratoire de Biométrie et Biologie Evolutive
Lyon, France

5. Vincent Berry

vberry@lirmm.fr

Université Montpellier 2, Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier
Montpellier, France

Models, algorithms, and programs for phylogeny reconciliation

JP. Doyon, V. Ranwez, V. Daubin, V. Berry

Abstract. Gene sequences contain a gold mine of phylogenetic information. But unfortunately for taxonomists, this information does not only tell the story of the species from which it was collected. Genes have their own, complex histories, which record speciation events of course, but also many other events. Among them, gene duplication, transfer and loss are especially important to identify and differentiate from speciation events. These events are not only crucial to account for when reconstructing the history of species, but they also play a fundamental role in the evolution of genomes, the diversification of organisms and the emergence of new cellular functions.

We present a state of the art of reconciliations between gene and species trees, which is a rigorous approach to identify the nature of the various events that mark the evolution of a gene family. Existing models and algorithms of reconciliations are reviewed, and difficulties to model gene transfer are discussed. We also compare different reconciliation programs together with their advantages and disadvantages and summarize reconciliation results obtained on numerous gene families.

Keywords: phylogeny; gene duplication, loss, horizontal gene transfer; parsimony; probability; reconciliation.

1 Introduction

The systematic reconstruction of gene phylogenies from a wide variety of organisms reveals an unforeseen diversity of histories, which are difficult to understand through a simple pattern of species evolution. These differences underline the importance of modelling the factors that specifically affect gene evolution. Population genetics has brought a powerful framework in which to understand the pattern of evolution of alleles through a succession of speciations. Under these multispecies coalescent models, the differences among gene phylogenies turn out to be informative for estimating population size, structure, and speciation times [1]. On another side, models of genome evolution, which account for the mechanisms of gene duplication, gene loss and Lateral Gene Transfer (LGT), are emerging. The development of such models is crucial especially to understand the evolution of unicellular organisms where LGT has played a major role [2], and more generally to clarify homology relationships among genes. The combination of duplication, transfer and loss may have been such in the history of life that no single phylogenetic marker can be considered reliable to readily infer the history of species. The following studies arguably rely on the development of models of duplication, transfer and loss [3]: reconstructing the tree of life; understanding the principles of genome evolution, the role of transfer in species adaptation, and the contribution of duplication and transfer to the evolution of new functions.

As in the multispecies coalescent model, reconciliation models consider a species tree within which a gene can evolve (Figure 1). Leaves of the species tree and the gene tree are associated and specific events are invoked to allow the gene tree to evolve within the species tree with its own phylogeny. Partial models accounting for duplication and loss alone [4,5,6], or LGT and loss alone [7,8,9] have been described. These models are realistic in particular biological cases (resp. multicellular organisms where LGT is rare and gene families for which functional redundancy can be detrimental). Here, we focus on models that account for DL (Duplication and Loss) and DTL (Duplication, Transfer and Loss) events.

Figure 1

Reconciliation is a popular approach to infer orthology relationships [10–14]. It has been shown on real datasets that such phylogeny-based methods are relatively accurate [15]. Other methods that rely solely on sequence similarity can lead to false positives and false negatives [16]. Reconciliation approach has applications in other areas such as estimating DTL rates [17,18], gene tree inference based on molecular substitution and gene evolution [19–21], and genome phylogeny reconstruction from discordant gene trees [5,22]. Reconciliation can also be used to study coevolution between parasite and their hosts (parasitology) and between organisms and their living areas (biogeography) [23–25].

Sections 2.1 and 2.1 review the main approaches of reconciliation with DL and DTL events, respectively. Section 2.1 also explains why transfers are hard to consider and describes several methods to handle them. Section 2.1 presents the pros and cons of available reconciliation software. Section 2.1 presents results of reconciliation analyses on a real dataset that covers the three domains of life.

2 Models and algorithms for reconciliation

As for phylogenetic reconstruction, parsimony and probabilistic frameworks have been developed for reconciliation inference. The former are based on explicit discrete models of gene evolution and search for a reconciliation of minimal cost given costs for individual evolutionary events. Probabilistic methods rely on continuous models and seek a reconciliation with maximum likelihood or maximum posterior probability. Parsimony methods are faster but use less realistic models than probabilistic ones.

2.1 Evolutionary scenarios with duplications and losses

Historically, DL models are the first ones to have been considered and remain relevant when only considering eukaryotes.

Parsimony models

Numerous reconciliations can exist between a gene tree G and a species tree S. For instance, Figure 1 depicts two reconciliations R1 and R2 for the same trees, which differ by the location in S of the duplication u of an ancestral gene of G (and by the induced losses). Several ways to represent reconciliations have been proposed. One of the most widespread is through a tree called a *Reconciled Tree* [4,6,26], here denoted RT, and defined as follows: (1) the clade of each node of RT has to be present in S; (2) for each internal node of RT, the clades of its children are either equal (duplication) or disjoint (speciation) and (3) G has to be obtained from RT by pruning some of its subtrees (see Figure 2 below).

Figure 2

The model of reconciled tree (Figure 2; RT) is sometimes less intuitive than a drawing of G embedded into S (Figure 2; R1). A model that leads to such a representation is defined in [27], where each ancestral gene of G is mapped either on a vertex (speciation) or a branch (duplication) of S.

Based on the reconciled tree formalism, [28] have introduced an architecture that allows to describe and explore the whole set of reconciliations between G and S. They proved that a simple polynomial time algorithm, called LCA mapping, allows to find one of the *Most Parsimonious Reconciliations* (MPR). This well-known algorithm [4,6,29] can be implemented to run in time linear w.r.t. the number of nodes in G [30]. The LCA mapping maps each gene u of G onto the most recent species x of S such that each contemporary gene that descends from u belongs to a contemporary species that descends from x (see Figure 3). According to this LCA mapping, a node u of G is a duplication if, and only if, it is mapped to the same vertex of S than one of its children. Otherwise, u is a speciation. Based on this rule, the so-called LCA reconciliation is computed as follows. For each node u of G, mapped on a vertex x of S, if u is a duplication, it is located on the branch immediately above x, otherwise u is a speciation placed on x. For instance, R1 of Figure 1 is the LCA reconciliation, while R2 is not.

Figure 3

Given a cost for each event (i.e. duplication, loss and speciation), the parsimony score of reconciliations can be defined in various ways. It can be either the sum of costs for duplications or the sum of costs for duplications and losses. The LCA reconciliation provides an MPR for the two scores, and is even the unique one in the latter case [28,31]. Given a choice for event costs, numerous reconciliations may exist with near-optimal scores. Some of them could have been optimal if slightly different event costs were used [32]. However, most reconciliation analyses focus on the LCA reconciliation and ignore such near-optimal reconciliations.

All branches of a phylogenetic tree are not equally reliable, bootstrap and posterior probabilities being usual support measures. Uncertainties can have

an important effect on reconciliation analyses. Reconciliation methods are biased when the inferred gene tree is not correct: duplications and losses tend to be placed toward the root and the leaves, respectively, of the species tree [33]. This uncertainty concerning parts of the gene tree can be taken into account by collapsing weakly supporting branches, thus creating polytomous nodes. An MPR can still be computed in polynomial time when S contains polytomies [34], but only approximate algorithms exist when G is polytomous [35,36].

Probabilistic models

On the basis of a Birth-and-Death process [37], a probabilistic model of gene evolution along a species tree S has been developed [38], where each branch has fixed length (in time) and associated duplication and loss rates. This model is used to compute the probability that the evolution of a gene which belongs to the root of S gives rise to the gene tree G and to the reconciliation R. This probability is the likelihood of R, denoted P(G,R), and can be computed in time $O(n_G n_S)$, where n_G and n_S are the number of nodes in G and S. The reconciliation of maximum likelihood can also be computed in time $O(n_S n_G \log^3 n_G)$. This reconciliation model has been extended into mixed models which additionally consider the probability that G gives rise to the observed sequences. Such extensions assume either a strict molecular clock [38] or a relaxed one [19].

An important breakthrough is due to [39], who developed an efficient algorithm to compute the probability P(G), which is the sum of P(G,R) over all reconciliations R. This algorithm allows to compute the posterior probability P(R|G) = P(G,R)/P(G) in time $O(n_G^2 n_S)$. This posterior probability is useful to evaluate the reliability of the most likely scenario with respect to other reconciliations, particularly those having close likelihood. A similar algorithm computes the probability that a given ancestral gene of G is a speciation in time $O(n_G^2 n_S)$. By sampling duplication and loss rates, [40] developed an MCMC that uses the latter algorithm to estimate posterior probabilities of orthology relationships among sequences of a gene family.

[27] have proposed a similar algorithm as [28] to explore the space of reconciliations. The idea is to go from one reconciliation to another thanks to elementary changes in the reconciliation. This algorithm can be used in order to compute or efficiently estimate P(G) and thus the posterior probability P(R|G) [41]. Their analysis of the probabilistic landscape of the space of reconciliations shows that the immediate neighbourhood of the MPR (i.e. the LCA reconciliation) contains the most likely ones and covers most of the probability mass of P(G). This provides an efficient way to obtain very precise posterior probabilities of the most probable reconciliations. These results emphasize the strong relationship between the parsimony and probabilistic paradigms.

When considering a discrete distribution model of duplications parameterized by the branch lengths of S, a maximum likelihood

reconciliation can be computed in time $O(n_G{}^4 n_S)$ [42]. However, this is achieved without taking into consideration events that left no trace [1] nor loss rates, in contrast with the above Birth-and-Death models. Such limitations can be problematic when losses are prevalent [43]. This approach is computationally more simple but less realistic than the above models.

2.2 Evolutionary scenarios with duplications, losses, and transfers

Computing an MPR becomes a computationally difficult problem when transfers are considered [44], although it can be solved in polynomial time with realistic constraints [45] (e.g. bounding the number of transfers, of genes per species lineage, etc.). This strong contrast in complexity is due to the chronological constraints induced by transfers. A transfer has to be locally consistent, which means that it occurs between two coexisting species. Two (or more) consecutive transfers have also to be globally consistent (see Figure 4). If these constraints are omitted, time inconsistent scenarios can ensue. [46] solves in time $O(n_S{}^2 n_G)$ such a variant of the MPR problem, where losses are considered a posteriori.

Figure 4

A recent promising approach to handle time constraints is to accept a dated tree S as input. Time consistency can then be ensured locally by checking that donor and receiver branches of a transfer have intersecting time intervals (see Figure 4). This approach has been used in five reconciliation models: Merkle et al. [47], Hadas et Charleston [48], Tofigh [32], Doyon et al. [49] and David et Alm [18]. These approaches differ in their way to handle global consistency and in the degree of generality of their model.

Global time consistency can be ensured with the following approaches: (1) alter the position of the proposed transfers a posteriori; (2) check that all branches involved in a succession of transfers share a sub-interval of time; or (3) use a subdivision of the branches of S into time slices and allow for transfers only between branches of a same time slice. The two-steps strategy of (1) does not guarantee to find an optimal reconciliation. Approach (2) only considers a subset of scenarios due to over-restrictive rules. For instance, reversing the direction of transfers T1 and T2 in Figure 4 leads to a globally consistent scenario that is here however rejected, since the three involved branches do not share a common time subinterval. Approach (3) ensures to find an optimal reconciliation [48,49].

Models have to be general enough to encompass the variability of all possible scenarios involving transfers. In particular, they have to consider the following two possibilities: (a) Transfers where the donor branch Looses its gene copy[2] (TL event for short; R1 Figure 5); and (b) scenarios where

[1] A gene duplication immediately followed by a loss of one of the copies.

[2] That is the gene copy of the donor left no trace in the contemporary species (i.e. it goes extinct).

speciations and duplications are located below the LCA vertex of S. Due to transfers, constraining such events to be located on or above the LCA is no longer guaranteed to be optimal. For instance in Figure 5, R1 is more parsimonious than R2, due to points (a) and (b) above. In R1, the gene lineage (u,b) follows a TL event from (x,A) toward (z,B) and node w is a speciation located below its LCA mapping (vertex y of S).

Figure 5

According to the features introduced above, Table 1 below summarizes the pros and cons of the five reconciliation models presented in [18,32,47–49,52].

	Input Gene/Species Trees		Model characteristics		Algorithm	
	Tree G	Tree S	Transfer with Loss	Location of Spec. / Dup.	Global consistency	Time complexity
Merkle et al. [47,52]	Binary and polytomous; time interval	Binary; dated	No	Only on/above the LCA	Not guaranteed (considered a posteriori)	$O(\max(n_S, n_G)^3)$
Hadas et Charleston [48]	Binary	Network; dated;	No	Anywhere	Guaranteed, with time slices	Polynomial in ,
Tofigh [32]	Binary	Binary; dated	Yes	Anywhere	Guaranteed, with time slices	$O(n_S{}^5 n_G)$
Doyon et al. [49]	Binary	Binary; dated	Yes	Anywhere	Guaranteed, with time slices	$O(n_S{}^2 n_G)$
David et Alm [18]	Binary	Binary; dated	No	Only on/above the LCA	Guaranteed, with simple rules	$O(n_S{}^3 n_G)$

Table 1 [Comparison of five reconciliation models accounting for duplications, losses and transfers]. The models of [32] and of [18,47–49,52] are continuous and discrete, respectively. and are the size of S and G.

In contrast to scenario with duplication and loss, where the LCA reconciliation is the unique MPR, several optimal reconciliations are possible in the presence of transfers. This multiplicity of optimal scenarios may well lower the confidence we have in a single reconciliation drawn at random by a program from the set of all possible reconciliations. In this context, algorithms that enumerate all MPRs have been proposed [46,49,50].

3 Available programs

TreeMap [51] was the first program proposed for reconciling a gene tree G with a species tree S. A graphical interface is provided with a number of options. However, this program does not deal with dates for nodes of S, and as such cannot ensure time consistency of transfers. *Notung* [20] reconciles G and S according to the DL model, where at least one of the trees is binary. It has an interface that displays orthology relationships and a command line

version. It can also root or resolve polytomies of G (i.e. nodes with low support) by minimizing the parsimony score.

The reconciliation approach of [47,52] has been implemented in *CoRe-Pa*. This software includes a reconciliation viewer, an editor for modifying the G and S trees, as well as resampling facilities to evaluate the statistical relevancy of an MPR. It does not require costs for individual evolutionary events as it heuristically seeks those that best fit the reconstructed event frequencies.

The model of [48] has been implemented in a program called *Jane* [53], which also includes resampling facilities. It additionally allows a visual edition of a reconciliation (its cost is updated accordingly), and can be run from the command-line (for large-scale experiments). Reconciliations are built for a dated tree S, whose dates can be provided by the user. Alternatively, *Jane* uses a genetic algorithm to find optimal dates (w.r.t. reconciliation costs). Jane also allows controlling the maximal distance between two species that can exchange genes. This latter option is especially relevant when *Jane* is used for coevolution studies [24].

The reconciliation approach of [49] is implemented in a command-line program called *Mowgli* [54]. Besides computing an MPR, it also computes the number of optimal reconciliations for G and S trees. This provides an alternative (and much faster) way to measure the statistical significance of a single MPR. The method of [18] is also implemented in a command-line program called *AnGST*. It deals with phylogenetic uncertainties by inferring G as a combination of bootstrap subtrees that yields the “best reconciliation”.

Last, we note that some of the above software can differ from the models presented in the associated paper. For instance, on several datasets the reconciliations proposed by *CoRe-Pa* and *AnGST* have speciations located below the LCA mapping, while they are not supposed to [18,47,52].

In order to compute reconciliations, it is usually necessary to estimate duplication and loss rates. These can be estimated with *Cafe* [55], which takes as input a dated tree S and the number of genes per species for several gene families. For consistencies of transfers, dates for nodes of S can be obtained by relaxed molecular clock techniques working from molecular sequence [56,57].

4 Experimental results

To compare the main programs cited above, we performed some experiments on the dataset of [18]. It consists of 3983 gene family trees (with an average of 33.0 ± 26.5 contemporary genes) globally covering 90 genomes (11 eukaryotic, 12 archaeal, and 67 bacterial) together with a dated species tree. The average cost of reconciliations computed by Mowgli on these families is 64.3 ± 53.0, and in 42% of the cases these reconciliations are strictly more parsimonious than those provided by *AnGST*.

Jane was only tested on one of these gene families as this software accepts dates through its graphical interface only. The dates used by *AnGST* and

Mowgli were then entered manually together with the same event costs. Figure 6 displays the reconciliations proposed by the three software. For this gene family, *Mowgli* finds a reconciliation that is more parsimonious than those inferred by the other two software.

Figure 6

The three reconciliations differ according to the number and kind of events, the different models allowing to optimize parsimony at different degrees. When transfers are considered, the above results show the pitfalls in the design of reconciliation models and algorithms. Moreover in this case, an important aspect is that several most parsimonious reconciliations can exist. The prevalence of this effect has not yet been measured, nor the effect of the event costs in this respect. Last, we note that the probabilistic approach has been less developed here than when only considering duplications and losses.

Acknowledgments

This work was supported by the French Agence Nationale de la Recherche "Domaines Emergents" [ANR-08-EMER-011, "PhylAriane"] and by the Languedoc-Roussillon's "Chercheur d'Avenir" program. This publication is the contribution No 2011-YYY of the Institut des Sciences de l'Evolution de Montpellier (UMR 5554 - CNRS).

Key points

Reconciliation is an approach used to depict the evolution of a gene family with respect to the evolution of the species.

Several reconciliation models based on parsimony and probabilistic criteria have been proposed.

There is no agreement on a reconciliation model that deals with Horizontal Gene Transfers.

Bibliographie

1. Degnan J, Rosenberg N. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution* 2009*, 24* (6), 332-340.

2. Treangen T, Rocha E. Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes. *PLoS Genetics* 2011, *7* (1), e1001284.

3. Boussau B, Daubin V. Genomes as documents of evolutionary history. *Trends in Ecology & Evolution* 2009, 1-9.

4. Goodman M, Czelusniak J, Moore G et al.. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Systematic Zoology* 1979, *28* (2), 132-163.

5. Guigo R, Muchnik I, Smith T. Reconstruction of ancient molecular phylogeny. *Molecular phylogenetics and evolution* 1996, *6* (2), 189-213.

6. Page R. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Systematic Biology* 1994, *43* (1), 58.

7. Abby S, Tannier E, Gouy M et al.. Detecting lateral gene transfers by statistical reconciliation of phylogenetic forests. *BMC Bioinformatics* 2010, *11* (1), 324.

8. Beiko R, Hamilton N. Phylogenetic identification of lateral genetic transfer events. *BMC evolutionary biology* 2006, *6* (1), 15.

9. Nakhleh L, Ruths F, Wang L.S. RIATA-HGT: A Fast and Accurate Heuristic for Reconstructing Horizontal Gene Transfer. Proceedings of the 11th International Computing and Combinatorics Conference 2005 (LNCS 3595), 84-93.

10. Dufayard J.-F, Duret L, Penel S et al.. Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics* 2005, *21* (11), 2596-603.

11. Storm C, Sonnhammer E. Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics* 2002, *18* (1), 92.

12. Van Der Heijden R, Snel B, Van Noort V et al.. Orthology prediction at scalable resolution by phylogenetic tree analysis. *BMC Bioinformatics* 2007, *8* (1), 83.

13. Wapinski I, Pfeffer A, Friedman N et al.. Natural history and evolutionary principles of gene duplication in fungi. *Nature* 2007, *449* (7158), 54-61.

14. Zmasek C, Eddy S. RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics* 2002, *3* (1), 14.

15. Chen F, Mackey A, Vermunt J et al.. Assessing Performance of Orthology Detection Strategies Applied to Eukaryotic Genomes. *PLoS ONE* 2007, *2* (4).

16. Koski L, Golding G. The closest BLAST hit is often not the nearest neighbor. *Journal of Molecular Evolution* 2001, *52* (6), 540-542.

17. Rivera A, Pankey M, Plachetzki D et al.. Gene duplication and the origins of morphological complexity in pancrustacean eyes, a genomic approach. *BMC evolutionary biology* 2010, *10* (1), 123.

18. David L, Alm E. Rapid evolutionary innovation during an Archaean genetic expansion. *Nature* 2010, 1-4.

19. Akerborg O, Sennblad B, Arvestad L et al.. Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. *Proceedings of the National Academy of Sciences of the United States of America* 2009, *106* (14), 5714-9.

20. Durand D, Halldórsson B, Vernot B. A hybrid micro-macroevolutionary approach to gene tree reconstruction. *Journal of Computational Biology* 2006, *13* (2), 320-335.

21. Rasmussen M, Kellis M. A Bayesian Approach for Fast and Accurate Gene Tree Reconstruction. *Molecular Biology and Evolution* 2011, *28* (1), 273-290.

22. Sanderson M, McMahon M. Inferring angiosperm phylogeny from EST data with widespread gene duplication. *BMC evolutionary biology* 2007, *7* (Suppl 1), S3.

23. Page R, Charleston M. Trees within trees: phylogeny and historical associations. *Trends in Ecology & Evolution* 1998, *13* (9), 356-359.

24. Nieberding C, Jousselin E, Desdevises Y. The use of co-phylogeographic patterns to predict the nature of interactions, and vice-versa, in The geography of host-parasite interactions by Serge Morand and Boris Krasnov (eds). 2010. Oxford University Press.

25. Brooks D, Ferrao A. The historical biogeography of co-evolution: emerging infectious diseases are evolutionary accidents waiting to happen. *Journal of Biogeography* 2005, *32* (8), 1291-1299.

26. Bonizzoni P, Vedova G, Dondi R. Reconciling a gene tree to a species tree under the duplication cost model. *Theoretical computer science* 2005, *347* (1-2), 36-53.

27. Doyon JP, Chauve C, Hamel S. Space of gene/species trees reconciliations and parsimonious models. *Journal of Computational Biology* 2009, *16* (10), 1399-1418.

28. Górecki P, Tiuryn J. DLS-trees: a model of evolutionary scenarios. *Theoretical computer science* 2006, *359* (1-3), 378-399.

29. Zmasek C, Eddy S. A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics* 2001, *17* (9), 821.

30. Mirkin B, Muchnik IB, Smith TF. A Biologically Consistent Model for Comparing Molecular Phylogenies. Journal of Computational Biology 1995; 2.4 : 493-507

31. Chauve C, El-Mabrouk N. New perspectives on gene family evolution: losses in reconciliation and a link with supertrees. *Research in Computational Molecular Biology* 2009, 46-58.

32. Tofigh A. Using Trees to Capture Reticulate Evolution, Lateral Gene Transfers and Cancer Progression. PhD thesis, KTH Royal Institute of Technology, Sweden , 2009.

33. Hahn M. Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. *Genome Biology* 2007, *8* (7), R141.

34. Vernot B, Stolzer M, Goldman A et al.. Reconciliation with Non-Binary Species Trees. *Journal of Computational Biology* 2008, *15* (8), 981-1006.

35. Berglund-Sonnhammer AC, Steffansson P, Betts M et al.. Optimal Gene Trees from Sequences and Species Trees Using a Soft Interpretation of Parsimony. *Journal of molecular evolution* 2006, *63* (2), 240-250.

36. Chang W, Eulenstein O. Reconciling gene trees with apparent polytomies. *Computing and Combinatorics* 2006, 235-244.

37. Novozhilov A, Karev G, Koonin E. Biological applications of the theory of birth-and-death processes. *Briefings in Bioinformatics* 2006, *7* (1), 70.

38. Arvestad L, Berglund A, Lagergren J et al.. Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. *Proceedings of the eighth annual international conference on Resaerch in computational molecular biology* 2004, 326-335.

39. Arvestad L, Lagergren J, Sennblad B. The gene evolution model and computing its associated probabilities. *Journal of the ACM (JACM)* 2009, *56* (2), 1-44.

40. Sennblad B, Lagergren J. Probabilistic orthology analysis. *Systematic Biology* 2009, *58* (4), 411.

41. Doyon JP, Hamel S, Chauve C. An efficient method for exploring the space of gene tree/species tree reconciliations in a probabilistic framework. LIRMM technical report 2010, RR-10002 2010.

42. Górecki P, Burleigh GJ, Eulenstein O. Maximum likelihood models and algorithms for gene tree evolution with duplications and losses. *BMC Bioinformatics* 2011.

43. Csűrös M, Miklós I. Streamlining and large ancestral genomes in Archaea inferred with a phylogenetic birth-and-death model. *Molecular Biology and Evolution* 2009, *26* (9), 2087.

44. Ovadia Y, Fielder D, Conow C et al.. The Cophylogeny Reconstruction Problem Is NP-Complete. *Journal of Computational Biology* 2011, 191-223.

45. Charleston M, Perkins S. Traversing the tangle: algorithms and applications for cophylogenetic studies. *Journal of Biomedical Informatics* 2006, *39* (1), 62-71.

46. Tofigh A, Hallett M, Lagergren J. Simultaneous Identification of Duplications and Lateral Gene Transfers. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2011; 517-535.

47. Merkle D, Middendorf M. Reconstruction of the cophylogenetic history of related phylogenetic trees with divergence timing information. *Theory in Biosciences* 2005, *123* (4), 277-299

48. Libeskind-Hadas R, Charleston M. On the computational complexity of the reticulate cophylogeny reconstruction problem. *Journal of Computational Biology* 2009, *16* (1), 105-117.

49. Doyon JP, Scornavacca C, Szöllősi GJ et al.. An Efficient Algorithm for Gene/Species Trees Parsimonious Reconciliation with Losses, Duplications and Transfers. *Proceedings of the 14th International Conference on Research in Computational Molecular Biology (RECOMB)* 2011, *Volume 6398 of LNCS*: 93-108.

50. Charleston M. Jungles: a new solution to the host/parasite phylogeny reconciliation problem. *Mathematical Biosciences* 1998, *149* (2), 191-223.

51. Charleston MA, Page RDM. TreeMap 3 program. 2002. http://sydney.edu.au/engineering/it/~mcharles/software/treemap/treemap3.html

52. Merkle D, Middendorf M, Wieseke N. A parameter-adaptive dynamic programming approach for inferring cophylogenies. *BMC Bioinformatics* 2010, *11* (Suppl 1), S60.

53. Conow C, Fielder D, Ovadia Y et al.. Jane: a new tool for the cophylogeny reconstruction problem. *Algorithms Mol Biol* 2010, *5*, 16.

54. Doyon JP, Scornavacca C, Szöllősi GJ et al.. *Mowgli* program. 2011. http://www.atgc-montpellier.fr/Mowgli/.

55. De Bie T, Cristianini N, Demuth J et al.. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* 2006, *22* (10), 1269-71.

56. Lartillot N, Poujol R. A Phylogenetic Model for Investigating Correlated Evolution of Substitution Rates and Continuous Phenotypic Characters. *Molecular Biology and Evolution* 2011, *28* (1), 729-744.

57. Sanderson M. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 2003, *19* (2), 301.