Multiscale phylogenetic approaches for the evolution of the holobiont

Hugo Menet

N°d'ordre NNT :
2022LYSE1122

THESE de DOCTORAT DE L'UNIVERSITE DE LYON
opérée au sein de
l'Université Claude Bernard Lyon 1

Ecole Doctorale N° 341
Evolution Ecosystèmes Microbiologie Modélisation

Spécialité de doctorat :

Bioinformatique, Biomathématiques

Soutenue publiquement le 04/07/2022, par :
Hugo Menet

Approches phylogénétiques multi-niveaux pour l'évolution de l'holobionte

---

Devant le jury composé de :

Arvestad, Lars Professeur, Université de Stockholm, Suède	**Rapporteur**
Matias, Catherine Directrice de Recherche, CNRS	**Rapporteure**
Szöllősi, Gergely, Chercheur, Université Loránd Eötvös, Budapest, Hongrie	**Rapporteur**
Peres, Sabine Professeure des Universités, Lyon 1	**Présidente**
Tannier, Eric Directeur de Recherche, INRIA	**Directeur de thèse**
Daubin, Vincent Directeur de Recherche, CNRS	**Co-directeur de thèse**

Université Claude Bernard – LYON 1

Président de l'Université	M. Frédéric FLEURY
Président du Conseil Académique	M. Hamda BEN HADID
Vice-Président du Conseil d'Administration	M. Didier REVEL
Vice-Président du Conseil des Etudes et de la Vie Universitaire	Mme Céline BROCHIER
Vice-Président de la Commission de Recherche	M. Petru MIRONESCU
Directeur Général des Services	M. Pierre ROLLAND

COMPOSANTES SANTE

Département de Formation et Centre de Recherche en Biologie Humaine	Directrice : Mme Anne-Marie SCHOTT
Faculté d'Odontologie	Doyenne : Mme Dominique SEUX
Faculté de Médecine et Maïeutique Lyon Sud - Charles Mérieux	Doyenne : Mme Carole BURILLON
Faculté de Médecine Lyon-Est	Doyen : M. Gilles RODE
Institut des Sciences et Techniques de la Réadaptation (ISTR)	Directeur : M. Xavier PERROT
Institut des Sciences Pharmaceutiques et Biologiques (ISBP)	Directeur : M. Claude DUSSART

COMPOSANTES & DEPARTEMENTS DE SCIENCES & TECHNOLOGIE

Département Génie Electrique et des Procédés (GEP)	Directrice : Mme Rosaria FERRIGNO
Département Informatique	Directeur : M. Behzad SHARIAT
Département Mécanique	Directeur M. Marc BUFFAT
Ecole Supérieure de Chimie, Physique, Electronique (CPE Lyon)	Directeur : Gérard PIGNAULT
Institut de Science Financière et d'Assurances (ISFA)	Directeur : M. Nicolas LEBOISNE
Institut National du Professorat et de l'Education	Directeur : M. Pierre CHAREYRON
Institut Universitaire de Technologie de Lyon 1	Directeur : M. Christophe VITON
Observatoire de Lyon	Directrice : Mme Isabelle DANIEL
Polytechnique Lyon	Directeur : Emmanuel PERRIN
UFR Biosciences	Administratrice provisoire : Mme Kathrin GIESELER
UFR des Sciences et Techniques des Activités Physiques et Sportives (STAPS)	Directeur : M. Yannick VANPOULLE
UFR Faculté des Sciences	Directeur : M. Bruno ANDRIOLETTI

Résumé

Les systèmes biologiques sont constitués d'entités à plusieurs niveaux d'organisation (macro-organismes, micro-organismes, gènes...), qui partagent une histoire commune par leur dépendance, mais sont aussi guidés par leurs intérêts individuels. La réconciliation phylogénétique est une façon d'aborder l'évolution d'un tel système en décrivant la coévolution de deux niveaux différents, gènes et espèces, ou hôtes et symbiotes par exemple. La réconciliation est cependant limitée à deux niveaux, et s'inscrit ainsi soit dans un contexte d'évolution moléculaire avec des arbres de gènes et d'espèces, soit dans un contexte écologique d'évolution des associations hôte-symbiote. Le concept d'holobionte, la prise en compte comme un tout d'un macro-organisme (plante ou animal notamment) et de tous les micro-organismes qui vivent et fonctionnent avec lui, est l'occasion de rassembler toutes ces échelles en modélisant des interdépendances à plusieurs niveaux. Le but de cette thèse est d'explorer et d'étendre la réconciliation pour modéliser de tels systèmes.

La réconciliation est une méthode phylogénétique née à l'intersection de deux communautés, l'une qui s'intéresse à la coévolution d'espèces en symbiose, et l'autre qui compare des arbres de gènes et d'espèces. Malgré ce développement initial commun, ces deux communautés ont tendance à peu interagir, même si elles ont beaucoup à apprendre l'une de l'autre. Nous proposons dans cette thèse un état de l'art de la réconciliation phylogénétique en adoptant un point de vue générique et en soulignant les avancées vers des modèles plus intégratifs, qui tendent vers des méthodes intégrant plus de deux niveaux.

Parmi ces nouvelles méthodes, certaines proposent de modéliser ensemble l'évolution des espèces, des gènes et des domaines de gènes, ou encore d'imposer des contraintes géographiques à un système hôte symbiote. Cependant, aucune ne s'est pour le moment intéressée aux niveaux qui sont au cœur du concept d'holobionte : hôte, symbiote et gènes. D'un point de vue méthodologique, aucune ne fait appel à un cadre probabiliste permettant des transferts horizontaux. Nous avons réimplémenté ALE, un logiciel probabiliste de réconciliation modélisant les événements de Duplication, Transfert horizontal et perte (Loss) (modèle DTL), et l'avons étendu pour considérer la réconciliation de trois niveaux : hôte, symbiote et gène. Ce nouveau modèle probabiliste, qui prend en entrée trois arbres, combine un modèle DTL pour la coévolution de l'hôte et du symbiote, et un modèle DTL pour l'évolution des gènes du symbiote. Nous avons conçu un algorithme de Monte Carlo pour construire des scénarios couplés et calculer leurs probabilités dans le modèle, en tenant notamment compte de la dépendance des taux de transfert de gènes à la réconciliation entre symbiotes et hôtes, ainsi que de l'impact des lignées fantômes sur ces taux.

Comme avec ALE, nous utilisons l'amalgamation pour tenir compte de l'incertitude dans les arbres de gènes, mais aussi pour inférer l'arbre de symbiote en utilisant les arbres des familles de gènes universels et unicopies comme une distribution de la topologie de l'arbre de leur génome. Nous avons évalué cette méthode sur un jeu de données simulées, sur lequel nous avons montré qu'il était possible de distinguer les modèles de coévolution à 2 et 3 niveaux en utilisant la vraisemblance. La méthode est également capable sur des phylogénies de pucerons et de leurs entérobactéries de mieux retrouver les transferts de gènes que la méthode ignorant l'arbre d'hôte.

Il peut être difficile d'interpréter la sortie d'une méthode de réconciliation, notamment lorsque l'on considère plusieurs scénarios échantillonnés, plusieurs familles de gènes ou lorsqu'on s'intéresse à des systèmes à plusieurs niveaux. Peu de logiciels proposent une sortie graphique générique, utilisable avec plusieurs logiciels de réconciliation. Un premier pas dans cette direction est l'utilisation de RecPhyloXML, un format de scénario adopté par une partie importante de la communauté gène espèce. C'est ce format que nous utilisons dans notre implémentation. Thirdkind, un logiciel que nous avons développé, est capable de produire une sortie graphique en SVG à partir d'un scénario de réconciliation en RecPhyloXML. Il est facile à utiliser et à installer, il peut afficher différentes vues représentant la réconciliation de trois niveaux, et résumer l'évolution de plusieurs familles de gènes ou de plusieurs scénarios échantillonnés dans une seule figure en agrégeant les transferts redondants.

Un exemple fascinant d'histoire complexe de coévolution est la relation entre *Helicobacter pylori* et son hôte humain. *Helicobacter pylori* est une bactérie pathogène qui aurait suivi *Homo sapiens* lors de ses migrations ancestrales : colonisation de l'Afrique, de l'Asie, de l'Europe, de l'Océanie et de l'Amérique. Les souches bactériennes sont structurées en populations dont la répartition géographique est le plus souvent congruente avec celle de leur hôte. L'une des exceptions significatives est la population européenne, qui semble résulter de l'introgression entre deux populations ancestrales, l'une apparentée à une population africaine moderne, l'autre à une asiatique. Les études précédentes de ce système reposent sur des modèles bayésiens d'attribution de SNP à des populations, pour des génomes entiers ou un sous-ensemble de gènes via une approche MLST. J'ai pris un point de vue phylogénétique sur cette question, en utilisant un jeu de données construit dans l'équipe. Ce jeu de données est constitué de la phylogénie de 120 souches, comprenant la souche ancienne séquencée chez Ötzi, une momie trouvée dans les Alpes et datée à plus de 5 000 ans, et de 1 034 arbres de gènes. Nous avons appliqué la réconciliation aux arbres de gènes et aux arbres de population pour mieux comprendre les origines des gènes de la population européenne. Cette nouvelle approche, qui repose sur l'appariement uniforme de certaines feuilles des arbres de gènes (ici les européennes) à toutes les feuilles de l'arbre du dessus puis sur la probabilité a posteriori d'appariement obtenue en échantillonnant des scénarios, pourrait être facilement transposée à d'autres problèmes. Nous avons également utilisé notre approche de réconciliation à 3 niveaux pour comparer différents arbres de population.

Mot clés : phylogénie, réconciliation phylogénétique, coévolution, symbiosis, *Helicobacter pylori*

Abstract

Biological systems like holobionts are made up of entities at many scales (macro-organisms, micro-organisms, genes...), which are, on the one hand, bound to a common history because they all function together and depend on each other, and on the other hand, driven by their individual interests. The evolution of such systems is approached by phylogenetic reconciliation, which describes the coevolution of two different levels, genes and species, or hosts and symbionts, for example. The limit to two levels has confined the use of reconciliation either to molecular studies on genes and species trees or to ecological studies on host-symbiont associations. The holobiont concept provides an opportunity to bring all of these scales together by modeling multi-level inter-dependencies. In this thesis, we explore and extend reconciliation to model such multi-level systems.

Phylogenetic reconciliation is a phylogenetic method born of the interaction of two communities, one interested by the coevolution of host and symbiont, and the other by the comparison of gene and species trees. Lately, despite this initial development, these two communities tend not to interact much, even if they have much to learn from each other. We review the development of these methods, take a generic approach, and highlight new advances that propose more integrative models, tending toward multi-level reconciliation.

In recent years, these advances proposed to integrate species, gene and gene domain evolution, or geography, host and symbiont evolution, but none have yet studied the levels central to the holobiont: host, symbiont and genes, and none in a probabilistic framework and with horizontal transfers. I reimplemented ALE, a probabilistic DTL reconciliation software, and extended it to consider the reconciliation of three levels: host, symbiont, and gene. This new probabilistic model of nested three-level coevolution allows gene transfer, host switch, gene duplication, symbiont diversification within a host, and gene or symbiont loss. Given three phylogenetic trees, we design a Monte Carlo algorithm capable of inferring joint scenarios and calculating their likelihood in the model, accounting for gene transfer rates' dependence on host symbiont reconciliation as well as the impact of ghost lineages on these rates. As in ALE, we use amalgamation to take into account uncertainty in the gene trees, but also to infer the symbiont tree using universal unicopy genes as a topology distribution for the symbiont tree. This method was evaluated using a simulated dataset on which we showed its capacity to distinguish models of 2-level and 3-level coevolution using the computed likelihood. In a aphids and enterobacteria system, it is able to retrieve transfers better than the host unaware method.

The output of reconciliation can be hard to interpret, especially when we want

to consider multiple sampled scenarios or multiple gene families. Few graphical software exists, and none are generic and can use RecPhyloXML, a common format endorsed by the gene species community. Graphical representation of multi-level reconciliations is an added layer to this question. We propose Thirdkind, a software we developed, that produce a graphical representation of a reconciliation scenario as an SVG file. It is easy to use and install. It can handle the embedment of three trees which is the output of our 3-level reconciliation framework and can resume the evolution of multiple gene families or scenarios in a single figure by aggregating redundant transfers.

A fascinating example of a complex coevolutionary history is the relationship between *Helicobacter pylori* and its human host. *Helicobacter pylori* is a pathogenic bacteria that is believed to have followed its human host during its ancestral migrations, during the colonization of Africa, Asia, Europe, Oceania, and the Americas. The bacterial strains are structured in populations whose geographical distribution most often correspond to that of their host. One significant divergence is the European population, which appears to be the result of introgression between two ancestral populations, one related to modern African and the other to modern Asian. These hypotheses are based on Bayesian models of SNPs attribution to populations, for whole genomes, or a small subset of genes via Multi Locus Sequence Typing. We took a more phylogeny-oriented approach using a dataset constructed in the team, with a phylogeny and the gene trees of 1034 gene families for 120 strains, including the ancestral strain found in Ötzi, a natural mummy of a man who lived in the Alps five thousands years ago. We applied reconciliation to gene trees and population trees to better understand the mixed origins of genes in the European population. This new approach, which relies on matching certain leaves of the gene trees (here the European ones) uniformly to all leaves of the upper tree and then looking at the posterior probability of matching, could be easily transposed to other problems. We also tested different population trees using our 3-level reconciliation framework. These new analyses help get a clearer understanding of the population structure of *Helicobacetr pylori* by contrasting the divergent histories of its gene families.

Keywords: phylogeny, phylogenetic reconciliation, coevolution, symbiosis, *Helicobacter pylori*

Résumé étendu

En biologie évolutive, la phylogénétique est une discipline ayant pour but de représenter la diversification des espèces sous la forme d'un arbre. À l'inverse d'un arbre généalogique, on place les espèces actuelles aux feuilles et les espèces ancestrales dans les nœuds internes de l'arbre. Étant donné un ensemble d'espèces, on peut ainsi se demander quelles sont les relations de parenté entre ces espèces, et quel est l'arbre qui représente le mieux leur diversification. On peut même chercher à reconstruire l'arbre de l'ensemble des espèces connues, que l'on nomme l'arbre de la vie. Depuis un ancêtre commun, et en quelques milliards d'années d'évolution, on obtient l'ensemble de la biodiversité actuelle. Cette diversité actuelle est le fruit d'une histoire faite de multiples diversifications et durant laquelle une grande partie de la biodiversité ancienne s'est éteinte sans laisser de descendants. En évolution moléculaire, on s'appuie sur des données moléculaires, notamment des séquences d'ADN ou protéiques, pour reconstruire cet arbre de parenté. Chaque espèce a un génome, dont l'une des briques est le gène, lui-même constitué de base nucléiques A, T, C et G. Grâce à des méthodes de laboratoire et de bioinformatique, il est possible d'identifier les gènes d'un individu et d'obtenir les séquences nucléiques ou protéiques correspondant à chacun de ces gènes. On peut ensuite aligner les séquences identifiées comme homologues, c'est-à-dire issues d'une même séquence ancestral, et construire l'arbre de relations entre des gènes issus de différentes espèces. On appelle cet arbre, arbre de gène. Pour construire la phylogénie d'un ensemble d'espèces, on peut concaténer les séquences alignées de plusieurs gènes de ces espèces et construire l'arbre de ce concaténat comme on le ferait pour un arbre de gène. Une autre approche, dite de superarbre, repose sur la prise en compte explicite de plusieurs arbres de gènes et la construction d'un arbre d'espèces qui soit cohérent avec ces arbres de gènes. Les approches de concaténat et de superarbre font toutes deux l'hypothèse que les histoires de gènes reflètent celle des espèces et porte un signal moyen permettant de reconstruire celle-ci. La réconciliation est une façon de se passer de cette hypothèse, voire de la tester, en modélisant plus finement les ressemblances et les dissemblances entre ces histoires.

En phylogénétique, la réconciliation est une approche pour relier l'histoire de deux ou plusieurs entités biologiques qui coévoluent. L'idée générale est qu'un arbre phylogénétique représentant l'évolution d'une entité peut être dessiné à l'intérieur d'un autre arbre phylogénétique représentant une entité englobante pour révéler leur interdépendance et les événements évolutifs qui ont marqué leur histoire commune. Le développement des approches de réconciliation a commencé dans les années 1980, principalement pour dépeindre la coévolution d'un gène et d'un génome, et d'un

hôte et d'un symbiote, qui peut être mutualiste, commensaliste ou parasitaire. Un exemple célèbre d'application hôte symbiote est la comparaison des phylogénies des rongeurs de la famille des Géomydés et de leurs poux. Différentes espèces de poux sont spécifiques à chacune des espèces de ces rongeurs, et la réconciliation a permis de tester des hypothèses de correspondance entre les différentes espèces ancestrales de ces deux groupes. La continuité des relations entre symbiote et hôte lors de leurs diversifications est ce que l'on appelle la coévolution de l'hôte et du symbiote. La réconciliation, avec une méthode informatique très similaire, voire identique, est aussi utilisée pour expliquer les différences entre les arbres d'espèces et les arbres de leurs gènes. Ainsi, un gène peut être dupliqué et se retrouver en deux copies dans un génome. Il peut aussi être perdu, ou être transféré horizontalement entre deux espèces, ce qui arrive fréquemment chez les bactéries ou les virus, par recombinaison par exemple, mais qui peut aussi être observé chez les eukaryotes. C'est cette méthode qui va nous intéresser dans le cadre de cette thèse, et que l'on va explorer et étendre pour modéliser des systèmes biologiques à plusieurs niveaux, notamment pour considérer au sein d'un même modèle l'évolution d'un hôte, de ses symbiotes et de leurs gènes.

Un des modèles biologiques que l'on étudie dans cette thèse est la bactérie *Helicobacter pylori*. *Helicobacter pylori* est un pathogène de l'homme, présent dans l'estomac d'un individu sur deux. D'un point de vue génétique les différentes souches de cette bactérie présentent une forte structure géographique, qui rappelle beaucoup la répartition des populations humaines. Il est admis que *Helicobacter pylori* est associée à l'homme depuis sa naissance en Afrique, et l'a suivi au cours de ses migrations à travers la planète, hors d'Afrique, puis en Asie, avant l'Europe, l'Océanie, et l'Amérique. Les bactéries *Helicobacter pylori* se sont également révélées hautement recombinantes. La recombinaison de deux *pylori* est un indice de la rencontre des populations humaines hébergeant chacune des souches, l'information contenue au niveau des gènes pourrait ainsi nous aider à raconter une histoire au niveau écologique de la relation hôte et symbiote. On pourrait même se demander comment cette information au niveau des gènes du symbiote pourrait nous permettre d'examiner les migrations de leur hôte entre zones géographiques ? Ces évolutions entremêlées à plusieurs niveaux, ici des gènes, des symbiotes bactériens, un hôte animal et des aires géographiques, sont à la base des modèles que nous avons voulu développer. À partir de la réconciliation phylogénétique, déjà capable d'expliquer la coévolution de deux niveaux, nous avons tenté de comprendre la coévolution de plusieurs niveaux, notamment avec la modélisation explicite des transferts horizontaux de gènes et des changements d'hôte dans un modèle hôte, symbiote et gène.

Un exemple biologique un peu plus simple, mais tout aussi intéressant, est donné par une symbiose impliquant des pucerons du genre *Cinara* et des bactéries capable de synthétiser des vitamines et des acides aminés. Les pucerons sont des insectes qui se nourrissent essentiellement de sève. Cependant cette alimentation est carencée en certains acides aminés qu'ils ne peuvent pas eux-même synthétiser. Pour survivre avec un tel régime, les pucerons ont développé des relations de symbiose avec des bactéries capables de produire ces acides aminés nécessaire à la survie de leur hôte. Cette relation est endosymbiotique : les symbiotes se trouvent dans des compartiments spéciaux à l'intérieur des insectes, les bactériocytes. Cet exemple nous intéresse pour plusieurs raisons. Il part d'un cas de symbiose plutôt bien connu, avec un hôte et un symbiote transmis verticalement (une relation nettement plus simple

que les relations entre les mammifères et leur microbiome intestinal par exemple). L'insecte hôte possède une bactérie *Buchnera* avec, dans son génome, la machinerie nécessaire pour synthétiser les acides aminés qui complètent le régime alimentaire des pucerons. D'un point de vue phylogénétique la phylogénie de l'hôte et celle du symbiote sont le plus souvent parfaitement congruentes. L'histoire devient ensuite plus complexe. Une nouvelle entérobactérie, un symbiote apparenté à *Erwinia*, est observée dans les lignées de *Cinara* dans un bactériocyte séparé. Il compense le rôle de *Buchnera* qui a perdu, chez ces pucerons, les voies de production de la biotine. Les gènes observés ici et hébergés dans les génomes d'*Erwinia* ne sont cependant pas ceux perdus par *Buchnera*, mais se rattachent à ceux d'une autre entérobactérie, *Sodalis* et ont certainement été acquis par transfert horizontal de gènes. En plus de compenser pour *Buchnera*, le génome d'*Erwinia* présente également des gènes pour de nouvelles voies de synthèse de la thiamine, également acquis dans ce transfert de gènes depuis *Sodalis*. De plus, dans deux lignées sœurs, les gènes de *Erwinia* ont été transférés à de nouveaux symbiotes apparentés à *Hamiltonella* alors que ces gènes ne sont plus présents dans les génomes d'origine. D'un point de vue fonctionnel, les gènes nécessaires pour compléter la nutrition de l'hôte sont toujours quelque part à l'intérieur celui-ci, mais jamais dans le génome de l'hôte et pas toujours dans celui de la même bactérie. On assiste à l'acquisition de deux nouveaux symbiotes et à deux séries de transferts horizontaux de gènes. Dans un cas, les gènes sont acquis à partir d'une bactérie extérieure (*Sodalis* à *Erwinia*), et dans l'autre, l'échange de gènes semble se produire à l'intérieur de l'hôte (*Erwinia* à *Hamiltonella*). Ce système présente une histoire à plusieurs niveaux, mais au sein d'une relation assez simple d'un point de vue phylogénétique et avec des événements coévolutifs bien identifiés. C'est donc un exemple idéal pour comprendre et tester de nouveaux modèles.

Le chapitre 1 de cette thèse comporte une version documentée et étendue de cette introduction à la phylogénie dans le cadre de l'évolution moléculaire. Il propose également un état de l'art et une introduction à la réconciliation phylogénétique, que nous avons écrit avec Vincent Daubin et Eric Tannier dans le cadre d'une review en cours de publication dans PLoS Computational Biology. Le format particulier de publication devrait ensuite donner lieu à l'ajout d'une page sur la réconciliation sur Wikipédia. Pour cette review, nous avons appliqué une approche essentiellement méthodologique, ce qui nous a permis de mettre en avant les liens qui existent toujours entre l'application aux modèles hôtes symbiotes et aux modèles gènes et espèces. Nous avons documenté des utilisations de la réconciliation ainsi que des études présentant des modèles biologiques à plusieurs niveaux qui motivent le développement de nouvelles méthodes, comme celle que nous présentons au chapitre 2.

Le chapitre 2 est dédié à notre nouveau modèle de réconciliation à plusieurs niveaux hôte, symbiote et gène. Ce travail est mis en forme dans un article, en cours de préparation. Cette nouvelle approche repose sur une extension de la méthode ALE, pour Estimation de vraisemblance (Likelihood) avec Amalgamation, qui a été développée au laboratoire il y a une dizaine d'années notamment par Gergely Szöllősi et des membres de l'équipe actuelle. ALE est une méthode DTL (Duplication Transfert et perte (Loss)), sans datation de l'arbre, qui utilise l'amalgamation pour parcourir efficacement plusieurs topologies possibles pour les arbres de gène. Dans notre extension, les arbres d'hôte et de symbiote sont d'abord réconciliés, et un scénario de coévolution est choisi en fonction de sa vraisemblance. Puis étant donné

ce scénario, on distingue deux types de transferts, selon que le donneur et le receveur sont ou non dans le même hôte. Le transfert est alors soit intra, soit inter. On introduit également une méthode de calcul mécaniste du taux de transfert inter à partir du taux de transfert intra et des paramètres de la réconciliation entre l'hôte et le symbiote. On suppose que pour qu'un transfert inter ait lieu, il faut dans tout les cas que les deux symbiotes impliqués se rencontrent. Cependant, lorsque l'on reconstitue l'histoire d'un gène, et notamment sa provenance dans un transfert, ce n'est pas le véritable donneur qui est identifié, mais l'espèce la plus proche du donneur dans la phylogénie considérée. Ainsi, il est possible de supposer que cette espèce inconnue, qui a transféré le gène, était présente dans le même hôte que le receveur, et qu'elle a simplement été perdue par la suite. On peut ainsi considérer tous les scénarios possibles qui mènent une espèce sœur du donneur vers l'hôte du receveur pour expliqué un transfert inter par un transfert intra. On peut ensuite considérer d'autres scénarios hôte symbiote, et recommencer, ce qui constitue une approche de Monte Carlo. Pour une estimation plus rapide, mais théoriquement moins robuste, on peut également choisir d'appliquer la méthode avec seulement le scénario hôte symbiote de vraisemblance maximale. On montre comment on peut utiliser cette méthode sur des données simulées pour une meilleure inférence des donneurs et receveurs des transferts de gènes. On montre également qu'il est possible de différencier les données simulées suivant un modèle deux niveaux ou un modèle trois niveaux, en comparant la vraisemblance de notre approche et celle de la réconciliation des gènes et du symbiote sans prendre en compte l'hôte. On applique également la méthode à nos deux modèles biologiques déjà présentés. D'abord les pucerons du genre *Cinara*, pour lesquels on retrouve les transferts de gènes précédemment identifiés alors que l'approche deux niveaux infère des transferts différents. Ensuite, pour *Helicobacter pylori*, on compare différents arbres de populations, et on utilise l'amalgamation pour inférer l'arbre de souche à partir des gènes universel unicopies. On présente également dans ce second chapitre, un second article, écrit par Simon Penel et portant sur un logiciel de visualisation qu'il a implémenté et que l'on a développé ensemble et avec Eric Tannier, Vincent Daubin et Théo Tricou. Ce logiciel permet d'obtenir des SVG depuis des scénarios de réconciliation en RecPhyloXML, format utilisé notamment par mon implémentation. Il permet de visualiser des réconciliations à trois niveaux, et de résumer l'information de plusieurs gènes et de plusieurs scénarios notamment au niveau des transferts.

Le chapitre 3 porte sur *Helicobacter pylori*. Il commence par une rapide bibliographie qui présente cette bactérie pathogène de l'homme, identifiée en 1984. On insiste notamment sur les différentes études qui ont tenté, notamment depuis 2003, d'établir une structure de population pour expliquer la diversité de l'espèce. Cette structure géographique ressemble fortement à la structuration géographique des populations humaines. L'hypothèse principale énonce que *Helicobacter pylori* est associé à l'homme depuis plus de 100 000 ans, et l'a notamment suivi lors de sa sortie d'Afrique, vers l'Asie, l'Europe, l'Océanie (l'ancien continent Sahul, qui regroupe Australie, Nouvelle-Guinée, et Tasmanie), et lors de l'expansion austronésienne (la colonisation de nombreuses îles du Pacifique, en partant notamment de Taïwan, avec une colonisation de la Nouvelle-Zélande il y a moins de mille ans) et depuis l'Asie en Amérique par le détroit de Béring. On voit également les migrations récentes, comme la présence de souches européennes et africaines en Amérique. Tous ces événements semblent correspondre fortement entre l'évolution de *pylori* et celle de l'homme.

Mais cette coévolution n'est tout de même pas parfaite. De nombreux événements diffèrent entre les deux espèces. Par exemple, la plupart des amérindiens portent aujourd'hui des *pylori* proches de ceux portés par les européens et les africains, et non celles de leurs plus proches cousins asiatiques. De la même manière, bien que le peuple Baka au Cameroun, ait une position basale dans l'arbre des populations humaines, leurs souches *pylori* sont proches de celles de leurs voisins arrivés là il y a quelques milliers d'années, représentatives de branches africaines proches de celles ayant colonisées l'Eurasie. Un dernier exemple de différence, qui nous intéresse plus particulièrement ici, est la question des *pylori* de la population européenne. En effet, ceux-ci semblent issus d'une introgression entre des *pylori* proches de ceux portés par la population d'Asie centrale et par ceux du Nord Est de l'Afrique. Comprendre cette introgression constitue le fil directeur de ce chapitre 3. Un indice supplémentaire nous est donné par Ötzi, l'homme des glaces, un humain retrouvé congelé dans les Alpes, et daté à plus de 5 000 ans. Il a été possible de séquencer les bactéries *pylori* présent dans son estomac et de montrer que cette souche est très similaire aux souches asiatiques. Les études présentes reposent, pour les premières sur l'étude de gène MLST, 7 gènes considérés comme marqueurs, et les suivantes sur des génomes complets, mais toujours avec des approches faisant appel à peu de phylogénie et basées sur l'attribution de population au niveau nucléotidique. Notre approche est de considérer des phylogénies, et de nous intéresser à un niveau intermédiaire, qui est celui des gènes, entre les nucléotides et les génomes. Partant d'un jeu de donnée construit par Alexia Nguyen Trung, et rassemblant 119 souches, et 1 034 gènes, pour lesquelles des arbres ont été construits, on a tenté d'apporter une réponse à cette question européenne. J'ai notamment développé une nouvelle méthode basée sur la réconciliation et qui permet d'assigner pour chaque famille de gènes et pour chaque souche, une population en fonction de l'assignation d'une partie importante du reste de l'arbre. La méthode utilise simplement le fait que l'approche probabiliste de la réconciliation permet de considérer plus d'une assignation *a priori* pour une feuille et de tirer par la suite des assignations avec les scénarios. Les résultats obtenus pour les branches européennes sont très encourageants, et indiquent une introgression ancienne, avec des gènes qui se branchent plutôt à la base des groupes d'origine, ou la présence de deux populations européennes disjointes.

Le chapitre 4 présente quatre problèmes ouverts autour de la réconciliation et de la phylogénie. Ce sont des questions simples à formuler, mais pour lesquels nous n'avons pas été capable dans le temps de cette thèse de répondre. Les questions sont les suivantes. En remarquant que les modèles non datés de réconciliation dans un cadre de parcimonie ou de probabilités ne considèrent pas les transferts de la même façon, et en remarquant que les deux sont inconsistants par rapport au modèle daté, plus réaliste, on peut se demander lequel est le plus proche de ce modèle daté ? Dans un processus de naissance mort de génération d'un arbre, si on prend une branche à un temps t, quelle est la distribution du temps de coalescence au premier ancêtre ayant des descendants au temps présent ? Le modèle dual de la réconciliation, en échangeant l'arbre du haut et l'arbre du bas, modélise-t-il un processus intéressant ? Peut-on échanger amalgamation et calcul de la vraisemblance, c'est-à-dire, par exemple, est ce que l'arbre d'espèces qui correspond au maximum de vraisemblance pour la réconciliation avec un ensemble de gènes universels et unicopies est le même que celui qui maximise la réconciliation avec l'amalgamation de ces gènes ?

Remerciements

Alors que ces années de thèse se terminent, des remerciements s'imposent.

First, I would like to thanks the thesis reviewers Lars Arvestad, Catherine Mathias and Gergely Szöllősi for their valuable insights and thorough examination of the manuscript. Merci également à Sabine Peres d'avoir bien voulu présider ce jury. En revenant un peu en arrière, merci à Annie Chateau, Séverine Bérard et Krister Swenson de m'avoir pris en stage en L3 alors que l'offre de stage était prévue pour quelqu'un d'autre, et de m'avoir ainsi un peu montré ce qu'était la bioinformatique, puis pour m'avoir orienté vers Cédric Chauve pour mon stage de M1. J'aimerais donc aussi remercier Cédric Chauve, qui m'a fait découvrir une bio informatique avec de l'application aux données en stage de M1. Je me souviens d'un mail envoyé à Eric Tannier pendant ce stage, seule personne capable de refaire tourner ALE pour qu'on obtienne de nouvelles données. Aujourd'hui je suis capable de lancer ALE (ma réimplémentation en tout cas) et il n'y a qu'une personne que je n'ai pas rencontré parmi les auteurs du papier MaxTiC, ma référence principal pour ce stage, et nombre d'entre eux ont joué un rôle de premier plan dans cette thèse. Merci à Céline Scornavacca d'avoir joué le jeu du second projet de thèse, et d'avoir suivi mon avancée pendant ces trois années. Merci à mon comité de suivi de manière plus général, Tristan Lefebure, Nicolas Lartillot, Laure Ségurel et Blerina Sinaimeri.

Merci ensuite à mes directeurs de thèse, Eric et Vincent qui m'ont tous les deux tant appris sur un grand nombre de sujets. Ça a toujours été très agréable de travailler avec vous pour votre bienveillance et votre expertise, et la grande liberté que vous m'avez laissé. Suite au stage avec Eric, quand Vincent a accepté de se joindre à la thèse, je ne m'attendais pas à ce que tu sois si présent, c'était un vrai 50%, et même sur les conversations plus informatiques ou mathématiques, tu étais même peut être celui avec le plus les pieds sur terre cherchant à vraiment comprendre ce qu'il se passait et contrant les arguments les plus bancals. Je crois commencer à voir ce qu'est le travail d'un biologiste. Merci Eric, d'avoir bien voulu me prendre en stage puis en thèse, d'avoir écrit si rapidement des dossiers pour des bourses. Après avoir discuté de plusieurs sujets possibles de stage, je me souviens que tu étais revenu vers moi par mail pour me présenter un autre sujet, "plus ambitieux". J'espère avoir pu vous aider à défricher cette nouvelle approche. En sortant pourtant d'un parcours avec déjà beaucoup de modélisation, à force de te côtoyer j'ai vu comme il me restait beaucoup à apprendre sur ce qu'était la science et les modèles, ou sur la rédaction scientifique. Il est rarement plus long d'écrire quelque chose d'exact. Merci à tous les deux pour toutes vos relectures, et pour toutes les discussions tout au long de ces 3 années.

La semaine précédant l'arrivée au labo, une bonbonne de gaz avait explosé sur le toit, un peu un signe, qu'entre les travaux et la pandémie cette thèse n'allait pas être de tout repos, mais grâce à tout le LBBE, ça a tout de même pu être une belle expérience. Merci à toutes les personnes du LBBE, un merci tout particulier à Alexia et Théo. Merci également aux gens avec qui j'ai eu l'occasion de travailler, notamment Damien et Simon, scientifiquement, et Lucas, Thibault, Etienne, Matthieu et Sébastien pour les enseignements. Merci à l'ensemble des doctorants qui rendent ce labo si vivant, notamment à l'étage du prabi: Thibault, Djivan, Alexandre, Florian, Julien, Antoine, Alice, Mélodie, Louis, Rémi-Vinh... , et à toute la team du Domus pour le midi. Merci à mes différents cobureaux, Nika, Anaïs, Vincent et Dominique. Un merci plus particulier à toute l'équipe du Cocon, fraîchement extraite de BPGE un peu avant mon arrivée, pour ceux que je n'ai pas encore cité Bastien, Annabelle, Laurent merci pour toutes les discussions et les retours. Merci au pôle administratif, si efficace, notamment, Nathalie, Odile, Sahra. Merci au pôle info, tout aussi efficace, notamment Bruno, Stéphane et Philippe.

Je ne sais pas si je dois remercier Lucy après plusieurs années de correspondances et toujours pas d'éditeur en vue.

Merci ensuite à mes amis lyonnais, ça commence à faire longtemps qu'on se suit. Adèle, Alice, Colin, Emile, Lilian, Loïs, Paul, Octave et les (plus ou moins) exilés Baptiste, Clément, Meven et Rédouane. Les confinements aurait été bien long. Si on mettait toutes nos thèse bout à bout on pourrait faire des choses magnifiques. Pourquoi pas une grosse machine codé avec de l'ADN, et alimentée par des explosions cellulaires (mais il faudra faire attention à prendre de l'ADN de plante ça n'a rien à voir sinon). Bien sûr il faudra du bayésien et de l'échantillonnage, on ne sait pas encore où, mais j'ai du mal à imaginer quelque chose qui puisse vivre sans, et quand nos experts auront fini de rendre les films Disney plus jolis et de mapper la faille de San Andreas, ils pourront s'y mettre. Il faudrait ensuite prouver que la machine marche correctement, c'est facile à côté des expressions transfinis, et un petit typage bidirectionnel et le tour est joué. Il y aura sûrement quelques questions juridiques à régler, mais on a quelqu'un sur le coup. Après avoir testé tout ça sur des souris, il nous restera à faire un peu de scheduling pour savoir quand la lancer, et avec l'argent récolté, on gardera un peu pour laisser quelqu'un faire de la théorie des nombres, il ne faudrait quand même pas qu'on arrête d'en faire un jour. Et pour garder pour toujours la machine en bonne état, il suffira de l'assimiler à un guide d'onde pour un petit contrôle non destructive en lançant des ondes aux fréquences de résonances (le problème est moins bien défini, mais c'est bien plus efficace). Et bien sûr tout cela n'aurait de sens si ce n'est à la lumière de l'évolution, et en prenant en compte la dépendance entre tous ces éléments. La vrai question restant de savoir si vraiment, on veut construire cette machine.

Merci à ma famille, mes parents qui ont toujours cherché à me montrer le monde, mon goût pour la science vient certainement un peu de là. Merci de m'avoir supporté dans mes études. Merci à ma sœur d'être toujours à l'écoute.

Merci enfin à Angèle, j'aurais presque oublié ce remerciement tant il me paraît maintenant naturel que tu sois là tous les jours à mes côtés. Je ferai tout pour que ça ne change pas.

Abbreviations and notations

Abbreviations

DL: Duplication Loss
DTL: Duplication Transfer Loss
HGT: Horizontal Gene Transfer
ILS: Incomplete Lineage Sorting
kya: kilo year ago
MC: Monte Carlo
MCMC: Monte Carlo Markov Chain
ML: Maximum Likelihood
MLST: Multi Locus Sequence Typing
SNP: Single Nucleotide Polymorphism

Essential notations

p^S, p^D, p^T, p^L: probability of speciation, duplication, transfer and loss
c^S, c^D, c^T, c^L: cost of speciation, duplication, transfer and loss
H, S, G: host, symbiont and gene trees
$|.|$: number of elements
$R(G, S)$: set of all reconciliation scenarios of tree G in tree S.
$P_{e,u}$: probability that the gene subtree rooted at u reconciles with the species subtree rooted at e
E: extinction probability
e, f, g, h: species node e and its two children f and g, and a parallel branch h.
u, v, w: gene node u and its two children v and w
$C, \bar{C}, C', C''$: a clade, its complement, and its two children clades.

Contents

Chapter 1

Introduction

I will start this manuscript with a description of two biological systems that will give a taste of the mathematical models and the interactions that interest us. I will present the relationship between the bacterial symbiont *Helicobacter pylori* and its host, *Homo sapiens*, and the genes and symbiont exchanges in a *Cinara* aphids and obligatory endosymbiont system. We used both datasets as test cases for the method we introduced in the second chapter, and our extensive study of *Helicobacter pylori* and human relationship is the focus of the third chapter.

SECTION 1.1

Two biological examples

1.1.1 *Helicobacter pylori*

Helicobacter pylori is a bacteria present in the stomach of half of the human population and which has been linked to pathogenicity (figure 1.1). Apart from its medical interest, *Helicobacter pylori* has been an important research subject to investigate the evolutionary relationship between humans and bacteria, and notably from a comparative phylogenetic point of view[148, 76].[1]

Helicobacter pylori genetic diversity is strongly structured by geography. Moreover, this population structure, and notably the phylogenetic arrangement of the different geographic populations, seems congruent with what we know of human migrations across the planet, out of Africa, then into Asia, before Europe and Oceania, and the Americas (Fig. 1.2).

The similarity between the geographic pattern of symbiont and host diversification hints at coevolution between the host and its symbiont.

Helicobacter pylori bacteria were also shown to frequently exchange genetic ma-

[1] An in-depth introduction to this system is given in chapter 3.

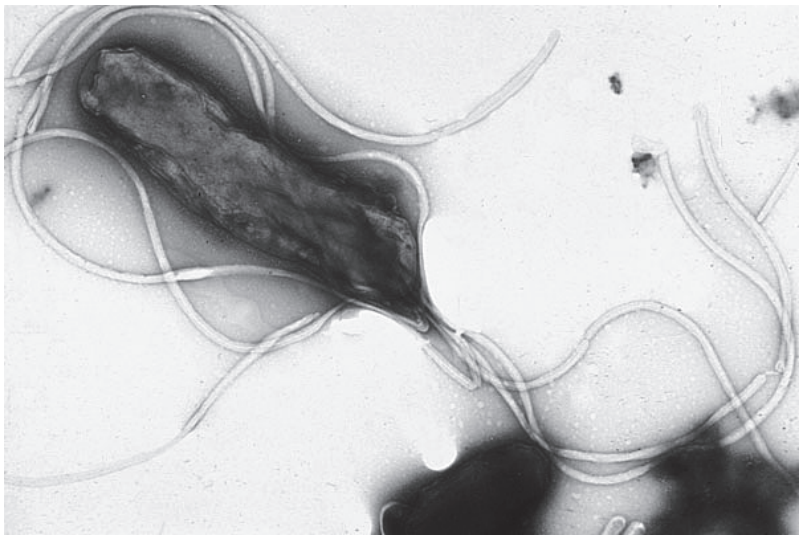


Figure 1.1: Electron micrograph of an *Helicobacter pylori*. The multiple flagella and helix shape makes it highly motile.
Picture present in the English-language Wikipedia page of *Helicobacter pylori*, with permission of copyrighted free use, from Yutaka Tsutsumi.

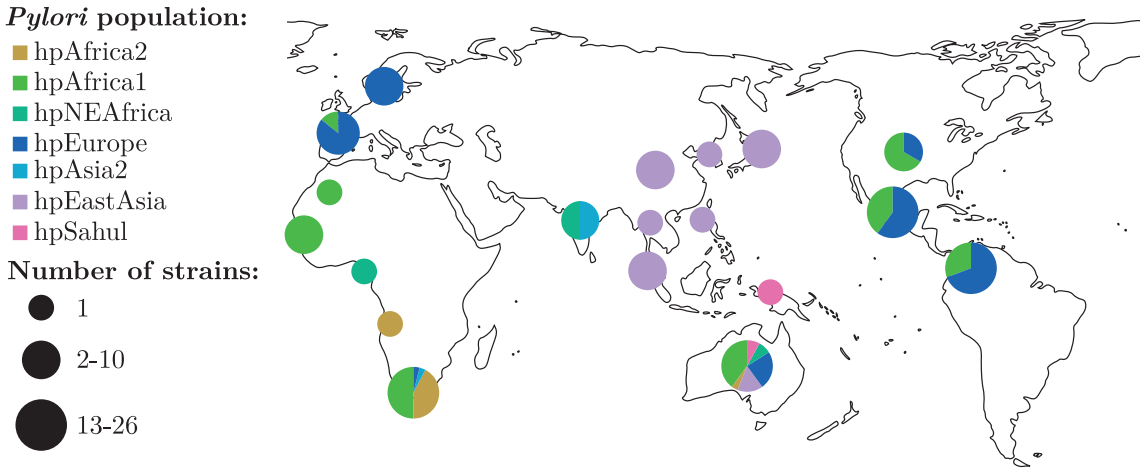


Figure 1.2: Distribution of *Helicobacter pylori* populations depending on the geographical sequencing area. Figure similar to Figure 3A in [76] but with the dataset constructed by Alexia Nguyen Trung in the team, and that we will use in this thesis, containing 119 strains.

terial with each other through recombination. What interest us is to see how these recombinations can give us information on the relationship between humans and *pylori*. As it is clear that for two *pylori* to recombine, the human populations hosting each of them have to meet, or at least the individuals hosting each of them, we want to see how the information contained at the genes' level can help us tell a history at the ecological level of the host and symbiont relationship. Or even, going one step higher, how this information at the symbiont's genetic level can enable us to examine the host migrations between geographical areas.

These embeddings of evolution at multiple levels, here genes, bacterial symbionts, animal host, and geographical ranges, is at the basis of the models we wanted to develop. Starting from phylogenetic reconciliation, an event-based model able to take two levels into account, we tried to understand the coevolution of multiple levels, notably with the explicit modeling of horizontal gene transfers and host switches in a host, symbiont, and gene model.

1.1.2 *Cinara* aphids

Aphids are small sap-feeding insects. Some are generalist and can feed on multiple host plants, while others are specific parasites. They are of particular interest to our agricultural society, as they are some of the most destructive insect pests on cultivated plants. From the aphids' point of view, though, their sap diet is one lacking essential amino acids that it cannot synthesize. To survive on such a diet, aphids have formed a symbiosis with bacteria able to produce them. These bacteria are deemed endosymbionts as they are found in special compartments in the insects called bacteriocytes.

The specific case of aphids we present here is one studied by Manzano-Marín *et al.* in [139], *Cinara* aphids.

This example is interesting to us for multiple reasons. First, it starts from a simple case of symbiosis, with one host and one symbiont, vertically transmitted. It is a simple system compared to relationships like the one between mammals and their gut microbiome, which coevolution is analyzed in [85] for instance. The insect host has a *Buchnera* bacteria with, in its genome, the necessary machinery to synthesize the amino acids to complement the aphids' diet. Moreover, from a phylogenetic point of view, because the symbiont is an endosymbiont and is vertically transmitted, we observe the phylogenies of the host and the symbiont to be perfectly congruent.

Nevertheless, this simple story gets new actors on two levels: new bacterial genes for the production of vitamins and new bacterial symbionts. A new enterobacteria, an *Erwinia*-related symbiont, is observed in the *Cinara* lineages in a separate bacteriocyte. It compensates for the *Buchnera* that lacks, in these aphids, the path-

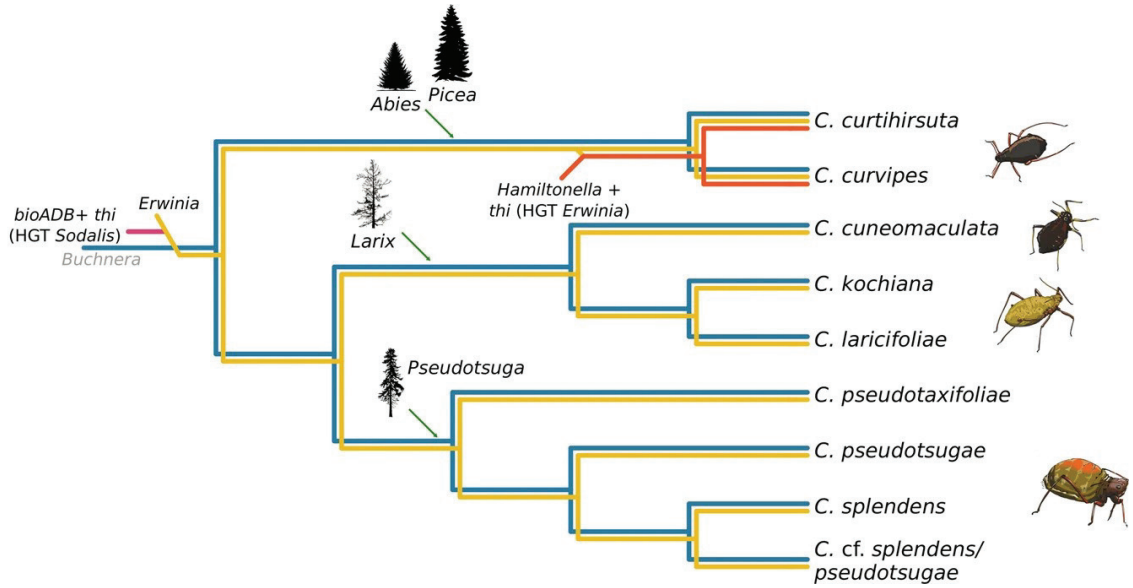


Figure 1.3: A coevolutionary scenario of *Cinara* aphids and their bacterial endosymbionts. The phylogenetic trees are: in blue *Buchnera*, which is believed to be congruent with the *Cinara* one, in yellow *Erwinia*, and in red *Hamiltonella*. Horizontal Gene Transfers are represented as reticulation between these trees, from *Erwinia* to *Hamiltonella* or from *Sodalis* to *Erwinia*. The tree hosts of the aphids are also depicted on top of this phylogeny.
Original figure from Manzano-Marín *et al.* [139], reproduced and modified according to Creative Commons Attribution 4.0 International License.

ways for the production of biotin. The genes observed here and harbored in the *Erwinia* partner genomes are not directly related to *Buchnera*'s but are closer to those of another enterobacteria, *Sodalis*, they were surely acquired through horizontal gene transfer. On top of this compensation, *Erwinia* also exhibits genes for new pathways to synthesize thiamin, also acquired in this gene transfer from *Sodalis*. Furthermore, in two sister lineages, the *Erwinia* genes have been transferred to new *Hamiltonella*-related symbionts and lost by their original host.

This scenario of gene and symbiont exchanges is depicted in Fig. 1.3 on top of the symbionts phylogeny. Note also a fourth level in this system, with the plants depicted on top of the phylogeny and which correspond to the aphids' hosts.

From a functional point of view, the genes necessary to complement the host nutrition are always somewhere inside the host but never in the host's genome and not always in the same bacteria. We witness the acquisition of two new symbionts and two sets of horizontal gene transfer. In one case, the genes are acquired from an exterior bacteria (from *Sodalis* to *Erwinia*), and in the other, the gene exchange seems to occur inside the host (from *Erwinia* to *Hamiltonella*).

This system presents a multi-level history but within a pretty simple relationship from a phylogenetic point of view and with well-identified coevolutionary events. It is thus an ideal example to understand and test new models.

1.1.3 Introduction plan announcement

My goal in this thesis is to present our efforts to model such multi-level systems from a coevolutionary and phylogenetic point of view.

The two biological examples that I had the opportunity to discuss raise several important notions that I still have to define further. The next part of the introduction briefly presents the two main concepts used in this thesis, phylogenetics and molecular evolution on one side and symbiosis on the other.

The rest of the introduction presents phylogenetic reconciliation, an event-based approach to cophylogenetic studies. I present a review of the field and give some biological motivation to the models. In the last part, I describe ALE undated, the specific reconciliation model and method we used and expanded in this thesis.

Contents

SECTION 1.2

Phylogeny and molecular evolution

1.2.1 Phylogenetics

Taxonomy groups extant biological organisms in a variety of scales of granularity, from domains to genus and species. Phylogenetics is the study of the question that follows: what are the relationships between those groups? Phylogenetics study the evolutionary history of organisms, and aim at reconstructing the history of diversification that takes us from one ancestral population, LUCA (last universal common ancestor), to all the organisms we can observe today, after some billions years of evolution [20].

The story of species diversification can be represented with a tree (from Darwin's first sketches to Lifemap complete tree of life viewer [230], Fig. 1.4). Extant species are represented by the leaves of the tree, and ancestral ones are represented by the internal nodes. When progressing from the root to the leaves, we go from the past to the present, from a single species at the root toward multiple extant species at the leaves through events of speciations. With this point of view, the tree represents a scenario of diversification. However, if we take the reverse route, going from the leaves to the root, we take a taxonomic approach. Species are grouped together in bigger and bigger groups that can represent diverse levels of classification.

This widely used model for the diversification of species is sometimes confronted, for instance, in favor of networks [99] that give the possibility for a species to have multiple parent species to account for the importance of horizontal mechanisms, such as horizontal transfer or introgressions in life evolutionary history.

1.2.2 Trees

Trees are a specific case of graphs with useful properties.

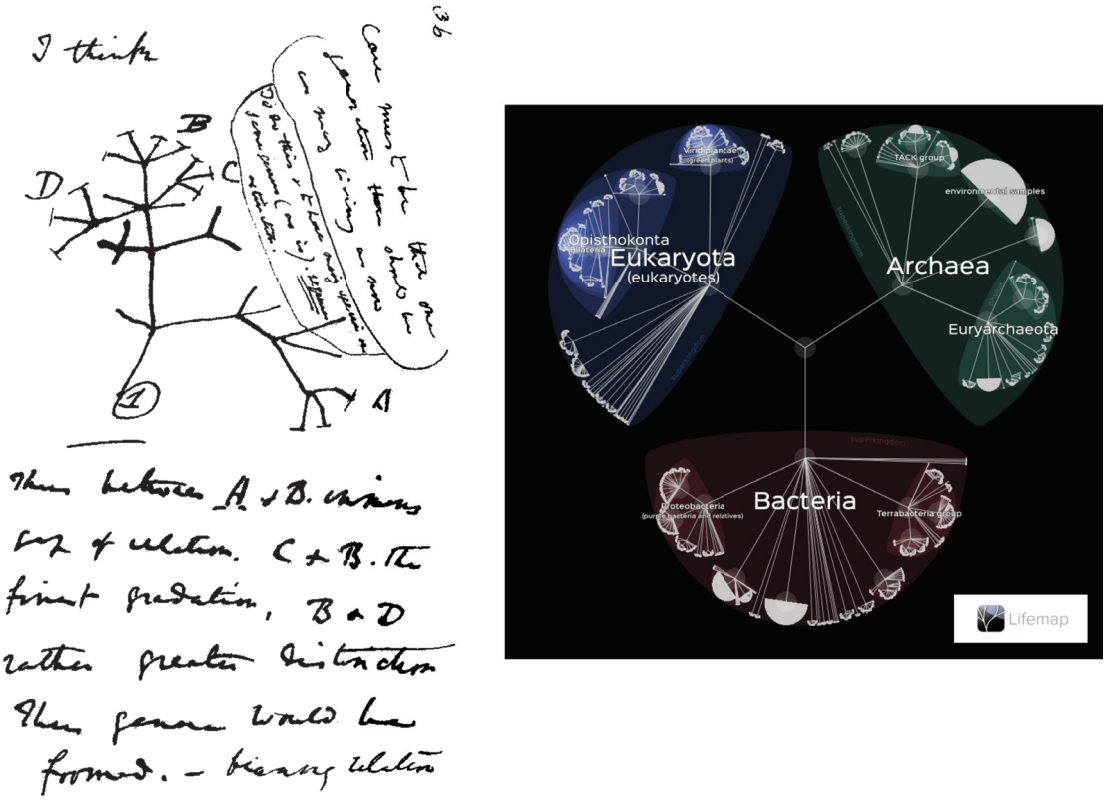


Figure 1.4: On the left Darwin famous sketch of a phylogenetic tree, on the right, the zoom out view in Lifemap [230], a viewer for the entire tree of life (presented unrooted), that makes it possible to explore the tree by zooming in, similarly to an online geographic map.

The science of trees

While an undergrad during my first year entirely devoted to the study of computer science, I assisted to a talk on text compression by a computer scientist, Stéphan Thomassé, who stated that computer science was the science of trees. This provocative statement seems, in fact, quite right to me now, though I am not quite sure why. Trees give more structure to data than a list while being easily traversed and translated to a list of actions for an algorithm. It is the basis for multiple data structures useful for algorithmics, like binary search trees or red and black trees [49], as well as representing practical abstract ideas like decision trees in machine learning[186] or the branch and bound paradigm [115]. It is also a central part of logic and proof theory with proof trees in natural deduction[187]. So could computer science be the science of trees? As commented by a biology student during that talk, "sorry, but the science of trees is biology", and "trees should be drawn with the root at the bottom." What is sure is that this thesis is about trees, both computer science and biological ones, and specifically phylogenetic trees.

In this section, I will use figures to give useful characteristics of trees regarding phylogeny and this thesis. Figures from 1.5 to 1.10 give an overview of what trees are and the notation we will use:

- Fig 1.5 describes the different parts of a tree and how we identify branches and nodes together.
- Fig 1.6 recall the difference between rooted and unrooted trees and how to go from one to the other.
- Fig 1.7 differentiates multifurcating tree from binary trees.
- Fig 1.8 gives a link between trees and systematics notions like ancestors, descendants, clades, monophyletic or paraphyletic groups, and last common ancestor.
- Fig 1.9 shows a post-order traversal of a tree, which is the one we will use the most in this thesis, as it enables us to propagate information from the leaves to the root.
- Fig 1.10 presents what we mean by a matching between two trees (restricted to the leaves or complete), which corresponds to a coevolutionary relationship between the entities in those two trees.

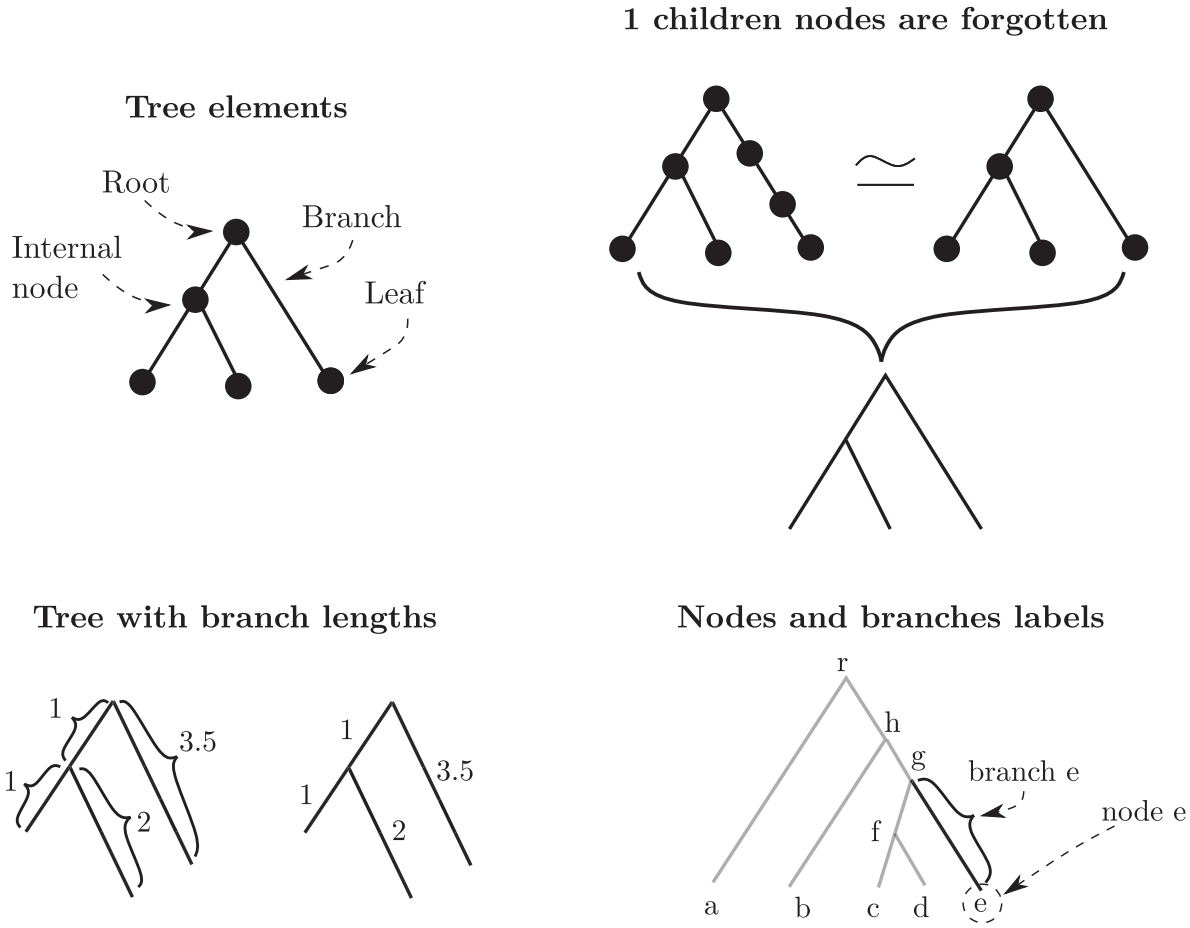


Figure 1.5: A rooted tree is a connected undirected acyclic graph, it has leaves, internal nodes, branches, and a root. In phylogenetics, extant species are represented by leaves, and the internal nodes represent the ancestor species to those species. Branches can be labeled with a length, for instance, that can be represented graphically. In this thesis, as I will mostly use undated models, I will make the two following assumptions on trees: (a) a node with only one child will be seen as equivalent to its parent node, as we see one branch as always representing the same species. In a way, a species only changes when it gets a sister with extant descendants. (b) The branch between a children node and its parent node will be identified to the children node and given the same label.

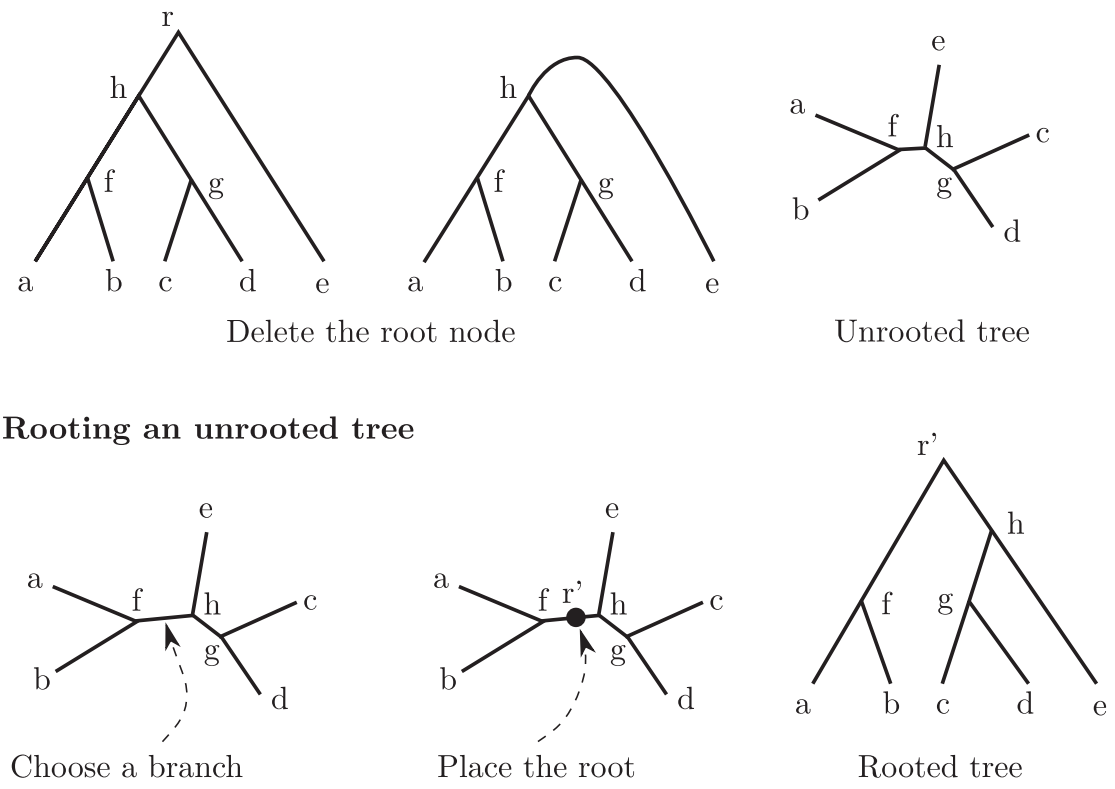


Figure 1.6: A tree can be rooted or unrooted. We can go from rooted to unrooted by deleting the root node. To go from an unrooted to a rooted tree, we choose a branch to be the place of the new root. In this example, we use (a,b) to root the tree. In practice, in phylogeny, an outgroup can be used to find a proper position for the root.

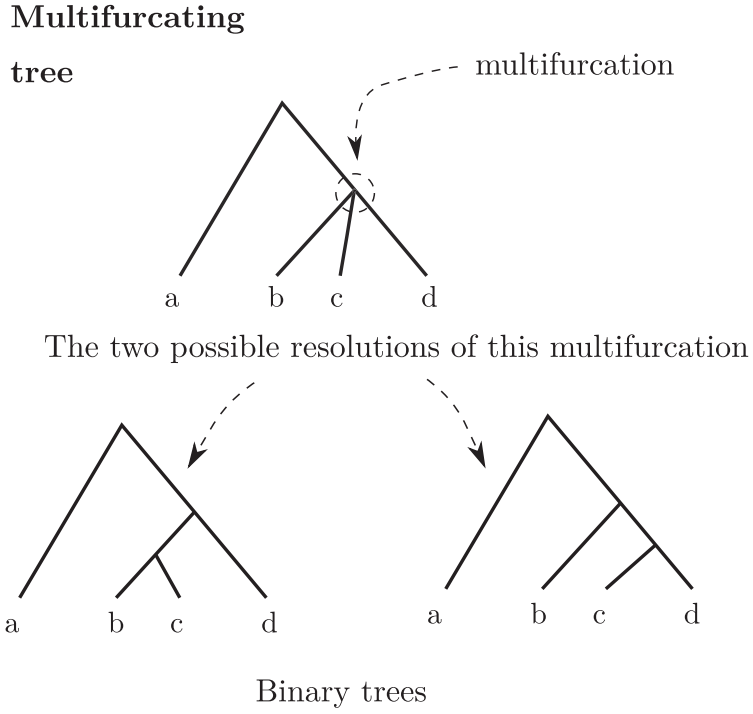


Figure 1.7: A node with more than two children is called a multifurcation (by contrast with a bifurcation). A tree with only bifurcations is called a binary tree. In this thesis, we will mostly consider binary trees. In practice multifurcation can be useful to represent uncertainty, for example in cases where the data or the method do not give enough information to decide between the different resolutions of that multifurcation, and in some cases it might even correspond to a biological event of multiple simultaneous diversifications. A "uncertainty" multifurcation can be resolved to get a binary tree compatible with the multifurcating one. For computational reasons, we will consider uncertainty through other means, for instance by taking a sample of binary trees. In a binary unrooted tree, all nodes are of degree 3, except for the leaves that have degree 1. If the number of leaves is n, then the total number of branches is $2n-2$, and the number of nodes is $2n-1$.

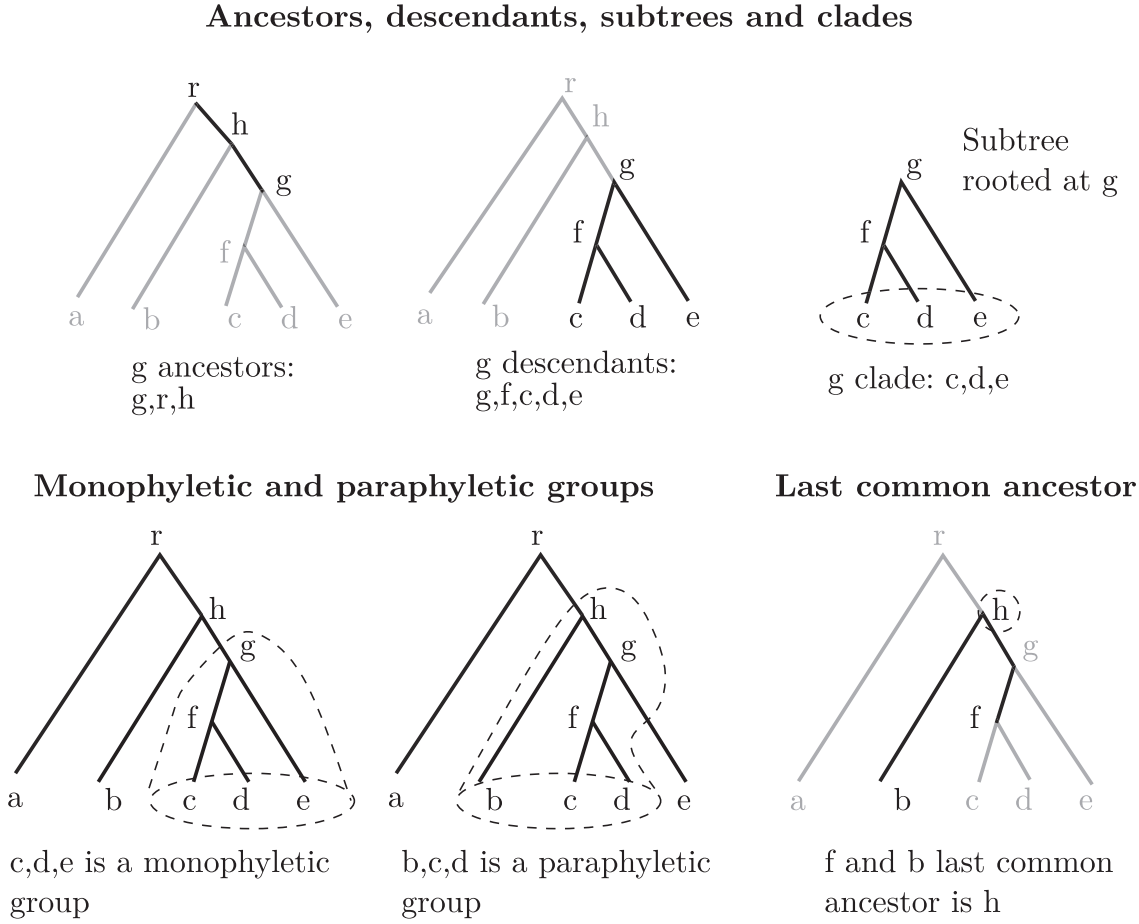


Figure 1.8: In a rooted tree, we can define subtrees by taking a node, all its descendants, and the branches between them. The set of leaves of this subtree is a clade. Ancestors of a node are all the nodes between itself and the root, including itself and the root. Descendants are the nodes of its subtree. A monophyletic group is a set of leaves that is a clade, *i.e.* that is the leaves of some subtree in the tree. We call a non-monophyletic group paraphyletic. If we take all ancestors of a node and all of another, the intersection is not empty, and the first element that is common to the paths from each node to the root is their last common ancestor.

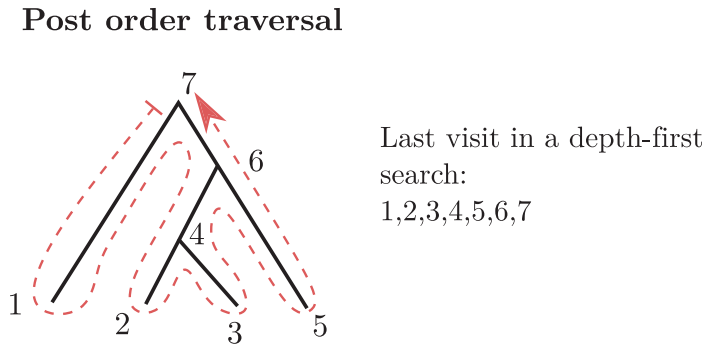


Figure 1.9: We can visit nodes using a post-order traversal of the tree. We add each node the last time we see it in a depth-first traversal starting from the root. This traversal is particularly interesting because it goes from the leaves toward the root, and never visits a node before having visited all its descendants. It thus can be used to propagate information from the leaves toward the internal nodes when we have dependencies between a node and its children, for instance to reconstruct ancestral characters from extant ones.

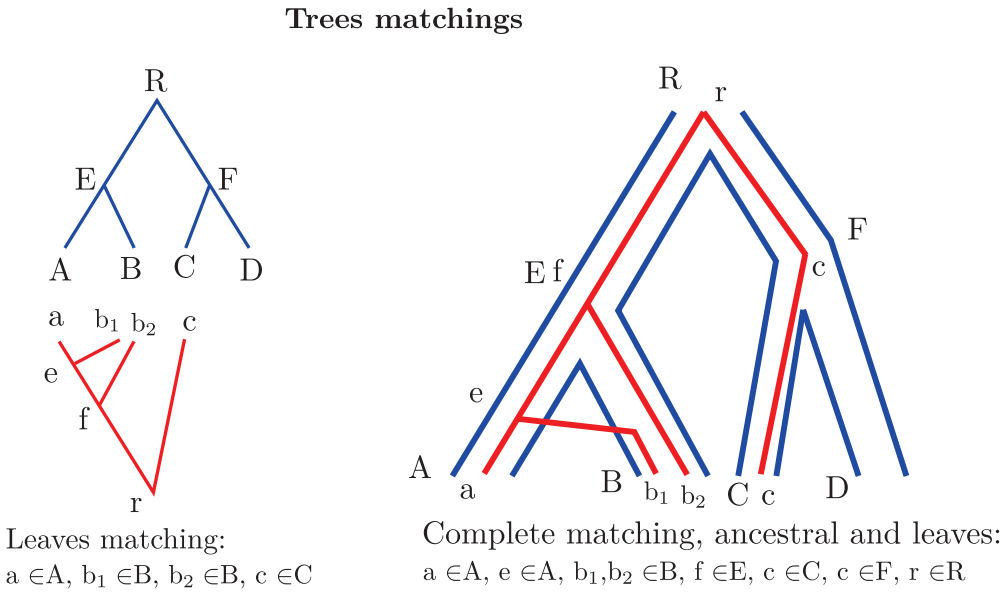


Figure 1.10: As our questions are around the relation between multiple trees, we will often use matchings between two trees, notably host and symbiont trees, or gene and species ones. The leaves of two trees can be matched together. For instance, if we have a symbiont and a host tree, we can match the symbiont leaves to the host they are associated with. Then we can propose scenarios with a complete matching between the two trees. We will represent the matching by making two trees facing one another by the leaves. We will denote coevolution scenarios of the two trees by one as a tube with the other inside. Such a scenario implies a complete matching between the two trees.

1.2.3 Molecular evolution

A tree can represent the diversification of species. The question that follows is how to construct a tree depicting the diversification of a given set of extant species. To do so we can rely on traits exhibited by the species and common to multiple species, for instance, morphological ones. Species exhibiting the same version of some traits will be deemed part of the same clade, as their common ancestor would have displayed the trait. This approach assumes that the trait observed in the extant species descends from the trait of their ancestor. Instead of the presence/absence of some characters, we can also try to reconstruct the history of more complex character patterns. With the assumption that if some characters are inherited, i.e., reproduced with only slight errors from one generation to the next, then the character history must be congruent with the species one. These histories can then support one of the possible trees as the species' phylogeny.

One such character is the genome of living beings. This perspective on evolution is deemed molecular evolution.

Living individuals carry a genome, a sequence of DNA bases, A, T, C, and G. This sequence is not just a list of these four bases, but it also exhibits a structure. For example, it can be organized in multiple linear chromosomes in eukaryotes or in a single circular chromosome in most prokaryotes. A particularly interesting type of subsequences in a genome are the genes. Genes can be transcribed into RNA, that can be directly functional or used to code for proteins. Two individuals have two different genomes, as genomes change over generations through events at multiple scales. Single bases can undergo substitutions, deletions, and additions, while mechanisms such as duplications can impact the sequence at a gene scale; chromosomal rearrangements involve multiple genes; whole-genome duplications impact, as the name suggests, the whole-genome. However, these changes are "slow" so that two species that have diverged recently share an important part of their genomes.

Animal genomes

What is exactly an animal genome? If we take a human being, the nucleus of its cells contains 23 pairs of chromosomes. However, the cells also contain mitochondria, vertically transmitted from the mother to the child, that also has a genome, denoted the mitochondrial genome as opposed to the nuclear genome. And what about the genomes of the multiple symbionts present in the human body? Taking all that into account is the hologenome, from the ancient greek *holos* "whole". We will discuss this question further in the Symbiosis section.

Today, molecular evidence is the character of choice for the inference of species

trees. The place of morphological traits, and more classic cladistics aspects, is still debated [156]. Complementing molecular phylogenetics with a morphological traits-based approach could help construct more robust phylogenies, for instance, to use fossil evidence for dating [119].

1.2.4 The inference of gene trees

Genomes are composed of genes and non-coding DNA. A genome is a sequence of base. It is obtained from the sequencing of the DNA contained in the cells of an organism, complemented by a bioinformatics assembly process to transform the results of the sequencing into a complete genome. The first step to infer phylogenetic trees from DNA sequences is genome annotation which determines the position of the genes in that sequence.

Two genes that descend from a common ancestral gene are said to be homologous and are part of the same gene family. Due to specific events, such as duplication or horizontal transfers, two genes from the same family can be found in the same genome. Before inferring species trees, we will see how to infer trees for gene families.

To retrieve the evolutionary history of a gene family, we first have to identify the genes that belong to the same family. This process is called homology detection. As addition and deletion can happen, two genes in the same family may not have the same length (in base number). The next step, multiple sequence alignment, is to match the homolog bases together, *i.e.* the bases that come from a common ancestral base in that sequence.

Exact methods

Methods are not exact.

- The model, the question we ask, is only an approximation of the real.
- The methods often rely on heuristics, rarely giving the exact answer to the mathematical question we ask them.
- Even if one part of an inference pipeline gives a good view of the uncertainty of its output, it is not evident to propagate this uncertainty to the next part of the pipeline, even more to do so until the end of the pipeline.

From this multiple sequence alignment, we can start the proper tree inference. What is important to keep in mind for this thesis is that all these processes, sequencing, genome assembly and annotation, gene homology detection, and multiple sequence alignment, are complex and deliver an uncertain output (with that uncertainty known or not). It makes it important to keep a way to correct inputs or take

into account uncertainty in the subsequent methods.

Now that we have aligned sequences, each element of the alignment is called a site, and for each gene and site, we have a base, A, T, C, or G, or a gap, when the base is absent. This is the case for a nucleotide alignment although many of the trees we use are derived from protein alignments. Given a potential tree, we will use some criteria to see how well the tree is compatible with the base at each site. In maximum parsimony, we minimize the number of base-pair substitutions in the tree, while bayesian and maximum likelihood rely on a probabilistic model of base-pair substitution to return a tree.

Likelihood

Likelihood is a concept central in statistics and phylogenetics. The likelihood of a hypothesis or a model is the probability of generating the data given that hypothesis. We can go from likelihood to probability using Bayes theorem and introducing priors. With data D and two hypotheses H_1 and H_2, with L the likelihood and P the probability:

$$\frac{L(H_1)}{L(H_2)} = \frac{P(D|H_1)}{P(D|H_2)} = \frac{P(H_1|D)P(D)P(H_2)}{P(H_2|D)P(D)P(H_1)} = \frac{P(H_1|D)P(H_2)}{P(H_2|D)P(H_1)} \tag{1.1}$$

The data prior $P(D)$ can be discarded when we want to compare two models, so the only prior important for us is the one on the hypothesis $P(H_1)$ and $P(H_2)$. What interests us is the probability of the hypotheses given the data, but in most cases, what we can compute directly from the data is the likelihood.

Bayes maximum likelihood and parsimony

Bayesian, maximum likelihood, and parsimony are three possible approaches to an inference problem. The simplest one, parsimony, relies on minimizing the number of some events deemed unlikely. When applying maximum likelihood we try to get the tree that is the most likely to have generated the data. In Bayesian, we want to access the posterior probability, $P(H|D)$ (instead of the likelihood $P(D|H)$). To do so, we need to have a prior (and how to choose it is kind of the controversial question of Bayesian). We then can search the space of possible trees, often using Monte Carlo Markov Chains in its Metropolis Hasting version, such that we visit elements proportionally to their posterior probability in our model. Sampling the elements of this traversal makes it possible to suppress the autocorrelation of the chain and access an estimation of the posterior distribution. This estimation can then be returned or summarized by a single tree (for instance, by looking at the

frequency of presence of some bifurcation node).

This inferred gene tree represents the relationships between the genes in that gene family. We will see that gene trees are of interest in their own right, notably to better understand the mechanisms that make genes evolve inside species, but we first have to talk about how we can finally get a tree for the evolution of species.

1.2.5 The inference of species tree

The first proxy for the phylogeny of species, *i.e.* the phylogeny of complete genomes, is to take one universal and unicopy gene - a gene present in one and only one copy in the genomes of each of the species at hand - construct its tree and say: that is the tree of the species. There are three problems with that approach:

- From an information theory point of view, genes have a finite length, and as each site can only support a finite number of nodes in the gene tree, the more species we want to consider, the more base we need to construct a tree.
- Statistical methods and models do not infer real scenarios.
- Even if the information contained in a gene was enough to get the exact history of this gene, we could not be sure it is the same history as the species.

Genes do not always follow the same history as their genomes, and the observed differences between the phylogenies of different gene families on the same species set are not only the results of methodological variability or lack of information. The evolutionary processes that make a gene escape its genome are one of the main subjects of this thesis and will be discussed in section 1.4. Even with this caveat, a gene tree can be used as a proxy for a species tree when we think the gene has followed precisely its genome, that it evolved slowly, is well conserved, and did not transfer. For instance, 16s ribosomal RNA genes are often used with this aim, notably for their slow rate of evolution [241].

If we want to consider multiple genes to construct a species tree, they may (and with a good chance) not all be identical but disagree on some points. Multiple methods exist to do so that can be attached to two main groups, concatenate and supertrees [31] [233]. In concatenate methods, the aligned sequences of all gene families considered are concatenated, giving a single big alignment. A tree is then inferred using this multiple sequence alignment. On the other hand, supertree methods rely on explicit gene trees inferred beforehand. The information contained in these trees is aggregated to get a single species tree, depending on philosophical principles (like consensus and majority votes) or explicit mechanistic and statistical models.

Sequencing (and computational) capacities make it possible today to consider all the genes in whole-genome sequences analyses and construct species trees from this complete information[72]. However, many studies still rely on a smaller subset of genes in practice. Housekeeping genes, responsible for essential cellular function, are used for identifying isolates of microbial species, notably human pathogens, in multilocus sequence typing (MLST) - as opposed to 16s typing, which rely only on one gene - using predetermined genes' locus to identify strains [137].

For instance, in most studies we present on *Helicobacter pylori*, analyses rely on MLST, with seven to nine genes (from a set of seven housekeeping genes and two pathogenicity associated genes). Similarly in the *Cinara* aphids study we already presented [139], two of the symbionts are placed in an enterobacteria phylogeny using concatenate of all genes, while the last one (*Hamiltonella*) is placed using seven genes.

A significant discussion is the one between Ciccarelli *et al.* [46] and Dagan and Martin [53] in 2006. The first authors present a method to construct trees with species from all parts of the tree of life. They identified 31 orthologs proteins (homolog sequences that result from species diversification and not duplication at the gene level) universal across 191 species in the "three" domains of life and suitable to construct a phylogeny of these species. The virulent answer states that this is only the tree of 1%, as only a small portion of genes are kept.

Being able to take into account not only the universal genes and to explicitly model the evolutionary events that make genes disagree with their species is the goal of phylogenetic reconciliation. It can enable us to consider more diverse stories of gene and genome coevolution than strict coevolution and, in doing so, be a part of a more reliable species tree inference method. It is the main subject of this thesis, and I will introduce it after giving an overview of another biological framework where we can consider coevolution: host and symbionts.

Here, and in a significant amount of this thesis, I will discuss constructing more integrative and more complex models using more data. However, the more genes we take into account, the more complex models are used for tree reconstruction, and the more costly the construction of the tree is in terms of time, money, and impact on the environment (measured in carbon footprint for instance). Furthermore, the advances in reducing these costs are often used in more costly analyses than in cheaper ones, in a rebound effect manner. See [208] for a discussion on the current limits of phylogenomics, including a discussion on the carbon footprint increase between two analyses on the same dataset with different methodologies.

SECTION 1.3

Host and symbionts

1.3.1 Symbiosis definition

Symbiosis is a close interaction between the individuals of two distinct species. The standard denomination for the two species is symbionts. However, when studying this kind of relation, we will often lose the symmetry of the definition to denote one of the partners as host and one as simply symbiont. Usually, we will call the "big" one host, notably in cases where the symbiont is carried by the host on the "outside," like for toucan and chewing lice [234] or sloth and algae [79], or "inside" for enterobacteria in symbiosis with aphids [139], for mammals microbiomes [85].

Symbiosis and symbiont are generic terms for any close interaction, *i.e.* not only mutualistic interaction (beneficial for both), but also commensal (neutral for one and beneficial for the other), neutral, or parasitic (benefit for one and harmful for the other).

1.3.2 Host symbiont coevolution

Studying symbiosis and all the innovative ways two species can collaborate or prey on one another is a fascinating subject. In this thesis, however, we will have to focus only on the link between symbiosis and phylogeny, *i.e.* how the interaction between symbionts evolve and how that information can help us to construct a better evolutionary history for each symbiont.

We say two species coevolve when they evolve in interaction, often leading to phylogenies that are not independent but congruent. In fact, from a phylogenetic point of view, it is improper to say that two species coevolve, we cannot really say that two leaves of our two symbionts species trees are dependent. Coevolution is more defined on the long-term as the coevolution of two families of species.

Gene species coevolution

As we define coevolution between host and symbiont, we can more generally define coevolution between two biological entities. For instance gene and genome coevolve, as one gene evolution is strongly dependent on the species it belong to, and inversely the genome can be seen as the sum of all its genes. Genes can also coevolve together, as a results of a common coevolution with a genome they both are part of. They can also coevolve more strongly together than with this genome, for instance in cases where the genes are part of a common function, and thus can often be seen escaping the species evolution

together, for instance through segmental transfers. Those different levels of coevolution are exemplifies in the *Cinara* system, where endosymbionts coevolve with their hosts and genes of the symbiont coevolve with their host, and together, for the ones that are responsible for the thiamin and biotin synthesis, more than with the symbionts genomes.

An extreme answer to the coevolution question of symbiont species is Farenholz's rule[2] which states that host and symbiont phylogenies mirror each other. At the other end of the spectrum, the alternative hypothesis for the coevolution of two species, even if involved in symbiosis, is independent evolution. We can see this as accurate for very generalistic symbionts, for instance, parasites that can feed on a diversity of hosts. Still, it could be the case for most symbiosis [228].

We thus need a way to evaluate the coevolution of host and symbiont and be able to reconstruct the ancestral relationship between two families. Testing coevolution, and cospeciation, the common diversification of both partners at the same time to adapt to one another (as genes cospeciate with species), can be done with topology and distance-based methods - with no explicit modeling of the coevolution - or with event-based methods, that will also retrieve ancestral correspondences [228]. One of these event-based methods is phylogenetic reconciliation, which we already mentioned for the coevolution of gene and genome, and that I review in the next section, after telling a bit more about holobionts.

1.3.3 Holobiont: a thought-provoking word

The holobiont is a quite recent concept around the host/symbiont, and organisms paradigms [140]. It stipulates that complex organisms such as animals or plants can be considered as a system that takes into account the various organisms that live inside or around them and the genes of all these organisms, notably the ones not present in the "host" - the main species - nucleus [21, 197, 255]. Holobionts are complex systems at any time t, but what interested us was to confront their evolution. How do holobionts evolve? How can each part coevolve or escape one another or the host? And how can we test for such complex coevolution?

What is important to me is the idea of a multiplicity of individuals working together, some of them coevolving, with some links stronger than others, and not the one we were thinking of. In a way, trying to decipher these links is the final goal of our work.

An excellent example of that idea of a level escaping another is the *Cinara* aphids example we presented earlier. The genes coding for biotin and thiamin synthesis,

[2]The often-cited article from Farenholz dates back to 1913 and is only available in German.

beneficial to the host nutrition but which belong to the symbiont genomes, coevolved more strongly with their aphid host than with their own genome, *i.e.* their phylogeny is more similar to the one of the aphids than to that of their genome.

The use of this word in the thesis title was more thought-provoking than really at the center of our methodological work, but it is found here and there as a big picture that led us during these three years and that sometimes we went back to. The second part of our review on reconciliation discusses systems and ideas that go along with that holobiont concept.

SECTION 1.4

Phylogenetic reconciliation

In this part, I propose a review article on reconciliation, with particular attention to a specific set of events duplication, horizontal transfer, and loss (DTL). It is in two parts, the first is a history and explanation of DTL reconciliation, while the other presents datasets, studies, and the first methodological advances that tackle multi-level systems, such as the ones we presented during the first pages of this introduction with genes, bacteria, and host.

"Phylogenetic reconciliation" review article

This section is based on a review article we wrote on phylogenetic reconciliation. The article was prepared for Plos Computational Biology Topic Pages format, which propose to write a review article that will also be included in Wikipedia. As phylogenetic reconciliation was not featured in the collaborative encyclopedia, we decided it would be an excellent opportunity to review our field and propose a simple as possible entry in it. Strange enough, on Wikipedia, though no pages existed in English, a small one was present in French, with a single reference, an article written by Eric Tannier, Bastien Boussau, and Vincent Daubin in a French science popularization magazine, *Pour la science*.

We have been in discussion for more than one year with PLoS Computational Biology for this Topic Pages format, and have two publicly available reviews online that we answered. In this introduction, I propose a revised version, taking the reviewers' comments into account. If you read this thesis on a computer with an internet connection, the nicest way to read the following section, from page 40 to page 71, might be to check the page online on PLoS Wiki. You can also look at the insightful comments of the two reviewers, Ross Mounce and Mukul Bansal, and our responses in the discussion.

If you want an idea of a published PLoS Computational Biology topic page on a related subject, you can look at the *Inferring horizontal gene transfer* one, on Wikipedia, PLoS wiki, or as a PLoS Computational Biology article.

The goal of our approach regarding the thesis I was starting was two kinds. First, get a better understanding of the field, with a bit of a bigger picture than only looking at articles directly on my subject. That, I tried to do exhaustively, though I know I did not achieve such a goal. Second, I then constructed a list, this time of examples, looking more toward diversity than completeness, that had anything to do with 3-level approaches, both on a methodological

and biological basis. I found our first biological test case while working on that review, the *Cinara* aphids one.

Figures 1.14 and 1.25 are meant to be used as a visual summary of the different methods presented in this section. The first is about 2-level reconciliation methods, the second more about multiple levels reconciliation frameworks. The cells of these figures are also used to illustrate the paragraphs.

1.4.1 Definition

In phylogenetics, reconciliation is an approach to connect the history of two or more coevolving biological entities. The general idea of reconciliation is that a phylogenetic tree representing the evolution of an entity (e.g. homologous genes, symbionts...) can be drawn within another phylogenetic tree representing an encompassing entity (respectively, species, hosts) to reveal their interdependence and the evolutionary events that have marked their shared history. The development of reconciliation approaches started in the 1980s, mainly to depict the coevolution of a gene and a genome, and of a host and a symbiont, which can be mutualist, commensalist or parasitic. It has also been used for example to detect horizontal gene transfer, or understand the dynamics of genome evolution.

Phylogenetic reconciliation can account for a diversity of evolutionary trajectories of what makes life's history, intertwined with each other at all scales that can be considered, from molecules to populations or cultures. A recent avatar of the importance of interactions between levels of organization is the holobiont concept, where a macro-organism is seen as a complex partnership of diverse species. Modeling the evolution of such complex entities are one of the challenging and exciting direction of current research on reconciliation.

1.4.2 Phylogenetic trees as matryoshka dolls

Phylogenies have been used for representing the diversification of life at many levels of organization: macro-organisms [87], their cells throughout development [164], micro-organisms through marker genes [242], chromosomes [60], proteins [256], protein domains [9], and can also be helpful to understand the evolution of human culture elements such as languages [84] or folktales [218]. At each of these levels, phylogenetic trees describe different stories made of specific diversification events, which may or may not be shared among levels. Yet because they are structurally nested or functionally dependent, the evolution at a particular level is bound to others.

Phylogenetic reconciliation is the identification of the links between levels through the comparison of at least two associated trees. Originally developed for two trees, reconciliations for more than two levels have been recently constructed. As such, reconciliation provides evolutionary scenarios that reveal conflict and cooperation among evolving entities. These links may be unintuitive, for instance, genes present in the same genome may show uncorrelated evolutionary histories while some genes present in the genome of a symbiont may show a strong coevolution signal with the host phylogeny. Hence, reconciliation can be a useful tool to understand the constraints and evolutionary strategies underlying the assemblage that makes an holobiont.

Because all levels essentially deal with the same object, a phylogenetic tree, the same models of reconciliation, in particular those based on duplication-transfer-loss events, which are central to this article, can be transposed, with slight modifications, to any pair of connected levels [237]: an "inner", "lower", or "associate" entity (gene, symbiont species, population...) evolves inside an "upper", or "host" one (respectively species, host, geographical area...) (Figure 1.12). The upper and lower entities are partially bound to the same history, leading to similarities in their phylogenetic trees, but the associations can change over time, become more or less strict or switch to other partners (Figure 1.11).

In the following part of this text, we will give a review of DTL reconciliation methods and models, starting by an historical and methodological approach to the construction of the model. Two-level reconciliation methods, have been reviewed several times, but generally focusing on a particular pair of levels, e.g. gene/species or host/symbiont [26, 217, 63, 166, 40, 39, 143], the following parts are written with a generic voice and to confront models constructed in different frameworks. The last part of the article focus on efforts toward reconciliation with more than two levels, and a description of some biological studies that look at such models.

1.4.3 History

The principle of phylogenetic reconciliation was introduced in 1979 [81] to account for differences between genes and species phylogenies. In a parsimonious setting, two evolutionary events, gene duplication and gene loss were invoked to explain the discrepancies between a gene tree and a species tree. It also described a score on gene trees knowing the species tree and an aligned sequence by using the number of gene duplication, loss, and nucleotide replacement for the evolution of the aligned sequence, an approach still central today with new models of reconciliation and phylogeny inference[158].

The name *reconciliation* has been used by Maddison, 1997 [135], as a reverse

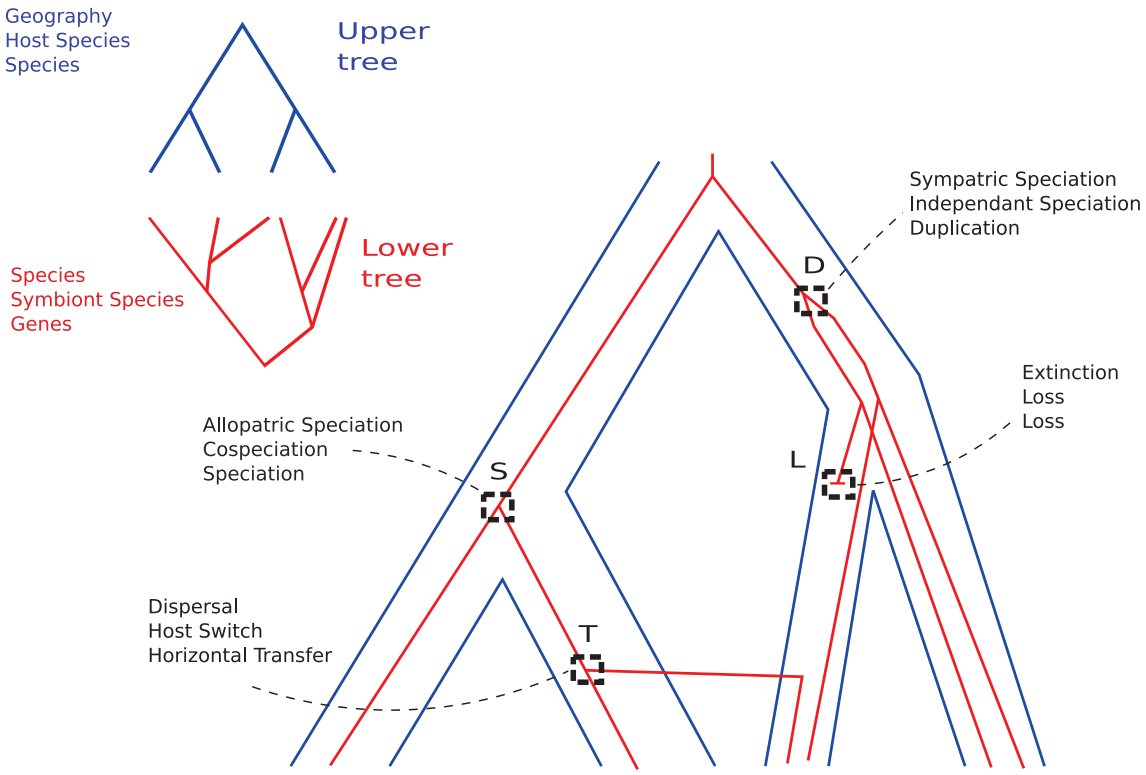


Figure 1.11: A phylogenetic reconciliation between an upper, blue, and a lower, red, tree, with the most often used evolutionary events (S,D,T,L), and their name in phylogeography, host/symbiont and gene/species frameworks. For instance S event is called allopatric speciation when reconciling geographical areas and species, cospeciation between host and symbiont, and speciation for gene and species, but always correspond to the same co-diversification pattern.

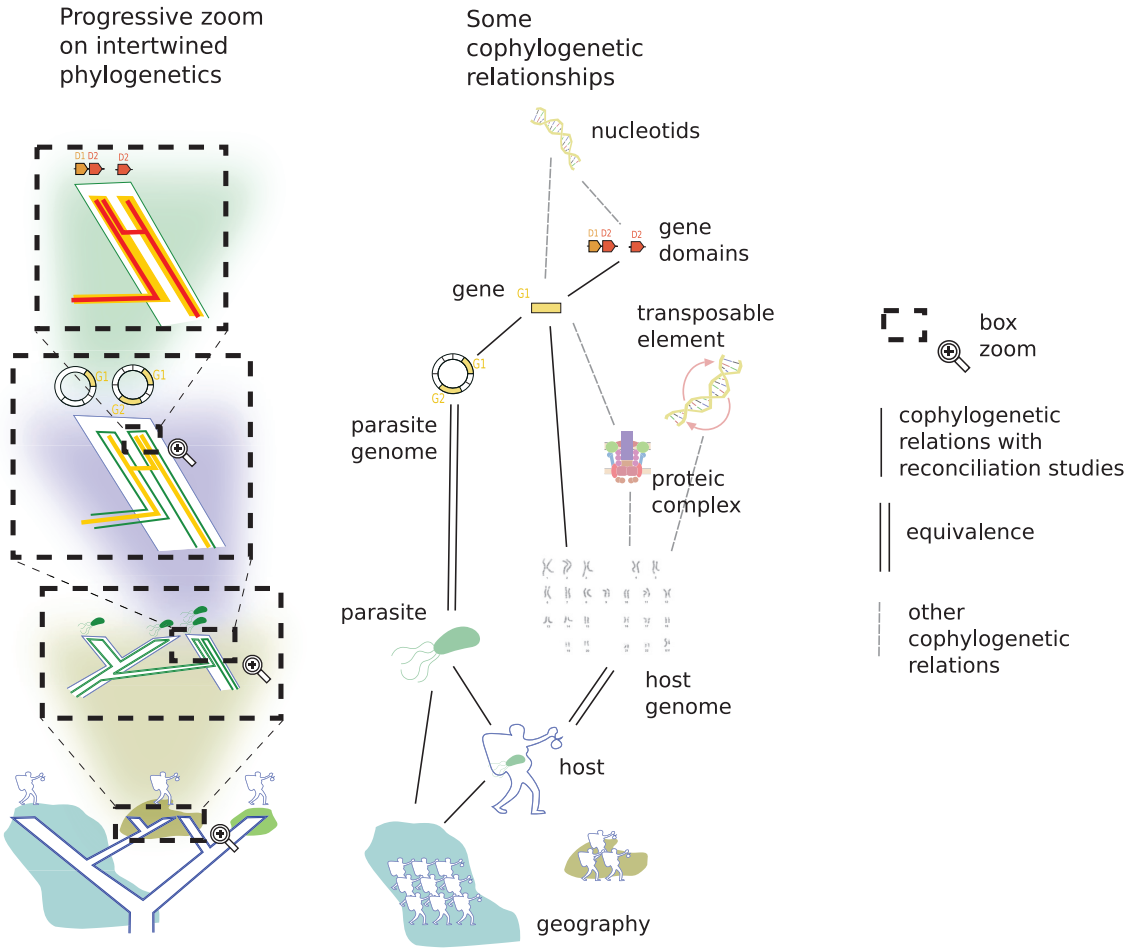


Figure 1.12: Phylogenetic trees are intertwined at all levels of organization, integrating conflicts and dependencies within and between levels. Macro-organism populations migrate between continents, their microbe symbionts switch between populations, the genes of their symbionts transfer between microbe species, and domains are exchanged between genes (left third). This list of organization levels is not representative or exhaustive, but give a view of levels where reconciliation methods have been used. As a generic method, reconciliation could take into account numerous other levels, for instance it could consider the syntenic organization of genes [69, 257], the interacting history of transposable elements and species [131], the evolution of protein complex among species [58]. The scale of evolutionary events considered can go from population events such as geographical diversification to nucleotides levels one inside genes[78], including for instance chromosome levels events inside genomes such as whole genome duplication [257].

image of "phylogenetic discord" resulting from gene level evolutionary events.

Reconciliation was then developed jointly for the coevolution of host and symbiont and the diversification of species on geography. In both settings, it was important to model a horizontal event that implied parallel branches of the host tree: host switch for host and symbiont and species dispersion from one area to another in biogeography. Unlike genes and genomes, the coevolution of host and symbiont and the explanation of species diversification by geography are not always the null hypothesis. A visual depiction of the two phylogenies in a tanglegram can help assess such coevolution, although it has no statistical obvious interpretation[229].

Character methods, such as Brooks Parsimony Analysis [30], were proposed to test coevolution and reconstruct scenarios of coevolution. In these methods, one of the trees is forgotten except for its leaves, which are then used as a character evolving on the second tree.

First models for reconciliation, taking explicitly into account the two topologies and using a mechanistic event-based approach, were proposed for host and symbiont and biogeography [168, 196]. Debates followed, as the methods were not yet completely sound but integrated useful information in a new framework[176].

Costs for each event and a dynamic programming considering all pairs of host and symbiont nodes were then introduced in a host and symbiont approach, both of which still underlies most of the current reconciliation methods for host and symbiont, and species and genes[38]. Reconciliation returned to the framework it was introduced in, gene and species. After character models were considered for horizontal gene transfer [93], a new reconciliation model, following and improving the dynamic programming approach presented for host and symbiont, effectively introduced horizontal gene transfer to gene and species reconciliation on top of the duplication and loss model[90].

The progressive development of phylogenetic reconciliation was thus possible through exchanges between multiple communities, the host and symbiont, gene and species, and biogeography one. This story and its modern developments have been reviewed several times, generally focusing on specific pairs of levels, with a few exceptions [178, 237]. New developments start to bring the different frameworks together with new integrative models.

1.4.4 Pocket Gophers and their chewing lices: a classic example

Pocket gophers (*Mammalia : Rodentia*) and their chewing lice (*Insecta : Phthyraptera*) is a well studied system of host and symbiont coevolution[88]. The phylogeny of host and symbiont and the matching of their leaves are depicted on

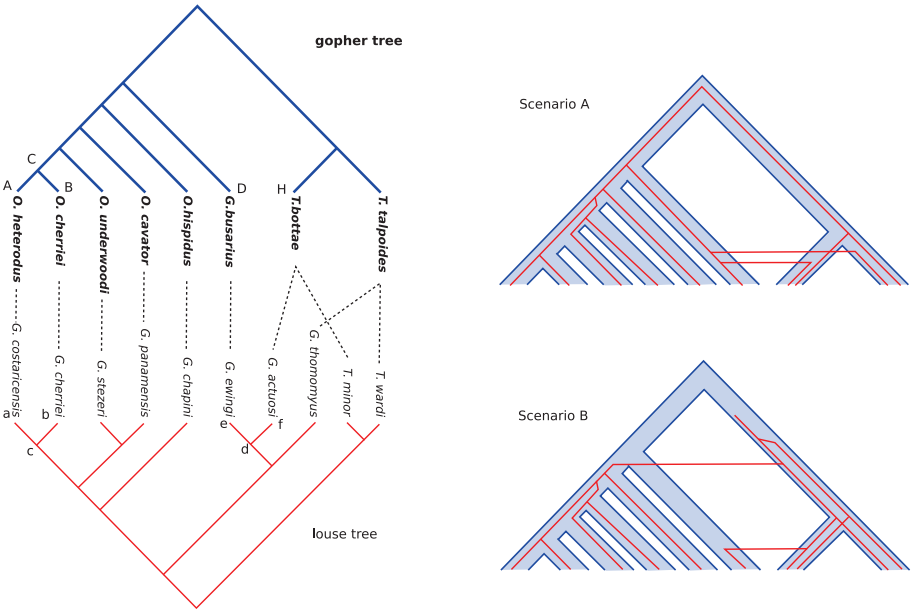


Figure 1.13: Tanglegrams and two proposed reconciliation scenario for pocket gophers and their chewing lices symbionts. For the host, O. stands for *Orthogeomys*, G. for *Geomys* and T. for *Thomomys*; for the symbiont G. stands for *Geomydoecus* and T. for *Thomoydoecus*.

the left of figure 1.13. Reconciling the two trees consists in giving a scenario with evolutionary events and matching on the ancestral nodes depicting the coevolution of the two trees. The events considered in this system are the events of the DTL model: duplication, transfer (or host switch), loss, and cospeciation, the null event of coevolution. Two scenarios were proposed in two studies[179] [195], using two different frameworks which could be deemed as pre-dynamic programming DTL reconciliation. In modern DTL reconciliation frameworks, costs are assigned to events. The two scenarios were then showed to correspond to maximum parsimonious reconciliation with different cost assignments [38]. The scenario A uses 6 cospeciations, 2 duplications, 3 losses and 2 host switchs to reconcile the two trees, while scenario B uses 5 cospeciations, 3 duplications, 3 losses and 2 host switchs. The cost of a scenario is the sum of the cost of its events. For instance with cost of 0 for cospeciation, 2 for duplication, 1 for loss and 3 for host switch, scenario A has a cost of $6\times0+1\times2+3\times1+1\times3=8$ and scenario B of $5\times0+1\times2+3\times1+2\times3=11$, and so according to a parsimonious principle, scenario A would be deemed more likely (scenario A stays more likely as long as the cost of cospeciation is less than the cost of duplication).

1.4.5 Development of phylogenetic reconciliation models

Models and methods used today in phylogeny (Figure 1.14) are the result of several decades of research, made of a progressive complexification, driven by the nature

of the data and the quest for biological realism on one side, and the limits and progresses of mathematical and algorithmic methods on the other. See Figure 1.14 for an illustration of the models and methods presented.

Computational complexity and NP-hardness

In computer science, computational complexity is an abstraction of the needed computing resources time for one problem, depending on the size of the input. It is often denoted with big O notations, to keep only the most important factors. Acceptable complexity depends a lot on the problem at hand and the usual size of the input. One important complexity class is NP, which is the class of problems for which we can check a solution in polynomial time, which mean that in exponential time we can give a solution, as we can enumerate all possibilities in that time and check each. One of the results that I myself find fascinating is Cook theorem and the definition of NP completeness, which say that one problem is harder than all others in the NP class, meaning if we can solve it fast, we can solve all others fast. These problems are called NP-complete, and we know only exponential algorithms to solve them. It is thus possible to show that a problem is NP hard, by showing that is harder than a known NP hard problem. NP hardness is quite interesting, as with any problem, solving an NP hard problem is long, as exponential grows really fast. When a problem is shown to be NP hard, it is not the end of the work on it, but we at least stop to search for a polynomial exact solution for all cases. We can look for heuristics, approximation (heuristic with a known distance to the exact solution), cases for which the problem would not be NP hard, or a better comprehension of the part of the input that is important (Fixed Parametr Tractable solutions).

Pre-reconciliation models: characters on trees.

Character methods can be used when there is no tree available for one of the levels, but only values for a character at the leaves of a phylogenetic tree for the other level. A model defines the events of character value change, their rate, probabilities or costs. For instance the character can be the presence of a host on a symbiont tree [30], the geographical region on a species tree [240], the number of genes on a genome tree [51], or nucleotides in a sequence [78]. Such methods thus aim at reconstructing ancestral characters at internal nodes of the tree [85].

Although these methods have produced results on genome evolution, the utility of a second tree appears with very simple examples. If a symbiont has recently acquired the ability to spread in a group of species and thus it is present in most

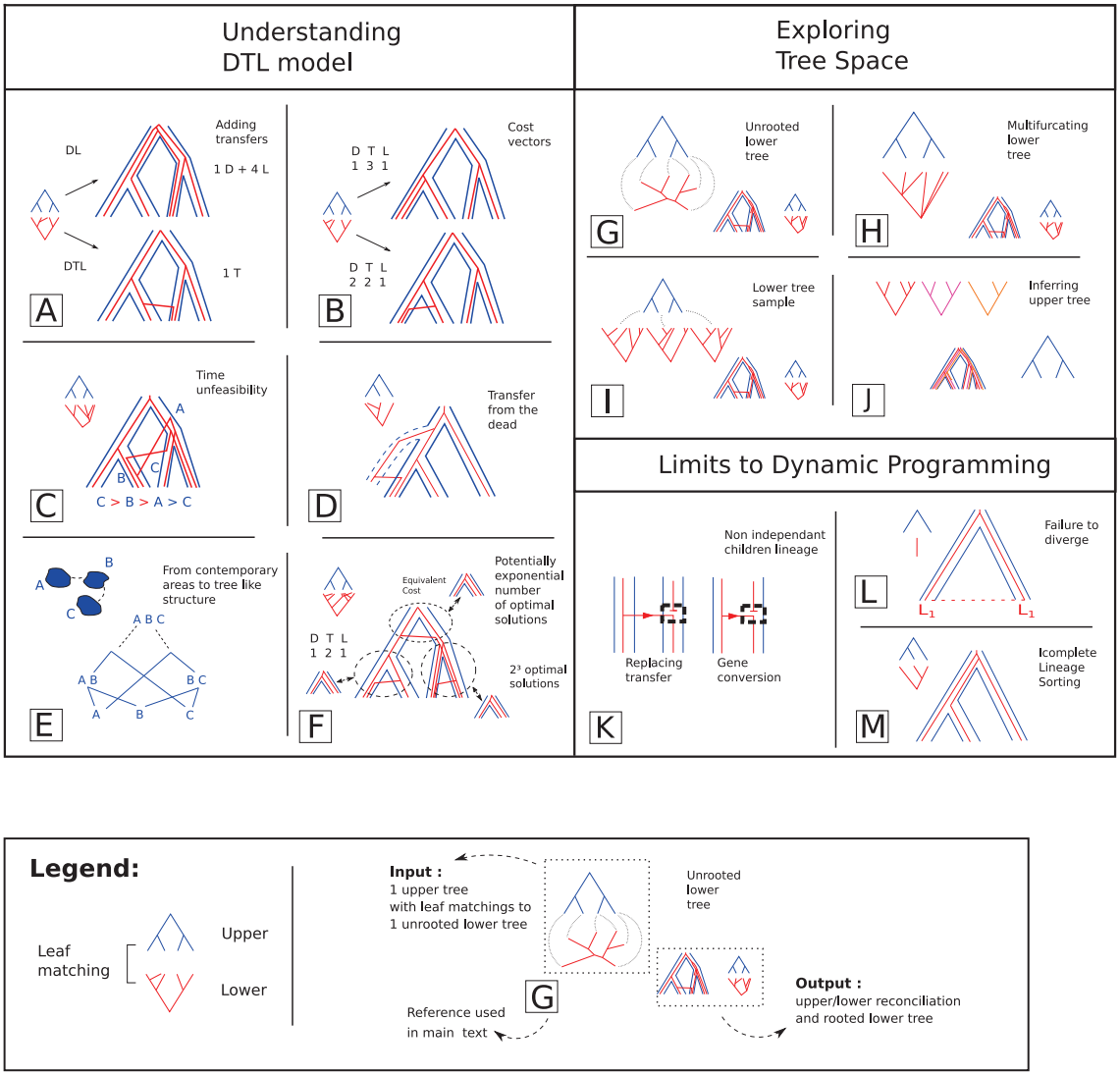


Figure 1.14: Illustration of reconciliation events, inputs, outputs, and computational difficulties. This table is intended to serve as illustration to section 1.4.5 and can be read along it. Inputs are on the left of entries, output on the right. Upper trees are drawn in blue, lower trees in red. Adding the horizontal Transfer event add new more parsimonious solutions compared to the previous DL model (A). With this new event, costs must be assigned to D,T and L events, and different costs give different solutions (B). Not all scenarios including transfers are time feasible. Some might include time constraints incompatible with the upper tree (C). Transfer can go from a species to one of its descendant via a sister lineages that went extinct (D). In biogeography, a tree like structure can be constructed to account for the possible migrations between different geographical areas (E). In some cases, an exponential number of scenarios might be most parsimonious, for example when two equivalent patterns have the same cost (F). The lower tree can be unrooted (G), multifurcating (H), or given as a sample of potential trees (I) and reconciliation can be used to resolve those uncertainties to get a binary rooted lower tree. Reconciliation score can also be used to help construct an upper tree (J). The dynamic programming is limited, by the fact it assume independence between sister lineages, that makes it unable to consider replacing transfers or gene conversion (K), as well as Failure to diverge (L) and Incomplete Lineage Sorting (M), two population level events.

of them, characters methods will wrongly indicate that the common ancestor of the hosts already had the symbiont. In contrast, a comparison of the symbiont and host trees would show discrepancies revealing horizontal transfers.

The origins of reconciliation: the Duplication Loss model and the Lowest Common Ancestor mapping.

Duplication and loss were invoked first to explain the presence of multiple copies of a gene in a genome or its absence in certain species [256]. It is possible with those two events to reconcile any two trees [81] *i.e.* to map the nodes and branches of the lower and upper trees, or equivalently to give a list of evolutionary events explaining the discrepancies between the upper tree and lower tree. A most parsimonious Duplication and Loss (DL) reconciliation is computed through the Lowest Common Ancestor (LCA) mapping: proceeding from the leaves to the root, each internal node is mapped to the lowest common ancestor of the mapping of its two children.

A Markovian model for reconciliation.

The LCA mapping in the DL model follows a parsimony principle: no event should be invoked if it is not necessary. However the use of this principle is debated[78] and it is commonly admitted that it is more accurate in molecular evolution to fit a probabilistic model as a random walk, which does not necessarily produce parsimonious scenarios. A birth and death Markovian model is such a model that can generate a lower tree "inside" a fixed upper one from root to leaves [7]. Statistical inference provides a framework to find most likely scenarios, and in that case, a maximum likelihood reconciliation of two trees is also a parsimonious one. In addition, it is possible with such a framework to sample scenarios, or integrate over several possible scenarios in order to test different hypotheses, for example to explore the space of lower trees. Moreover probabilistic models can be integrated in larger models as probabilities simply multiply when assuming independence, for instance combining sequence evolution and DL reconciliation [8].

Introducing horizontal transfer.

Host switch, *i.e.* inheritance of a symbiont from a kin lineage, is a crucial event in the evolution of parasitic or symbiotic relationships between species. This horizontal transfer also models migration events in biogeography and became of interest for the reconciliation of gene and species trees when it appeared that many discrepancies could not simply be explained by duplication and loss and that horizontal gene transfer (HGT) was a major evolutionary process in micro-organisms evolution. This switching, or horizontal transfer, pattern can also model admixture or introgression

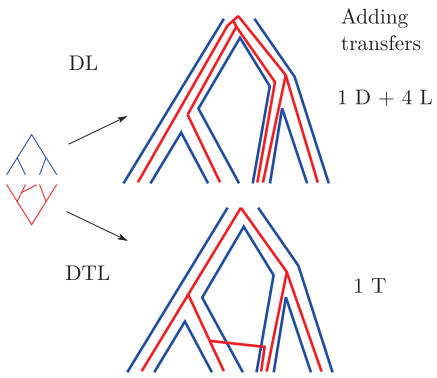


Figure 1.15: Phylogenetic reconciliations in Duplication Loss and Duplication Transfer Loss.

[250]. It is considered in character methods, without information from the symbiont phylogeny [30, 50]. On top of the DL model, horizontal transfer enables new very different reconciliation scenarios (Figure 1.14A).

The simple yet powerful dynamic programming approach

The LCA reconciliation method yields a unique solution, which has been shown to be optimal for the problem of minimizing the weighted number of events, whatever the relative weights of duplication and loss [43]. In contrast, with Duplication, horizontal Transfer and Loss (DTL), there can be several equally parsimonious reconciliations. For instance a succession of duplications and losses can be replaced by a single transfer (Figure 1.14 B). One of the first ideas to define a computational problem and approach a resolution was, in a host/symbiont framework, to maximize the number of co-speciations with a heuristic algorithm [179]. Another solution is to give relative costs to the events and find a scenario that minimizes the sum of the costs of its events [38]. In the probabilistic model frameworks, the equivalent task consists in assigning rates or probabilities to events and search for maximum likelihood scenarios, or sample scenarios according to their likelihood. All these problems are solved with a dynamic programming approach.

Dynamic programming

Dynamic programming is an algorithmic paradigm which goal is to not compute the same thing twice. This technique is often used alongside a datastructure, often a table (possibly multidimensional), and an induction definition for the cells of the table. The answer to the problem is the last cell, which can be computed once all the other cells are computed, and the induction often call to more than one of the other cells to be computed. It is quite useful in cases when we try to construct scenarios. Once the table is filled, we can backtrack and sample a scenario.

This dynamic programming method consists in traversing the two trees in a postorder. Proceeding from the leaves and then going up in the two trees, for each couple of internal nodes (one for each tree), the cost of a most parsimonious DTL reconciliation is computed [38].

In a parsimony framework, costs of reconciling a lower subtree rooted at l with a upper subtree rooted at U is initialized for the leaves with their matching:

$$c(U,l) = 0 \text{ if } l \in U \text{ else } c(U,l) = \infty \tag{1.2}$$

And then inductively, denoting $l', l"$ the children of l, $U', U"$ the children of U, c^S, c^D, c^T, c^L the costs associated to speciation, duplication, horizontal transfer and loss, respectively (with c^S often fixed to 0),

$$c(U,l) = \min \begin{cases} c^S + \min(c(U',l') + c(U",l"), c(U",l') + c(U',l")) \\ c^S + c^L + \min(c(U',l) + c^L, c(U",l) + c^L) \\ c^D + c(U,l') + c(U,l") \\ c^T + \min(\min_V(c(V,l')) + c(U,l"), \min_V(c(V,l")) + c(U,l')) \end{cases} \tag{1.3}$$

The costs $\min_V(c(V,l'))$ and $\min_V(c(V,l"))$, because they do not depend on U, can be computed once for all U, hence achieving quadratic complexity to compute c for all couples of U and l. The cost of losses only appears in association with other events because in parsimony, a loss can always be associated with the preceding event in the tree.

The induction behind the use of dynamic programming is based on always progressing in the trees toward the roots. However some combinations of events that can happen consecutively can make this induction ill-defined. One such combination consists in a transfer followed immediately by a loss in the donor lineage (TL). Restricting the use of this TL event [64] repairs the induction. With an unlimited use it is necessary to use or add other known methods to solve systems of equations like fixed point methods [216], or numerical solving of differential equations [192]. In 2016, only two out of seven of the most commonly used parsimony reconciliation programs did handle TL events [100] although its consideration can drastically change the result of a reconciliation [63].

Unlike LCA mapping, DTL reconciliation typically yields several scenarios of minimal cost, in some cases an exponential number. The strength of the dynamic programming approach is that it enables to compute a minimum cost of coevolution of the input upper and lower tree in quadratic time [12], and to get a most parsimonious scenario through backtracking. It can also be transposed to a probabilistic

framework to compute the likelihood of coevolution and get a most likely reconciliation, replacing costs with rates, minimums by sums and sums by products [214]. Moreover the approach is suitable, through multiple backtracks, to enumerate all parsimonious solutions or to sample scenarios, optimal and sub-optimal, according to their likelihood.

Estimation of event costs and rates.

Dynamic programming *per se* is only a partial solution and does not solve several problems raised by reconciliation. Defining a most parsimonious DTL reconciliation requires giving costs to the different kind of events (D, T and L). Different cost assignations can yield different reconciliation scenarios (Figure 1.14B), so there is a need for a way to choose those costs. There is a diversity of approaches to do so. CoRe-PA [151] explores in a recursive manner the space of cost vectors, searching for a good matching with the event frequencies in reconciliations.

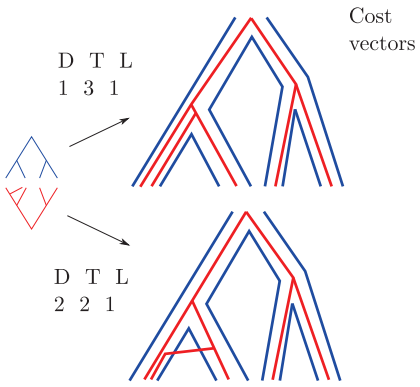


Figure 1.16: Different cost assignments can give different most parsimonious solutions.

ALE [214] uses the same idea in a probabilistic framework to estimate the event rates by maximum likelihood. Alternatively COALA [18] is a pre-process using approximate bayesian computation with sequential Monte Carlo: simulation and statistic rejection or acceptance of parameters with successive refinement.

In the parsimony framework it is also possible to divide the space of possible event costs in areas of costs which lead to the same Pareto optimal solution [125]. Pareto optimal reconciliations are such that no other reconciliation has a strictly inferior cost for one type of event (duplication, transfer or loss), and less or equal for the others.

It is also possible to rely on external considerations in order to choose the event costs. For example the software Angst [54] chooses the costs that minimize the variation of genome size, in number of genes, between parent and children species.

The problem of temporal feasibility.

The dynamic programming method works for dated (internal nodes are totally ordered) or undated upper trees. However with undated trees there is a time feasibility issue. Indeed a horizontal transfer implies that the donor and the receiver are contemporary, therefore implying a time constraint on the tree. In consequence two

horizontal transfers may be incompatible, because they imply contradicting time constraints (Figure 1.14C). The dynamic programming can not easily check for such incompatibilities. If the upper tree is undated, finding a time feasible most parsimonious reconciliation is NP-hard [90, 222, 173]. It is fixed parameter tractable, which means that there are algorithms running in time bounded by an exponential of the number of transfers in the output scenarios [222].

Some solutions imply integer linear programming [238] or branch and bound exploration [237]. If the upper tree is dated, then there is no incompatibility issue because horizontal transfers can be constrained to never go backward in time. Finding a coherent optimal reconciliation is then solved in polynomial time [222], or with a speed-up in RASCAL [65, 66], by testing only a fraction of nodes mapping. Most of the software taking undated trees do not look for temporal feasibility, except Jane [48] which explores the space of total orders via a genetic algorithm, or, in a post process, Notung [71] and Eucalypt [61], which search inside the set of optimal solutions for a time consistent ones. Other methods work as supplementary layers to reconciliations, correcting reconciliations [134] or returning a subset of feasible transfers [44], which can be used to date a species tree [44, 56].

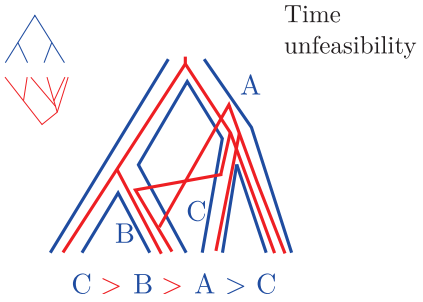


Figure 1.17: Not all scenarios including transfers are time feasible, some might include time constraints incompatible with the species tree.

Expanding phylogenies: Transfers from the dead.

In phylogenetics in general, it is important to keep in mind that the species, extant and ancestral which are represented in any phylogeny are only a sparse sample of the species that currently exist or have existed. This is why one can safely assess that all transfers that can be detected using phylogenetic methods have originated in lineages that are, strictly speaking, absent from a studied phylogeny (Figure 1.14 D) [213]. Accounting for extinct or unsampled biodiversity in phylogenetic studies can give a better understanding of these processes [55]. Originally, DTL reconciliation methods did not recognize this phenomenon and only allowed for transfer between contemporaneous branches of the tree, hence ignoring most plausible solutions. However methods working on undated upper trees can be seen as implicitly handling the unknown diversity by allowing transfers "to the future" from the point of view of one phylogeny, that is, the donor is more ancient than the recipient. A transfer to the future can be translated into a speciation to unknown species, followed by a transfer from unknown species.

ALE [213] in its dated version explicitly takes the unknown diversity into account by adding a Moran process of speciation/extinctions of species to the dated birth/death model of gene evolution. Transfer from the dead are also handled in a parsimonious setting by Tera and eccetera [202, 100], showing that considering these transfers improve the capacity to reconstruct gene trees using reconciliation, and with a more explicit model in [236] and in probabilistic setting, in ALE undated [215].

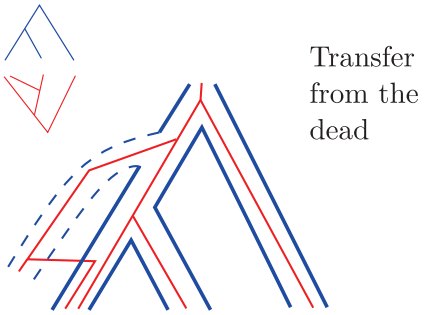


Figure 1.18: Transfer can go from a species to one of its descendant via a sister lineages that went extinct.

The specificity of biogeography: a tree like structure for the "evolution" of areas.

In biogeography, some applications of reconciliation approaches consider as an upper tree an area cladogram with defined ancestral nodes. For instance the root can be Pangea and the nodes contemporary continents. Sometimes internal nodes are not ancestral areas but the unions of the areas of their children, to account for the possibility of species evolving along the lower tree to inhabit one or several areas. In this case, the evolutionary events are migration, where one species colonizes a new area, speciation Allopatric speciation, or vicariance, equivalent to co-speciation in host/symbiont comparisons (Figure 1.14E). Despite this does not always give a tree (if the unions AB and BC of leaves A, B, C exist, a child can have several parents) and this structure is not associated with time (it is possible for a species to go from A to AB by migration, as well as from AB to A by extinction), reconciliation methods, with events and dynamic programming, can infer evolutionary scenarios between this upper geographical structure and lower species tree. Diva [194] and Lagrange [191, 192] are two reconciliation models constructing such a tree-like structure and then applying reconciliation, the first with a parsimony principle, the second in a probabilistic framework. Additionally BioGeoBEARS [146] is a biogeography inference package that reimplemente DIVA and Lagrange models and allows for new options, like distant dependent transfers [227] and discussion on statistical model selection [145].

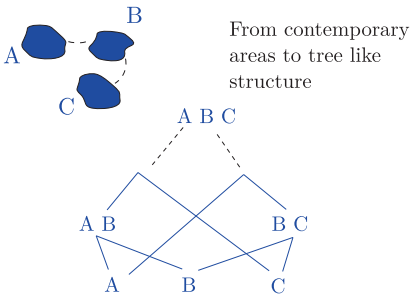


Figure 1.19: In biogeography, a tree like structure can be constructed to account for the possible migrations between different geographical areas.

Graphical output

With two trees and multiple evolutionary events linking them to represent, viewing reconciled trees is a challenging but necessary question in order to make reconciliation studies more accessible. Some reconciliation software include annotation of the evolutionary events on the lower trees [71], while others [48, 201, 61, 151] and specific packages, in DL [206] or DTL[45], trace the lower tree embedded in the upper one. One difficulty in this regard is the variety of output format for the different reconciliation software, however recently a common standard, recphyloxml [70], has been established and endorsed by part of the community with available viewer.

1.4.6 Addressing additional practical considerations

Applying DTL reconciliation to biological data raises several problems related to uncertainty and confidence levels of input and output. Concerning the output, the uncertainty of the answer calls for an exploration of the whole solution space. Concerning the input, phylogenetic reconciliation has to handle uncertainties in the resolution or rooting of the upper or lower trees, or even to propose roots or resolutions according to their confidence.

Exploring the space of reconciliations.

Multiple DTL reconciliation scenarios can have equal cost or tight probabilities (Figure 1.14E). Dynamic programming makes it possible to sample reconciliations, uniformly among optimal ones [13] or according to their likelihood. It is also possible to enumerate them in time proportional to the number of solutions [61], a number which can quickly become intractable (even only for optimal ones) (Figure 1.14F). Finding and presenting structure among the multitude of possible reconciliations has been at the center of recent methodological developments, especially for host and symbiont aimed methods. Several works have focused on representing a set of reconciliations in a compact way, from a uniform sample of optimal ones [13] or by constructing a graph summarizing the optimal solutions [204]. This can be achieved by giving support values to specific events based on all optimal (or sub-optimal) reconciliations [170], or with the use of a consensus reconciled tree [111, 133]. In a DL model it is possible to define a median reconciliation, based on shared events and to compute it in polynomial time [98].

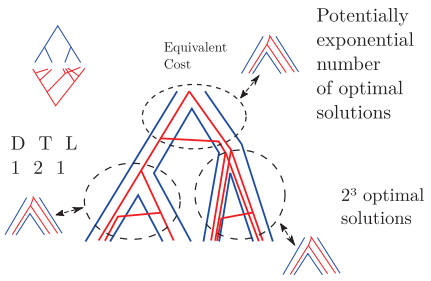


Figure 1.20: An exponential number of scenarios might be most parsimonious, for example when two equivalent patterns have the same cost.

EMPRess [201] can group similar reconciliations through clustering [147], with all pairwise distance between reconciliations computable in polynomial time (independently of the number of most parsimonious reconciliations) [200]. With the same aim, Capybara [232] defines equivalence classes among reconciliations, efficiently computing representative for all classes, and outputs with linear delay a given number of reconciliations (first optimal ones, then sub optimal). The space of most parsimonious reconciliation can be expanded or reduced when increasing or decreasing horizontal transfer allowed distance [61], which is easily done by dynamic programming.

Inferring phylogenetic trees with reconciliation

Reconciliation and input uncertainty Reconciliation works with two fixed trees, a lower and an upper, both assumed correct and rooted. However, those trees are not first hand data. The most frequently used data for phylogenetics consists in aligned nucleotidic or proteic sequences. Extracting DNA, sequencing, assembling and annotating genomes, recognizing homology relationships among genes and producing multiple alignments for phylogenetic reconstruction are all complex processes where errors can ultimately affect the reconstructed tree [25]. Any topology or rooting error can be misinterpreted and cause systematic bias. For instance, in DL reconciliations, errors on the lower tree bias the reconciliation toward more duplication events closer to the root and more losses closer to the leaves [89].

On the other hand, reconciliation, as a macro evolutionary model, can work as a supplementary layer to the micro evolutionary model of sequence evolution, resolving polytomies (nodes with more than two children) or rooting trees, or be intertwined with it through integrative models in order to get better phylogenies.

Most of the works in this direction focus on gene/species reconciliations, nevertheless some first steps have been made in host/symbiont, such as considering unrooted symbiont trees [226] or dealing with polytomies in Jane [48].

Exploring the space of lower trees with reconciliation. Reconciliation can easily take unrooted lower trees as input (Figure 1.14G), which is a frequently used feature because trees inferred from molecular data are typically unrooted. It is possible to test all possible roots, or a thoughtful triple traversal of the unrooted tree allows to do it without additional time complexity [64]. In a duplication-loss model the set of roots minimizing the costs are found close to one another, forming a "plateau", [82] a property which does not generalizes to DTL [226, 111].

Reconciliation can also take as input non binary trees (Figure 1.14H), that is, with internal nodes with more than two children. Such trees can be obtained for example by contracting branches with low statistical support. Inferring a binary

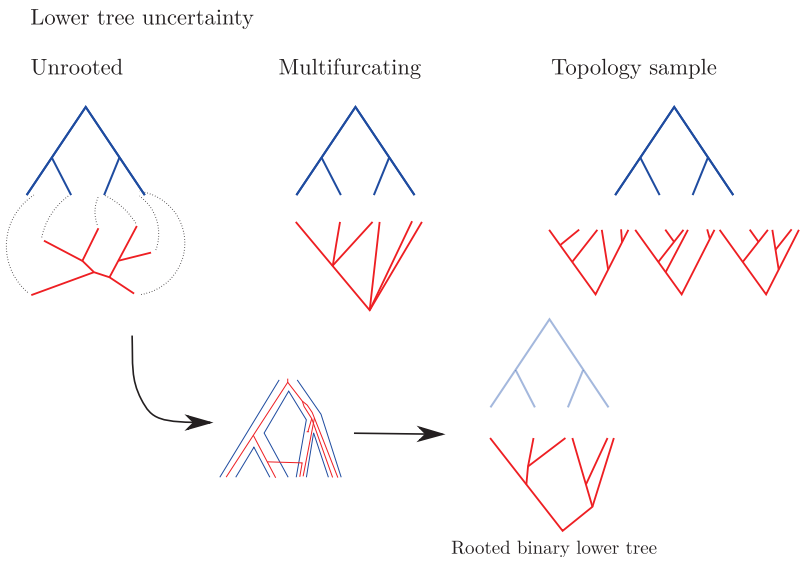


Figure 1.21: The lower tree can be unrooted, multifurcating, or given as a sample of potential trees and reconciliation can be used to resolve those uncertainty to get a binary rooted lower tree.

tree from a non binary tree according to reconciliation scores is solved in DL with efficient methods [71, 211, 113, 47, 254]. In DTL, the problem is NP hard [107]. Heuristics [114] and exact fixed parameter tractable algorithms [107, 106] [101] are possible resolutions.

Another way to handle uncertainty in lower trees is to take as input a sample of alternative lower trees instead of a single one. For example in the paper that gave reconciliation its name [81] it was proposed to consider all most likely lower trees, and choose from these trees the best one according to their DL costs, a principle also used by TreeFix-DTL [17].

The sample of lower trees can also reflect their likelihood according to the aligned sequences (Figure 1.14I), as obtained from bayesian Markov chain Monte Carlo methods as implemented for example in Phylobayes [117]. AngST [54], ALE[216] and EcceTERA [202] use "amalgamation", a extension of the DTL dynamic programming that is able to efficiently traverse a set of alternative lower trees instead of a single tree.

A local search in the space of lower trees guided by a joint likelihood, on the one hand from multiple sequence alignments and on the other hand from reconciliation with the upper tree, is achieved in Phyldog with a DL model [27] and in GeneRax with DTL [158]. In a DL model with sequence evolution and relaxed molecular clock the lower tree space is explored with an MCMC in [4]. MowgliNNI [169] can modify the input gene tree at poorly supported nodes to increase DTL score, similarly TreeSolve resolve the multifurcations added by collapsing poorly supported nodes [108].

Finally, integrative models, mixing sequence evolution and reconciliation, can compute a joint likelihood via dynamic programming (for both reconciliation and gene sequences evolution) [216], use Monte Carlo Markov Chain to include molecular clock to estimate branch lengths, in a DL model [7] or with a relaxed molecular clock [4], and in a DTL model [209]. These models have been applied in gene/species frameworks, not yet in host/symbiont or biogeography.

Inferring upper trees using reconciliation. Inferring an upper tree from a set of lower trees is a long standing question related to the supertree problem[233]. It is particularly interesting in the case of gene/species reconciliation where many (typically thousands of) gene trees are available from complete genome sequences. Supertree methods attempt to assemble a species tree based on sets of trees which may differ in terms of contemporary species sets and topology, but usually without consideration for the biological process explaining these differences. However some supertree approaches are statistically consistent for the reconstruction of the species tree if the gene trees are simulated under a DL model. This means that if the number of input lower trees generated from the true upper tree via the DL model grows toward infinity, given that there are no additional error, the output upper tree converges almost surely to the true one. This has been shown in the case of a quartet distance [120], and with a generalized Robinson Foulds multicopy distance [153], introduced in [42], with better running time but assuming gene trees do not contain bipartitions contradicting the species tree, which seems rare under a DL model.

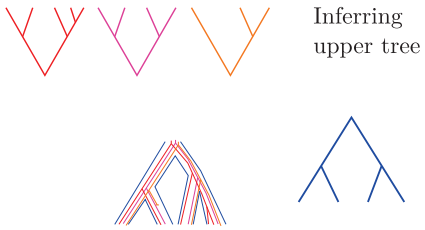


Figure 1.22: Reconciliation score can be used to help construct an upper tree.

However, reconciliation can also be used for the inference of upper tree. It is a computationally hard problem: already resolving polytomies in a non binary upper tree with a binary lower one, minimizing a DL reconciliation score, is NP-hard [253]. In particular, reconstructing the species tree giving the best DL cost for several gene trees is called the Gene Duplication problem or more generally Gene Tree parsimony. The problem was seen as a way to detect paralogy to get better species tree reconstruction [86, 177]. It is NP-hard, with interesting results on the problem complexity [132, 16] (Figure 1.14J) and the behavior of the model with different input size, structure and ILS presence [136]. Multiple solutions exists, with ILP [37] or heuristics [174, 235], and with the possibility of a deep coalescence score [41].

ODTL [214] takes as input gene trees and searches a maximum likelihood species

tree according to a DTL model, with a hill-climbing search. The approach produces a species tree with internal nodes ordered in time ensuring a time compatibility for the scenarios of transfer among lower trees (see paragraph 1.4.5).

Addressing a more general problem, Phyldog [27] searches for the maximum likelihood species tree, gene trees and DL parameters from multiple family alignments via multiple rounds of local search. It thus performs the exploration of both upper and lower trees at the same time. MixTreEM [224] presents a faster solution.

1.4.7 Limits of the two-level DTL model

A limit to dynamic programming: non independent evolution of children lineages.

The dynamic programming framework, like usual birth and death models, works under the hypothesis of independent evolution of children lineages in the lower tree. However this hypothesis does not hold if the model is complemented with several other documented evolutionary events, such as horizontal transfer with replacement of an homologous gene in the recipient lineage, or gene conversion. Horizontal transfer with replacement is usually modeled by a rearrangement of the upper tree, called Subtree Prune and Regraft (SPR) (Figure 1.14 K left). Reconciling under SPR is NP-hard, even in dated trees, and fixed parameter tractable regarding the output size [22, 91].

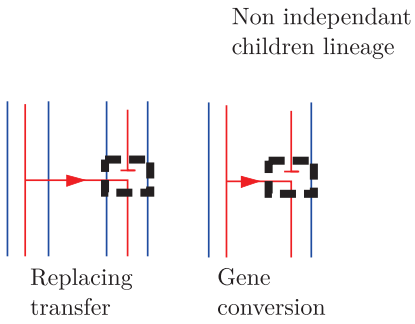


Figure 1.23: Events such as replacing transfer or gene conversion can not be modeled with independent children lineages.

Another way to model and infer replacing horizontal transfers is through maximum agreement forest, where branches are cut in the lower and upper trees in order to get two identical (or statistically indistinguishable [1]) upper and lower forests. The problem is NP-hard [94], but several approximations have been proposed [193]. Replacing transfers can be considered on top of the DL model [109]. In the same vein gene conversion can be seen as a "replacing duplication" (Figure 1.14K right). In this latter case, a polynomial algorithm which does not use dynamic programming and is an extension of the LCA method, can find all optimal solutions including gene conversions [92].

Integrating population levels: failure to diverge and Incomplete Lineage Sorting.

In host/symbiont frameworks, a single symbiont species is sometimes associated to several hosts species. This means that while a speciation or diversification has been observed in the host, the populations are indistinguishable in the symbiont. This is handled for example by additional polytomies in the symbiont tree, possibly leading to intractable inference problems, because polytomies need to be resolved. It is also modeled by an additional evolutionary event "failure to diverge" (Jane [48], Amocoala [225]) (Figure 1.14L). Failure to diverge can be a way to allow "free" host switch in a population, a flow of symbionts between closely related hosts. Following that vision, host switch allowed only for close hosts is considered in [61]. This idea of horizontal flow between close populations can also be applied to gene/species frameworks, with a definition of species based on a gradient of gene flow between populations [141].

Failure to diverge is one way of introducing population dynamics in reconciliation, a framework mainly adapted to the multi-species level, where populations are supposed to be well differentiated. There are other population phenomena that limit this framework, one of them being deep coalescence of lineages, leading to Incomplete Lineage Sorting (ILS), which is not handled by the DTL model [211, 217]. The multi species coalescent is a classic model of alleles evolution along a species tree, with birth of alleles and sorting of alleles at speciations, that takes into account population sizes and naturally encompass ILS [189, 57, 136, 128, 188]. In a reconciliation context, several attempts have been made in order to account for ILS without the complex integration of a population model. For example, ILS can be seen as a possible evolutionary pattern for the gene tree (Figure 1.14M). In that case children lineages are not independent of one another, leading to intractability results. ILS alone can be handled with LCA, but ILS + DL reconciliation is NP hard, even without transfers[23].

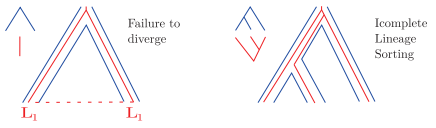


Figure 1.24: Failure to diverge and Incomplete Lineage Sorting are two population level events resulting in a particular reconciliation pattern.

Notung [211] handles ILS by collapsing short branches of the species tree in polytomies and allowing ILS as a free diversification of gene trees on those polytomies. EcceTERA [35] bounds the maximum size of connected parts of the species tree where ILS can happen, proposing a fixed parameter tractable algorithm in that parameter.

ILS and DL can be considered on an upper network instead of tree. This models

in particular introgression, with the possibility to estimate model parameters [67].

More integrative reconciliation models accounting for ILS have been proposed including both DL and multispecies coalescent [190], with DLCoal. It is a probabilistic model with a parsimony translation [244], proposing two sequential LCA-type heuristics handled via an intermediate locus tree between gene and species. However outside of the gene/species reconciliation framework ILS seems, for no particular reason, never considered in host/symbiont, nor in biogeography.

1.4.8 Reconciliation in models with more than two levels

A striking aspect of reconciliation is the common methodology handling different levels of organization: it is used for comparing domain and protein trees, gene and species trees, hosts and symbiont trees, population and geographic trees. However, now that scientists tend to consider that multi-level models of biological functioning bring a novel and game changing view of organisms and their environment [219], the question is how to use reconciliation to bring phylogenetics to this holobiont era (Figure 1.12).

Coevolution of entities at different scales of evolution is at the basis of the holobiont idea: macro-organisms, micro-organisms and their genes all have a different history bound to a common functioning in a single ecosystem. Biological system like the entanglement of host, symbionts and their genes imply functional and evolutionary dependencies between more than two levels.

Examples of multi-level systems

Genes coevolving beyond genome boundaries The holobiont concept [140] stresses the possibility of genes from different genomes to cooperate and coevolve [21, 197, 255]. For instance, certain genes in a symbiont genome may provide a function to its host, like the production of a vital compound absent from available feeding sources. An iconic example is the case for blood-feeding or sap-feeding insects, which often depend on one or several bacterial symbionts to thrive on a resource that is abundant in sugar, but lacks essential amino-acids or vitamins [157]. Another example is the association of Fabaceae with nitrogen-fixing bacteria. The compound beneficiary to the host is typically produced by a set of genes encoded in the symbiont genome, which throughout evolution, may be transferred to other symbionts, and/or in and out of the host genome. Reconciliation methods have the potential to reveal evolutionary links between portions of genomes from different species. A search for coevolving genes beyond the boundaries of the genomes in which they are encoded would highlight the basis for the association of organisms in the holobiont.

Horizontal gene transfer routes depend on multiple levels In intracellular mutualistic symbiont insect systems, multiple occurrence of horizontal gene transfers have been identified, whether from host to symbiont, symbiont to host or symbiont to symbiont [130].

Transfers of endosymbiont genes involved in nutrition pathways beneficiary to the insect host have been shown to occur preferentially if the donor and recipient lineages share the same host [182, 172, 139]. This is also the case in insect with bacterial symbionts providing defensive protein [165] or in obligate leaf nodule bacterial symbionts associated with plants [183]. In the human host, gene transfers has been shown to occur preferentially among symbionts hosted in the same organs [102].

A review on horizontal gene transfers in host/symbiont systems [239] stresses the importance of supporting HGTs with multiple evidence. Notably it is argued that transfers should be considered better supported when involving symbionts sharing a habitat, a geographical area, or a same host. One should however keep in mind that most of the diversity of hosts and symbionts is unknown and that transfers may have occurred in unsampled closely related species, hosts or symbionts.

The idea that gene transfer in symbionts is constrained by the host can also be used to investigate hosts history. For instance, based on phylogeographical studies, it is now accepted that the bacteria *Helicobacter pylori* has been associated with Human populations since the origins of the human species [154, 3]. Analysis of the genomes of *Helicobacter pylori* in Europe suggests that they are issued from a recombination between African and Asian *Helicobacter pylori*. This strongly implies early contacts between the corresponding human populations.

Similarly, an analysis of HGTs in coronaviruses from different mammalian species using reconciliation methods has revealed frequent contact between viruses lineages which can be interpreted as frequent host switches [80].

Cultural evolution The evolution of elements of human culture, for instance languages and folktales, in association with human population genetics, has been studied using concepts from phylogenetics. Although reconciliation has never been used in this framework, some of these studies encompass multiple levels of organization, each represented by a tree or the evolution of a character, with a focus on the coevolution of these levels.

Language trees can be compared with population trees in order to reveal vertically transmitted folktales, via a character model on this language tree [207]. Variants in each folktales family, languages, genetic diversity, populations and geography can be compared two by two, to link folktales diversification with languages on one side and with geography on the other side [198]. As in genetics with symbionts sharing host promoting HGTs, linguistic barriers can foreclose the transmission of

folktales or language elements [24].

Investigating three-level systems using two-level reconciliation

Multi level reconciliation is not as developed as two-level reconciliation. One way to approach the evolutionary dependencies between more than two levels of organization is to try to use available standard two-level methods to give a first insight into biological system's complexity.

Multi-gene events: implicit consideration of an intermediate level. At the gene/species tree level, one typically deals with many different gene trees. In this case, the hypothesis that different gene families evolve independently is made implicitly. However this needs not be the case. For instance, duplication, transfer and loss can occur for segments of a genome spanning an arbitrary number of contiguous genes. It is possible to consider such multi-gene events using an intermediate guide for lower trees inside the upper one. For instance one can compute the joint likelihood of multiple gene tree reconciliations with a dated species tree with duplication, loss and whole genome duplication [257] or in a parsimonious setting [86, 175, 32, 15], and one definition of the problem is NP-hard [77] (Figure 1.25A). Similarly the DL framework can be enriched with duplication and loss of chromosome segments instead of a single gene (Figure 1.25B). However DL reconciliation becomes intractable with that new possibility [62].

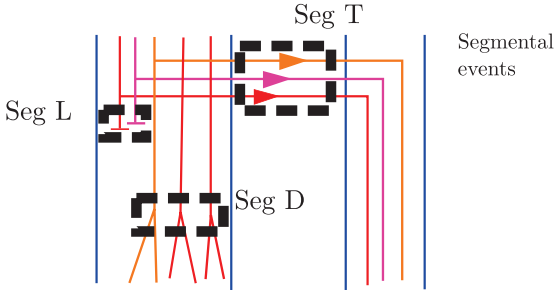


Figure 1.26: Multiple gene lineages can undergo joint events like segmental duplication, transfer or loss, or even whole genome duplication.

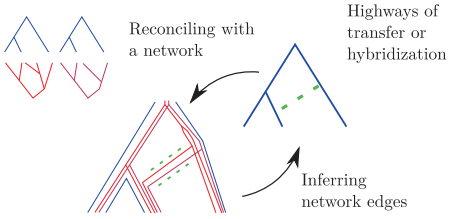


Figure 1.27: Reconciliation can help identify highways of transfers and hybridizations.

The link between two consecutive genes can also be modeled as an evolving character, subject to gain, loss, origination, breakage, duplication and transfer [69]. The evolution of this link appears as an additional level to species and gene trees, partly constrained by the gene/species tree reconciliation, partly evolving on its own, according to genome organization. It thus models the synteny, or proximity between genes. At another scale it can as well model the evolution of the belonging of two domains to a protein.

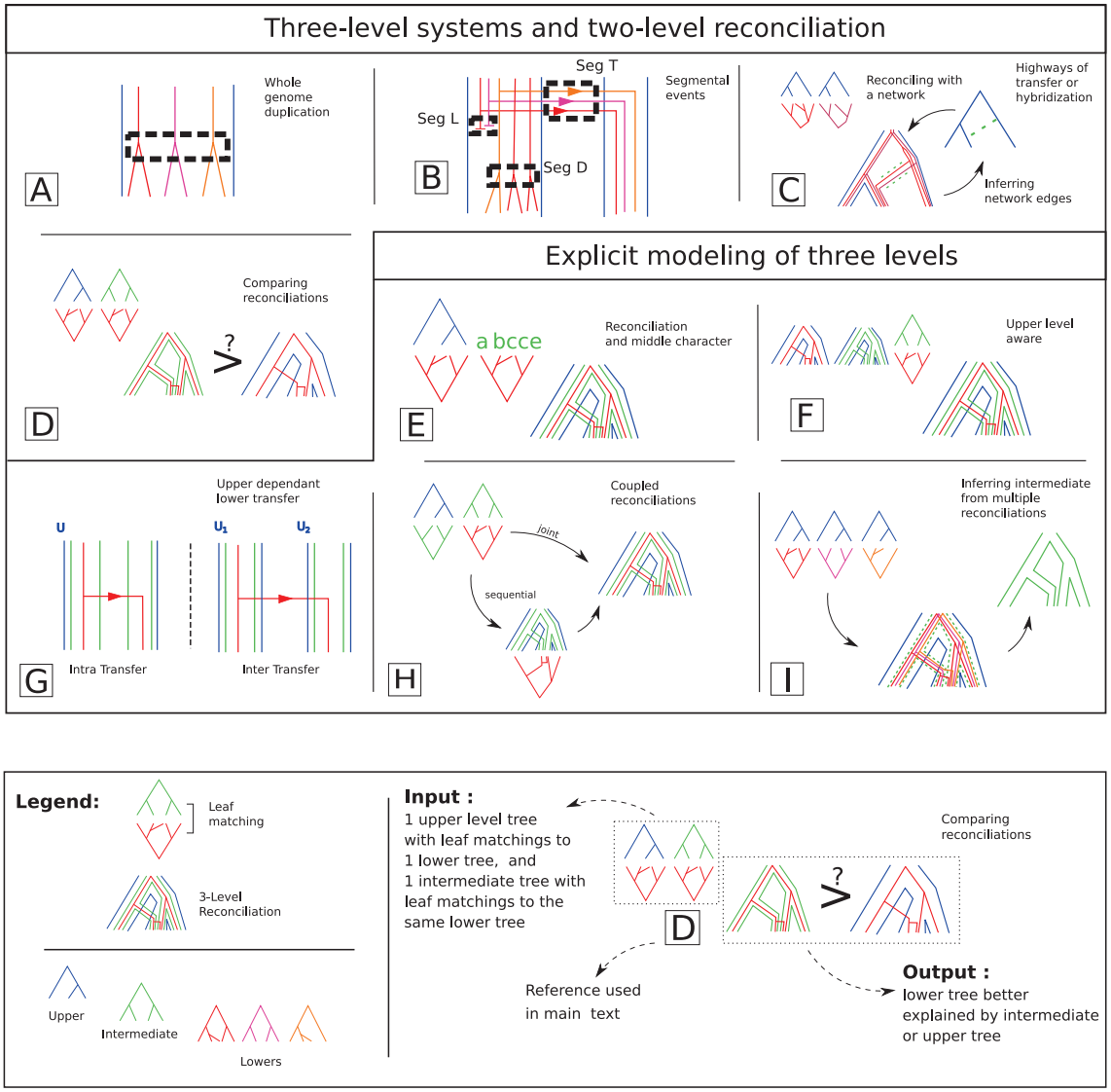


Figure 1.25: Illustration of input, output and events, of published methods which can be identified with 3-level methods. The formalism is similar to the one on Figure 1.14. Multiple gene lineages can undergo joint events like whole genome duplication (A) or segmental events (B), some events might be more probable than others, like specific horizontal transfers with highway of transfers or hybridization (C). Cophylogenetic patterns can be compared, to see for instance if the common pattern of a host and a symbiont are not just the common pattern of the symbiont and the geography (D). Characters can evolve on reconciled phylogeny, like gene synteny (E), or two levels can be reconciled with the constraint of an upper one (F). Transfers can be upper dependent, more likely between two intermediate entities that belong to a same upper one (G). Three levels can be reconciled together, sequentially, the intermediate in the upper before adding the lower, or trying to find a joint most parsimonious scenario for the two reconciliations (H). These multi-level models can also be used to reconstruct the intermediate phylogeny (I).

The detection of "highways of transfers", the preferential acquisition of groups of genes from a specific donor, is another example of non-independence of gene histories [14], similarly multi-gene transfers can be detected [105]. It has also lead to methodological developments such as reconciliations using phylogenetic networks, seen as a tree augmented with transfers edges, which can be used to constrain transfers in a DTL model [203]. Networks can also be used to model introgression and Incomplete Lineage Sorting [251, 249, 250] (Figure 1.25C).

Detecting coevolution in multiple pairs of levels. It is a central question to understand the evolution of an holobiont to know what are the levels that coevolve with each others, for instance between host species, host genes, symbionts and symbiont genes. It is possible to approach the multiple inter-dependencies between all levels of evolution by multiple pairwise comparisons of two evolving entities.

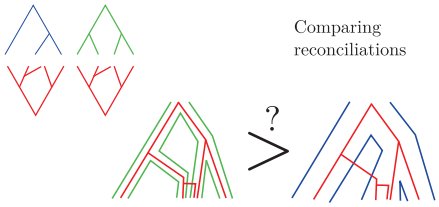


Figure 1.28: With more than two levels the reconciliation of the lower and intermediate levels can be compared to the reconciliation of the lower and upper.

Reconciliation of host and symbiont on one side and geography and symbiont on the other side, can also help to identify patterns of diversification of host and symbiont that reflect coevolution on one side, and patterns that can be explained by a common geographical diversification on the other [171, 144, 234, 79] (Figure 1.25D). Similarly, a study used reconciliation methods to differentiate the effect of diet evolution and Phylogenetic inertia on the composition of mammalian gut microbiomes. By reconstructing ancestral diets and microbiome composition onto a mammalian phylogeny, the study revealed that both effects contribute but at different time scales [85].

Explicit modeling of three or more levels

In a model of a multi-level system as host/symbiont/genes, horizontal gene transfers should be more likely between two symbionts of a same host. This is invisible to a two-level gene tree/species tree or host/symbiont reconciliation: in some cases looking at any combination of two levels can lead to miss an evolutionary scenario which can only be the most likely if the information from the three trees are considered together (Figure 1.29).

Trying to face the limitation of these use of standard two-level reconciliations with systems involving inter-dependencies at multiple levels, a methodological effort has been done in the last decade to construct and use multi-level models. It requires the identification of at least one "intermediate" level between the upper and the

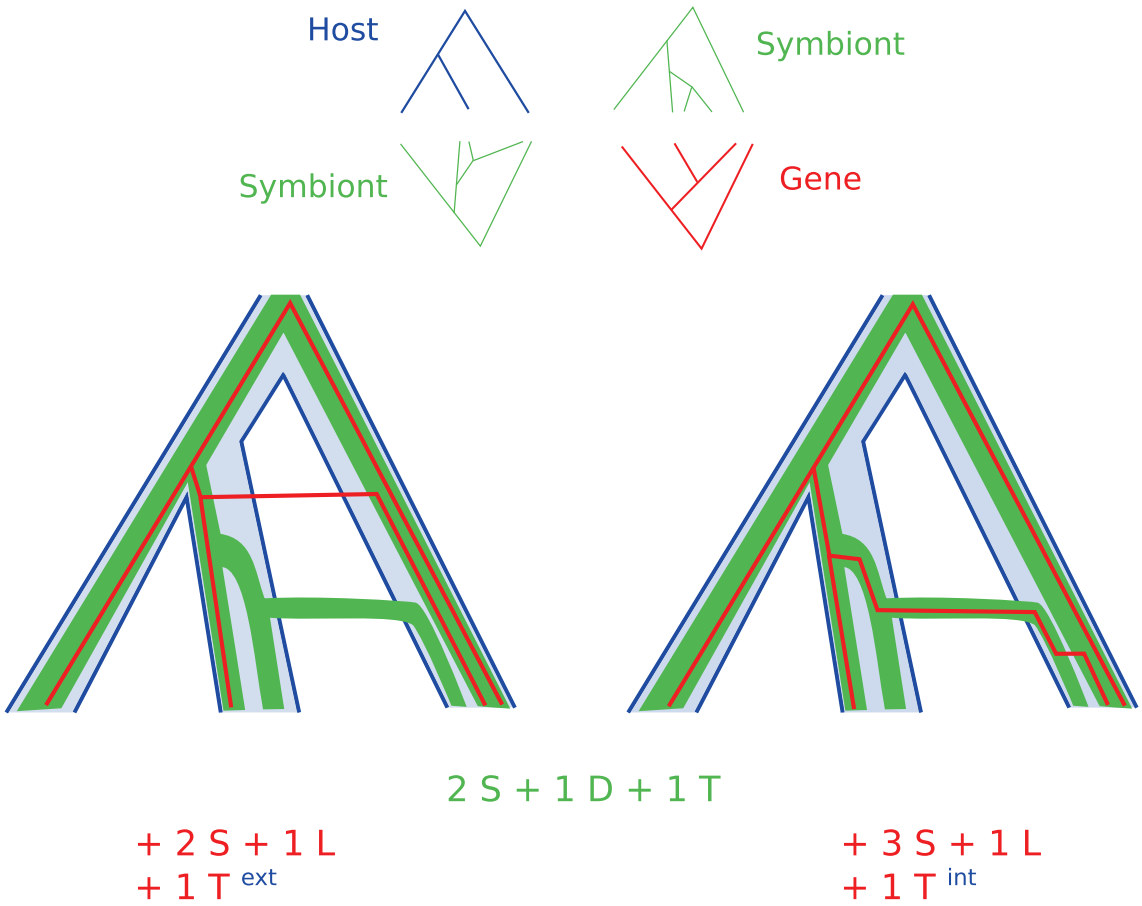


Figure 1.29: Higher level of organization can shed light on lower levels reconciliation. In this example, the goal is to reconstruct the history of a gene present in a symbiont genome. A single transfer and a single loss of gene is the most parsimonious scenario for the reconciliation of the gene tree with either the host or the symbiont tree. Yet when considering the reconciliation of the symbiont and host trees, this scenario implies a gene transfer between two symbionts across branches of the host tree (left). Such an inter-host transfer should be considered unlikely because a series of hidden events are necessary for the gene to come in contact with its next recipient symbiont. Considering the three levels together puts forward a new scenario without inter-host transfer (right) which is slightly less parsimonious in two-level reconciliations, but implies a more likely event of gene transfer within host.

lower one.

Pre-reconciliation: characters onto reconciled trees. A first step towards integrated three levels model is to consider phylogenetic trees at two levels and another level represented only with characters at the leaves of one of the trees (Figure 1.25E). For instance a reconciliation of host and symbiont phylogenies can be informed by geographic data [19]. Ancestral geographic locations of host and symbiont species obtained through a character inference method can then be used to constraint the host/symbiont reconciliation: ancestral hosts and symbionts can only be associated if they belong to the same geographical location (Figure 1.25F).

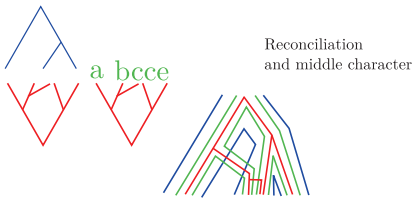


Figure 1.30: Characters can evolve on reconciled phylogenies, like gene synteny on a gene/species reconciliation.

At another scale the evolution at the subgene level can be approached with a character method [245]. Here, parts of genes (e.g. the sequence coding for protein domains) is reconciled according to a DL model with a species tree, and the genes they belong to are mentioned as characters of these parts. Ancestral genes are then reconstructed a posteriori via merge and splits of gene parts.

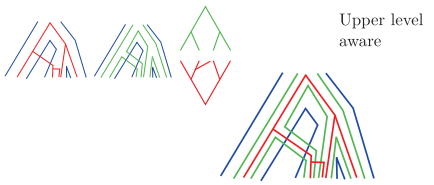


Figure 1.31: Two levels can be reconciled with the constraint of an upper one, for instance host and symbiont with geography.

Two-level reconciliations informed by a third level. As pointed by several studies (see paragraph 1.4.8), an upper level can inform a reconciliation between an intermediate and lower one, notably for horizontal transfers. Three level models can take into account these assumptions to guide reconciliations between an intermediate and lower trees with the knowledge of an upper tree. The model can for example give higher likelihoods to reconciliation scenarios where horizontal gene transfers happen between entities sharing the same habitat. It has been achieved for the first time with DTL gene/species reconciliations nested with a DTL gene domain and gene reconciliation [212]. Different costs for inter and intra transfers depend on whether or not transfers happen between genes of the same genomes (Figure 1.25G,H sequential).

Note that this model explicitly considers three levels and three trees, but does not yet define a real three level reconciliation, with a likelihood or score associated [212]. It relies on a sequential operation, where the second reconciliation is informed

by the result of the first one.

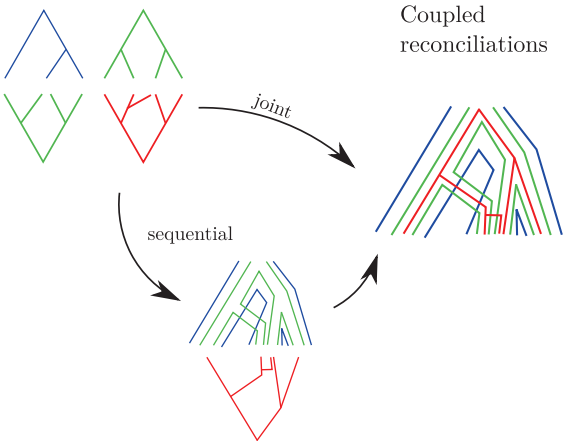


Figure 1.32: Three levels can be reconciled together, sequentially, the intermediate in the upper before adding the lower, or trying to find a joint most parsimonious scenario for the two reconciliations.

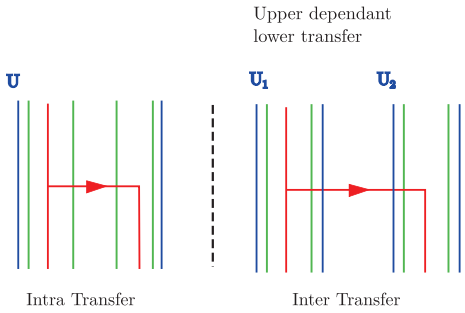


Figure 1.33: Transfers can be upper dependant, more likely between two intermediate entities that belong to a same upper one.

The reconciliation problem in multi-level models. The next step is to define the score of a reconciliation consisting of three nested trees and to compute, given the three trees, three-level reconciliations according to their score. It has been achieved with a species/gene/domain system, where genes evolve within the species tree with a DL model and domains evolve within the gene/species system with a DTL model, forbidding domain transfers between genes of two different species (Figure 1.25G) [122]. Inference involves candidate scenarios with joint scores (Figure 1.25H joint). Computing the minimum score scenario is NP-hard, but dynamic programming or integer linear programming can offer heuristics [122, 123]. Variation of the problem when multiple domains are considered [124] and a simulation framework [112] is available.

Inferring the intermediate tree using model of 3-level lower/intermediate/upper reconciliation. Just like two-level reconciliation can be used to improve lower or upper phylogenies, or to help constructing them from aligned sequences, joint reconciliation models can be used in the same manner. In this vein a coupled gene/species DL, domain gene DL and gene sequence evolution model in a bayesian framework improves the reconstruction of gene trees [161] (Figure 1.25 I).

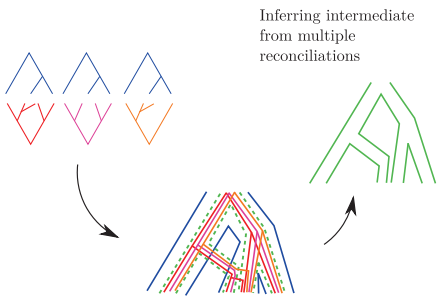


Figure 1.34: With 3-levels reconciliation models, the intermediate tree can be inferred from the lower and upper trees.

1.4.9 Software

Multiple software have been developed to implement the various models of reconciliation. The following table does not aim for exhaustivity but present a consequent number of software aimed at reconciling trees to infer reconciliation scenarios or for other usage such as correcting or inferring trees, or testing coevolution. The levels of interest section detail the levels for which the software was implemented, even though it is entirely possible, for instance, to use a software made for species and gene reconciliation to reconcile host and symbionts [10]. Parsimony or probability is the underlying model that is used for the reconciliation.

1.4.10 Future directions

Reconciliation is now mature as a methodological research subject, a network of researchers and labs working together is emerging, with an active research, a good diversity of available software, and cooperative initiatives like RecPhyloXML, a common standard of output of reconciliations [70]. In the future methodological advances which sustain the development of new models will certainly play an important part in the possibilities of studies surrounding reconciliations. Notably, new approaches may depart from the dynamic programming solution for DTL which progresses along a rather narrow road: almost each new constraint or event on top of it yields intractability results.

In this article we progressed from two to three embedded trees, and there is potentially an infinity of interacting and coevolving levels to study (see four levels examples in [52, 207, 198, 139, 183, 11]). Current quantitative methods obviously cannot yet handle such a complexity. In order to compare hypotheses, and assess them in a statistically grounded framework, they are still to be developed and generalized to help the understanding of multi-level evolving systems, including protein domains, genes, protein complexes, micro and macro organisms, and their ecology.

We showed that there have been multiple first steps in the modeling and methods for the embedding of three trees with lower/intermediate and intermediate/upper reconciliations. Methodological efforts could propose new hints for a joint optimization with horizontal transfers for each levels, and moreover offer a probabilistic framework.

Three level reconciliations have only been applied to domain/gene/species combinations while they could handle the classic holobiontic combination gene/symbiont/host. Models could allow the identification of the coevolving entities inside an ecosystem or a holobiont. For example, the parts of a symbiont tree which follow its hosts, while other parts escape this host but follow geography. Or, at another level, the parts of gene trees evolving with symbiont genomes, and the parts evolving with

Name	Levels of interest	Platform	Command line or Graphical User Interface	Usage	Probability or parsimony	Software License
Diva (https://sourceforge.net/projects/diva/)	Geography and species	Unix, Mac, Win	Command line	Reconciliation inference	Parsimony	GNU GPLv2
Lagrange	Geography and species	Linux, Mac, Win		Reconciliation inference	Probability	GNU GPL 2
BioGeoBEARS (http://phylo.wikidot.com/biogeobears)	Geography and species	R package	R package	Reconciliation inference, statistical model test	Probability	GPL-2 GPL-3
Jane (https://www.cs.hmc.edu/~hadas/jane/)	Host and symbionts	Unix, Mac, Win	GUI or command line	Reconciliation inference, tree uncertainty	Parsimony	Proprietary, registration to download
eMPRess (https://sites.google.com/g.hmc.edu/empress/home)	Host and symbionts	Unix, Mac, Win	GUI or command line	Reconciliation inference, costs estimation, solution space study	Parsimony	GPL 3
Eucalypt (https://team.inria.fr/erable/en/software/eucalypt/)	Host and symbionts		Command line, graphical output with included viewer CophyTrees	Reconciliation inference, solution space study	Parsimony	Cecill
Capybara (https://capybara-doc.readthedocs.io/)	Host and symbionts	Linux, Mac, Win, and python package	GUI, python package	Reconciliation inference, solution space study	Parsimony	code available on github (https://github.com/Helio-Wang/capybara)
CoRe-PA (http://pacosy.informatik.uni-leipzig.de/49-1-CoRe-PA.html)	Host and symbiont	Linux, Mac, Win	GUI, Command line, graphical svg output	Reconciliation inference, cost estimation, dated tree, statistical test	Parsimony	-
CoRe-ILP (http://pacosy.informatik.uni-leipzig.de/217-0-CoRe-ILP.html)	Host and symbiont	Linux, Mac, Win	Command line	Reconciliation inference, temporal feasibility, dated tree	Parsimony	require IBM ILOG CPLEX Optimizer (academic license can be obtained for free)
Rascal	Host and symbiont	-	-	Reconciliation inference, dated tree	Parsimony	-
Treemap (https://sites.google.com/site/cophylogeny/treemap/)	Host and symbiont	Java	Command line	Reconciliation inference, statistical test	Parsimony	-
Mowgli (http://www.atgc-montpellier.fr/Mowgli/)	Species and genes, Host and symbiont	Linux, Mac	Command line, graphical output with compatible viewer [www.sylvx.org Sylvx]	Reconciliation inference, tree uncertainty, ILS, geographical constraints, dated tree input	Parsimony	-
RANGER-DTL (https://compbio.engr.uconn.edu/software/ranger-dtl/)	Species and genes	Linux, Mac, Win	Command line	Reconciliation inference, tree uncertainty, solution sampling, replacing transfers	Parsimony	GPL 3
Notung (http://www.cs.cmu.edu/~durand/Notung/)	Species and genes	Linux, Mac, Win < 7	GUI	Reconciliation inference, tree uncertainty, gene, gene domain, species model	Parsimony	Proprietary
AnGST (https://web.mit.edu/almlab/angst.html)	Species and genes	Python 2	Command line	Reconciliation inference, cost estimation, dated tree input, tree uncertainty	Parsimony	code available on github (https://github.com/almlab/angst)
ecceTERA (https://github.com/celinescornavacca/ecceTERA)	Species and genes	Linux, Mac, built from code	Command line, compatible with Sylvx viewer, and recphyloxml output	Reconciliation inference, cost estimation, dated, partially dated or undated species tree input, tree uncertainty, reconciliation space study, species network	Parsimony	Cecill
ALE (https://github.com/ssolo/ALE)	Species and genes	Linux, Mac	Command line	Reconciliation inference, cost estimation, dated or undated species tree input, tree uncertainty	Probability	GPL 3
PrIME (http://prime.scilifelab.se/)	Species and gene	Linux, Mac	Command line, graphical output (PrIMETV)	Reconciliation inference, gene and species tree inference from reconciliation and aligned sequences, orthology analysis	Probability	(Open source)
JPrIME (https://github.com/arvestad/jprime)	Gene and species	Java lib	Command line	Reconciliation inference, gene and species tree inference from reconciliation and aligned sequences	Probability	New BSD

Figure 1.35: Reconciliation software that aim at inferring reconciliation scenarios.

Name	Levels of interest	Platform	Command line or Graphical User Interface	Usage	Probability or parsimony	Software License
iGTP (https://genome.cs.iastate.edu/igtp/home)	Species and genes	Linux, Mac, Win	GUI	Gene tree correction in DL or deep coalescence	Parsimony	Source code on request
TreeSolve (https://combbio.engr.uconn.edu/software/treesolve/)	Species and genes	Linux, Win	GUI	Gene tree correction in DTL	Parsimony	Source code on request
TreeFix (https://www.cs.hmc.edu/~yjw/software/treefix/), TreeFix-DTL (https://www.cs.hmc.edu/~yjw/software/treefix-dtl/)	Species and genes	Linux	Command line	Gene tree correction in DL and DTL	Parsimony	GNU GPLv3
Treerecs (https://project.inria.fr/treerecs/)	Species and genes	Linux, Mac, Win	GUI, integrated to Seaview (http://doua.prabi.fr/software/seaview)	DL tree correction	Parsimony, Probability	GNU Affero GPL Version 3.0-or-later
Phyldog (https://pbil.univ-lyon1.fr/software/phyldog/)	Species and genes	Linux, docker, vm	Command line	Gene and species tree inference from reconciliation and aligned sequences	Probability	Cecill
MixTreEM (http://prime.scilifelab.se/mixtreem/index.html)	Species and gene	Linux, Mac, Win (build from source)	Command line	Gene and species tree inference from reconciliation and aligned sequences	Probability	(Open source)
GeneRax (https://github.com/BenoitMorel/GeneRax)	Species and genes	Linux, Mac	Command line, graphical output with recphyloxml and thirdkind	Gene tree inference from reconciliation and aligned sequences, species tree inference	Probability	GNU Affero GPL v3.0
Coala (https://team.inria.fr/erable/en/software/coala/)	Host and symbionts	Linux, Mac	Command line	Costs estimation	Parsimony	Cecill
Sylvx (http://www.sylvx.org/)	Species and genes, Host and symbiont	Linux, Mac, Win	GUI	Viewer, compatible with Mowgli, ecceTERA	-	-
Thirdkind (https://github.com/simonpenel/thirdkind/wiki)	Species and genes	Linux, Mac, Win	Command line	Viewer, compatible with recphyloxml	-	Cecill
ARTra (https://combbio.engr.uconn.edu/software/ARTra/)	Species and genes	Linux, Mac	Command line	Additive and replacing transfers inference	Parsimony	GNU GPL
DLCoal (http://combbio.mit.edu/dlcoal/)	Species and genes	Linux, Mac, Win	Command line	Reconciliation inference with ILS	Parsimony	GNU GPL
SEADOG (https://combbio.engr.uconn.edu/software/seadog/)	Species and genes and domains	Linux, Mac	Command line	3-level reconciliation inference	Parsimony	GNU GPL version 3

Figure 1.36: Reconciliation software which primary goal is not to infer reconciliation scenarios. Most of them are used for tree correction using reconciliation score, some are used for rates inference or graphical visualization of scenarios.

hosts, indicating at which level they are selected.

SECTION 1.5

An undated probabilistic model of reconciliation

In this section, I will take an in-depth look at ALE undated, a probabilistic method for inferring DTL reconciliation. ALE stands for Amalgamated Likelihood Estimation, which makes prominent two parts of ALE. It is a probabilistic framework and, as such, can be seen as a successor to previous probabilistic DTL reconciliation models [221]. It uses amalgamation, which is a way to account for uncertainty in a gene tree topology by using a sample of topologies for each gene tree. Amalgamation is used in some parsimony frameworks, like Angst [54] or ecceTera [100]. ALE was developed by Gergely Szöllősi at the LBBE, Lyon, around ten years ago, with the people I am working with now. The implementation of ALE is available on Github, and it is that implementation I will refer to in this section.

ALE exists in two versions, an undated and a dated one, I will introduce both here, but I will only use the undated one in our extensions.

A speed-up version of ALE undated is used in Generax [158] and Speciesrax [159], two efficient methods using a shared model that mixes gene and species coevolution via reconciliation and gene tree inference from aligned sequences, to give better gene trees (Generax) or species trees (Speciesrax).

1.5.1 An undated model of evolution

Dated models are the more intuitive and natural ones, and they seem reasonable, considering the states at each time in a forward or backward manner. However, we can also consider undated models. The interest of undated models is twofold. First, they do not require reliable dated data, which can be challenging to obtain. Second, they are often faster and simpler computational methods, as they do not account for branch length or time. The downside is that their underlying model is less realistic, and they can easily be shown to produce biased results compared to the more realistic dated models. For instance, the models underlying the parsimony framework of phylogeny inference to compute the cost of a tree given an alignment are undated, as they do not consider branch lengths, but they were shown to be inconsistent[78].

ALE undated uses a probabilistic undated model, which is both different from parsimony methods and from its dated version. In this subsection, we describe the dated and undated models, as well as the simpler birth death model to generate a tree. These models are presented in figure 1.37. We will call scenario an instance

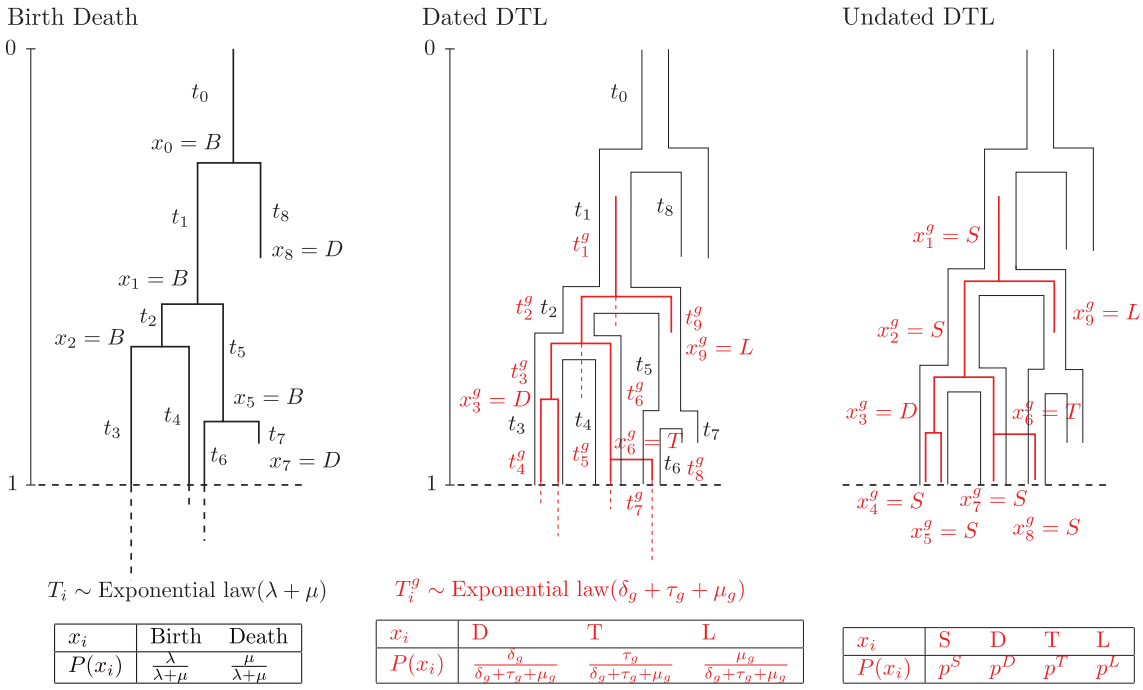


Figure 1.37: The Birth-Death, dated and undated DTL models. The DTL models take the tree generated by a birth-death process as their species tree. In birth-death and DTL dated models, we draw time to the next event from an exponential distribution (t_i), and then we choose an event among the possible ones (birth and death, or duplication, transfer, and loss) (x_i). When the time drawn exceeds specific limits, specific events occur. If we go past the fixed time of simulation in the birth-death process, we stop that lineage's evolution with an extant leaf. If we go past the species diversification time in the DTL, the gene cospeciate with the species. In the undated model, there are no times, only events are drawn, and we reach a cospeciation when a speciation event is drawn, same for reaching the end of the simulation (also by drawing a speciation).

of these models. With the reconciliation model, multiple scenarios can generate the same tree. A scenario is thus a list of events.

Birth death model

The model underlying dated ALE is a classic model for generating trees, the birth-death process. It generates the evolution of dated trees using continuous-time Markov chains. Here we present its fixed-rate version.

The model has two parameters, λ and μ the per lineage rates of birth and death respectively. For one lineage, the birth death model is a Poisson process. The waiting time to the next event is drawn from an exponential distribution of rate $\lambda+\mu$, so of density function $f(t) = \begin{cases} (\lambda+\mu)e^{-(\lambda+\mu)t} \text{ if } t \geq 0 \\ 0 \text{ if } t < 0 \end{cases}$, of cumulative distribution

function $F(t) = \begin{cases} 1 - e^{-(\lambda+\mu)t} \text{ if } t \geq 0 \\ 0 \text{ if } t < 0 \end{cases}$ and expectancy $E(T) = \frac{1}{\lambda+\mu}$.

The dated tree is constructed in a forward manner, as follow: if we are at time t_0 and a time t is drawn from the exponential distribution, then we get to time $t_0 + t$ and choose between birth and death using the rates: $\frac{\lambda}{\lambda+\mu}$ for birth probability, $\frac{\mu}{\lambda+\mu}$ for death probability. If death is chosen, the lineage stops. If birth is chosen, the lineage splits into two, and each lineage is continued using an independent birth-death process with the same rates for the rest of the time. We stop the generation of a lineage when the next time is beyond a fixed time chosen at the beginning. The resulting trees are binary, rooted, and dated.

Forward and backward

In phylogenetics, we will often denote models to be forward or backward. Forward models, such as the birth-death model, generate trees from the roots to the leaves, while backward models start from the leaves and construct the tree from these leaves. Forward models are often more intuitive and seem to be more realistic as happening in the same direction than time, while backward models are simpler to use for inference and straightforward to condition on the leaves. An interesting question is to show the equivalence or interchangeability between a forward and a backward model [141].

DTL dated model

The dated model used in ALE is named ODT (Origination, Duplication, Transfer, and loss) and is described in [214]. It is an adaptation of the birth-death model to the case of a gene tree evolving inside a given species tree. In the remaining part of this section I will use a gene and species vocabulary, though the model can as well describe the coevolution of host and symbiont as it was used for in [10].

At a time t_0, a gene lineage is associated with a species branch. It undergoes an event if the time to the next event is less than the time of speciation of the species branch it belongs to. The rate is $\mu+\delta+\tau$ the rates corresponding to loss, duplication and transfer. As for the birth-death model, the three events are chosen depending on their relative rates. If the time to the next event is more than the time of the speciation, the gene diversifies, and we restart the process at the start of this new branch. If the species is a leaf, we then stop the process. If the next event's time is before the species diversification, an event is chosen depending on its rate. If a loss is chosen, the gene lineage is lost. The two other events make the gene lineage diversifies. If a duplication is chosen, we get two new copies in the same species as their parent. If a transfer is chosen, we then chose uniformly among the species

branches alive at the time one receiver species. The gene diversify and one of the two children is transferred to that receiver while the other stay in the donor species. We then continue the process independently for each child lineage.

For the start of the process, the gene tree can originate anywhere in the species tree with different possible distributions for this origination. One is to choose a time uniformly, then a branch uniformly at that time, another is to choose a branch uniformly (depending on their lengths). Bias can be added toward an origination at the root.

As for the birth-death model, outcome is a rooted dated binary tree; however, this time, there is one more rate, and a species tree guides the evolution.

This model is the one for which ALE dated computes the likelihood, and is the one we use for simulation, in our framework, and a similar one is implemented in the simulation framework we used for our 3-level studies, Sagephy [112]. It is similar to simulation models discussed, for instance, in [64].

DTL undated model

From this part on, we will use the same notation as in the undated pdf on ALE github, which were then used in the Generax paper [158].

This part describes the model underlying ALE undated inference equations, even though this model itself is not presented in the accompanying literature.

The gene tree generated by this model is a rooted binary tree. The model takes as parameters event probabilities (more so than rates) and the species rooted undated phylogeny. We denote these probabilities p^S, p^D, p^T, p^L, and they are such that $p^S + p^D + p^T + p^L = 1$.

- p^S is the probability of undergoing speciation if the gene is in an ancestral species, the children of the gene are then matched to the children of the species or ending the evolution of the gene lineage if the gene is present in an extant species.

- p^D is the probability of undergoing a duplication in that species.

- p^L is the probability of being lost in that species.

- p^T is the probability of being horizontally transferred to another species that is not an ancestor of the current one. The receiver species is chosen uniformly among the possible ones.

1.5.2 Computing likelihood: the equations

Here we present the equations underlying the undated ALE model. As far as I know, there are no proof that these equations correspond to the likelihood according to the undated model we described. These equations are classic dynamic programming ones, and as it is our usual way to compute reconciliation, they seem natural. For proof of an analog dynamic programming system, you can see the Felsenstein tree likelihood for a site computation[78], or the dated DTL model, in ALE [214], or in a previous dated reconciliation model [221]. The equations are written on ALE github, and they are now presented in Generax paper[158].

The data in our model is a set of gene trees. Gene trees are undated binary rooted trees with leaves labeled by one of the leaves of the species tree. Each gene leaf is matched to a single leaf of the species tree, but the leaf of the species tree can be linked to multiple gene leaves in one family (in case of duplicative events) or none (in case of losses).

Gene trees are supposed to be independent once conditioned on the species tree, *i.e.* all their dependence is contained in the species tree. Thus, the likelihood of the species tree given multiple gene family trees is simply the product of the likelihood of the species tree knowing each of the families: $L_{G^*}(S) = P(G^*|S) = \prod_{G \in G^*} P(G|S)$ with S the species tree and G^* the set of gene family trees.

The likelihood of the species tree given a gene tree corresponds to the sum of the probabilities of all possible scenarios that generate this gene tree inside that species tree: $P(G|S) = \sum_{r \in R(G,S)} P(r)$ with $R(G,S)$ the set of reconciliation scenarios generating the gene tree G in the species tree S. The probability of a reconciliation scenario is simply the product of the probabilities of all events in the scenario. We cannot compute the likelihood sum directly, but we can use dynamic programming to compute it efficiently. The probabilities $P(G|S)$ sum to 1 over all possible gene trees.

We will use the following notations. $P_{e,u} = P(u \in e)$ is the probability that the gene tree node u is matched to the branch e of the species tree. It corresponds to the reconciliation of the subtree rooted at u of the gene tree with the subtree rooted at e of the species tree, a reconciliation with origination at the root. E_e is the extinction probability in branch e. S is the species tree.
Events rates : $p^S + p^D + p^T + p^L = 1$

Figure 1.38: The input.

With f, g children of e, v, w children of u, S set of species tree nodes.

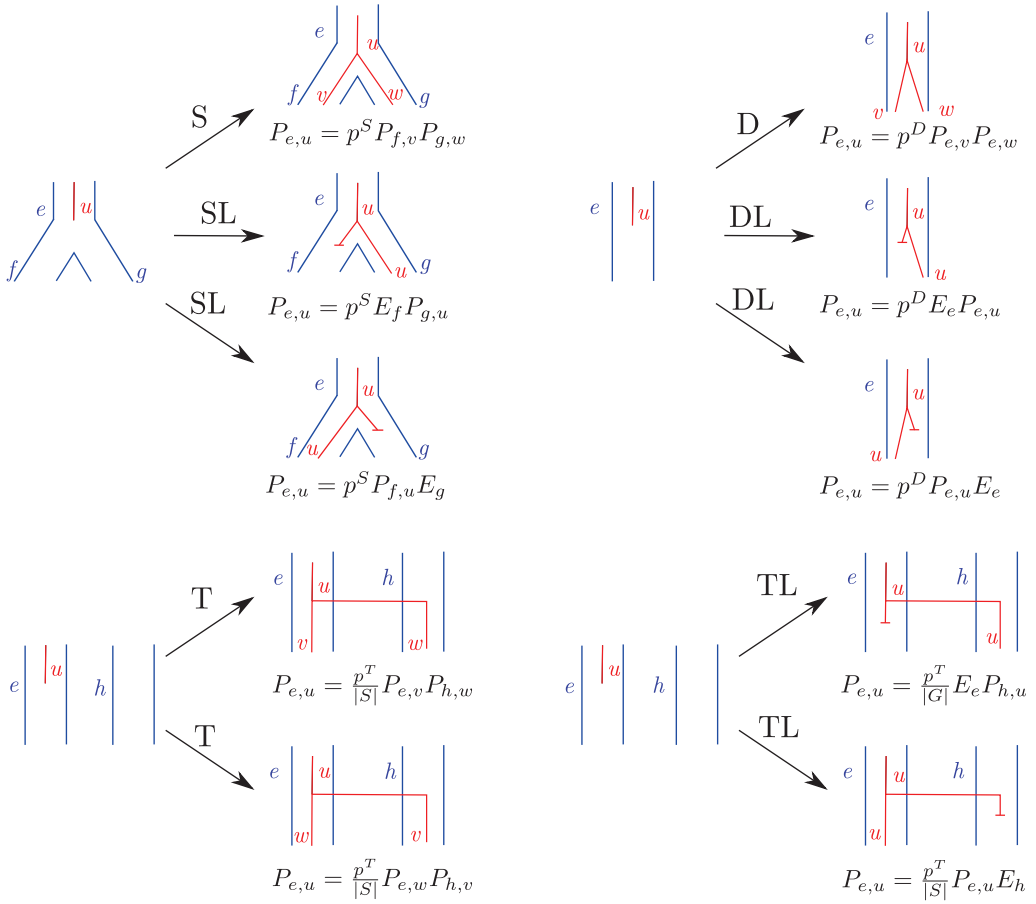


Figure 1.39: The different possibilities of coevolution for a gene u in a species e in the undated DTL model and the probabilities for each.

The equation of undated ALE are:

$$E_e = p^L + p^S E_f E_g + p^D E_e E_e + \frac{1}{|S|} \left(\sum_{h \in S} p^T E_h \right) E_e \tag{1.4}$$

$$\begin{aligned} P_{e,u} = {} & p^S \left(P_{g,v} P_{f,w} + P_{g,w} P_{f,v} + E_f P_{g,u} + P_{f,u} E_g \right) \\ & + p^D \left(P_{e,v} P_{e,w} + 2 P_{e,u} E_e \right) \\ & + \frac{1}{|H|} \left(\sum_{h \in S} p^T P_{h,w} \right) P_{e,v} + \frac{1}{|H|} \left(\sum_{h \in S} p^T P_{h,v} \right) P_{e,w} \\ & + \frac{1}{|H|} \left(\sum_{h \in S} p^T E_h \right) P_{e,u} + \frac{1}{|H|} \left(\sum_{h \in S} p^T P_{h,u} \right) E_e \end{aligned} \tag{1.5}$$

An illustration of the possible events and associated probabilities is given in figure 1.39.

1.5.3 Computing the likelihood: solving the equations

In the ideal case where $P_{e,u}$ would only depend on P_{e,u_d} where u_d is a strict descendant of u, $P_{e_d,u}$ where e_d is a strict descendant of e, and P_{e_d,u_d}, we could compute all $P_{e,u}$ with post order traversals of the gene and species trees:

for *all genes u in a post order traversal of G* **do**
  for *all species e in a post order traversal of S* **do**
    Compute $P_{e,u}$
  end
end

The equations behind most reconciliation models use this approach, however it is not compatible with the TL event, a transfer followed by a loss of the gene in the donor (see figure 1.39 for an illustration). With the TL event, the induction in a post-order traversal is ill-defined, as the probability that a gene u matches with e depends on the probability that u matches with the species h that are not descendants or ancestors to e (that themselves depend on $P_{e,u}$). For a list of software that consider this TL event see [100], for a discussion on the importance of considering these events, see [63], these questions are also discussed in our review in subsubsection 1.4.5.

ALE search for a solution using a point fix method, with a fixed number of iterations of this algorithm, five in the implementation. Even though the equations require to compute $\sum_{h \in S} p^T P_{h,u}$, this sum is computed once and for all as it does not depend on e, at the start of the loop. This gives a quadratic complexity to the likelihood estimation.

for *5 iterations* **do**
  for *all genes u in a post order traversal of G* **do**
    for *all species e in a post order traversal of S* **do**
      Compute $P_{e,u}$
    end
    Compute $\sum_{h \in S} P_{h,u}$
  end
end

In the implementation, a correction is added to $\sum_{h \in S} p^T P_{h,u}$ when computing $P_{e,u}$ to forbid transfers toward the ancestors of e.

1.5.4 TL counter, the solution I implemented

In this subsection, I propose another solution than the fix point method to compute the likelihood. The modification gives very similar results, and is more than 2 times faster. It also has the advantage of not depending on a fix point method which convergence we're not certain of. The idea is one we found in ecceTERA [100], to allow only one Transfer Loss (TL) per gene, inducing equations that depend only on values already computed and with one unknown value. Note that the TL counter method is different to one iteration of the fix point method, as the different steps of our method rely on giving the exact solution to a (simplified) linear equation, while the fix point method iterate over the complete equation.

We denote $P^{TL}_{e,u}$ the probability of gene u in species e after gene u has done a TL, and $P_{e,u}$ the probability of gene u in species e whether or not gene u has done any TL, we then have:

$$\begin{aligned} P^{TL}_{e,u} &= p^S_e \left(P_{g,v}P_{f,w} + P_{g,w}P_{f,v} + E_f P^{TL}_{g,u} + P^{TL}_{f,u}E_g\right) \\ &+ p^D_e \left(P_{e,v}P_{e,w} + 2P^{TL}_{e,u}E_e\right) \\ &+ \left(\frac{1}{|S|}\sum_{h\in S} p^T_h P_{h,w}\right) P_{e,v} + \left(\frac{1}{|S|}\sum_{h\in S} p^T_h P_{h,v}\right) P_{e,w} \\ &+ \left(\frac{1}{|S|}\sum_{h\in S} p^T_h E_h\right) P^{TL}_{e,u} \end{aligned} \tag{1.6}$$

$$\begin{aligned} P_{e,u} &= p^S_e \left(P_{g,v}P_{f,w} + P_{g,w}P_{f,v} + E_f P_{g,u} + P_{f,u}E_g\right) \\ &+ p^D_e \left(P_{e,v}P_{e,w} + 2P_{e,u}E_e\right) \\ &+ \left(\frac{1}{|S|}\sum_{h\in S} p^T_h P_{h,w}\right) P_{e,v} + \left(\frac{1}{|S|}\sum_{h\in S} p^T_h P_{h,v}\right) P_{e,w} \\ &+ \left(\frac{1}{|S|}\sum_{h\in S} p^T_h E_h\right) P_{e,u} + \left(\frac{1}{|S|}\sum_{h\in S} p^T_h P^{TL}_{h,u}\right) E_e \end{aligned} \tag{1.7}$$

And we can compute the values following this process:

for *all genes u in a post order traversal* **do**
  for *all species e in a post order traversal* **do**
    Compute $P_{e,u}^{TL}$
  end
  for *all species e in a post order traversal* **do**
    Compute $P_{e,u}$
  end
end

So when we compute P^{TL} we already know every other terms in the equation, as for P, and we only have to solve a linear equation to get it. The sums are arranged as in undated with averages computed only once. Both a version with ancestral correction (to prevent direct transfer to the past) and without have been implemented and tested.

As for complexity, if we did 5 iterations for the fix point method (as in the implementation of ALE), it is straightforward to see that the time gain will be at least 2. Furthermore, most of the terms in the equation for P^{TL} and P are the same, so only a small part of the computations have to be done again, and for instance in number of product or sum, we have a gain of a bit more than 2.4, which is close to what we found in the implementation.

I've implemented this solution in Python, and tested it on a simulation I've also implemented (dated DTL model). The likelihood obtained is not exactly the same but is quite similar. The values are really close, using as witness a version with no TL, and the reconciliation induced by those likelihoods are really often the same, using as witness the simulated reconciliation (more than 0.95 agreement for the root against 0.8 against the simulated reconciliation). As can be expected if the loss rate is too important (bigger than birth and death rate), the approximation of only one TL is insufficient, and the TL counter method is not as good as the fix point method (we can see that in term of agreement to the simulated reconciliation).

It is also possible to modify the computation of E to get rid of the fix point method for the whole computation. The values are similar and the computation is a bit faster. However the gain in computation time with this modification is negligible in front of the computation time for P (one has linear complexity, the other quadratic). The method is to do some kind of asymptotical development of E, order 3 and 5 presented here both give good results.

$$E_e = p_e^L + p_e^S E_f E_g + p_e^D E_e^2 + E_e \frac{1}{|S|} \sum_{h \in S} p_h^T p_h^L \tag{1.8}$$

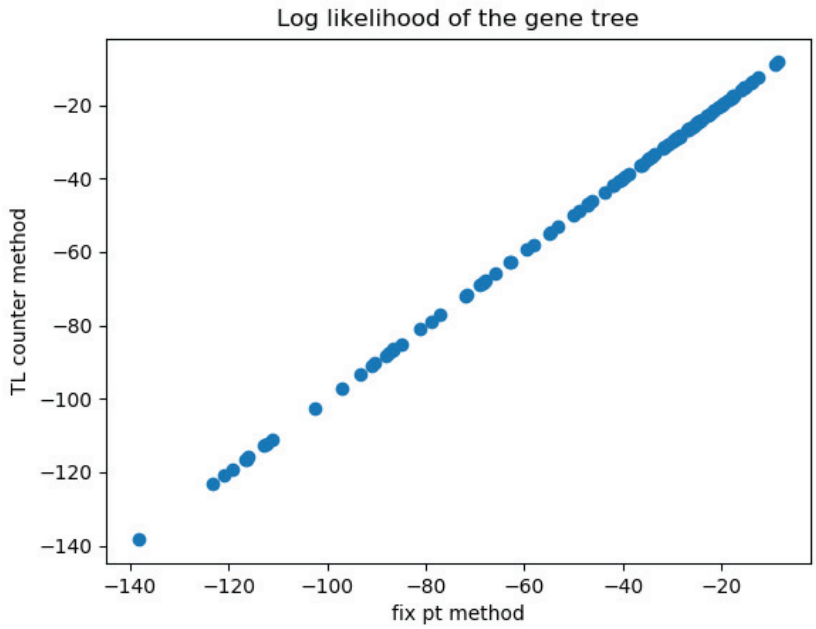


Figure 1.40: Log likelihood for 100 samples (gene and species tree simulated), with 70 nodes (leaves included) in the species tree, and birth rate 0.5, transfer rate 0.5, death rate 0.1 for the gene.

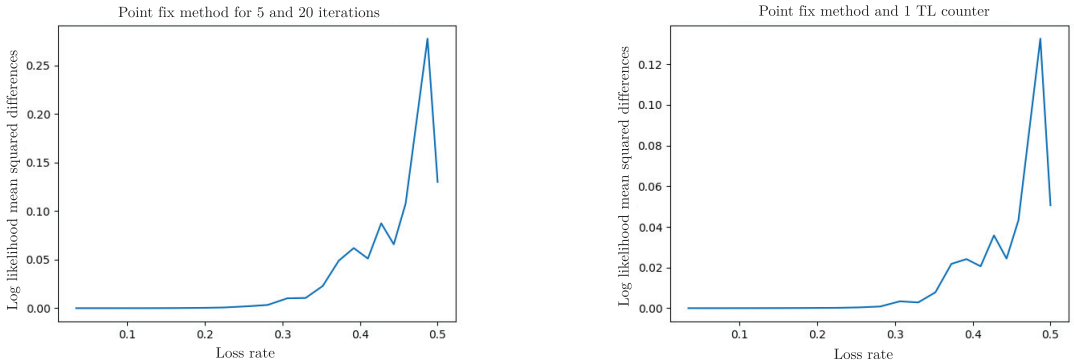


Figure 1.41: Average squared difference between the log likelihoods for varying loss rates. Left, comparing 5 and 20 iterations of the fix point method. Right, comparing the fix point method with 5 iterations and our proposed solution with only one TL.

$$E_e = p_e^L + p_e^S E_f E_g + p_e^D E_e^2 + E_e \frac{1}{|S|} \sum_{h \in S} p_h^T \left(p_h^L + p_h^S p_{h_i}^L p_{h_j}^L + p_h^D (p_h^L)^2 + p_h^L \left(\frac{1}{|S|} \sum_{k \in S} p_k^T p_k^L \right) \right) \tag{1.9}$$

1.5.5 Clade prior computation and amalgamation

Amalgamation enables us to take as input multiple potential trees instead of just one for the gene tree. The final goal is to use the information contained in these

multiple trees in the reconciliation process and produce a most likely gene tree. A list of trees instead of a single one can be obtained with bayesian tree inference methods from aligned sequences, or with bootstrap approaches. A simple way to consider all these potential topologies would be to do the product of the likelihood of the reconciliation of each of these topologies with the species tree. However, this would not be computationally efficient, and it would not allow the construction of scenarios using some clades coming from one tree and some clades coming from another in the list.

Amalgamation starts by estimating conditional clade probability before constructing a tree-like structure with these probabilities and reconciling it with the species tree, to get a chimeric gene tree constructed from the observed clades. Conditional clade probability was introduced in [96] and [116], and is used in multiple reconciliation softwares [100, 216]. Angst [54] also reconstructs chimeric trees but from bootstraps instead of conditional clade probability. We first define and give a way to compute conditional clade probability.

Let us begin by defining conditional clade probability. A clade is a group of leaves. In a rooted tree we say we observe a clade if its group of leaves is monophyletic, *i.e.* if they are the leaves of some subtree in that tree. For an unrooted tree, a clade is observed if it is observed in some rooting of this tree. For a clade C we will look at the different ways it can be split into two disjoint clades, C' and C''. The idea is that we will construct a new tree-like structure, where the nodes are clades and couples of clades, and we will put a link between a clade and a couple of clades if it is the union of both, and we will weight branches with the probability to observe that split of clade C in the data: $P(C', C''|C)$.

To estimate these conditional probabilities, we look at bipartition and tripartition in the sample trees, and assume that non-overlapping clades are independent. To get a bipartition, we remove a branch in an unrooted tree and thus get two clades. Similarly, we remove an internal node and obtain three clades to get a tripartition (see an illustration in figure 1.42). Let us count the number of clades in an unrooted tree. For each internal branch in this tree, we have two clades (defined by the bipartition), and we also have to add the clade that is comprised of all the leaves. Denoting n the number of leaves, we have $2n-3$ internal branches in the unrooted tree, and $2 \times (2n-3) + 1 = 4n - 5$ clades.

A clade C will have C' and C'' as couple children if there is a tripartition $(C', C'', \bar{C})$, where $\bar{C}$ is the set of leaves from the tree that are not in C. We will then look at the frequency at which C has (C', C''), by comparing this number of tripartitions, with the number of bipartitions between C and $\bar{C}$. So we look at the frequency $\frac{|(C',C'',\bar{C})|}{|(C,\bar{C})|}$, where $|(C', C'', \bar{C})|$ is the number of trees for which we observe the tripartition $(C', C'', \bar{C})$, and $|(C, \bar{C})|$ the number of trees where we observe the

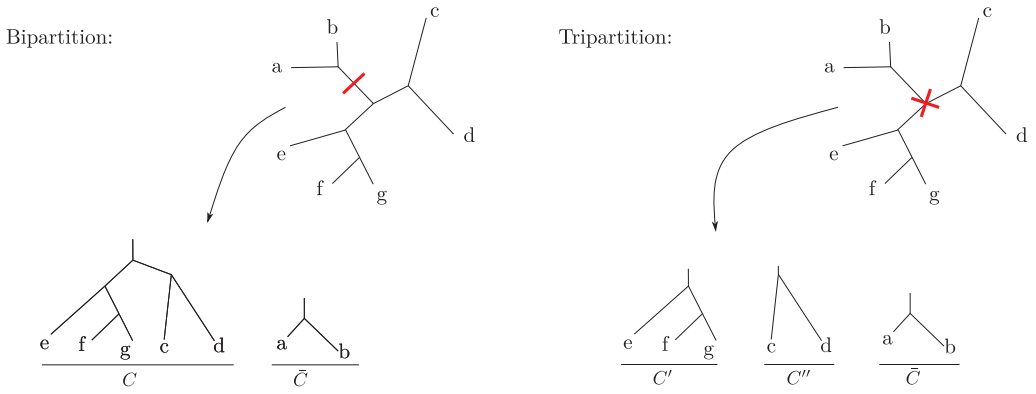


Figure 1.42: Bipartition are obtained by taking out one branch. Tripartition by taking out one internal node. We then look at the clades, the set of leaves of the induced rooted trees. Here the bipartition induces two clades (c,d,e,f,g) and (a,b). The tripartition induces three clades (e,f,g), (c,d) and (a,b). We can then look at the proportion of sampled trees where the clade (c,d,e,f,g) split in (e,f,g) and (a,b) by looking at the ratio of the number of tripartitions and the number of bipartitions corresponding to these clades, see figure 1.43.

bipartition $(C, \bar{C})$, which is an estimation on our data of the probability $P(C', C''|C)$ (see figure for an example 1.43).

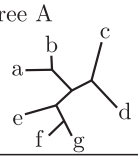
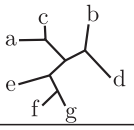
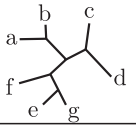
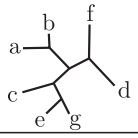
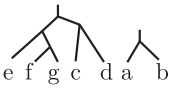
$$P(C', C''|C) \simeq \frac{|(C', C'', \bar{C})|}{|(C, \bar{C})|} \tag{1.10}$$

We use these frequencies to weight branches in a tree that link clades to their potential couple of children, and where the root is the clade of all leaves of the tree (see figure 1.44).

It is this tree we will reconcile with the species tree, summing over all possible couples of children of u weighted by their clade conditional probabilities when computing $P_{e,u}$, in place of the single couple children in the computation without amalgamation. For instance, if we look only at the speciation event, we now have

$$P(C, e) = p^S \sum_{(C', C'')} P(C, C'|C'')(P_{C',f} P_{C'',g} + P_{C'',f} P_{C',g}) \tag{1.11}$$

Amalgamation takes as input unrooted trees, and as such, if we use it with only one tree as input for each gene tree, it is a way to consider all possible roots of the tree in a computationally efficient way. If a tree has n leaves, it has $2n-3$ internal branches and as many possible roots, so we would need to do the reconciliation of a $2n-1$ nodes tree $2n-3$ times to consider all possible roots separately. With amalgamation we just have to consider all the clades obtained from the unrooted tree, so 1 reconciliation of a $4n-5$ nodes tree-like structure, which is a substantial speedup as reconciliation is linear in the number of nodes of the lower tree. However reconciliation with the amalgamated tree-like structure obtained via reconciliation

	Tree A	Tree B	Tree C	Tree D
Unrooted tree sample				
Trees with (a,b),(c,d,e,f,g)				
Trees with (e,f,g),(c,d) and (a,b),(c,d,e,f,g)				

$$P((e,f,g),(c,d)|(c,d,e,f,g)) \simeq \frac{2}{3}$$

Figure 1.43: For all clades, we look at its possible children. Here we want to compute the probability of having clade (c,d,e,f,g) splits in (e,f,g) and (c,d). We have a sample of four trees with different topologies. We first count the number of trees where we observe the bipartition between (a,b) and (c,d,e,f,g). Trees A, C and D display that bipartition. Then we look for the tripartition (a,b),(c,d),(e,f,g), that can only be found in trees that display the bipartition. Trees A and C have that tripartition. So on the 3 trees with the bipartition, 2 have the tripartition. We conclude to a probability of $\frac{2}{3}$ of this split. The other observed split with these tree for (c,d,e,f,g) is (c,e,g),(f,d) in Tree D, that will thus get a probability of $\frac{1}{3}$. By considering all bipartitions and tripartitions in each tree of our sample we find all the observed clades and the frequencies of their children. This information enable us to construct a tree-like structure that can be used for reconciliation, see figure 1.44 for an illustration.

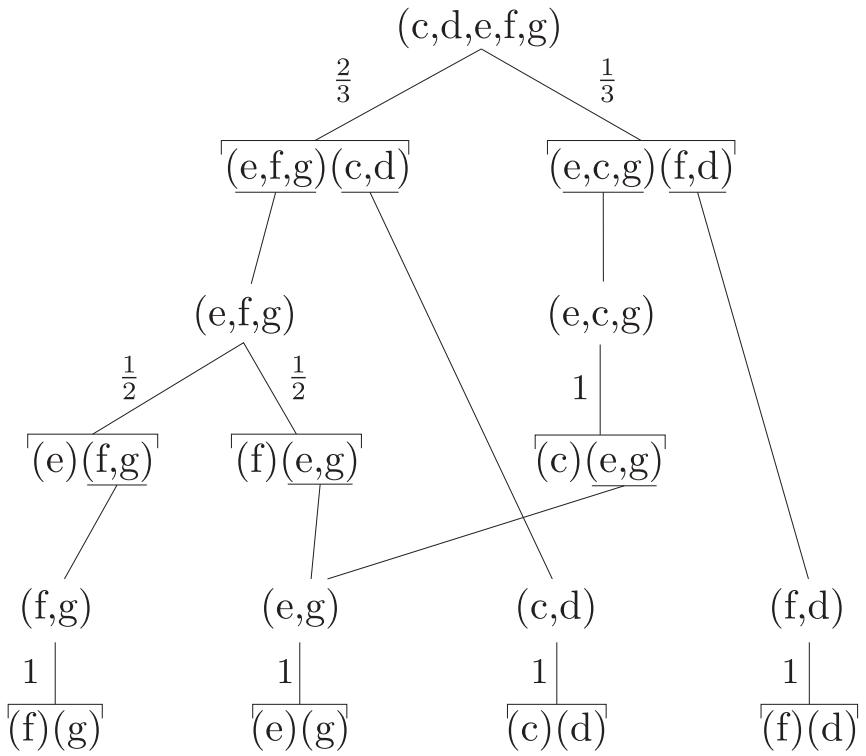


Figure 1.44: We look here at the tree-like structure obtained starting from clade (c,d,e,f,g) with the tree sample of figure 1.43. As we saw, this clade has two possible splits, (e,f,g),(c,d) with probability $\frac{2}{3}$ and (e,c,g),(f,d) with probability $\frac{1}{3}$. Then clade (e,f,g) also has two splits (e),(f,g) and (f),(e,g) with equal probability, while clade (c,d) only have one possible split (c),(d). A clade can have one children (like (c,d)), two children (like (e,f,g)), or even a higher number. What makes amalgamation efficient is that clades can have multiple parents, like (e,g) here that can be accessed from (e,f,g) and (c,e,g). With dynamic programming, multiple computations will rely on the same already computed elements.

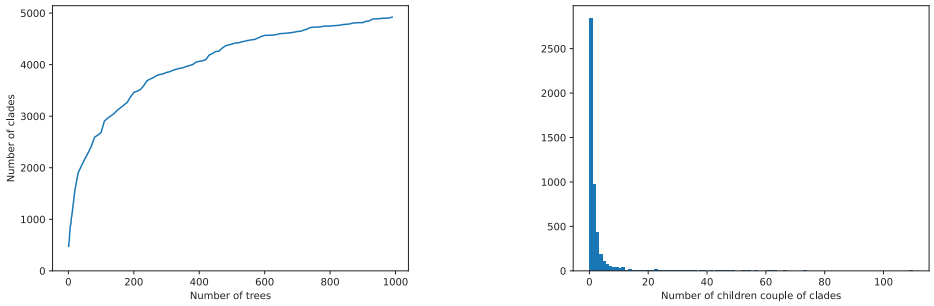


Figure 1.45: (Left) The evolution of the number of clades with the number of sampled trees for one of the gene families in our *pylori* dataset. The gene trees contain 119 leaves, which means 471 clades. At first the number of clades grows fast, and then each new tree add few new clades. But even from the first new elements, it is faster than just doing each tree independently. A thousand trees are considered by an object only 20 times bigger than one tree in regard to the number of nodes. (Right) The distribution of the number of children couple of clades in the tree-like structure, for the amalgamation of 1000 trees, the root clade is omitted for readability, but it has a large number of children (2466).

is not always linear in the number of clades, it depends on the number of couple clades children of each clades. Here with just one unrooted tree considered, only the root clade has multiple couple of children, it has $2n - 3$ such couples, so we have a $6n$ factor in the complexity ($4n + 2n$) when the reconciliation of one non-amalgamated tree is $2n$, so by the equivalent of 3 reconciliations we test all roots, instead of n reconciliations. This result is similar to another one that gives the same complexity, by computing, for each node in the unrooted tree, the three clades that it separates in every rooted trees (depending on the root, which one is $\bar{C}, C', C''$ changes). With multiple trees the number of clades with multiple couple children can grows to the order of n. Nevertheless if we consider that the number of possible resolutions of each nodes in the trees is bounded (we look only at the ones supported by the aligned sequences), we can assume that reconciliation is linear in the number of clades observed in the trees. See figure 1.45 for an example of the evolution of the number of clades with the number of sampled trees and of the distribution of the number of children.

1.5.6 Output, frequencies and chimeric trees

The first part of the computation for ALE undated is that of likelihood and of a table of $P_{e,u}$ the probabilities of all pairs of upper and lower subtrees to be matched. The likelihood is thus the first interesting information given: it measures the probability of generating the gene tree inside this species tree with the given evolutionary rates.

As is often done with likelihood, in practice, we will use the logarithm of the like-

lihood, as likelihood values are close to zero, for there are many possible gene trees and similarly likely ways to grow a tree inside a given species tree. Likelihood can be used to compare the topology of species trees or gene trees or to infer maximum likelihood rates. It is a measure of the coevolution of the two trees.

We can then use the dynamic programming table to backtrack and sample gene and species coevolution scenarios. The scenarios give a view of the possible events at the different branches. In ALE undated, multiple scenarios are sampled, and the frequency of observation of each event is computed. In the implementation, by default, a hundred scenarios are sampled. An event is in the form of a gene node matched to a species node and one event among D, T, S, SL, or TL. A scenario is a list of such events, a possible format for a scenario is Recphyloxml[70], and ALE output can be seen in that format using an outside transcription software.

The last thing we can get from ALE is the result of the amalgamation in the case where we put a list of potential topologies for each gene tree. Each scenario sampled comes with a reconstructed gene tree. As we saw previously, amalgamation creates chimeric trees that are not present in the input distribution but for which all clades are. Thus, we can get a clade posterior probability by aggregating multiple scenarios' results. We can also simply look at some of the sampled trees.

1.5.7 Rates estimation

One of the advantages of a probabilistic framework is that it provides a natural way to estimate parameters: maximum likelihood. In ALE undated, there are three rates of evolution, D, T, and L, estimated by a expectation-maximization algorithm. Starting from some initial probabilities, for instance, 0.01 for each (giving 0.97 to S), ALE reconcile the two trees and look at the event posterior probabilities through the observed frequencies of D, T, and L events in the multiple scenarios sampled. It then fixes the probability of each event to its relative frequency and start again. ALE repeats that process for a fixed number of iterations or until the likelihood stops improving.

1.5.8 Generax version

The likelihood estimation of generating a gene tree inside a given species tree implemented in ALE undated has also been implemented in Generax [158], a software meant to infer better gene trees from alignments by using both the alignment likelihood and the reconciliation likelihood of gene and species to guide a search for the best tree. As reconciliation is used in the evaluation step and is a computationally heavy method, a speedup was needed to evaluate a significant number of tree topologies. Generax paper describes these modifications.

The main speedup focuses on transfers. First, TL events are not considered, as it is the only event that makes the induction ill-defined, as we discussed in subsection 1.5.3. Then not all transfers are considered, making it possible not to consider all couples of gene and species nodes matching and fill a smaller dynamic programming table. A couple e, u is only considered if one of the descendants of u is matched to one of the descendants of e. Otherwise their matching is considered unlikely, and is ignored.

SECTION 1.6

Outline

In this introduction, I tried to give an idea of the objects we will use during this thesis. We presented our two main biological models[3]: *Helicobacter pylori* their genes and their human host, *Cinara* aphids, their endosymbionts, and their genes. I gave a brief introduction to molecular evolution and phylogenetics, as well as discussed symbiosis. Then I reviewed phylogenetic reconciliation and gave an overview of multi-level biological systems and the attempts to model them. Finally, I described ALE undated, the model we expand to construct a multi-level reconciliation framework able to consider the coevolution of genes, symbionts and hosts.

The following chapter is the main one of the thesis and consists of original material and ideas. It has two parts and is guided by articles we wrote with Vincent Daubin and Eric Tannier, one, with Alexia Nguyen Trung, our primary model of 3-level reconciliation, and the other, with Simon Penel and Théo Tricou, a graphical viewer for reconciliation.

The third chapter, *Helicobacter pylori*, is centered around a biological model and biological questions. This chapter contains its own introduction to *Helicobacter pylori* and its link to human hosts (from a biogeographical and evolutionary point of view, more than a medical one). I then present our work on this subject, mainly using reconciliation.

A short fourth chapter is devoted to some theoretical questions around reconciliation and phylogenetics that I stumbled upon during this thesis.

In terms of people, for these three years, I worked with Eric Tannier and Vincent Daubin to design the method and model for 3-level reconciliation, with help from Alexia Nguyen Trung for the application to *pylori*. I designed the graphical viewer with Simon Penel, Eric Tannier and Vincent Daubin, and Simon implemented it all, with valuable insights for the design and test from Théo Tricou. For the third

[3]Model can be an ambiguous term. In this thesis, I will precise *biological models* when writing about a model organism, a biological system of interest. Mathematical models will simply be noted as *model*.

chapter, it was guided by Alexia and Vincent that Eric and I tried to get a good understanding of what was happening with *pylori* and how we could find a method to consider the questions at hand.

Chapter 2

3-level reconciliation

Contents

In this chapter, I present two articles. One, the main of this thesis, is about a model and inference method for reconciliation with three levels, depicting coevolution between gene and symbiont, and symbiont and host, with a DTL model. The second is a graphical viewer that displays various useful new features for 2-level and 3-level reconciliation, figures throughout this section were generated using it. Some additional analyses are presented after the main methodological article, an application to a coronaviruses dataset as well as a section devoted to ideas around the inference of symbiont trees.

SECTION 2.1

3-level model and method

In this section, we present the main contribution of this thesis, a 3-level probabilistic host/symbiont/gene reconciliation framework. It is introduced in an article. Technical points, such as the comparison of the likelihood of the 3-level and 2-level models, additional analyses, such as the robustness of the Monte Carlo, and motivations to our choices are discussed after the article.

Implementation

I reimplemented ALE undated in python, to serve as a basis for a multi-level approach, and it was also the occasion to get a better understanding of the different components of the method. The code is available on GitHub: https://github.com/hmenet/TALE This new version of ALE has different interesting features compared to the original ALE, but possess most of the features of the original ALE, (notable exceptions include MCMC estimation of rates, and branch dependent rates), though it is less optimized. My implementation is easy to use and install, as it is a command line software, written in python, with usual python package, so no dependencies as for ALE and BIO++. It directly outputs RecPhyloXML that can be read by Thirdkind, our graphical viewer we present in this chapter, to get a graphical representation. In a way it could be use to test the method for someone willing to try, before going through with ALE installation. It has no dated version. It can give back the maximum likelihood reconciliation, instead of a sample. It can be used with uncertainty on the matching of the lower tree, which we used in our approach to pylori.

Main methodological paper: Host-Symbiont-Gene phylogenetic reconciliation

Our model and method is described, with results on simulated and biological datasets. This paper introduces a variety of new ideas and concepts that it only briefly discusses, and so the rest of this section is devoted to going deeper and reexplaining some of these ideas. Additional context and explanation on the *Helicobacter pylori* dataset is developed in the next chapter. The paper is presented as it was submitted and rejected by ISMB 2022. We are currently preparing it for a new submission.

2.1.1 Host-Symbiont-Gene phylogenetic reconciliation

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Host-symbiont-gene phylogenetic reconciliation

Hugo Menet [1,2,*], Alexia Nguyen Trung [1,2], Vincent Daubin [1,2] and Eric Tannier [1,2,3]

[1]Laboratoire de Biométrie et Biologie Evolutive, Centre National de la Recherche Scientifique, Unité Mixte de Recherche 5558, Université Lyon 1, F-69622 Villeurbanne, France,
[2]Université de Lyon, F-69000 Lyon, France and
[3]Centre de Recherche Inria (Institut National de Recherche en Informatique et en Automatique) de Lyon, Villeurbanne, France

*To whom correspondence should be addressed.

Abstract

Motivation: Biological systems are made of entities, organized at different scales (macro-organisms, symbionts, genes...) which evolve in interaction. These interactions range from cooperation and coevolution, which results in them having a common history, to independence or conflict. The evolution of such systems is approached by phylogenetic reconciliation, which describes the coevolution of two different levels, genes and species, or hosts and symbionts for example. The limit to two levels hides the multi-level inter-dependencies that characterize complex systems.
Results: We present a probabilistic model of evolution of three nested levels of organization which can account for the coevolution of hosts, symbionts and their genes. This model allows gene transfer as well as host switch, gene duplication as well as symbiont diversification inside a host, gene or symbiont loss. It handles the possibility of ghost lineages as well as temporary free-living symbionts.
Given three phylogenetic trees, we devise a Monte Carlo algorithm which samples evolutionary scenarios of symbionts and genes according to an approximation of their likelihood in the model. We evaluate the capacity of our method on simulated data. Then we show in a aphid enterobacter system that some reliable transfers detected by our method, are invisible to classical 2-level reconciliations. We finally evaluate different hypotheses on human population histories in the light of their coevolving *Helicobacter pylori* symbionts, reconciled together with their genes.
Contact: hugo.menet@univ-lyon1.fr
Availability and implementation: Implementation and supporting data are available on GitHub https://github.com/hmenet/TALE. Data are available upon request.

1 Introduction

The toolbox of evolutionary biology largely relies on the assumption of statistical independence of biological objects at any level of organization: organisms from different species are isolated from a biological system based on their genomes, genomes are cut into independent genes, and inside genes, nucleotides are evolving independently from each other (Felsenstein, 2004).

Yet the essence of living systems lies in dependence: constraint, cooperation or conflict (Sapp, 1994). Symbiotic micro-organisms coevolve with animals or plants (Sonnenburg and Sonnenburg, 2019). The ensemble they form is gathered under the holobiont concept. It allows to see genes as entities not only following their own interest, not only participating to the functioning of the genome it is hosted by, but also participating to, and probably evolving with, a larger biological system.

A powerful tool to study these inter-dependencies is phylogenetic reconciliation: an ensemble of models and methods explaining the differences and similarities between phylogenies of two coevolving entities. Gene/species systems have been studied by phylogenetic reconciliation, accounting for events of gene duplication, horizontal gene transfer and gene loss (DTL model) (Doyon *et al.*, 2011; Nakhleh, 2013; Szöllősi *et al.*, 2015b; Boussau and Scornavacca, 2020; Menet *et al.*, 2021). The same model can be applied with little or no modification to symbiont/host (Charleston and Libeskind-Hadas, 2014; Santichaivekin *et al.*, 2020; Donati *et al.*, 2015), protein domain/gene coevolution (Rasmussen and Kellis, 2012; Stolzer *et al.*, 2015), or biogeography (Martínez-Aquino, 2016; Ree and Smith, 2008; Ronquist, 1997). DTL models have also been used to reconstruct genome histories (Duchemin *et al.*, 2015), detect highways of lateral gene transfers in bacteria, archaea or eukaryota (Bansal *et al.*, 2011), assess the relative role of duplication and gene transfer in the evolution of genomes (Sjöstrand *et al.*, 2014), infer ancient symbiotic relationships (Bailly-Bechet *et al.*, 2017), reconstruct

histories of gene fusion and fission (Duchemin *et al.*, 2017), model endosymbiotic gene transfer (Anselmetti *et al.*, 2021), etc... They are the subject of active research to integrate concepts from population genetics (e.g., incomplete lineage sorting) (Stolzer *et al.*, 2012; Wu, 2012; Chan *et al.*, 2017).

A limitation of reconciliation methods is their separate application on molecular studies on one side (gene/species coevolution), and ecological studies on the other (host/symbiont coevolution). The striking methodological unity of the two (the same DTL model is applied on both the molecular and ecological systems) and the growing interest for multi-level systems integrating molecular and ecological inter-dependencies (e.g. the holobiont concept) calls for a unique model for host, symbiont, gene coevolution. In support of this claim, a number of empirical studies already rely on host symbiont histories when proposing horizontal gene transfers between symbionts (Penz *et al.*, 2012; Nikoh *et al.*, 2014; Manzano-Marín *et al.*, 2019; Nakabachi *et al.*, 2013), when often, only symbiont gene/species comparisons do not provide enough statistical support for them (Wijayawardena *et al.*, 2013; Ravenhall *et al.*, 2015).

Three level reconciliations have been introduced by Stolzer *et al.*, 2015 and applied to protein domain, gene and species. It consists of two embedded DTL models and an inference method by parsimony, applying one reconciliation after the other. Further efforts in this direction have been published by Li and Bansal, 2019a with a duplication/loss model between gene and species and a DTL model, forbidding inter species transfers, between protein domains and genes. They show NP-hardness of inferring the most parsimonious couple of nested reconciliations (Li and Bansal, 2019a) and propose different heuristics and problem variants (Li and Bansal, 2019b, 2018). A probabilistic model without transfers has been proposed by Muhammad *et al.*, 2018, which aims at inferring dated gene trees from protein domain alignments using Markov Chain Monte Carlo. These attempts prove that it is possible to jointly handle three nested levels in a single computational model, but none of them can yet handle host/symbiont/gene systems in a statistical framework, because of specific limitations of each of them (parsimony framework, no transfer or no inter-host transfer, no joint inference between levels of organization).

We propose a probabilistic model that describes the evolution of three nested coevolving entities at three different scales, adapted to a host/symbiont/gene system. In our model a symbiont tree is generated by a DTL model inside the host, with a possibility of evolving temporary outside the host phylogeny. A gene is generated by a DTL model inside the symbiont, where gene transfer is more likely between symbionts that share a common host (which we will call "intra" transfer) than for those that do not (which we will call "inter" transfer).

Based on this model we propose an inference method extending the two-level reconciliation "ALE" software (Szöllősi *et al.*, 2013, 2015a). It takes three trees as input, constructs joint scenarios and estimates event rates and likelihoods according to the model. Our implementation also features the possibility to take only the host tree and several gene trees, assuming some of them to be universal unicopy, and to infer a likely symbiont tree. A comparison of the likelihood of two-level and three-level reconciliations can be used as a test for multi-scale coevolution.

We report a benchmark test of the inference method on simulated data, using an external simulator (Kundu and Bansal, 2019), showing that under the hypothesis that gene transfers are more likely between symbionts of a same host, the three-level reconciliation represent a significant gain compared to the two-level one.

We use the inference method to identify horizontal gene transfers between *Cinara* aphid symbionts that are missed by two-level reconciliations.

Finally we show on genes of *Helicobacter pylori* from human populations how likelihood computations can be used to compare different hypotheses on the diversification of a host, given the genes of its symbionts, taking into account the coevolution between all three scales.

2 Two level reconciliation, definitions and preliminaries

We denote by G, S, H respectively the gene tree, species (or symbiont) tree and host tree. Given a tree T, $|T|$ the number of nodes of T.

We briefly describe in this section a two-level DTL reconciliation model, based on the undated version of the ODT model as implemented in ALE undated (Szöllősi *et al.*, 2015a). It is a birth and death like model generating a rooted phylogenetic tree G inside S, with speciation at all speciation nodes of S, and duplications, transfers and loss specific to G along the branches of S. We thus have three rates for duplication, transfer and loss events, concerning the evolution of genes inside their species. Gene tree can originate in any branch of the species tree with a uniform prior.

The input of 2-level reconciliation inference is one gene tree, and one species tree, with a many to one matching of the leaves. Both trees are assumed undated, binary and rooted.

We call reconciliation scenario a list of events (D,T,L,S) for each internal gene tree node, that can be the result of the birth and death process. It thus transcribes into a mapping of the gene tree nodes to the species tree nodes it evolves in. We note $R_{G,S}$ the set of all possible reconciliation scenarios between G and S.

We denote by p^S, p^D, p^T, p^L the probabilities for a gene to undergo each of the S,D,T,L events, with $p^S + p^D + p^T + p^L = 1$. When confusion is possible we add a S index for symbiont/gene reconciliation, and H index for host/symbiont one. If the event is a transfer, the probability to transfer to a specific branch is uniform: $\frac{p^T}{|S|}$. The likelihood of a scenario r, $P(r|S)$ is the product of the probabilities of all events. Summing over all possible scenarios for one gene tree and species tree we obtain the likelihood of coevolution of the two trees, $P(G|S)$. However we do not have to enumerate all scenarios to get that sum, instead we can compute this likelihood using a dynamic programming, considering matching all couples of gene and species sub-trees, starting from the leaves, and enumerating all possible events to get each match. This in return enables us to sample scenarios according to their likelihood, or finding the most likely scenario, by backtracking through the table constructed.

We will call such a reconciliation of a gene tree and symbiont tree, "2-level" reconciliation, in opposition to the symbiont/gene aware of an host 3-level reconciliation that we introduce in the following section.

3 Three level reconciliation, likelihood estimation and scenario inference

3.1 Elements of the probabilistic model

A rooted binary host phylogenetic tree H is first generated (we do not include the generation parameters in our model and consider instead the rooted tree as a parameter). Then we model the evolution of one or multiple symbiont trees S with an adaptation of the DTL model (Szöllősi *et al.*, 2015a). The adaptation consists in adding the possibility for a symbiont to live in an unknown host (this feature is described in more details later).

We then model the evolution of genes in the symbionts with duplication, loss and intra horizontal transfer, meaning that horizontal transfer is possible only between symbiont branches that are present in the same host branch. We thus have six rates in our model, three for the duplication, transfer and loss between host and symbiont, and three additional ones for duplication, intra-transfer and loss events concerning genes inside their species. An illustration of the realization of such a model,

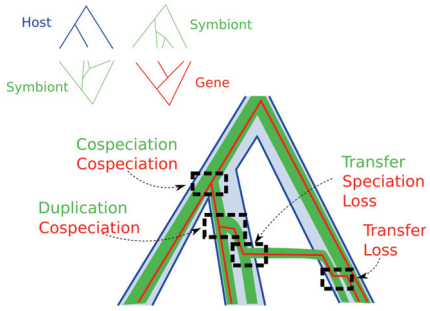


Fig. 1. One 3-Level reconciliation input (top left) and reconciliation scenario.

as well as the input for the inference part, is given in Figure 1.

This model can be immediately used for simulations, but we chose to use an external simulator for our tests (Kundu and Bansal, 2019), to minimize the possibilities of errors due to endogenous comparisons.

3.2 Monte Carlo approximation of the likelihood

The inference consists in computing the probability that some data have been generated by the model, and estimate the evolutionary rates, the scenarios, or the symbiont tree. We use an undated framework similar to the one implemented in ALE undated (Szöllősi *et al.*, 2015a) presented in previous section. All given trees are supposed to be binary, and branch length is not taken into account.

Given a rooted binary host tree H, and a (or a set of) rooted binary symbiont tree S, the parameters of our inference model are the DTL probability of evolutionary events for the two reconciliations, the host/symbiont one, and the symbiont/gene one. We compute the probability that a (or a set of) unrooted binary gene tree G has evolved inside host and symbiont, denoted $P(G|S,H)$.

Because a similar computation in a parsimonious framework is NP-hard (Li and Bansal, 2019a), it is probably not possible to exactly and quickly compute this number. We thus apply an approximation technique based on sampling reconciliations. The probability of a gene tree can indeed be decomposed by summing over all possible host/symbiont reconciliation scenarios $r_{S,H}$:

$$P(G|S,H) = \sum_{r_{S,H} \in R_{S,H}} P(G|S,H,r_{S,H}) P(r_{S,H}|S,H) \quad (1)$$

The number of reconciliations in this sum is at least exponential in the size of the input (and even the number of scenarios maximizing $P(r_{S,H}|S,H)$ can be exponential) (Donati *et al.*, 2015), so we use a Monte Carlo approach to estimate it, sampling a reasonable number N of symbiont/host reconciliations :

$$P(G|S,H) \simeq \frac{1}{N} \sum_{n=1}^{N} P(G|S,H,r_n) \quad (2)$$

where r_n is sampled in the set $R_{S,H}$ of all reconciliations according to its likelihood.

3.3 Reconciliation inference and ghost lineages

Sampling reconciliations in $R_{S,H}$ can be done with the dynamic programming algorithm implemented in "ALE undated" and is a two-level reconciliation problem (Szöllősi *et al.*, 2015b).

Given $r_n \in R_{S,H}$, the probability $P(G|S,H,r_n)$ can be computed with an adaptation of the same dynamic programming algorithm. It consists in checking, during the dynamic programming process, for all gene transfer possibilities, if the donor symbiont i and receiver one j share a host in r_n. If they do, then it is an "intra" transfer and the transfer has the probability defined by the transfer rate.

In our model we make the hypothesis that gene transfer can only occur between two symbiont species inside a same host. However transfer between two symbionts in different hosts is possible through ghost species. Indeed it is always reasonable to assume that a major part of species are extinct or unsampled and gene transfer are "from the dead" (Szöllosi *et al.*, 2013; Fournier *et al.*, 2009; Zhaxybayeva and Gogarten, 2004). In consequence in the model a transfer can occur from a donor that is now extinct. This transfer is traced back to an ancestor of this extinct donor that is not in the same host than the receiver. See in Figure 2 how an "inter" transfer between i and j (on the left) can be modelled (on the right) by an extinct sister lineage to symbiont i, that switched host, and transferred a gene to j while being in the same host. As the sister lineage goes extinct, in the inferred gene history it is transferred from i to j.

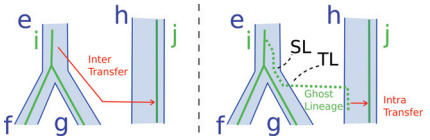


Fig. 2. Computation of inter transfer rate from intra transfer rate and ghost lineages: the left inter transfer can be modeled by multiple scenarios without inter gene transfer but implying ghost symbiont lineages, such as the one on the right.

We denote by $P_S^T(i \to j)$ the probability for a gene present in symbiont i to undergo a horizontal transfer to symbiont j, and $P_H^T(e \to h)$ the probability for a gene present in a symbiont associated to host e to transfer to a symbiont associated to host h. Let H_i (H_j) be the branch of the host tree which contains symbiont lineage i (resp. j).

$$P_S^T(i \to j) = \sum_{e \in H_i, h \in H_j} P_H^T(e \to h) \quad (3)$$

At fixed h we rewrite with $P_e = P^T(e \to h)$. Recall p_S^T are the probability of horizontal transfer in the symbiont/gene reconciliation, and $p_H^S, p_H^D, p_H^T, p_H^L$ the probabilities of speciation, duplication, transfer and loss in the host/symbiont reconciliation. Let E_e be the probability of extinction, that is, the probability that a gene is present in a branch e and absent from all the leaves. Let $|S_h|$ be the number of symbiont branches matched to host h in the host/symbiont reconciliation scenario.

$$\begin{cases} P_h &= \frac{1}{|S_h|} p_S^T \\ P_e &= p_H^S (P_f E_g + P_g E_f) + 2 p_H^D P_e E_e + \sum_{k \in H} \frac{p_H^T}{|H|} P_k E_e \end{cases} \quad (4)$$

This equation has a self dependency due to the Transfer/Loss event, already accounted in reconciliation methods (Jacox *et al.*, 2016; Szöllősi *et al.*, 2013). We forbid successions of several Transfer/Loss events to break this self dependency and solve this equation.

3.4 Time complexity

DTL reconciliation methods use a dynamic programming approach to compute the probability of the coevolution of two trees (and all of their subtrees) (Charleston, 1998), before backtracking to get a reconciliation scenario. In a gene/species reconciliation, if all transfers are done with the same rate *i.e.* the probability of transfer is independent from the couple

of donor, receiver species node, DTL reconciliation can be computed in quadratic time (Bansal *et al.*, 2012).

However, in our model, transfer rates depend on the donor receiver couple, and thus we cannot use the efficient computation trick used for uniform rates, and, for each couple of gene and symbiont subtree couple, we must explicitly consider transfers toward all symbiont nodes, thus getting at least a cubic complexity for our host aware, symbiont/gene reconciliation. The two other part of our algorithm are the classical 2-level reconciliation of host and symbiont, that can be done in quadratic time, before the backtrack to sample a scenario, and the computation of transfer probabilities between each couple of symbiont nodes (with equation 3).

Denoting by m, n, p the number of nodes of the host, symbiont, and gene tree respectively, we have a complexity of $O(mn + m^2n^2 + n^2p)$ and in the reasonable case where the number of symbiont nodes per host nodes (in the reconciliation scenario) is below a constant k, we get $O(mn + m^2k^2 + n^2p)$. Only the n^2p part is gene dependent and must be repeated for each gene tree, which is, in our experience, the bottleneck of the approach, as we often have one host and one symbiont tree, but multiple gene trees, all of similar sizes.

For the Monte Carlo, we have this complexity times the number of sampled host/symbiont scenarios. The different Monte Carlo samples can be executed in parallel, as they are independent, as can be the computation for all the gene trees.

3.5 Symbiont tree inference

Instead of considering that the symbiont tree is given, rooted and binary, we can infer or root the symbiont tree by amalgamation (David and Alm, 2011; Szöllősi *et al.*, 2013). Clade prior probabilities are computed from universal unicopy gene trees, and dynamic programming is used to compute the likelihood. In the backtrack a symbiont tree is sampled at the same time as the host/symbiont reconciliation scenario.

This amalgamation is also implemented for the symbiont/gene part, to account for gene tree being unrooted, and to be able to include uncertainty in gene tree topology, just like in 2-level reconciliationsJacox *et al.* (2016); Szöllősi *et al.* (2013).

3.6 Sequential and two level estimation of the likelihood

The Monte Carlo approach to the estimation of the likelihood can be computationally heavy, so we give a faster heuristic. Instead of sampling scenarios randomly like in the Monte Carlo, we select the one that maximises the marginal likelihood (Yang *et al.*, 2006). That is, at each step of the backtracking of the dynamic programming procedure we select the maximum likelihood position.

This approach is similar to the one of Stolzer *et al.*, 2015, but in a probabilistic setting, using marginal likelihood, and with a way of computing the inter transfer probabilities from the host/symbiont and symbiont/gene DTL reconciliation parameters.

3.7 Rates estimation and likelihood comparison

In our model, the data is the gene trees, and the free parameters are the three DTL probabilities of the symbiont/gene reconciliation. We consider the host/symbiont DTL parameters as fixed, *i.e.* estimated without knowing the data. This makes it possible to compare, based on the likelihood, our approach and a 2-level one (unaware of the host), because they have the same free parameters, and because they both define a probability distribution on the same space, the one of rooted binary trees whose leaves are the one of the symbiont tree, with multiplicity of each leaf being an integer (possibly zero).

So we first estimate the host/symbiont DTL parameters, as done in ALE (Szöllősi *et al.*, 2015b), with an expectation maximization method, and then once these parameters are fixed we take into account the genes and run our Monte Carlo or sequential approach multiple times to estimate rates for the symbiont/gene reconciliation with the same expectation maximisation method.

3.8 Free living symbionts

In host/symbiont reconciliation, it can happen that in the course of their evolutionary history, some symbiont have lived outside a host, or within an unknown host.

This is particularly important for us because we invoke unknown hosts in the case of inter host horizontal gene transfers. In order to consider these cases we added the possibility for a symbiont to be "Free living", meaning associated to no host.

We did that by adding the symbiont tree as a possible host tree, and matching the symbiont leaves with no host to themselves. In that way, we see transfer between free living as less likely than when a common host is known. A biological example where we need such a model is presented in the *Cinara* aphids example developed in the Results section (see Fig. 5).

3.9 Output format and solution visualization

Our implementation can output a sample of full scenarios, both for symbiont/genes and the corresponding host/symbiont reconciliations. The scenarios are given in RecPhyloXML, a common standard for reconciliation output endorsed by a significant part of the gene/species reconciliation community (Duchemin *et al.*, 2018). The scenarios can be visualised using Thirdkind https://crates.io/crates/thirdkind (Penel *et al.*, 2021), a reconciliation viewer that handles 3-level reconciliations. We also output events frequencies based on the reconciliation scenario sampling. Indeed we sample a number (100 by default) of symbiont/gene reconciliations and observe the frequency of each events over these replicates, thus getting an estimation of the posterior probability of events. It is this output we use when evaluating the capacity of our method to infer specific events in the following section, such as horizontal transfers receiver and donor symbionts, compared to the simulated scenario on simulated data, or to a previously proposed scenario in the *Cinara* aphids dataset.

4 Experimental results

4.1 Simulated dataset

Our probabilistic model can be used for simulation, however in order to test our method, we chose to use an exterior simulation framework. We used the available software Sagephy developed by Kundu and Bansal, 2019 for the case of protein domain, gene and species reconciliation, generating three embedded trees and that also support replacing transfers on top of additive ones. We used the parameters proposed by the same team in another article (Kordi *et al.*, 2019), as representative of small (D 0.133, T 0.266, L 0.266), medium (D 0.3, T 0.6, L 0.6) and high (D 0.6, T 1.2, L 1.2) transfer rates, without replacing transfers. The software enables to specify an inter transfer rate, corresponding to the probability for a gene transfer between different hosts. When a horizontal transfer is chosen during generation of the gene tree (inside a symbiont tree and knowing a host/symbiont reconciliation), the transfer is chosen to be an inter host one with the inter transfer rate. So an inter transfer rate of 0 corresponds to our inference model of only intra transfer, and of 1 corresponds to a case where transfers are only between symbionts in separate hosts.

We constructed two simulated datasets, one with a combination of the different rates for the DTL parameters, and one with only medium rates but with different rates of "inter" and "intra" transfers. For the first dataset, we used all 9 combinations of small, medium and high rates for the symbiont

generation and the gene generation, with only intra host gene transfer (*i.e.* an inter transfer rate of zero). For the second dataset, we used only medium rates for both symbiont and genes generation, but we used 6 inter transfer rates going from 0 to 1.

For both datasets, and for each set of rates, we generated 50 instances consisting of 1 host tree with 100 leaves, 1 symbiont tree and 5 gene trees, each generated in the pruned version of the other trees (branch that do not reach present are pruned before the generation of the next tree). We then kept only 8% of all host leaves to simulate unexhaustive sampling. This ended up to 399 instances for the first dataset and 226 instances for the second one, and at least 29 instances of 5 genes for each set of parameters.

We compared the results from three approaches. (1) The "2-level" heuristic which is a 2-level reconciliation between the gene and symbiont trees, ignorant of the host tree. (2) The "sequential" heuristic, which consists in computing the most likely host/symbiont DTL reconciliation and doing the symbiont/gene reconciliation, given that host/symbiont reconciliation. (3) The "Monte Carlo" method, summing the results of the gene reconciliations over 50 sampled host/symbiont reconciliation scenarios. We let our approaches estimate evolutionary rates.

We measured first the capacity of the three methods to infer the correct donor and recipient lineages of gene transfers (with precision and recall), and second, the likelihood they attribute to each symbiont/gene coevolution. Identifying the exact donor and recipient of simulated transfers is usually considered a hard task for reconciliation algorithms. Usually reconciliation studies are not evaluated with this strong criterion (Mykowiecka *et al.*, 2018), but look at the inference of ancestral characters (Wieseke *et al.*, 2015), the number of transfers (Szöllősi *et al.*, 2012), the ability to infer better trees (Bansal *et al.*, 2015), or the ability to map the correct event type to each gene node (Kordi *et al.*, 2019). We chose to look at the capacity to infer specific transfers because we feel that it is in this task that our model has the capacity to show its utility. It can infer more precise gene transfers because transfers are constrained by additional elements compared to other methods.

Our probabilistic reconciliation approaches output estimation of posterior probabilities of evolutionary events, so we used these probabilities as weight for our precision and recall definition. Denoting $L_{t,sim}$ the list of simulated transfers and $L_{t,obs}$ the list of observed transfers, and $P_{obs}(T)$ the estimation of our approach for the probability of transfer T.

$$\text{Precision} = \frac{\sum_{T \in L_{t,sim}} P_{obs}(T)}{\sum_{T \in L_{t,obs}} P_{obs}(T)} \text{ and Recall} = \frac{\sum_{T \in L_{t,sim}} P_{obs}(T)}{\sum_{T \in L_{t,sim}} 1}$$

Overall the Monte Carlo and sequential approaches give similar results, and better results (in particular for recall and to a lesser extent for precision) than the 2-level approach (Figure 3). Theoretically the Monte Carlo is more precise but computationally more costly, but in these simulations we could not discriminate them. However we can see that their results are qualitatively often different.

The difference between 2-level and 3-level methods are more important when the inter species gene transfer rate is low (Figure 4), which is expected because this rate reflects the dependence of the symbiont/gene reconciliation to the host/symbiont one. Note that when this rate is high (the gene transfers are less and less affected by the host), the 2-level reconciliation has a better likelihood. This shows that in a system with less multi-scale interdependence a 2-level reconciliation is more likely.

The relation between the behavior of the likelihood and the precision and recall of transfers, seen all along the simulations, is interesting: while the likelihood is accessible by inference on biological data, the precision and recall are not. So one can be used as a proxy for the other. We see that likelihood could be used to differentiate a dataset following a strong 3-level coevolution with no inter transfer, and one that does not, by comparing the results of the 2-level and 3-level approaches. With this idea it would even be possible to consider a parameter which would permit to go from the 2-level toward the 3-level model and estimate this parameter using likelihood, thus getting an index of coevolution for a given dataset.

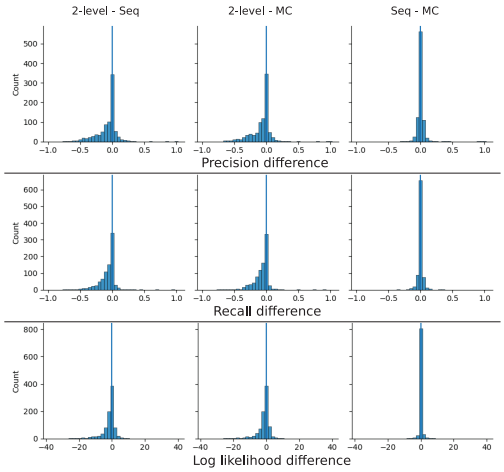


Fig. 3. Distribution of differences of precision, recall and likelihood for all combinations of two approaches, centered on 0. Distributions for all 874 gene families which undergo at least one horizontal transfer in the simulations, are drawn with histograms (no inter host gene transfer in the simulations).

4.2 *Cinara* aphids and symbiotic enterobacteria

A recent study on Cinara aphids enterobacteria systems (Manzano-Marín *et al.*, 2019) identified one host switch and two horizontal gene transfers, one intra-host from *Erwinina* to *Hamiltonella* and one inter-host from *Sodalis* to *Erwinia*. The genes transferred (thi) and some others (bioa,d,b) were first inherited through gene transfers, probably from Sodalis related symbionts. Moreover, those genes transferred are part of functions to complement the lack in the sap-feeding host nutrition. This shows coevolution between host and symbiont genes, more than host and symbiont, with a new endosymbiont acquiring the genes of an old one to sustain the host. We reproduced this scenario on Figure 5 (top left).

Gene trees including *Cinara* endosymbionts and some other enterobacteria's species (also including *Sodalis*, closest parent to the transferred genes), were available from the supplementary material (Manzano-Marín *et al.*, 2019), as were *Cinara* and their endosymbionts phylogenies, which phylogenies show exact correspondences. We chose a representative subset of the species present in the gene trees, and we used an exterior source for the phylogeny for the enterobacteria present in the gene trees using Annotree (Mendler *et al.*, 2019) (for the one that are not associated to *Cinara* aphids, and are thus "free living" in this setting). We used our three level reconciliation on the host tree and symbiont tree, using the possibility of our method to take into account these "free living" bacteria. As the host and symbiont (apart from the free living) are identical, we used the sequential heuristic.

We tested the capacity of the 3-level method compared to a 2-level one to detect the gene transfers. The intra transfer is retrieved in around 80 percent of the scenarios sampled by the 3-level method, and both are better retrieved than in the method that do not take the host into account (Figure 5 bottom right). An explanation is given in Figure 5 bottom left. An alternative transfer, in the other direction, from *Hamiltonella* to *Erwinia* is slightly more likely but the configuration of the host evolution supports the intra transfer.

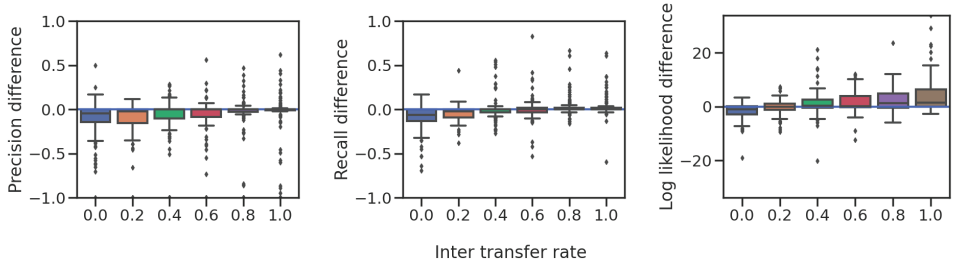


Fig. 4. The results of evaluation from simulations. We tested on a simulated dataset the sensitivity to the value of the inter host gene transfer probability in Sagephy. When the rate goes from 0 to 1 we go away from our model of only intra transfers toward a model with only inter transfer. Boxplots correspond to differences in precision, recall and log likelihood between the 2-level approach, unaware of the host, and the Sequential 3-level heuristic.

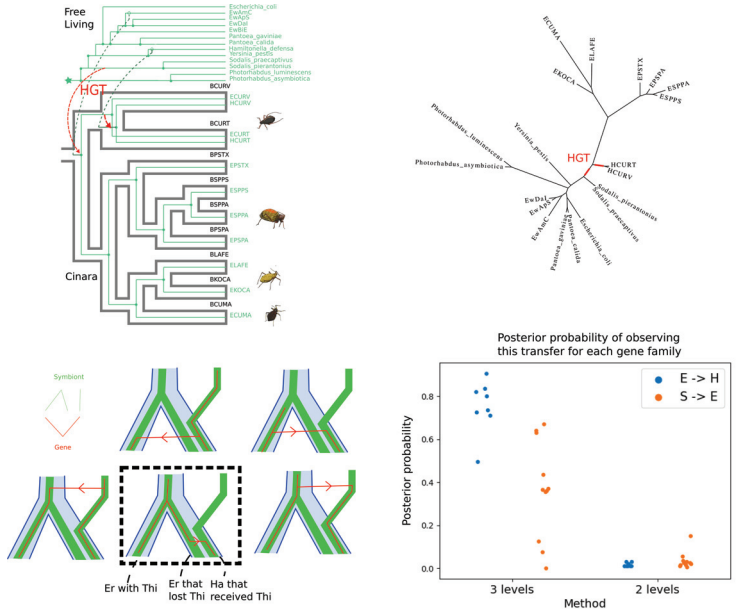


Fig. 5. The Cinara analyses. Top left: reconciliation of hosts (Cinara aphids) and symbionts (bacteria), with horizontal gene transfer positions (in red). Top right: the phylogenetic tree of one gene with the two transfers identified, illustrating the kind of evidence found for the two transfers. Bottom left: theoretical explanation of the difference between the results of a 2-level and a 3-level reconciliation method. The two top reconciliations are a bit more likely in a 2-level framework, as they require a single transfer while the bottom ones require a transfer and a loss, but one of the bottom one (with the dotted square) is better in a 3-level model, as it allows an intra-host transfer. Bottom right: support (a posteriori probability of the transfer, computed from its observed frequency in the reconciliation sample) for the identified HGTs, from Erwinia to Hamiltonella, and from Sodalis to Erwinia, for 3-level and 2-level reconciliations.

This exemplifies how multi-scale dependencies can only be captured by 3-level models.

4.3 Human populations and *Helicobacter pylori*

Helicobacter pylori is a bacterial symbiont of a significant proportion of humans, which has been supposed to be a marker of human migrations across the Earth (Achtman, 2016). Bacterial strains have been divided in different populations corresponding to geographical areas (Africa 1, Africa 2, Asia 2, East Asia, North East Africa, Europe) (Waskito and Yamaoka, 2019; Mégraud *et al.*, 2016).

The supposed coevolving complex made by humans, their bacterial symbiont and their genes makes it an ideal system for the host/symbiont/gene reconciliation method. In particular gene transfers should be more probable between *Helicobacter* strains if they are hosted by a same human population.

We collected available current strains of H. pylori from the NCBI which have a genetic population assigned by MLST allelic profile (Achtman *et al.*, 1999; Jolley *et al.*, 2018). A phylogenetic tree was built based on the concatenation of universal-unicopy genes (322 genes), and a sample of 113 strains representing the diversity of H. pylori in the old world (excluding strains from the Americas) was obtained using Treemmer (Menardo *et al.*, 2018).Then, 6 non pylori strains were added (*H. hepaticus, H. acinonychis, H. canadensis, H felis, H. bizzozeronii, H. cetorum*), as an external group.

In this study we considered the 1034 gene families, including 322 universal unicopy family, which displayed strains from the external group and from at least 3 continents.

We then considered four different population trees (host trees) containing the geographical areas as leaves, coherent with the scientific literature (Waskito and Yamaoka, 2019; Mégraud *et al.*, 2016). 322 universal unicopy gene trees were used, and the strain (symbiont) tree was amalgamated from gene trees with the population trees as a guide (see subsection 3.5). As strains were much more numerous than populations, and subject to a more complex diversification than DTL events, we allowed an additional event, named I, that consists in a duplication followed by a speciation and loss of one of the copies, with a specific rate, inferior to the combination of these three events. This event allows a strain to be present in a population and one of its descendants, and is used as one of the default events in biogeography reconciliation frameworks (Ree *et al.*, 2005).

We then applied our sequential approach and compared the likelihood of the gene/strains aware of the host reconciliation to compare the population trees. The results are depicted in Figure 6. In the array of the figure, the likelihood of the system is written, according to the population tree chosen (in columns), divided into two components: the likelihood of the population/strain comparison, and the likelihood of the gene/strain comparison. This means that this population tree is more likely given the model, the method and the used data. Assessing the robustness of the result would require a sensibility study which is out of the scope of this mainly methodological contribution.

We also present a reconciliation scenario in Figure 6 for the host tree with the maximal likelihood. We see the host tree and the amalgamated strain tree reconciled (I events are represented as transfers from a parent node to one of its child). On top of these two embedded trees red lines represent the aggregation of gene transfers depending on the host of the donor and receiver strains. The opacity of the transfer lines are proportional to the number of times a certain kind of transfer is observed across the 1034 gene families in one sampled scenario.

Interestingly in all our experiments the 2-level likelihood is higher than the 3-level likelihood (reconciling the amalgamated strain tree with the gene trees), a situation found in the simulations to correspond with genes coevolving with the symbiont but with a limited impact of the host on the coevolution. It seems that the 3-level model does not capture a great extent of coevolution between humans and *Helicobacter pylori*. Further investigations are needed to know if this comes from a default of the method, or an overestimation of the coevolutionary pattern between humans and their symbiont in the literature.

5 Discussion

In a review on horizontal gene transfer in host symbiont systems (Wijayawardena *et al.*, 2013) the authors highlight the need of plurality of evidence to robustly assess the existence of transfers. Evidence can be of multiple types, gene trees, donor receiver ecology, or host symbiont association. We provide a framework were these multiple evidence can be gathered, and the proof of concept that it can work, on *Cinara* aphids and their enterobacteria.

Our method uses a probabilistic framework that enables rate estimation, tree inference, tree comparison and model comparison. We also introduced a method to compute the inter transfer rate from the intra transfer one and the modeling of ghost lineages in the host symbiont reconciliation. We introduced a Monte Carlo approach that enables to estimate event probabilities and likelihood, by sampling through multiple host symbiont scenarios in a double DTL model.

While our intuition is that the Monte Carlo approach is more robust than the sequential one, notably in cases where gene events happen around uncertain host symbiont reconciliation nodes, our evaluation on simulated data did not show a big difference in most cases. We think that in biological data, we can expect more interaction between the events of the host symbiont reconciliation and the ones of the gene symbiont one, which are independent in our simulation.

All these features deserve further tests to know their domain of validity and to draw biological conclusions. In particular, the inference of the symbiont tree, with the use of amalgamation, from an input distribution of universal unicopy gene tree would deserve to be tested against other standard methods as concatenate or species tree reconstruction with 2-level reconciliation model as it is implemented in SpeciesRax (Morel *et al.*, 2021).

An interesting future direction in this line would be to construct, instead of a symbiont tree, compartment trees, which would depict the coevolution of genes that are not necessarily in the same species.

More generally, the model is not bound to host/symbiont/gene systems, but any set of three nested coevolving entities can be studied with it: species/gene/protein domain as it was done in previous studies (Stolzer *et al.*, 2015; Li and Bansal, 2018; Muhammad *et al.*, 2018), or geography/species/gene, and so on. As the scales of biological observation are probably infinite, so are the combination of three nested scales.

Biologically significant studies using these methods have yet to be undertaken, but we think that our, and others, methodological development are a step forward, and that the examples presented in this article can show the possibility of such methods. Biological systems that fit into this multi-scale coevolution framework are more and more numerous.

6 Acknowledgements

This work was performed using the computing facilities of the CC LBBE/PRABI.

7 Funding

This work was supported by the French National Research Agency (Grant ANR-19-CE45-0010 Evoluthon).

References

Achtman, M. (2016). How old are bacterial pathogens? *Proceedings of the Royal Society B: Biological Sciences*, **283**(1836), 20160990.

Achtman, M., Azuma, T., Berg, D. E., Ito, Y., Morelli, G., Pan, Z. J., Suerbaum, S., Thompson, S. A., van der Ende, A., and van Doorn, L. J. (1999). Recombination and clonal groupings within Helicobacter pylori from different geographical regions. *Molecular Microbiology*, **32**(3), 459–470.

Anselmetti, Y., El-Mabrouk, N., Lafond, M., and Ouangraoua, A. (2021). Gene tree and species tree reconciliation with endosymbiotic gene transfer. *Bioinformatics*, **37**(Supplement_1), i120–i132.

Bailly-Bechet, M., Martins-Simões, P., Szöllősi, G. J., Mialdea, G., Sagot, M.-F., and Charlat, S. (2017). How Long Does Wolbachia Remain on Board? *Molecular Biology and Evolution*, **34**(5), 1183–1193.

Bansal, M. S., Banay, G., Gogarten, J. P., and Shamir, R. (2011). Detecting Highways of Horizontal Gene Transfer. *Journal of Computational Biology*, **18**(9), 1087–1114.

Bansal, M. S., Alm, E. J., and Kellis, M. (2012). Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. *Bioinformatics*, **28**(12), i283–i291. Publisher: Oxford Academic.

Bansal, M. S., Wu, Y.-C., Alm, E. J., and Kellis, M. (2015). Improved gene tree error correction in the presence of horizontal gene transfer. *Bioinformatics*, **31**(8), 1211–1218.

Boussau, B. and Scornavacca, C. (2020). Reconciling Gene trees with Species Trees. In C. Scornavacca, F. Delsuc, and N. Galtier, editors, *Phylogenetics in the Genomic Era*, pages 3.2:1–3.2:23. No commercial publisher | Authors open access book.

Chan, Y.-b., Ranwez, V., and Scornavacca, C. (2017). Inferring incomplete lineage sorting, duplications, transfers and losses with reconciliations. *Journal of Theoretical Biology*, **432**, 1–13.

Pylori putative population tree				
Host/symbiont Log likelihood	- 308	- 320	- 322	- 320
Symbiont/gene Log likelihood	-653x10^3	-658x10^3	-658x10^3	-676x10^3

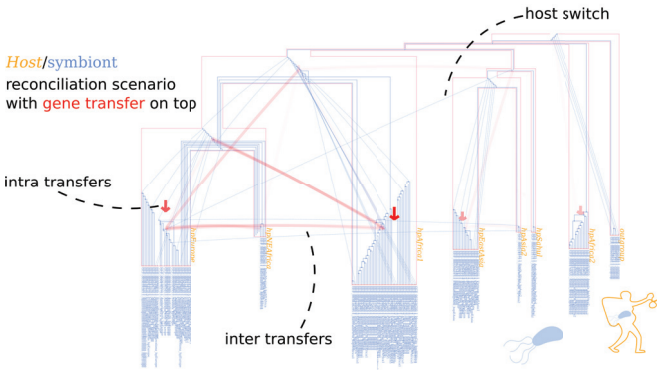


Fig. 6. Top: Log likelihood of the different population trees. Bottom: The representation with ThirdKind of one possible reconciliation scenario of Helicobacter pylori strain tree and the population tree maximizing likelihood. Aggregated gene transfers are depicted on top of the DTL reconciliation, with the opacity corresponding to the number of time the transfers were seen across the 1034 gene families.

Charleston, M. and Libeskind-Hadas, R. (2014). Event-based cophylogenetic comparative analysis. In *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology*, pages 465–480. Springer Berlin Heidelberg.

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

David, L. A. and Alm, E. J. (2011). Rapid evolutionary innovation during an Archaean genetic expansion. *Nature*, **469**(7328), 93–96.

Donati, B., Baudet, C., Sinaimeri, B., Crescenzi, P., and Sagot, M.-F. (2015). EUCALYPT: efficient tree reconciliation enumerator. *Algorithms for Molecular Biology*, **10**(1), 3.

Doyon, J.-P., Ranwez, V., Daubin, V., and Berry, V. (2011). Models, algorithms and programs for phylogeny reconciliation. *Briefings in Bioinformatics*, **12**(5), 392–400.

Duchemin, W., Daubin, V., and Tannier, E. (2015). Reconstruction of an ancestral Yersinia pestisgenome and comparison with an ancient sequence. *BMC Genomics*, **16**(10), S9.

Duchemin, W., Anselmetti, Y., Patterson, M., Ponty, Y., Bérard, S., Chauve, C., Scornavacca, C., Daubin, V., and Tannier, E. (2017). DeCoSTAR: Reconstructing the Ancestral Organization of Genes or Genomes Using Reconciled Phylogenies. *Genome Biology and Evolution*, **9**(5), 1312–1319.

Duchemin, W., Gence, G., Arigon Chifolleau, A.-M., Arvestad, L., Bansal, M. S., Berry, V., Boussau, B., Chevenet, F., Comte, N., Davín, A. A., Dessimoz, C., Dylus, D., Hasic, D., Mallo, D., Planel, R., Posada, D., Scornavacca, C., Szöllősi, G., Zhang, L., Tannier, E., and Daubin, V. (2018). RecPhyloXML: a format for reconciled gene trees. *Bioinformatics*, **34**(21), 3646–3652.

Felsenstein, J. (2004). *Inferring Phylogenies*. Oxford University Press.

Fournier, G. P., Huang, J., and Gogarten, J. P. (2009). Horizontal gene transfer from extinct and extant lineages: biological innovation and the coral of life. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **364**(1527), 2229–2239.

Jacox, E., Chauve, C., Szöllősi, G. J., Ponty, Y., and Scornavacca, C. (2016). ecceTERA: comprehensive gene tree-species tree reconciliation using parsimony. *Bioinformatics (Oxford, England)*, **32**(13), 2056–2058.

Jolley, K. A., Bray, J. E., and Maiden, M. C. J. (2018). Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Research*, **3**, 124.

Kordi, M., Kundu, S., and Bansal, M. S. (2019). On Inferring Additive and Replacing Horizontal Gene Transfers Through Phylogenetic Reconciliation. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '19, pages 514–523, Niagara Falls, NY, USA. Association for Computing Machinery.

Kundu, S. and Bansal, M. S. (2019). SaGePhy: an improved phylogenetic simulation framework for gene and subgene evolution. *Bioinformatics*, **35**(18), 3496–3498.

Li, L. and Bansal, M. S. (2018). An Integer Linear Programming Solution for the Domain-Gene-Species Reconciliation Problem. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '18, pages 386–397, New York, NY, USA. ACM. event-place: Washington, DC, USA.

Li, L. and Bansal, M. S. (2019a). An Integrated Reconciliation Framework for Domain, Gene, and Species Level Evolution. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **16**(1), 63–76.

Li, L. and Bansal, M. S. (2019b). Simultaneous Multi-Domain-Multi-Gene Reconciliation Under the Domain-Gene-Species Reconciliation Model. In Z. Cai, P. Skums, and M. Li, editors, *Bioinformatics Research and Applications*, Lecture Notes in Computer Science, pages 73–86. Springer International Publishing.

Manzano-Marín, A., D'acier, A. C., Clamens, A.-L., Orvain, C., Cruaud, C., Barbe, V., and Jousselin, E. (2019). Serial horizontal transfer of vitamin-biosynthetic genes enables the establishment of new nutritional symbionts in aphids' di-symbiotic systems. *The ISME Journal*, pages 1–15.

Martínez-Aquino, A. (2016). Phylogenetic framework for coevolutionary studies: a compass for exploring jungles of tangled trees. *Current Zoology*, **62**(4), 393–403.

Menardo, F., Loiseau, C., Brites, D., Coscolla, M., Gygli, S. M., Rutaihwa, L. K., Trauner, A., Beisel, C., Borrell, S., and Gagneux, S. (2018). Treemmer: a tool to reduce large phylogenetic datasets with minimal loss of diversity. *BMC Bioinformatics*, **19**(1), 164.

Mendler, K., Chen, H., Parks, D. H., Lobb, B., Hug, L. A., and Doxey, A. C. (2019). AnnoTree: visualization and exploration of a functionally annotated microbial tree of life. *Nucleic Acids Research*, **47**(9), 4442–4448.

Menet, H., Daubin, V., and Tannier, E. (2021). Phylogenetic reconciliation.

Morel, B., Schade, P., Lutteropp, S., Williams, T. A., Szöllősi, G. J., and Stamatakis, A. (2021). SpeciesRax: A tool for maximum likelihood species tree inference from gene family trees under duplication, transfer, and loss. Technical report. Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article.

Muhammad, S. A., Sennblad, B., and Lagergren, J. (2018). Species tree-aware simultaneous reconstruction of gene and domain evolution. *bioRxiv*, page 336453.

Mykowiecka, A., Muszewska, A., and Górecki, P. (2018). Inferring time-consistent and well-supported horizontal gene transfers. pages 79–83. IEEE Computer Society.

Mégraud, F., Lehours, P., and Vale, F. F. (2016). The history of Helicobacter pylori: from phylogeography to paleomicrobiology. *Clinical Microbiology and Infection*, **22**(11), 922–927.

Nakabachi, A., Ueoka, R., Oshima, K., Teta, R., Mangoni, A., Gurgui, M., Oldham, N. J., van Echten Deckert, G., Okamura, K., Yamamoto, K., Inoue, H., Ohkuma, M., Hongoh, Y., Miyagishima, S.-y., Hattori, M., Piel, J., and Fukatsu, T. (2013). Defensive Bacteriome Symbiont with a Drastically Reduced Genome. *Current Biology*, **23**(15), 1478–1484.

Nakhleh, L. (2013). Computational approaches to species phylogeny inference and gene tree reconciliation. *Trends in ecology and evolution*, **28**, 719–728.

Nikoh, N., Hosokawa, T., Moriyama, M., Oshima, K., Hattori, M., and Fukatsu, T. (2014). Evolutionary origin of insect–Wolbachia nutritional mutualism. *Proceedings of the National Academy of Sciences of the United States of America*, **111**(28), 10257–10262.

Penel, S., Menet, H., Tricou, T., Daubin, V., and Tannier, E. (2021). Thirdkind: displaying phylogenetic encounters beyond 2-level reconciliation. *To appear in Bioinformatics*.

Penz, T., Schmitz-Esser, S., Kelly, S. E., Cass, B. N., Müller, A., Woyke, T., Malfatti, S. A., Hunter, M. S., and Horn, M. (2012). Comparative Genomics Suggests an Independent Origin of Cytoplasmic Incompatibility in Cardinium hertigii. *PLOS Genetics*, **8**(10), e1003012.

Rasmussen, M. D. and Kellis, M. (2012). Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome Research*, **22**(4), 755–765.

Ravenhall, M., Škunca, N., Lassalle, F., and Dessimoz, C. (2015). Inferring Horizontal Gene Transfer. *PLOS Computational Biology*, **11**(5), e1004095.

Ree, R. H. and Smith, S. A. (2008). Maximum Likelihood Inference of Geographic Range Evolution by Dispersal, Local Extinction, and Cladogenesis. *Systematic Biology*, **57**(1), 4–14.

Ree, R. H., Moore, B. R., Webb, C. O., and Donoghue, M. J. (2005). A LIKELIHOOD FRAMEWORK FOR INFERRING THE EVOLUTION OF GEOGRAPHIC RANGE ON PHYLOGENETIC TREES. *Evolution*, **59**(11), 2299–2311.

Ronquist, F. (1997). Dispersal-Vicariance Analysis: A New Approach to the Quantification of Historical Biogeography. *Systematic Biology*, **46**(1), 195–203.

Santichaivekin, S., Yang, Q., Liu, J., Mawhorter, R., Jiang, J., Wesley, T., Wu, Y.-C., and Libeskind-Hadas, R. (2020). eMPRess: a systematic cophylogeny reconciliation tool. *Bioinformatics*, (btaa978).

Sapp, J. (1994). *Evolution by association*. Oxford University Press.

Sjöstrand, J., Tofigh, A., Daubin, V., Arvestad, L., Sennblad, B., and Lagergren, J. (2014). A Bayesian Method for Analyzing Lateral Gene Transfer. *Systematic Biology*, **63**(3), 409–420.

Sonnenburg, J. L. and Sonnenburg, E. D. (2019). Vulnerability of the industrialized microbiota. *Science*, **366**(6464), eaaw9255.

Stolzer, M., Lai, H., Xu, M., Sathaye, D., Vernot, B., and Durand, D. (2012). Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics*, **28**(18), i409–i415.

Stolzer, M., Siewert, K., Lai, H., Xu, M., and Durand, D. (2015). Event inference in multidomain families with phylogenetic reconciliation. *BMC Bioinformatics*, **16**(14), S8.

Szöllosi, G. J., Tannier, E., Lartillot, N., and Daubin, V. (2013). Lateral gene transfer from the dead. *Systematic Biology*, **62**(3), 386–397.

Szöllősi, G. J., Boussau, B., Abby, S. S., Tannier, E., and Daubin, V. (2012). Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proceedings of the National Academy of Sciences*, **109**(43), 17513–17518. Publisher: National Academy of Sciences Section: Biological Sciences.

Szöllősi, G. J., Rosikiewicz, W., Boussau, B., Tannier, E., and Daubin, V. (2013). Efficient Exploration of the Space of Reconciled Gene Trees. *Systematic Biology*, **62**(6), 901–912.

Szöllősi, G. J., Davín, A. A., Tannier, E., Daubin, V., and Boussau, B. (2015a). Genome-scale phylogenetic analysis finds extensive gene transfer among fungi. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **370**(1678), 20140335. Publisher: Royal Society.

Szöllősi, G. J., Tannier, E., Daubin, V., and Boussau, B. (2015b). The Inference of Gene Trees with Species Trees. *Systematic Biology*, **64**(1), e42–e62.

Waskito, L. A. and Yamaoka, Y. (2019). The Story of Helicobacter pylori: Depicting Human Migrations from the Phylogeography. In S. Kamiya and S. Backert, editors, *Helicobacter pylori in Human Diseases: Advances in Microbiology, Infectious Diseases and Public Health Volume 11*, Advances in Experimental Medicine and Biology, pages 1–16. Springer International Publishing, Cham.

Wieseke, N., Hartmann, T., Bernt, M., and Middendorf, M. (2015). Cophylogenetic Reconciliation with ILP. *IEEE/ACM transactions on computational biology and bioinformatics*, **12**(6), 1227–1235.

Wijayawardena, B. K., Minchella, D. J., and DeWoody, J. A. (2013). Hosts, parasites, and horizontal gene transfer. *Trends in Parasitology*, **29**(7), 329–338.

Wu, Y. (2012). Coalescent-Based Species Tree Inference from Gene Tree Topologies Under Incomplete Lineage Sorting by Maximum Likelihood. *Evolution*, **66**(3), 763–775.

Yang, Z. *et al.* (2006). *Computational molecular evolution*, volume 284. Oxford University Press Oxford.

Zhaxybayeva, O. and Gogarten, J. P. (2004). Cladogenesis, coalescence and the evolution of the three domains of life. *Trends in genetics: TIG*, **20**(4), 182–187.

2.1.2 Computation time and tractability

The datasets presented in the article give a good idea of the size of the data we can consider with our new method.

The computation on the *Cinara* aphids dataset, with a size of 25 leaves for the symbiont tree, 9 leaves for the host, and 13 gene families takes around 3 minutes on a single core on a laptop, including the rate estimation steps. It is a dataset on which it would be possible to use the Monte Carlo approach.

The *pylori* dataset is bigger, symbiont has 119 leaves, host 7 leaves, and there are 1034 genes, 322 of which have 119 leaves. The reconciliation, with fixed rates (without the rates estimation steps) took a bit less than a day in parallel with 8 cores.

2.1.3 Monte Carlo and Sequential heuristic

In the paper, we see that the Monte Carlo and the Sequential heuristic give similar results when evaluating the capacity to retrieve simulated transfers donors and receivers. However, we mention that the faster Sequential heuristic may not be as robust as the Monte Carlo one. The results are often different (see [124] for a similar discussion in a parsimony model, with an example where giving the best host symbiont reconciliation forbid any gene and symbiont scenarios, in a case where inter horizontal gene transfer are forbidden). In figure 2.1 we present another example, with this time an emphasis on the "not continuous" aspect of the Sequential heuristic in regard to the host and symbiont reconciliation events rates.

Heuristic	Gene transfer A	Gene transfer B
Fixed rates **T 0.006** D 0.1 L 0.1		
Monte Carlo	0.43	0.27
Sequential	0.90	< 0.05
2-level	0.18	0.21
Fixed rates **T 0.005** D 0.1 L 0.1		
Monte Carlo	0.35	0.33
Sequential	< 0.05	0.49
2-level	0.19	0.23

Table 2.1: Comparison of the support for the two gene transfer scenarios in the example presented in figure 2.1.

A small change in the transfer rate of the host and symbiont makes a big difference for the gene and symbiont reconciliation with the Sequential heuristic, but a small one for the Monte Carlo one, see the results in table 2.1.

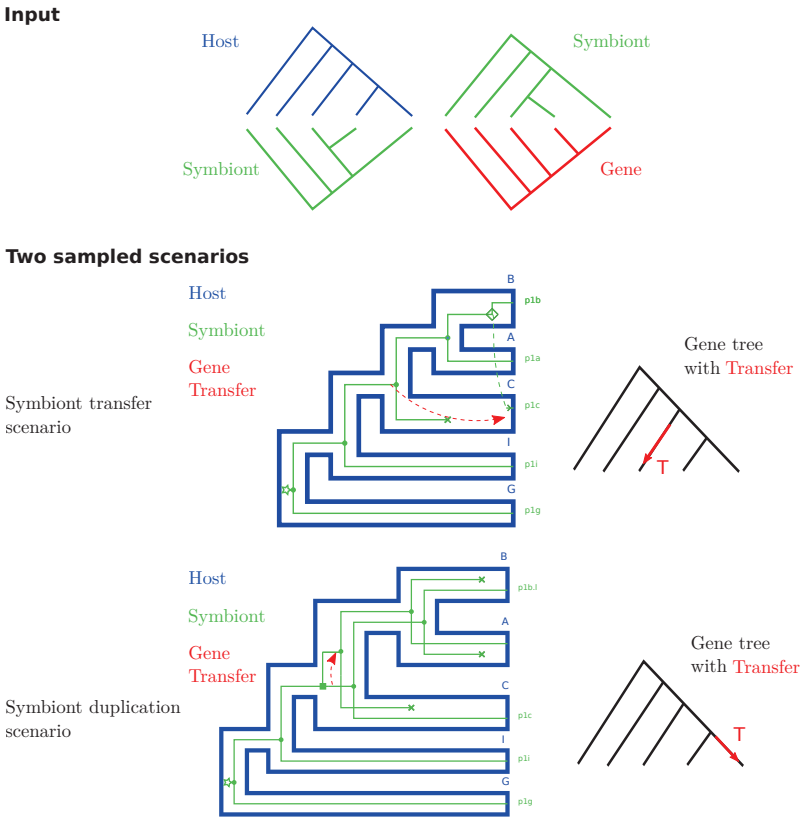


Figure 2.1: An example of input and two possible host symbiont and gene symbiont scenarios. We compare the support for two gene transfer scenarios with the Monte Carlo and the Sequential heuristic, depending on the rates of host and symbiont events. One of the gene transfer scenario is more likely with the first host and symbiont reconciliation, and inversely for the other. The results are presented in table 2.1

2.1.4 Likelihood comparison between 2-level and 3-level models

In the article we wrote that we could test our model and compare it notably to a 2-level one. It seems counterintuitive at first, as we have more information and more parameters in the 3-level case. To compare what is comparable, we consider two quantities, denoting the gene tree G, the symbiont tree S and the host tree H, the DTL rates for reconciliation in the host θ_H and in the symbiont θ_S, P_3 refers to probability in the 3-level model, P_2 in the 2-level one:

- $P_3(G|S, H, \theta_H, \theta_S)$
- $P_2(G|S, \theta_S)$

We compare the probability to generate the gene tree, our data, in two different frameworks. We have to identify the free parameters that will have to be estimated depending on the data. Symbiont and host trees are fixed beforehand. We estimate θ_H to maximum likelihood for the host and symbiont reconciliation, so without knowing the data, we can thus consider this parameter as fixed beforehand. Thus, we have the same set of free parameters in the two models: θ_S.

This is possible because the difference between intra and inter transfers depends on θ_H and not on an additional parameter. We then can compare these two probabilities by taking the maximum likelihood estimation of the same number of free parameters.

Likelihood test and free parameters

We have already seen how maximum likelihood and bayesian inference differ in the choice or not of a prior, but in practice, we rarely can compute directly $P(D|M)$ the probability of a data D in a model M, as M is often parameterized with a set of parameters θ. The problem with such a probability computation is that often we need all parameters fixed in θ to be able to compute the probability. To access the probability, we should then sum or integrate over all possible values for θ, which we cannot often do in practice. We also would have to deal with the prior on θ. $P(D|M) = \sum_\theta P(D|M, \theta)P(\theta|M)$ So what we do, is that we estimate that weighted sum by the maximum likelihood rates: $\max_\theta P(D|M, \theta)$.
If one model is included in another, for instance, with a parameter that makes it possible to go back to it, it seems certain that it will have a better maximum likelihood than the other. And more generally, the more we have parameters, the more we can fit data and have a high probability of having generated them. Some criteria take into account these differences to proceed

to likelihood tests. A well-known one is Akaike Information Criterion with the following formula: $2k - 2\ln(L)$ with L the maximum likelihood and k the number of free parameters. The main result associated to that formula is that the bias of the log likelihood as an asymptotical estimator of a measure of the divergence between the "true" distribution and the one given by the model (using Kullback-Leibler divergence), is, in some technical conditions, the number of parameters estimated in the model[33]. Corrections to small datasets, or other version based on other paradigms exists. Finding the number of free parameters can also be a difficult question, notably in our models where one parameter can be a tree.

It is important to notice that, at a fixed host and symbiont tree, the 3-level model does not include the 2-level one in a hierarchical manner, *i.e.* there does not exist a set of parameters of the 3-level model that reproduces the 2-level one. In the 3-level model, we have symbiont's branch-dependent transfer rates, depending on the host tree. Thus a model that would contain both our 3-level and 2-level is a 2-level one with branch-dependent transfer rates.

We also tried our likelihood 2-level against 3-level comparison on the *Cinara* aphids dataset. We restricted the dataset to the symbiont with a host to avoid issues with our model of free living symbiont. We have a better likelihood in the 2-level model (log likelihood of -102), even though the 3-level (log likelihood of -117) approach scenario is closer to the one identified in the paper. Figure 2.2 and figure 2.3 present the scenarios inferred by the two models.

The main problem here is that the intra transfer implies an additional loss (see the figure in the article). When testing only with genes for which all nodes are resolved the same way that the species tree, and the only difference is the transfers between *Erwinia* and *Hamiltonella*, the 2-level likelihood is better than the 3-level one. It is more likely to do an intra transfer in a 3-level model than a transfer in a 2-level model. The associated probabilities are $\frac{p^T}{|N_e|}$ and $\frac{p^T}{|S|}$ where N_e is the number of neighbors symbiont of the symbiont e in its host, and $|S|$ the total number of symbionts. However, it does not say that an intra transfer and another event required to get that transfer is more likely than the 2-level transfer. In our model of dated DTL, it seems in fact more likely to generate the data of this example in 2-level than in 3-level. Another model of losses could be useful, for instance to model losses as more likely when there are gene redundancy.

Another problem for the actual use of the likelihood comparison is that the cost of inter transfer is prohibitive, especially when transfers are used as way to correct errors in the input tree more than being real transfers. Using the difference between the 2-level and 3-level likelihoods as a marker of coevolution is not robust to un-

certainties in the gene trees. In figure 2.2 we see that we have some supplementary transfers, often between species that are closely related in the tree. Those transfers can be associated to poorly supported nodes in the gene trees. Uncertainty in the trees is a difference of real data compared to our simulated dataset, where we use directly the simulated trees, and for which we see that likelihood can differentiate 2-level and 3-level models. An idea to get over this problem could be to use amalgamation as it was developed to consider gene tree uncertainty.

2.1.5 Tree comparison and parameters estimation

If we look for both the host tree and the evolutionary rates, as we do with the *Helicobacter pylori* example, is it a good approach to take, for each tree, the maximum likelihood rates and then compare the likelihood associated? An issue with that approach is that small changes in rates can make big changes in likelihood. As in my experience with this dataset the host tree comparision is highly dependent on the rates, and I was not certain of the convergence and the robustness of the rates estimation process, I thought it was wiser to rely on fixed rates. When inferring a tree, (instead of comparing some fixed trees), in Generax for instance, topologies and rates are optimized alternatively. Topologies are compared at the same rates, and then good rates for the best topology are estimated, and then again a test of different topology, etc.

2.1.6 Marginal and joint maximum likelihood

In the paper, we write that we can choose to sample the maximum likelihood scenario. It is possible to compute that scenario, but it relies on a modification of the dynamic programming forward pass (not only in the backtrack). The resulting algorithm, with maxima in place of additions of probabilities, is similar to the parsimony version that searches for the solution of minimum cost.

A maximum likelihood scenario is one where the combination of the events is the most likely. However, it is not the scenario made up of the most likely events. A notion coming from ancestral state reconstruction (page 121 in Yang Computational Molecular Evolution [247]), defines scenarios associated with maximizing the probability of events instead of complete scenarios. We can maximize the score at a given position, for us, the matching of a given node, instead of the score of the complete scenario. These two kinds of likelihood are called marginal likelihood (given position) and joint likelihood (scenario).

To get the maximum marginal likelihood scenario, we first choose the maximum likelihood position for the root of the symbiont tree in the host tree, then we take the maximum likelihood position knowing that choice for its two children nodes, and

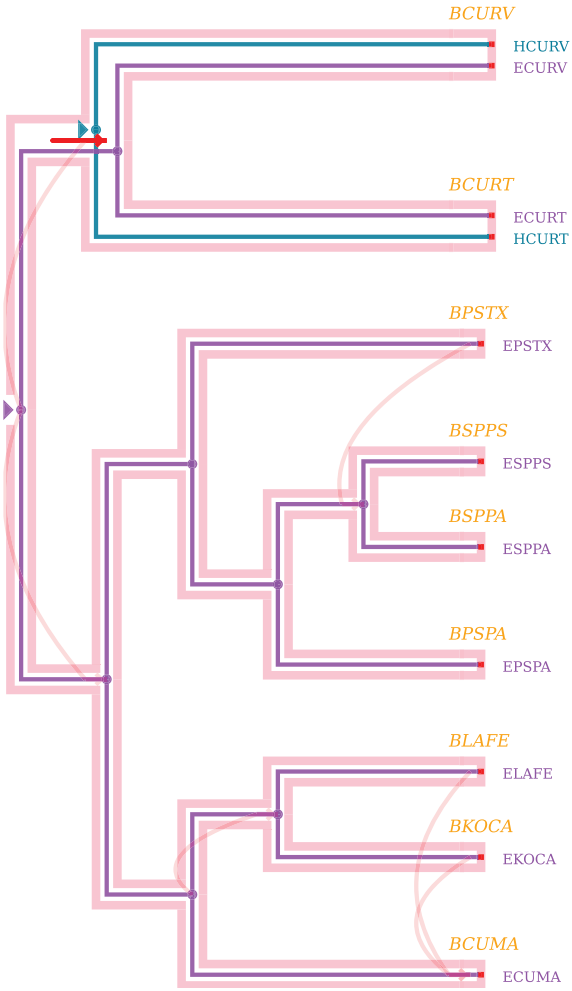
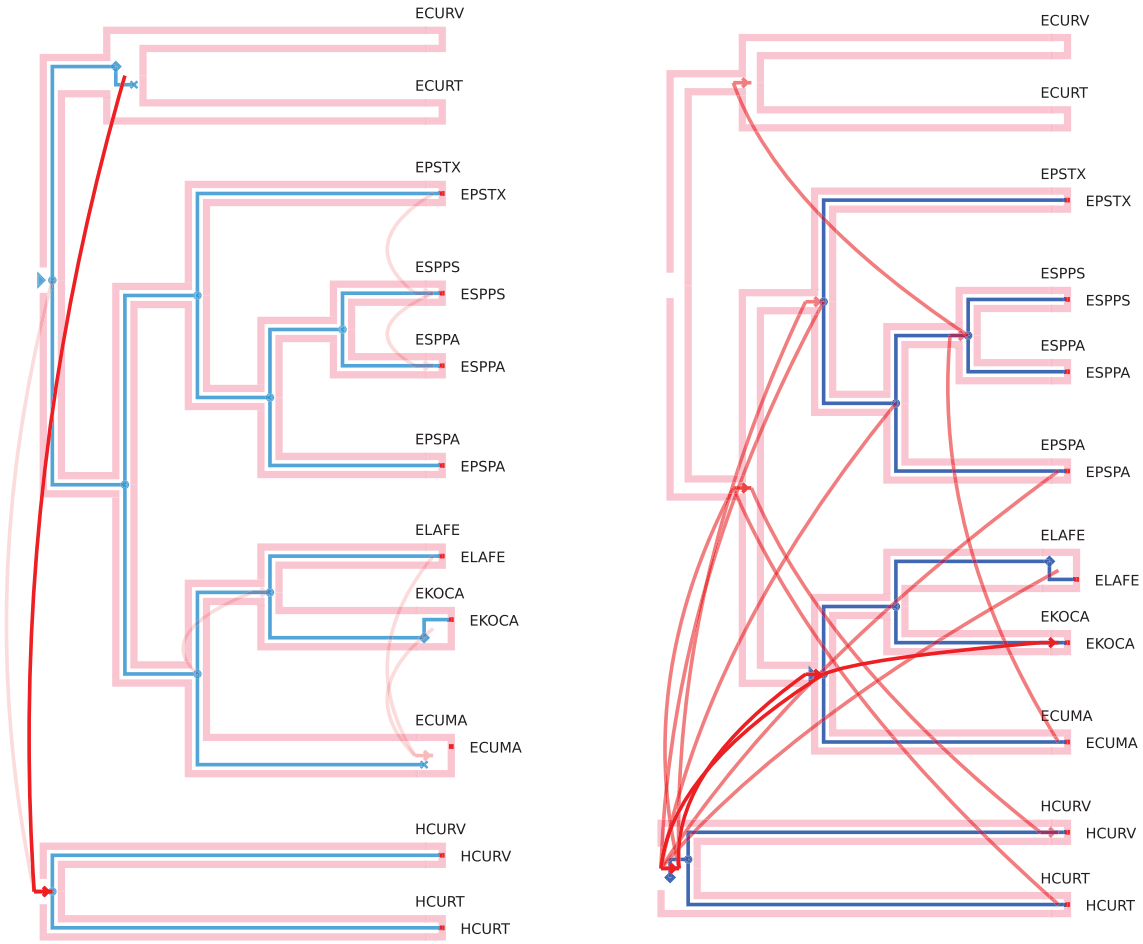


Figure 2.2: Results of a 3-level reconciliation. Host and symbiont scenario between the *Cinara* host, given by its *Buchnera* symbiont, and the *Erwinia* and *Hamiltonella* endosymbionts. Horizontal gene transfers are represented on top of this reconciliation, with opacity depending on the number of times the transfer is seen across the different gene families. The transfer between Erwinia and Hamiltonella inside their common host is represented as a straight line, depicting a transfer inside a host. Other transfers, specific to single gene families, are also inferred. The figure was generated using Thirdkind, a viewer we present in section 2.4.



3-level reconciliation

2 level reconciliation

Figure 2.3: Symbiont and gene reconciliation with the 3-level and 2-level models. One gene family is represented, and, as in figure 2.2 are aggregated over all gene families. The figure was generated using Thirdkind, a viewer we present in section 2.4.

so on. This approach uses the same dynamic programming as in the coevolution likelihood computation and sampling part, with modified backtracking. At each step in the backtrack, we choose the most likely event. In a way, it biases the reconciliation to give more importance to the most ancestral correspondences, as they are the first to be considered in the backtracking.

In our implementation we used the marginal likelihood. Using that likelihood is also interesting in that it is not the translation of the most parsimonious reconciliation.

Another possible approach to the transfer rate inference, that we did not present in the paper, and that does not rely on fixing the host/symbiont reconciliation one gene/symbiont reconciliation, is to assume all symbionts host matchings to be independent. We first compute symbionts' probability of matching with the host nodes by taking the frequency of match in a sample of host/symbiont reconciliation scenarios. Then we have:

$$P^T(i \to j) = p^T \sum_{h \in H} P_{i,h} P_{j,h} \tag{2.1}$$

SECTION 2.2

Coronaviruses and the disappearing prior

In this section I present the application of our method to a coronaviruses dataset, with a technical point, the behavior of the prior on the host/symbiont reconciliation compared to the gene reconciliations, presented in the preceding subsection.

2.2.1 The disappearing prior

We were puzzled at first about using 3-level likelihood to compare symbiont trees. The problem we had is that our probability was decomposed into two parts in our total probability sum: $P(r_{S,H}|H)$ and $P(G|r_{S_H}, H)$. The second term considers all the genes and will make the first one negligible, with the number of genes increasing.

It is an argument similar to the one developed in Felsenstein's Inferring phylogenies [78], page 249. When comparing two trees, the priors on the tree topologies are not too important as their difference will be crushed by the average site evolution likelihood difference at the power of the number of nucleotides.

It could be possible to use the agreement between the likelihood of the host symbiont scenarios and host symbiont and gene scenarios as a validation of the method. As the true one for both should be likely in the model.

2.2.2 Coronaviruses and host coevolution

Pandemic and research

This thesis was conducted during the global coronavirus pandemic for more than two-thirds of its time. Researchers from the academic world started to use their methods to understand the complex, multi-faceted problems that we were facing.
In our team for instance, some members applied bayesian methods to test the impact of maintaining the French election on the epidemic dynamic[68]. Congresses, such as the Complex System international one that took place in Lyon in 2021, dedicated whole symposia to models of the pandemic. The sudden availability of data on the subject was an interesting opportunity for us to work on such a multi-level system, with genes, viruses and hosts. Moreover, the proposed scenario for the emergence of the virus involves events captured by reconciliation: recombination and host switches[199].
Like in evolution and natural selection, at least as far as I know, the characters to be selected need to be present before the selective pressure. The methods researchers applied were firmly established ones, and no incredibly new method appeared out of nowhere in a week to give answers to our questions and, more importantly, an action plan. As a second global pressure is on us, we might need to keep that in mind.

We found two studies involving coronaviruses with phylogenetic reconciliation, the first one dating prior to the pandemic.

Anthony *et al.* [6] used reconciliation, Jane and Corepa software, to investigate the coevolution of bat and their coronaviruses, showing that host switch was the dominant force, but that cospeciation was also present and in a sufficient proportion to find a coevolutionary signal.

The second study used reconciliation to investigate host switch, but from a gene perspective, with an idea similar to our model. They assumed that an HGT between two viruses is a sign of host switch between different mammalian hosts [80]. Nevertheless, they did not propose a model or an automated approach.

I worked on a coronaviruses dataset from Jacques Van Helden's GitHub. The datasets consists of a coronaviruses phylogeny with a small number of strains, and a host for each strain, as well as phylogenies inferred from parts of the genome, and reffered to as "feature" trees. As coronaviruses are subject to recombination, instead of using genes as a basis for the lower level tree, the authors of [199] proposed to identify recombinant regions using Percent of Identical Positions (PIP) profiles, by alignment with a reference genome, and to separate these regions to infer "feature"

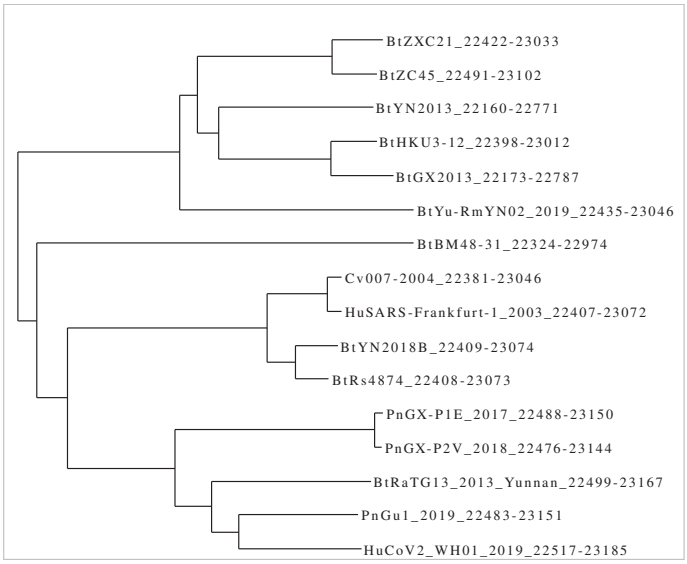


Figure 2.4: An example of a feature tree of coronaviruses, for the RBD domain of the spike protein. Bt refers to bat host, Cv civets, Hu humans, Pn pangolins. HuSARS is the human Sars Cov 1 virus, HuCoV2 is Sars-Cov 2.

trees. It is those trees we used as our lower level tree in our anlayses, Figure 2.4 give an example of one of these trees.

I used our 3-level reconciliation Monte Carlo method, sampling over multiple host and viruses reconciliation scenario and for each reconciling the genes with the viruses. I used a simplistic host tree consisting of five compartments: humans, bats, camels, civets, and pangolin. As the host does not constraint much the virus diversification the reconciliation is quite "chaotic", meaning that there are a lot of horizontal transfers, and the different scenarios are really differents from one another.

This model was our first try to use the 3-level reconciliation likelihood to choose between multiple host and symbiont reconciliation scenarios. It was also our first experience with what I evoked as the disappearing prior in the previous subsection. The most likely host/symbiont scenario in the 3-level model is the one that gives the most likely gene/symbiont reconciliation $P(G|r_{S,H}, H)$, and the host/symbiont reconciliation likelihood of $r_{S,H}$ participates to the 3-level likelihood in a negligible way.

We plotted the distribution of the likelihood of the virus and host knowing the gene, when sampling on the host and virus reconciliation. The likelihood of the complete model is the integral of this distribution (Figure 2.5). We looked at some elements of that distribution 2.6. At the right, the maximum likelihood scenario is a very chaotic and quite unlikely host and virus one, though it achieves maximum likelihood through the participation of the genes to that likelihood.

When the 2-level is better than the 3-level, a host symbiont reconciliation that

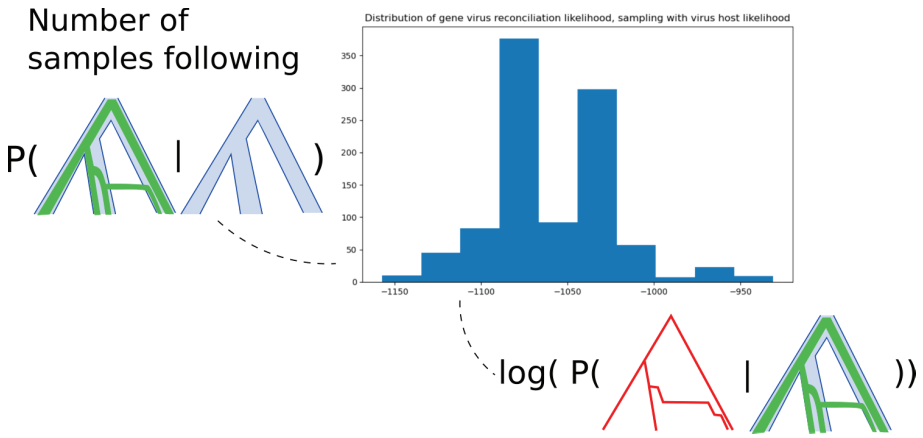


Figure 2.5: Distribution of gene and virus likelihood sampling over the host and symbiont reconciliation.

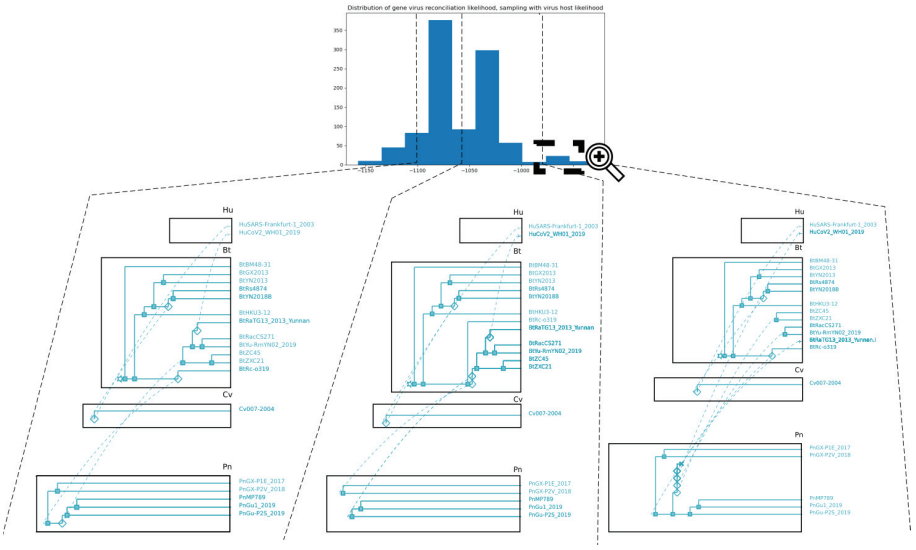


Figure 2.6: Distribution of gene and virus likelihood sampling over the host and symbiont reconciliation, with example host and symbiont reconciliation scenarios corresponding to different parts in the distribution.

will simulate the 2-level will have a good gene symbiont aware of the host likelihood. For instance, all symbionts can be matched to a single host before being transferred to the host they have to match. It is not a problem when computing the joint likelihood with the Monte Carlo approach, but it is to keep in mind when looking for maximum likelihood host and symbiont reconciliation.

It was an engaging dataset as it had a very contemporary subject and was valuable to test our approach, though we did not truly have a complete host tree. Further work on this dataset could begin with constructing a better host tree, particularly with greater granularity for bat species.

SECTION 2.3

Symbiont tree inference

The inference of tree topology in a 3-level framework is one of the primary extension of our 3-level scenario inference and likelihood computation framework that we discussed a lot during this Ph-D. It disappeared and reappeared multiple times in different iterations. Here I present the approach we decided to use to answer this question, amalgamation, and the various leads I followed.

The symbiont tree inference has two sides. It can be a way to correct a tree, to use the genes and the host to construct a better phylogeny for the symbiont. But it can also be a way to build compartments of common evolution, see figure 2.7. Given genes and host, we offer genes an intermediate compartment to account for reconciliation events common to multiple lineages or families.

2.3.1 Amalgamation

In our paper, we use amalgamation[1] to consider multiple symbiont topologies and construct a new topology through the reconciliation with the host. Amalgamation takes as input a sample of trees that are seen as an estimation of the distribution of the tree topology. One of the reasons for our approach here was not to use a concatenate to get a symbiont tree but start from the gene trees. We thus did not have access to samples of symbiont tree topologies. Instead, we used the universal unicopy gene trees as a distribution for the symbiont tree, as they are trees on the exact same set of leaves. The choice of amalgamation to construct this symbiont tree was motivated by the fact that it is a method we know well as it is already present in ALE. Moreover, it does not require an important time increase, and it was interesting to implement as it can be used, as in ALE, to consider uncertainty on a tree and choose a root. However, it is limited to the clades present in the

[1]See the introduction for a detailed description of amalgamation.

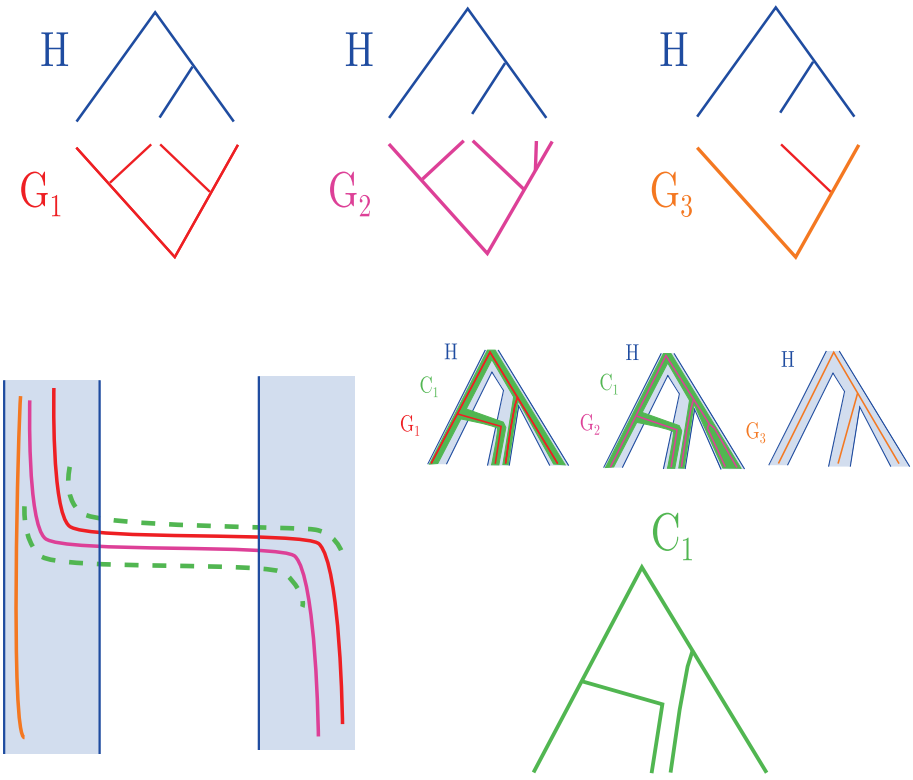


Figure 2.7: Given 3 genes G_1, G_2, G_3 and an host H, we can try to construct a compartment (or multiple ones) to account for the common evolution of these gene trees inside the host. Here for instance, we construct a compartment C_1 that account for the horizontal transfer common to G_1 and G_2.

sample.

The method, as it is presented in the paper, takes the maximum likelihood amalgamated symbiont tree from the host and symbiont model, but it could be interesting to sample multiple host/symbiont reconciliation scenarios and keep the one that maximizes the host/symbiont/gene reconciliation likelihood.

2.3.2 Clustering and supertree

Ideally, having built a method to evaluate the likelihood of 3-level reconciliation for three trees, we would like to be able, given the lower and upper trees, to generate the maximum likelihood intermediate tree (or set of intermediate trees). As we cannot just test all possible sets of trees, we need a way to propose intermediate trees.

We found two really interesting papers tackling similar questions. The first one [36] describes a distance between gene trees based on the observed and expected number of common events in a reconciliation with the species tree. The second [83] is about clustering gene trees. On simulated datasets, they compare different tree distances and clustering approaches to group genes that correspond to the same tree. They notably question how to know the number of clusters, as these clusters correspond to trees, and it is hard to parametrize.

Our idea was thus to start with clustering the gene trees, ideally, with a distance based on reconciliation [36]. Then for each cluster, we construct a tree. And we test the triplet of gene, compartment, and host tree using our 3-level framework.

We first thought about using supertree methods for the inference from the cluster. But amalgamation is a suitable tool, as we have a host tree on top. However, amalgamation cannot take trees that are not on the same leaves multiplicity. To account for the events of the DTL model, the intermediate tree may have different leaves than the species tree, which poses a methodological problem as a supertree method seeks to organize the leaves of the species tree, not to make new ones appear. For example, if we want to share a duplication that happens in the same leaf of the species tree, we have to add new labels to the leaves to somehow increase the granularity of the leaves. To do this, we could do clustering on the leaves to group them according to whether or not they have co-evolved with this same distance by reconciling but restricting ourselves to the lineages that descend to each of the two leaves that we are comparing.

We implemented the clustering part and played a bit with it, notably on the *Helicobacter pylori* dataset, but without obtaining interesting results or proceeding to test it with the supertree or amalgamation part.

2.3.3 Bayesian approach

Another possibility is to use a Bayesian approach by generating a tree inside the host tree and checking how good it is. To add a bit more information from the gene trees in the process, we could use the frequency of events in a reconciliation of the gene and the host to update the rates of DTL events used for the simulation with branch-dependent rates.

This approach would use intensive reconciliation computation, which is not really possible with our current version, but speedup could be implemented, just like ALE was adapted in Generax to be used in an evaluation step.

2.3.4 3-level evaluation of the number of clusters

Our 3-level framework has the advantage of penalizing the number of compartments, which is an interesting feature for clustering as it can be used to evade overfitting and choose a good number of clusters. To infer S, we consider the probability $P(S, G|H)$ (instead of the $P(G|S, H)$ that we use in the comparison of 2 and 3-level), and this probability takes into account the host and symbiont reconciliation, which probability will decrease with the number of symbiont tree.

2.3.5 Multiple prior matching of the leaves

One thing I let aside until now, but which is essential, is that what we are looking for is not only a tree but also a matching between this tree's leaves and the other levels' leaves. If we want multiple compartments, even if all trees are universal unicopy, then for each gene tree leave, we need to choose which compartment it matches. This seemingly hard problem is, in fact, easily answered with the dynamic programming of reconciliation. The equations do not assume a prior of the form 1 to one leaf and 0 to every others. We can define a prior matching a gene leaf to multiple compartment leaf uniformly and then look at the reconciliation scenario sampled to get a posterior on the matching.

We present an application of this approach in the *Helicobacter pylori* chapter, where we wanted to test different positions for a host branch.

SECTION 2.4

Graphical output

This section presents Thirdkind, a command-line software to output SVG from recPhyloXML input. Its home is on GitHub:

https://github.com/simonpenel/thirdkind/wiki.

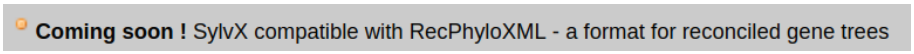


Figure 2.8: Still some work for recPhyloXML to become the standard.

First, I must talk about recPhyloXML. In 2018, Wandrille Duchemin, one of Eric and Vincent's previous Ph-D students, attempted to gather the community around a standard output format to facilitate exchange and benchmarking. Many gene/species reconciliation contributors cosigned the paper introducing the format.

The format is based on XML and phyloXML and is described in the introductory paper[70] and on a website[2]. Scripts are available on GitHub to transcribe to recPhyloXML the output of various reconciliation software. The introduction of the new format was also motivated by the possibility to have a common viewer, and an in-browser viewer was developed by Guillaume Gence.

Four years after the introduction, the compatibility with viewer Sylvx is still "to come soon" (see figure 2.8), Bansal's team's new simulation software Sagephy uses its own format, but the transcription script for their reconciliation inference Ranger-DTL is available directly on the software page. Reconciliation inference EcceTERA, simulation software Zombi, and the gene and species tree inference frameworks based on reconciliation Treerecs, GeneRax, and SpeciesRax use the format.

Thirdkind, a viewer for recPhyloXML and multi-level reconciliation was implemented by Simon Penel in Rust. It is exceptionally easy to install using Cargo (one command line to install Cargo, one to install Thirdkind).

Thirdkind

At first, Thirdkind was part of DL reconciliation and tree correction software Treerecs, a project developed in Eric Tannier's INRIA team. It was included as a viewer for DL reconciliations and was developed by Simon Penel. As few methods were dedicated to general visualization of reconciliation, Sylvx constituting the main exception, and none took recPhyloXML as input (a call was even present on GeneRax wiki for such a viewer), Simon Penel continued the implementation to consider DTL reconciliations. We also took this opportunity to propose visualizations of the reconciliation of three levels and add features to visually capture the reconciliation of multiple possible scenarios or multiple gene families. Simon Penel did all the implementation, and we designed together and with Eric Tannier, Vincent Daubin, and Théo Tricou the three-level viewer and the features to interpret transfers.
The following paper was published in Bioinformatics as an application note [181], as we thought it could be helpful for multiple usages. We also reproduce

[2]When the site is not down, which is not so often.

the supplementary material after the article.

In the remaining of this section, I discuss two features of this new software.

2.4.1 Thirdkind: displaying phylogenetic encounters beyond 2-level reconciliation.

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

Thirdkind: displaying phylogenetic encounters beyond 2-level reconciliation.

Simon Penel [1]*, Hugo Menet [1], Théo Tricou [1], Vincent Daubin [1] and Eric Tannier [1,2]

[1] Laboratoire de Biométrie et Biologie Evolutive/UMR5558, CNRS/UCBL, Villeurbanne, 69622, France.
[2] Inria Lyon, Villeurbanne, 69622,France.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Reconciliation between a host and its symbiont phylogenies or between a species and a gene phylogenies is a prevalent approach in evolution, however no simple generic tool (*i.e.* virtually usable by all reconciliation software, from host/symbiont to species/gene comparisons) is available to visualise reconciliation results. Moreover there is no tool to visualise 3-levels reconciliations, *i.e.* to visualise 2 nested reconciliations as for example in a host/symbiont/gene complex.
Results: Thirdkind is a light and easy to install command line software producing svg files displaying reconciliations, including 3-levels reconciliations. It takes a standard format recPhyloXML as input, and is thus usable with most reconciliation software.
Availability: https://github.com/simonpenel/thirdkind/wiki
Contact: simon.penel@univ-lyon1.fr
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Phylogenetic reconciliation consists in linking the history of two (or more) co-evolving biological entities, at two different levels of organization. It is achieved by embedding a phylogenetic tree into another, pointing at the dependencies between evolutionary histories and the level-specific events. For example, gene phylogenetic trees can be mapped into species trees, explaining the topological differences by speciation, duplication, loss, horizontal transfer. This can also apply for the comparison of host and symbiont co-evolution, or the evolution of protein domains inside genes, etc... Several methods and available software are dedicated to constructing reconciliations (for a review, see Menet *et al*., 2021). The XML format "recPhyloXML" has been proposed as a standard to describe phylogenetic reconciliations (Duchemin *et al*., 2018) and is now produced directly or via available translation scripts by a majority of reconciliation software. Visualisation of phylogenetic reconciliations are proposed by various programs and interfaces as NOTUNG (Chen *et al*., 2000), SylvX (Chevenet *et al*., 2016), Treerecs (Comte *et al*., 2020), Jane (Conow *et al*., 2010), eMPRess (Santichaivekin *et al*., 2021) and Capybara (Wang *et al*., 2020). However at the exception of SylvX, all are integrated in a specific reconciliation software and cannot visualise reconciliations produced by others. None of these software is handling recPhyloXML input files[1], and none of them is generic to any kind of reconciliation (for example SylvX does not allow temporary free living symbionts, as it is not allowed for genes to live outside a genome) nor can handle multiple horizontal transfer (*i.e.* several genes transfered with the same donor and recipient) and the consideration of numerous possible scenarios. DoubleRecViz (Kuitche *et al*., 2021) uses a derived version of recPhyloXML, adding a transcript level to gene and species format but without support for horizontal transfers. Eventually there is no software able to combine two nested reconciliations *i.e.* to get in a single representation the gene/symbiont reconciliation and the symbiont/host reconciliation, despite the recent interest of methodologists in these 3-levels systems Stolzer *et al*., 2015; Kundu *et al*., 2019; Muhammad *et al*., 2018 answering current and significant questions in biology Menet *et al*., 2021. Thirdkind is a light command-line software allowing the user to generate a svg from recPhyloXML files with a large choice of options (orientation, police size, branch length, multiple trees, redundant transfers

[1] at the exception of the unpublished Recphylovisu web interface (http://phylariane.univ-lyon1.fr/recphyloxml/recphylovisu) of which Thirdkind can be considered as an update

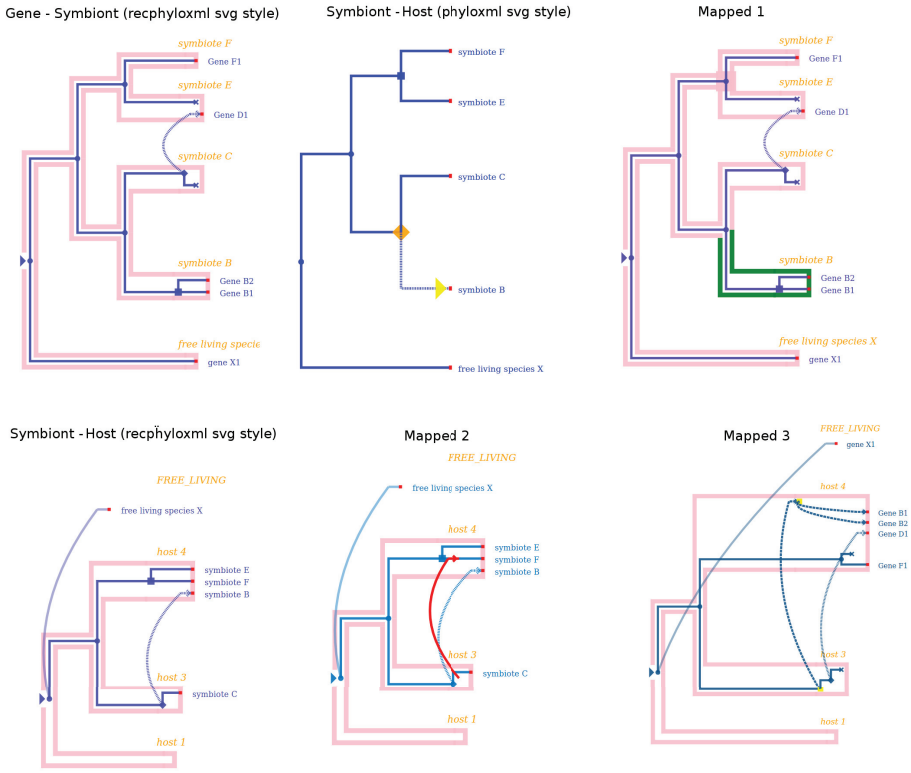


Fig. 1. The 6 main outputs of a 3-level gene/symbiont/host reconciliation. one “recphyloxml style” svg for each of the two input, a “phyloxml style” svg of the reconciled symbiont tree (from -f file) and three “mapped” svg files describing the gene/symbiont/host reconciliation The first “mapped” svg is a modified version of the "recphyloxml style" svg of the gene/symbiont reconciliation: the drawing of the symbiont tree as a tube presents features describing its reconciliation with the host (a big square for a duplication node, an additional tube branch colored in black for a loss and the tube segments between the start and end of a transfer are colored in green). The second “mapped” svg is a modified version of the "recphyloxml style" svg of the symbiont/host reconciliation in which gene transfers are mapped to the host nodes and displayed in red: a gene transfer between the symbiont “C” present in host “3” and the symbiont “E” present in host “4" is displayed as a red Bezier line between host “3” and host “4” in the tube host tree. The third “mapped” is a mapping of the genes trees over the host tree through the symbiont. If genes “B1” and “B2” are associated to the symbiont “B”, and the symbiont B is associated to host “4”, the genes B1, B2 are associated to host “4” in the svg. If a gene is transferred between hosts via a symbiont transfer, the transfer start with a yellow diamond and the stippling is different. A gene transfer across symbionts which is not affected by a transfer of the symbiont across hosts is displayed as a classic gene transfer

handling, etc.) and to handle the visualisation of 2 nested reconciliations. Trees can be dated via their branch lengths or undated.

2 Installation

Thirdkind is written in Rust (https://crates.io/crates/thirdkind) and thus very easy to install: install *cargo* and then type ‘cargo install thirdkind’. Source code is available at https://github.com/simonpenel/thirdkind.

3 Usage

3.1 Input files (option -f)

Thirdkind is dedicated to read recPhyloXML format files, but it can read newick or phyloXML files if needed. The option -f is used to indicate the name of the input file, whatever the format. The format is guessed from the extension of the file, or it can be chosen with the -F option.

Newick is a simple parenthesed tree format, phyloXML is a xml format dedicated to phylogeny, recPhyloXML derives from phyloXML and is dedicated to reconciliations.

A phyloXML file contains only 1 tree (reconcilied or not). A recPhlyloXML file contains at least one "upper" tree (the species tree in the species/gene complex, or the host in a host/symbiont complex) and one "lower" reconciled tree (respectively the gene, or symbiont) mapped to (one of) the upper tree(s). A clade presents several tags: a name, a location, a type of event, etc. Each node of the lower tree(s) has a “location”, the value of which should be the same as the value of the “name” of one of the clades in the upper tree(s). It is possible to have multiple lower trees, and multiples upper trees in a single file.

3.2 Output styles

Thirdkind allows to generate svg ouput files according to 2 different styles:

1) The "recphyloxml style" where the lower trees and upper trees are displayed, the lower tree being embeded into the upper trees. The output consists of one ore several reconciled lower trees drawn as lines inside

or outside one or several upper trees, drawn as tubes. Lower trees have symbols at their nodes (a square for a duplication, a circle for a speciation and a cross for a loss). Transfers (or host switches) are Bezier spotted lines ending with an arrow. If there is more than one lower tree, they are displayed with different colors (Figure 1, above and suppl.mat.).

2) The "phyloxml style" where a unique tree is displayed, with symbol within the tree representing evolution events: a square for a duplication, a circle for a speciation, a cross for a loss, transfer is a spotted line (Figure 1, above and suppl.mat.).

If the input format is phyloXML or newick, the style of the output is "phyloxml", if the input format is recPhyloXML, the style of the output is "recphyloxml", or "phyloxml" with the options -S (display the upper tree only), -G (display one of the reconcilied lower tree).

Note that if the file format is recPhyloXML but the file extension is phyloxml, the output style will be the upper tree in "phyloxml" style.

3.3 Dated trees (option -l)

The -l option allows to use the branch lengths to display the tree. It applies to the "upper" tree in “recphyloxml" style context, and to the current tree otherwise.

3.4 Minimising transfer crossings (option -O)

Since minimising transfer crossings is a NP-hard problem (Klavitter and Stumpf, 2020) we used a simple heuristic : explore the "upper" tree from the start node and the end node of a transfer until the ancestor of these 2 nodes, giving to each node on the way a score reflecting how it should be oriented to reduce the distance between start and end nodes. This is mainly useful when studying a single tree with few transfers (with 1 transfer the heuristic finds the best solution).

3.5 Multiple recPhyloXML input files (option -m)

It is possible to use a list of recPhyloXML files instead a single recPhyloXML file. This option is useful to handle large sets of gene histories inside species histories as the ones generated by GeneRax (Morel *et al*., 2020).

3.6 Dealing with redundant transfers (options -t , -T and -J)

In case of multiple gene histories it may be useful to enlighten redundant gene transfers, when trying to identify highways of transfer for instance. Option -t draws only one gene history without the transfers and then in red the transfers of all the histories according to their abundance: only the transfers with an abundance higher that the threshold are drawn and the opacity reflects the abundance. The option -T allows to choose the gene history to display. The option -J displays the abundance of the transfer (Figure 2, suppl.mat).

3.7 Dealing with ‘free living’ symbionts (options -e)

In the history of a micro-organism, some taxa may be free living species and some others may have evolved to be a symbiont of a host. In this case, free living organism should have a “location” indicating “FREE LIVING” instead of the name of a host. Thirdkind draws the free living part of the symbiont path tree outside the host pipe tree (Figure 3, suppl.mat).

3.8 Nested recPhyloXML files (options -g and -f)

It is possible to combine two reconciliations as for example a gene/species reconciliation and a symbiont/host reconciliation, in which the symbiont of the second reconciliation is the species of the first one. This is valid for any variant of a 3-levels co-evolution, as geography/species/genes, or species/gene/domains. For clarity of the exposition we adopt the host/symbiont/gene vocabulary, keeping in mind the genericity of the method. The -g option indicates the gene/species file, -f indicates the symbiont/host file. The software generates several svg files: one “recphyloxml style" svg for each of the two input, a “phyloxml style" svg of the reconciled symbiont tree (from -f file), a “phyloxml style” svg of each reconciled gene trees (from -g file) and three “mapped” svg files describing the gene/symbiont/host reconciliation (Figure 1, above and supl.mat.).

4 Execution time and readability

Thirdkind was able to process 5,000 reconcilied trees of 50 nodes in 2 seconds and to process a tree of 7,000 nodes in 1 second. The readibility depends on the ability to handle the resulting svg (desktop computer, mobile, poster, etc).

Acknowledgements. Grant from Agence Nationale pour la Recherche, ANR-19-CE45-0010.

References

Chen,K. *et al*. (2010) NOTUNG: a program for dating gene duplications and optimizing gene family trees. *J. Comput. Biol.*, **7**, 429–447.

Chevenet,F. *et al*. (2016) SylvX: a viewer for phylogenetic tree reconciliations. *Bioinformatics*, **32**, 608–610.

Comte,N. *et al*. (2020) Treerecs: an integrated phylogenetic tool, from sequences to reconciliations. *Bioinformatics*, **36**, 4822–4824.

Conow,C. *et al*. (2010) Jane: a new tool for the cophylogeny reconstruction problem. *Algorithms Mol Biol.*, DOI: 10.1186/1748-7188-5-16.

Duchemin,W. *et al*. (2018) RecPhyloXML: a format for reconciled gene trees. *Bioinformatics*, **34**, 3646–3652.

Klavitter,J. and Stumpf, P. (2020) Drawing Tree-Based Phylogenetic Networks with Minimum Number of Crossings, https://arxiv.org/abs/2008.08960

Kuitche,E. *et al*. (2021) DoubleRecViz: a web-based tool for visualizing transcript–gene–species tree reconciliation. *Bioinformatics*, **37**, 1920–1922.

Kundu,S. and Bansal,M.S. (2019) SaGePhy: an improved phylogenetic simulation framework for gene and subgene evolution. *Bioinformatics*, **35**, 3496–3498.

Menet,H. *et al*. (2021) Phylogenetic reconciliation. https://hal.archives-ouvertes.fr/hal-03258402.

Morel,B. *et al*. (2020) GeneRax: A Tool for Species-Tree-Aware Maximum Likelihood-Based Gene Family Tree Inference under Gene Duplication, Transfer, and Loss. *Molecular Biology and Evolution*, **37**, 2763–2774.

Muhammad,S.A. *et al*. (2018) Species tree-aware simultaneous reconstruction of gene and domain evolution. *bioRxiv*, https://doi.org/10.1101/336453.

Santichaivekin,S. *et al*. (2021) eMPRess: a systematic cophylogeny reconciliation tool. *Bioinformatics*, **37**, 2481–2482.

Stolzer,M. *et al*. (2015) Event inference in multidomain families with phylogenetic reconciliation. *BMC Bioinformatics*, **16**, S8.

Wang,Y. *et al*. (2021) Capybara: equivalence ClAss enumeration of coPhylogenY event-BAsed ReconciliAtions. *Bioinformatics*, **36**, 4197–4199.

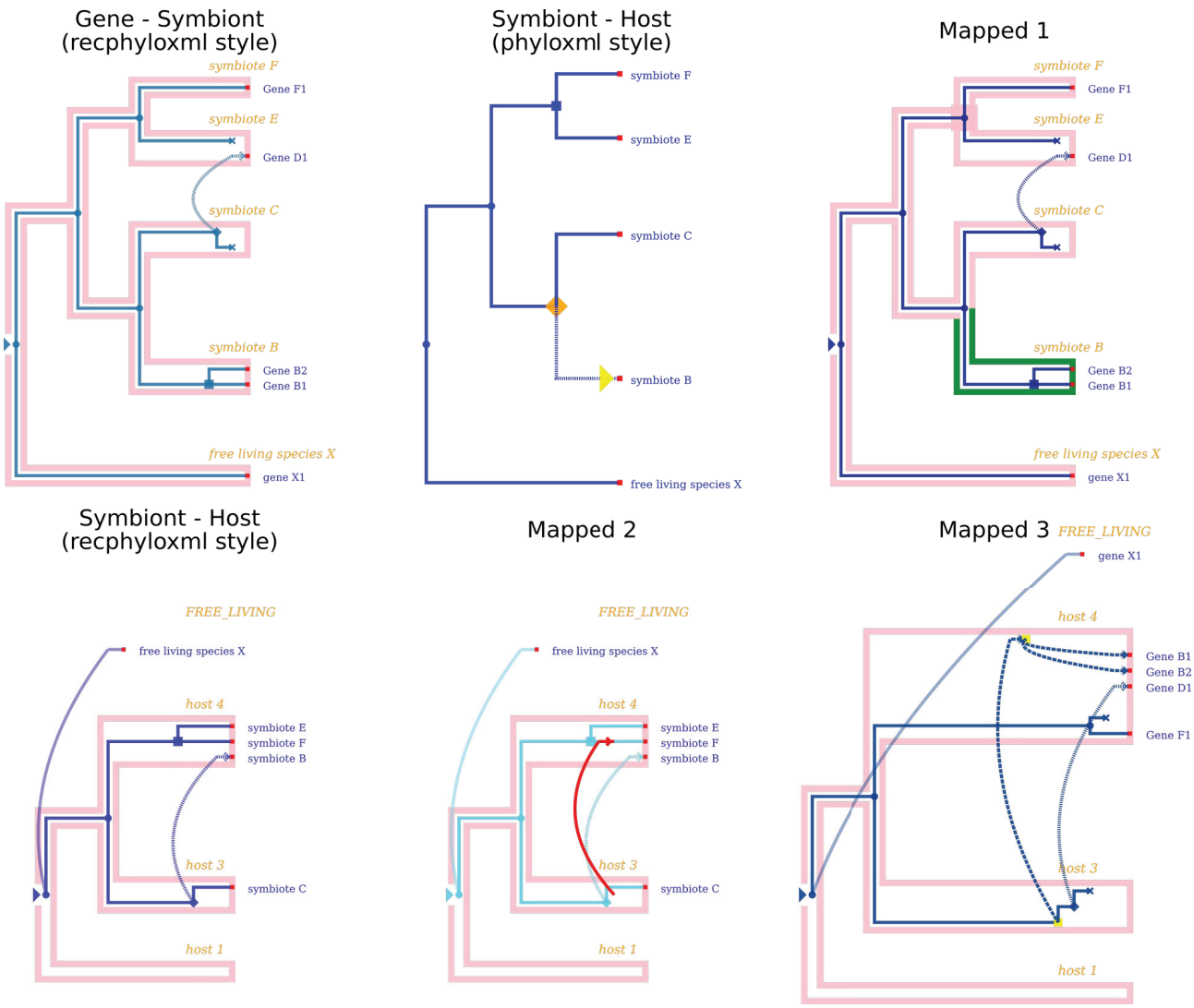


Figure 1. The 6 main outputs of a 3-level reconciliation gene/symbiont/host.

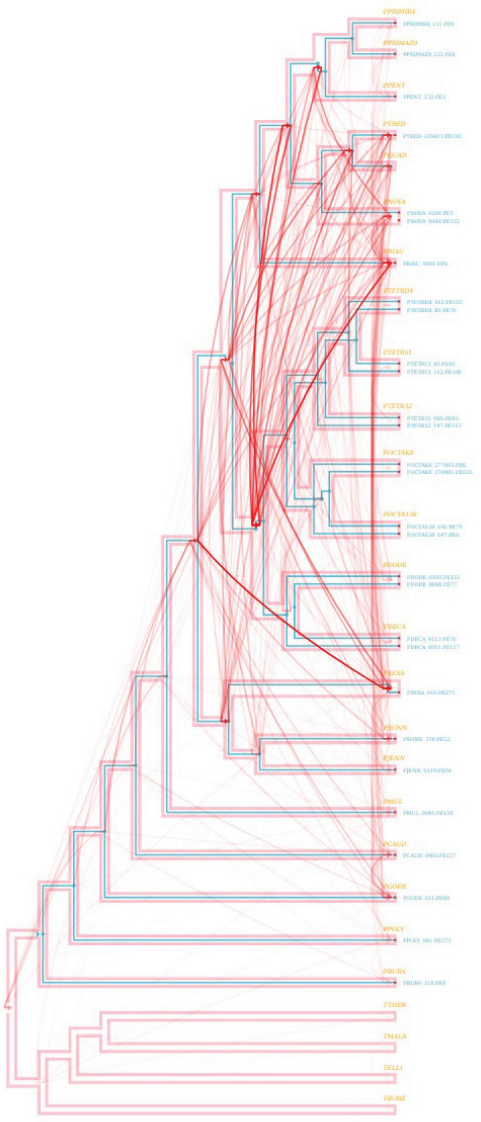
Mapped 1 : modified version of the recphyloxml style gene/symbiont reconciliation: the drawing of the symbiont tree as a tube presents features from phyloxml style symbiont/host reconciliation : a big square for a duplication node and the tube segments between the start and end of a transfer are colored in green.

Mapped 2 : modified version of the recphyloxml style symbiont/host reconciliation in which gene transfers are mapped to the host nodes and displayed in red: the gene transfer between the symbiont “C” present in host “3” and the symbiont “E” present in host “4" is displayed as a red Bezier line between host “3” and host “4” in the tube host tree.

Mapped 3 : mapping of the genes trees over the host tree through the symbiont. The genes “B1” and “B2” are associated to the symbiont “B”, and the symbiont B is associated to host “4”, thus the genes B1, B2 are associated to host “4” in the svg. When the genes are transferred between hosts via a symbiont transfer, the transfer representation start with a yellow square and the stippling is different. A gene transfer across symbionts which is not affected by a transfer of the symbiont across hosts is displayed as a classic gene transfer.

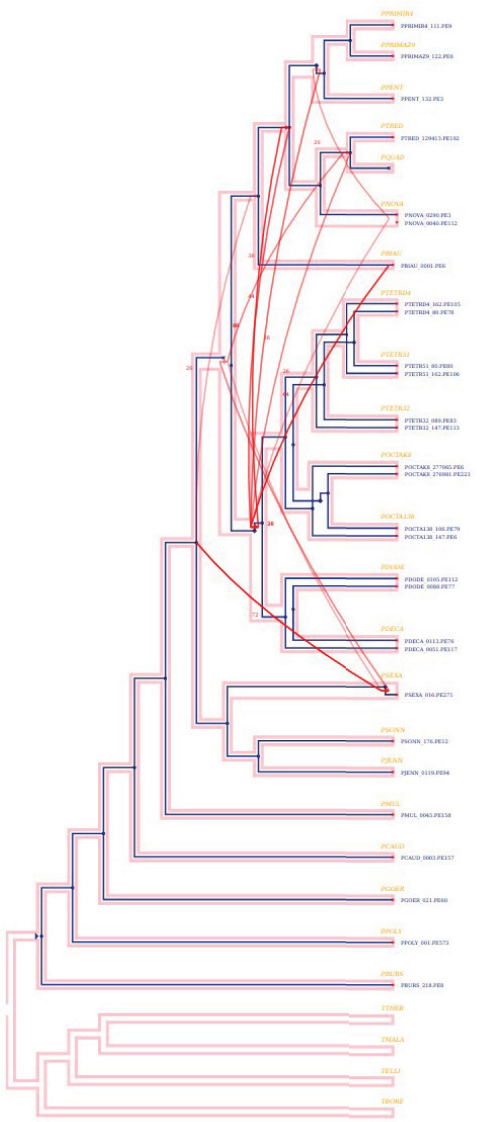
Command :

thirdkind -f example_wiki/publi/parasite_hote.recphylo -g example_wiki/publi/gene_parasite.recphylo -e -b



thirdkind -f paramecium_data/liste.txt -t 1 -m -b

Transfer redundancy in 1000 gene histories



thirdkind -f paramecium_data/liste.txt -t 25 -m -b -J

Transfer redundancy in 1000 gene histories, displaying only transfers with an abundance higher than 25

Figure 2. Using -m and -t options to display redundant gene transfers.

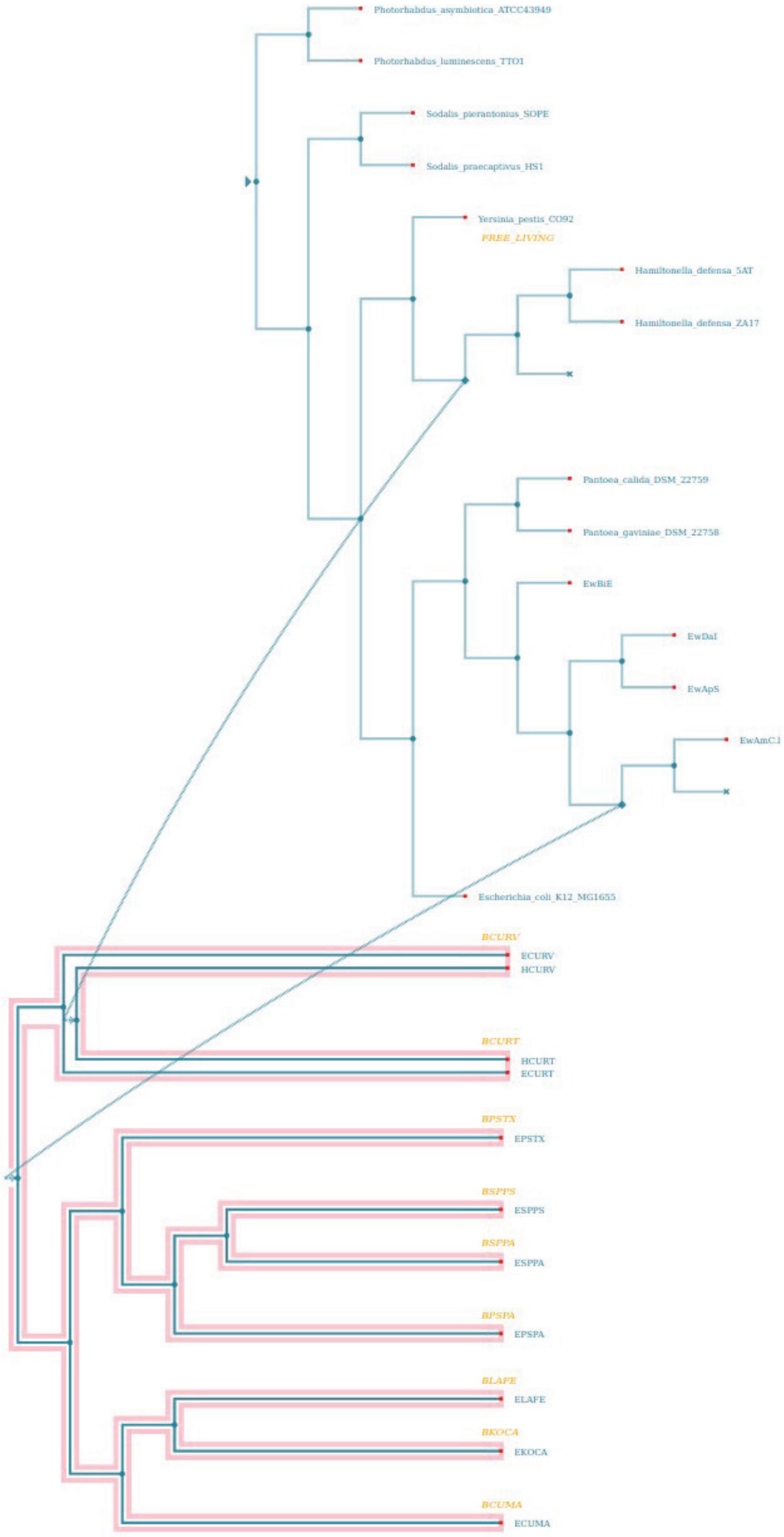


Figure 3. Using -e option to display free living species and symbionts.

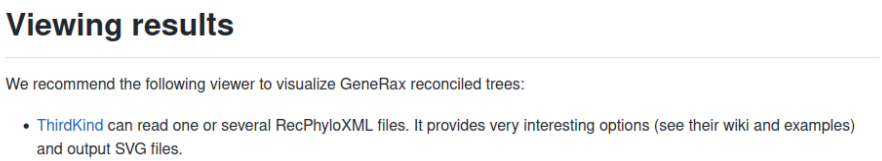


Figure 2.9: Thirdkind has been recommended on the GeneRax wiki to visualize reconciliations.

2.4.2 3-level viewer

One of the feature of Thirdkind is the possibility to investigate multi-levels reconciliations. Sylvx have some features to add the information of an upper level (used for geography on top of a host and symbiont reconciliation in [19]). Doublerecviz [110] aims at the visualization of three trees but with a modified version of recPhyloXML and without support for horizontal transfers in a model where genes are assigned to a location: nuclear or mitochondrial and an additional event of endosymbiotic gene transfer is considered. Conversely, Thirdkind simply use two recPhyloXML, one for host and symbiont, and one for gene and symbiont. It is one of the output of my implementation of 3-level reconciliation. We first thought about displaying all three trees together, but to gain in readability we chose to display only two trees together but with additional information coming from the remaining level. Three views, denoted mapped 1, 2 and 3 are available. Figure 1 in the supplementary material of the article (in the previous pages) illustrates those.

The mapped 1 output presents the gene and symbiont reconciliation with the symbiont colored depending on its reconciliation events with the host tree, in a manner reminiscent of phyloXML. It could be useful to look at congruence between the gene events and the symbiont ones. For instance, are symbiont host switches accompanied by subsequent gene exchanges? Or does the speciation of a symbiont in a host result in gene losses?

The mapped 2 output presents this time the host and symbiont reconciliation but with the gene transfers represented on top of it. With our model of host aware gene transfers, it is the view we used the most, examples are presented in this thesis for *Cinara* aphids (figure 2.2) or *pylori* and human coevolution (in the method article at the beginning of this chapter).

The mapped 3 view is useful, especially with the holobiont idea, to look at coevolution between levels that are not directly adjacents, here the genes from the symbionts' genome, and the host of the symbionts. It is a representation of the host and gene reconciliation constructed by hiding the symbiont that is the intermediate between them. For instance, for *Cinara* aphids, we see that the genes that transfer between the two symbionts inside the host follow exactly the phylogeny of the host (figure 2.10).

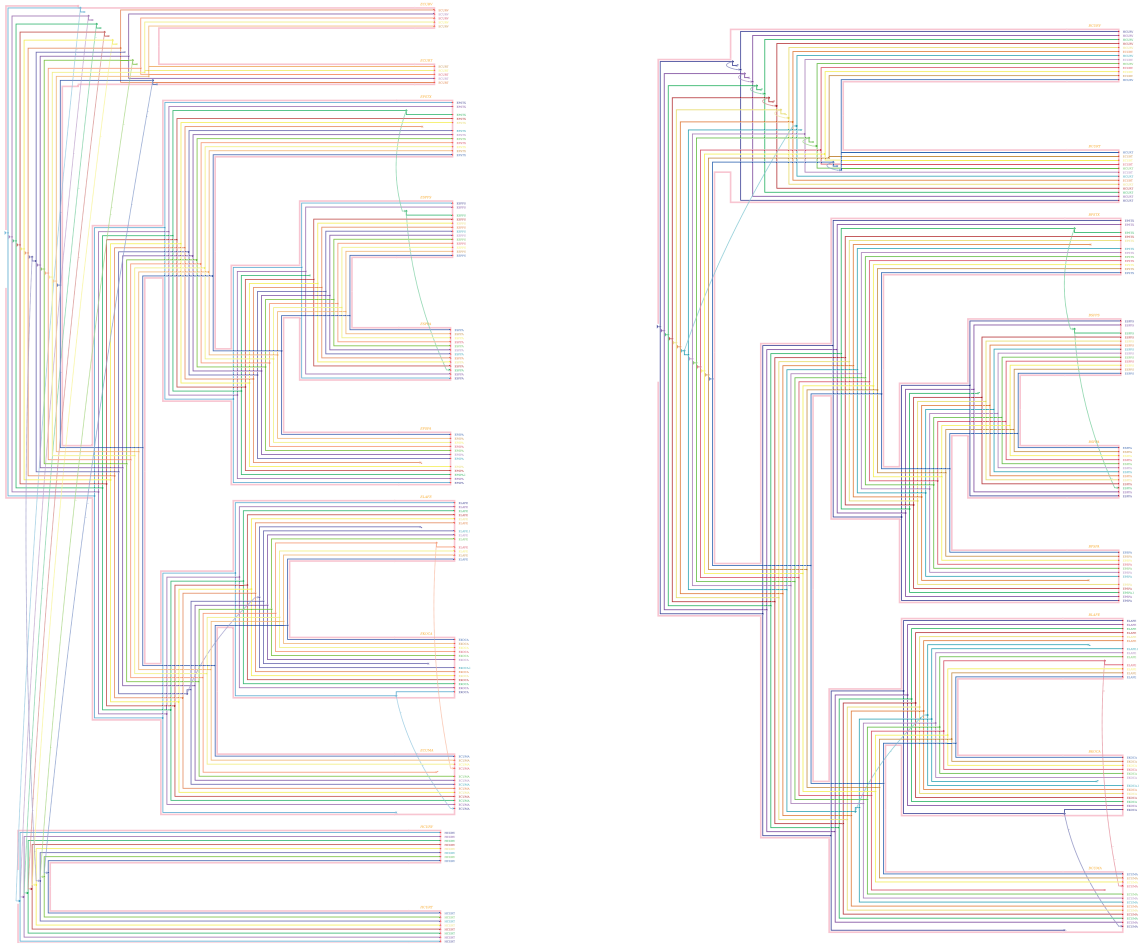


Figure 2.10: Views of *Cinara aphids*, endosymbionts and genes. Left the symbiont and genes reconciliation (only for the symbionts in the host), and right, the mapped 3 view of symbiont's genes inside the host, we see the genes follow almost exactly the host while there are important transfers between symbionts.

2.4.3 Redundant transfers and possible uses

Another feature specific to Thirdkind is the representation of redundant transfers. It is a usage deeply rooted in the probabilistic approaches of reconciliation of the team. As they are many ways to reconcile two trees together, from a probabilistic point of view, what makes sense is to output a sample of scenarios depending on their likelihood. These scenarios can then be aggregated in a text file to give observed frequencies of the different events found in them, as do ALE. However another approach is to aggregate these scenarios in a picture. This feature is only available for horizontal transfers, but it could be adapted for other events as well.

Thirdkind can be given a recPhyloXML with multiple gene families, or multiple recPhyloXML corresponding to different scenarios sampled, and will output the species tree reconciled with one of the gene trees, and on top of that the transfers that are seen more time than a given threshold, with an opacity depending on that number of times. It is a way to have a visual idea of highways of transfers, and also a way to view multiple scenarios, or multiple families at the same time, it makes probabilistic output of reconciliation truly user friendly. The approach is highly scalable and can be used with hundreds of scenarios and gene families.

2.4.4 A software to resume all meaningful data from reconciliation

An idea that could be developed around Thirdkind and the possibility to read multiple files and aggregate transfers would be to propose a recap of all important information from a reconciliation, for instance, from a sample of reconciliation scenarios.

How to give the best information to the user is something thought a lot by host/symbiont reconciliation efforts. What we designed with horizontal transfers in Thirdkind can easily be done with other potential events. We could look for segmental events and use statistical approaches to seek deviations from the expected number of common events on a given branch (with an approach similar to a coevolution score [36]). Such scores could be displayed on the SVG.

SECTION 2.5

Supplementary discussion: on biological models

In this chapter, I presented our main contribution to this thesis, our 3-level host, symbiont, gene reconciliation framework, and a graphical viewer.

Aside from these methodological aspects, we also introduced multiple datasets: a

simulated one using an exterior simulation framework and multiple biological ones. *Cinara* aphids dataset offers a small example that fits precisely the idea we had in mind when constructing our model. In contrast, *Helicobacter pylori* and their human host is a larger dataset with multiple unanswered questions on the structure of the evolutionary relationship between the two entities. We will delve back into *Helicobacter pylori* and human relations as it is the subject of the next chapter. The last dataset, the one with coronaviruses, gives temporality to this thesis, reminding us that it happened during a strange time for us all. We also witnessed a burst in the vulgarization of phylogenetic methods to study and present the development of variants of the coronaviruses that took man as hosts. Applying our method to this dataset shows that viruses can be studied with reconciliation methods for genetic material exchange or host switch.

Nevertheless, during the introduction, we motivated our goal to consider three levels with various biological studies. Aside from the biological models I just mentioned, that we investigated with our method, we also considered a multiplicity of other biological models and collaborations, though they were not meant to be, for now. To illustrate our approach and remind us that when an application is presented in a final result, chances are multiple ones have not been pursued, I give a view of some of them.

At a local scale, in our lab, the LBBE in Lyon, other insects, and bacterial symbionts models are studied. We discussed multiple times with Fabrice Vavre, around ticks dataset, with *Coxiella*, *Midichloria*, *Francisella*, *Rickettsia* symbionts, or with Sylvain Charlat, about flies and their *Wolbachia* partner. A biological model similar to the *Helicobacter pylori* one would be to consider mammals and their microbiomes, with the difficulty of a more loose relation. This model was considered through exchanges with Mattieu Groussin. We also had the opportunity to talk with Laura Eme about red and green algae's multiple levels of symbiosis.

These diverse possibilities show that our model is versatile and can be useful for a multiplicity of use. Even though we did not have the time, or the resources, to study more biological models, the future might hold more!

Chapter 3

Helicobacter pylori

Contents

A natural collaboration

When I started my Ph-D thesis, Alexia Nguyen Trung and Vincent Daubin were working on *Helicobacter pylori* and its relation with its human host. This system comprised a host population migrating between different geographical areas, a pathogen with sequenced genomes, frequent genetic exchanges through recombinations between pathogens, and a population structure echoing the hosts' one. Alexia was constructing a dataset of complete genomes, along with their assignation to a population linked to the host one. That dataset, with host, symbiont, and gene levels, was adapted and available to apply our 3-level method.

I spent extensive time during my second year working on that dataset. I applied the 3-level method, tinkered with it, and tried ad-hoc methods based on reconciliation. Though I was not able to tell biologically significant stories using the methods I worked with, it was a test dataset nonetheless.

The comprehension of the populational structure of *Helicobacter pylori* and the application of reconciliation to access the information at the gene level to get a deeper understanding and dating of the host/symbiont one are fascinating questions. Even if our 3-level model was not completely able to capture these interactions, I am sure Alexia and Vincent's further works will be able to shed some light on them and make this system enter the phylogenomic era.

I briefly presented the *Helicobacter pylori* and human system in the introduction, and the application of the 3-level reconciliation method to *Helicobacer pylori* is one of the test cases of the article in section 2.1. Here I will give more context and get into more details about the study of the population structure. I will then present the specific questions that interested us and our methods to tackle them.

SECTION 3.1

Context

For a review on *Helicobacter pylori* with a phylogeographic and paleomicrobiology point of view (as opposed to a medical one), see [148]. See also this review [104] for a table summarizing the populations and subpopulations levels, complemented with a review on the virulence factor. For another summary of the shared history of humans and *pylori* see [3], which also presents the very different cases of two other human bacterial pathogens *Mycobacterium tuberculosis*, the cause of tuberculosis, and plague inducing *Yersinia pestis*. Furthermore, for a complete, but not so recent, introduction to the species, see this book from 2001 [152].

3.1.1 *Helicobacter pylori*: a bacteria

Helicobacter pylori is a bacterial pathogen of humans colonizing their stomach, with a presence in around 50% of the population. It is one of the only bacteria able to colonize the human stomach, which is a hostile environment due to its acidity. More precisely, it colonizes the gastric mucosa. *Helicobacter* genera bacteria can be found in other animals with a specific or not symbiosis. For instance, *pylori* sister species *acinonychis*, is a parasite to cheetahs and other large felines. It is believed to have been acquired from a host switch originating from humans [73]. *Helicobacter* bacteria genera are not all found in the stomach, but one clade has acquired mechanisms to survive in this environment; notably, it can synthesize urease, an enzyme that catalyzes the hydrolyze of urea which byproduct is an increase in pH [152].

Helicobacter pylori was first described in 1984 by Marshall and Warren [142]. It was linked to gastric diseases, such as gastritis, ulcer, and stomach cancer. However, infection by *pylori* is rarely symptomatic (around 15% of the time). Some studies even suggest that *pylori* might have positive effects on its host [28].

The easy-to-remember one-in-two infection rate hides a more fragmented reality. The prevalence varies strongly between countries, from less than 20 % in Switzerland, to higher than 80% in Nigeria [97]. Moreover, a study on the evolution of prevalence between two time periods (1970 to 1999 and 2000 to 2016) showed a decrease in industrialized countries though no significant changes were observed in developing countries.

The transmission of *pylori* is preferentially intrafamilial [248]. However, the mechanism of transmission is still unclear, with evidence for and against oral-oral, gastro-oral, or fecal-oral transmissions [152].

Helicobacter pylori recombination rate is extraordinarily high compared to other bacteria such as *Escherichia coli*, *Neisseria meningitidis* or *Yersinia pestis* [74]. Studies estimated it to be more important than the mutation rate to explain nucleotides substitution [160]. The intra-species diversity is also very important [152].

3.1.2 Human migrations

We are interested in the relationship between *Helicobacter pylori* and its human host. The primary source of information on the long-term structure of this association is to compare the phylogeographic repartition of *pylori* with humans'. Extant phylogeographic repartition is deeply linked with the ancient migration of the populations.

Determining the history of the human peopling of the world relies on multiple sources. It goes from investigating archeological artifacts or skulls to identifying cultural movements or studying genetic data. One of the primary genetic information

is to look at the geographical or ethnic extant repartition of the haplogroups defined from mitochondrial DNA or the Y chromosome. See for instance a table giving the type of data and the methods used for the study of the Out of Africa event in [129].

Homo sapiens, also deemed as anatomically modern humans, appeared in Africa, while the exact origin, southern, northern, or eastern Africa, is still debated. The oldest identified human fossil remains date back to 190 to 210 kya (kilo year ago) and from 160 to 154 kya. *Homo sapiens* then proceeded toward the rest of Africa. Multiple groups of anatomically modern humans might have left Africa, but what is called "Out of Africa" is the waves that led to continuous colonization of territory exterior to Africa. Figure 3.1, reproduced from [129] recap the main out of Africa hypothesis. The dating of Out of Africa goes from 50 - 60 kya to 100 - 130 kya and might have happened in one or multiple waves [129]. In the one wave model, a population migrated to the Arabic peninsula into central Asia, then split, with one part going into South East Asia before heading toward the Sahul continent (the ancestral landmass constituted of Australia, New Guinea, and Tasmania). The other part diverged again, with one part colonizing Asia and America through the Bering Strait and the other colonizing Europe. The multiple waves model stipulates that the population that migrates to South East Asia and Sahul, and the one that migrates to Asia, America, and Europe, are part of two waves with separate Out of Africa events.

These migrations often denoted as early human migrations, model the expansion of anatomically modern humans out of Africa and across the different continents, but they only explain part of the geographical repartition of extant *Homo sapiens*. More recent migrations have to be considered, even though, as part of written history, they are better documented. For instance, to understand the human populations present in America, it is crucial to consider the migrations from Europe and Africa that followed Christopher Columbus' voyage in 1492.

3.1.3 Population structure from MLST

When looking at *Helicobacter pylori* phylogeny, depending on the place of sequencing and the population to which belong its host, a strong correspondence occurs between *pylori* and host phylogeny.

The first paper documenting this link between *Helicobacter pylori* and human populations dates back to 1999 [2]. This study was performed using Multi Locus Sequence Typing (MLST). MLST uses housekeeping genes to identify bacterial strains. MLST was first introduced to study the human pathogen *Neisseria meningitidis* using six loci, in 1998 [137]. A variety of other human pathogens were since studied using this technique. For *pylori*, the MLST genes are atpA, efp, mutY,

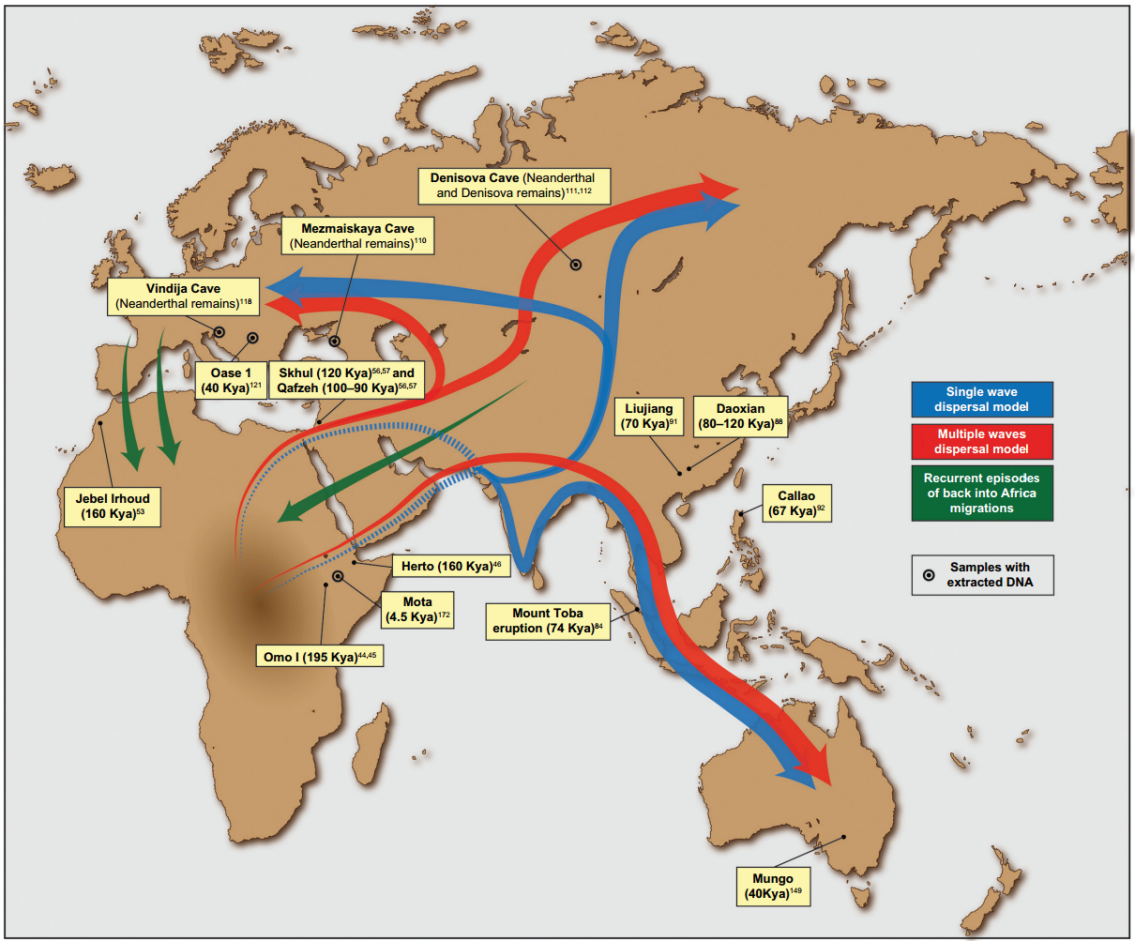


Figure 3.1: Two concurrent models explain the *Homo sapiens* peopling of areas outside Africa, with one or multiple population waves. The location of some ancient human remains and archeological sites are indicated on the map. The placement of arrows is indicative.
Reproduced from Lopez et al [129] under Creative Commons CC-BY 3.0 License

ppa, trpC, ureI, and yphC. These genes were complemented by caga and vaca, two pathogenicity genes. Twenty strains, sequenced in different places across the world, were included in that first study.

To get a more detailed view of *pylori* population structure, a particular method, named Structure, was developed to infer population structure from MLST genes. The Bayesian model considers K populations (K might be unknown) characterized by their allele frequency at each locus. Individuals are then assigned to one population or more, showing admixture (a genome can have multiple "parents"). The inference method relies on an MCMC, [185], and adds possibility for linkage between loci [75].

In [76], first *pylori* populations are defined hpAf1, hpAf2, hpEurope, and hpEastAsia, as well as subpopulations hspSAfrica, hspWAfrica of hpAf1, and hspAmerind, hspEAsia, hspMaori of hpEastAsia. Those assignation are the results of Structure, with the model assigning strains to populations. The admixture version of Structure is then used, where a strain can be assigned to multiple populations (each SNP is assigned while the previous model assigned strains). The interpretation of this model is that it attributes SNPs to ancestral populations, that might have introgressed to produce the current ones. Five populations are identified, three are direct ancestors to respectively hpAf1, hpAf2 and hpEastAsia and SNPs in the same strains are assigned to the same populations. However, the SNPs in hpEurope strains are assigned to two ancestral populations, AE1, an hpAf1 sister clade, and AE2, an hpEastAsia sister clade. The extant European population would be the result of the introgression between those two ancestral ones. The study assigns 370 bacterial strains to these populations, seeing how the populations corroborate geographical origins.

A subsequent study identified new populations, hpAsia2 and hpNEAfrica, and recovered the previous populations [126], with 769 strains. The two new populations appear to be the descendants of the two ancestral European populations, Ancestral Europe 1 for the hpAsia2, Ancestral Europe 2 for hpNeAfrica. For the 769 strains, the association to the different ancestral populations, except for the distinct hpAfrica2, is mostly continuous. Thus, the authors of [126] investigate whether the discrete clusters are due to method artifacts or to a significant biological structure. Looking at pairwise F_{ST}[1], they show that variance is similarly accounted for by geography or by the clusters. Nevertheless, more of the index is explained when adding the clusters to the geography, contrary to similar studies for humans. Clines[2] corresponding to the geographical position of the ancestral populations are

[1]Fixation index, a measure of population differentiation from genetic markers, for instance SNPs. It was introduced in a book from Cavalli-Sforza *et al.* [34] for the study of the genetic diversity of *Homo sapiens*.

[2]Clines are gradients in a biological trait across its geographical range.

examined.

The last identified population is the hpSahul one, presented in an article about the peopling of the Pacific [155].

Figure 3.2 shows the geographical distribution of *Helicobacter pylori* at different time periods, as well as the population tree of the bacteria.

3.1.4 Population (Fine)Structure from genomic analyses

The previous articles rely on the use of MLST with 7 to 9 genes. However, subsequent studies used genome wide data for *Helicobacter pylori*. A new model was developed, fineStructure [118]. As was done with Structure for genes, it paints chromosomes in chunks, which sizes range from a few to tens of SNPs (when Structure assigned individual SNPs). The method does not infer ancestral populations but uses the other strains as populations to paint a strain genome. It then constructs a coancestry matrix that shows the proportion of given and received chunks between all pairs of strains and from which it is possible to deduce a population structure. An advantage of fineStructure is that the number of populations needs not be small and fixed in advance as in Structure.

Yahara *et al.* apply fineStructure to *pylori* [246]. The result is a finer but congruent population structure than when using Structure with the MLST genes. New subgroups and singleton strains are identified in Europe, Amerind, and East Asia. The method also estimates the gene flow between the different populations, showing signs of admixture between Africa, Europe, and part of Asia. FineStructure was also used on a dataset focusing on strains sequenced in the Americas, finding new subpopulations in the same old world populations [220].

The population structure in the Americas, resulting from the recent migrations from Europe and Africa and the early migrations from Asia through the Bering Strait, is at the center of recent developments in the study of *pylori* population structure [163, 162]. From a methodological point of view, the first of these studies is compelling, [163], as it investigates the differences between whole-genome and MLST analysis. The authors draw phylogenies from MLST and whole genomes, with 113 *pylori* strains from Latin America and 54 from other parts of the world. The MLST analysis, unlike the whole genome one, is not able to differentiate the European strains from the Latin American ones. When inferring a phylogeny based on the two techniques, in MLST, European and Latin American strains are mixed, while in the whole genome analysis the European strains cluster together (in a comb-like shape on the branch leading to the Asian clade). Some countries' strains are grouped with both approaches (Nicaragua and Colombia), while a structure emerges for some Mexican strains with information from the whole genome. The

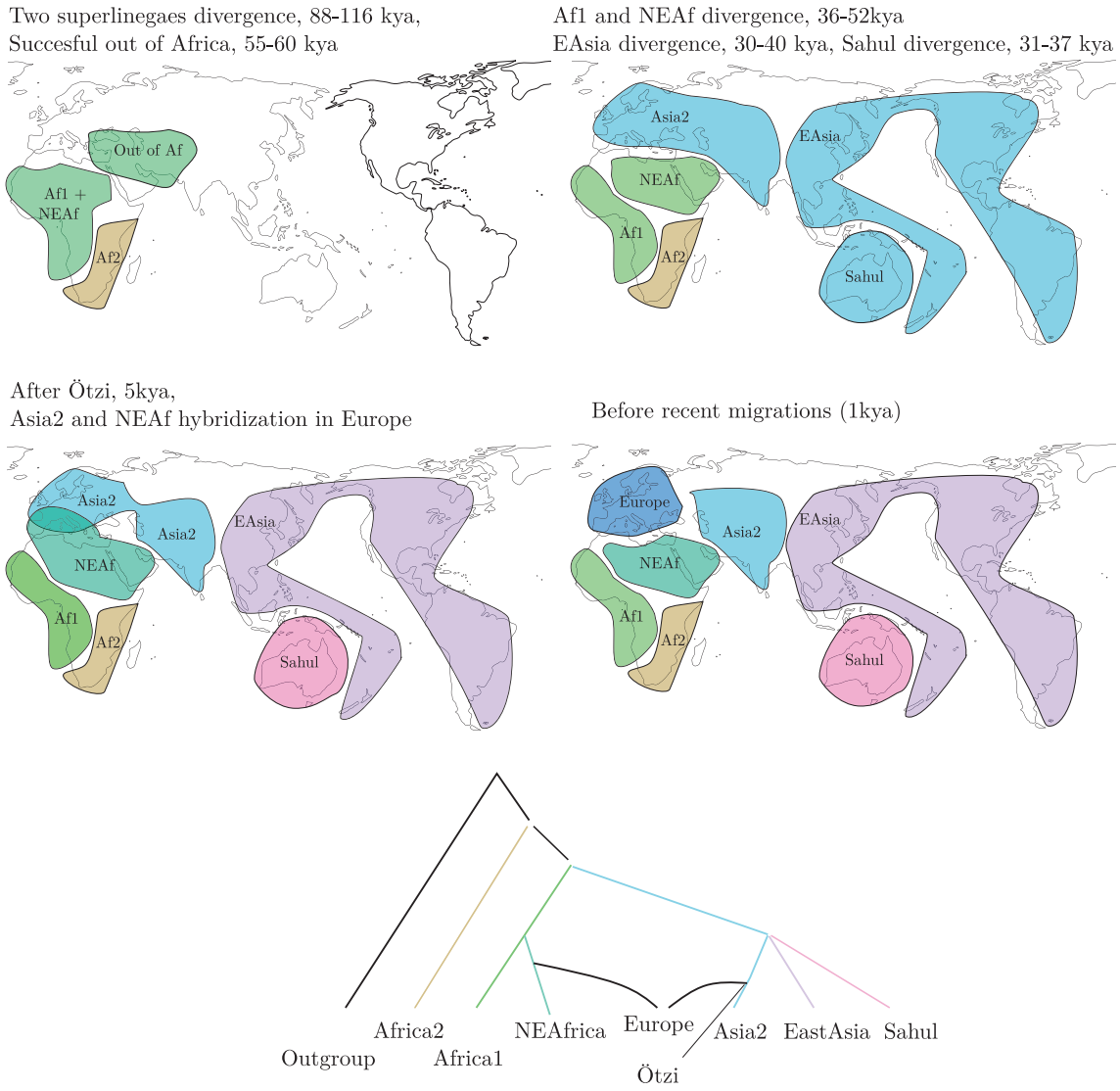


Figure 3.2: A potential story of *Helicobacter pylori* diversification. The estimated divergence dates between the populations are taken from [154], where Moodley *et al.* use human populations divergences to calibrate a model using genetic data to date *Helicobacter pylori* populations divergences. The estimation of the geographic ranges are taken from isolation by distance models, that aims at explaining genetic diversity by geographical distances, from [126]. The study of the ancestral strain of Ötzi is presented in [138]. This representation is schematic and indicative but it gives an idea of the major hypothesis for this story. At the bottom is depicted the resulting diversification tree of these populations. The modern repartition of *Helicobacter pylori*, for instance in the case of our dataset presented in figure 3.3, is quite different notably due to the important human migrations starting from the 16th century towards the Americas and Australia from Europe and Africa.

methodology of Muñoz-Ramirez *et al.* is similar to ours as it works with both whole genome and phylogenetic trees. Understanding *pylori* patterns of diversification in Latin America might be key to understanding the European population structure, as it is an example of population mixing for which we have a better understanding of human migrations.

3.1.5 Dating *pylori* population tree

Moodley et al., in 2009 [155] and 2012 [154], use previously estimated dates for some human population splits to calibrate two models of diversification for *pylori*, first in an application to the peopling of the Pacific, and then with a dataset with strains from Africa, Asia, and Europe. The first method they use, ClonalFrame, is a coalescent approach for bacterial evolution [59], while the second, IMa, is based on fitting an isolation model with migration to haplotype data [95], a method often used for eukaryotic populations. Both are used with the same set of calibrations and with MLST data. The two methods give similar results. The dates obtained and resulting scenarios are presented in figure 3.2.

The calibration events they use, corresponding to splits between *pylori* population or subpopulations, are the following (while the time estimates are based on the corresponding human population splits): Out of Africa, Split between Central and East Asia, peopling of the Americas, first humans in Taiwan, start of the Austronesian expansion, expansion to South America.

3.1.6 Ötzi the iceman

Ötzi, the iceman, is a mummy found in the Alps, dating from 5000 years ago. It was shown to harbor a *pylori* symbiont that was then sequenced. The strain was found to be close to the hpAsia2 population, using FineStructure, with few admixtures, while current European strains display substantial admixture of hpAsia2 and hpNeAfrica ancestral populations [138].

It is a stunning example of paleomicrobiology and a rare way to look into the past of a host and symbiont relationship. It is an additional element that must be taken into account when devising scenarios explaining *pylori* distribution, and it could be used to calibrate the dating of a strains tree using molecular evolution models without depending on dates estimated at the host level. However it is a single strain and we have no idea if it is really representative of the strains present in Europe at the time, though using a parsimony principle, we should assume it is.

SECTION 3.2

Our approach

3.2.1 Investigate the European introgression

The main question we were interested in with this dataset and in the scope of this thesis was to get a better understanding of *pylori* population structure in Europe. More generally, we wanted to get a better grasp on the relationship between humans and *pylori* using a more integrative approach.

3.2.2 Multi-level explanations of horizontal transfers

A mechanism at the center of *pylori* population structure and of our reconciliation model is horizontal gene transfer. When we find in *pylori* strains from Europe that a substantial part of the genome comes from one part of the population tree and the rest from a different one, we must wonder how these two got there. This is a challenging question as multiple events could be responsible for the emergence of such a pattern.

It is a question of horizontal transfers at multiple levels. To explain a genetic transfer between two *pylori* strains, we can invoke events at each one of the multiple levels that constitute this system. Humans can migrate between geographical areas, *pylori* can move between human hosts, and genes can go from one *pylori* to another by recombination between them. How can we differentiate those events or their combination adding up to a mixed *pylori* genome, notably in Europe?

As the primary hypothesis about *pylori* distribution is coevolution with its host, there are many examples of human migration to explain *pylori* movements. For instance, the same population in East Asia and Amerinds are signs of human migrations between those two areas, as does the current Af1 and Europe populations in the Americas, for more recent migrations. Interestingly, there is a population-level structure between East Asia and Amerinds though the isolation between the two human populations dates back to 19-23 kya [154], showing that the population structure is not only a signal of recent geographic uniformity.

On the other hand, *pylori* switching host is the main factor used to explain that most Amerindians bear European or Africa1 *pylori*, presumably obtained from African or European migrants post-1492 [76]. It is also the hypothesis for the presence of hpAfrica1 and hpNEAfrica *pylori* strains in the Baka pygmies, obtained from their Bantu neighbors who migrated to the area around 3 to 6 kya [167], while the Baka pygmies branch close to the root of human population trees, together with other populations associated to hpAfrica2.

Finally, the mix of ancestral populations that constitutes the current populations, at diverse degrees, results from recombination between bacteria from these different ancestral populations. Furthermore, for Europe, these significant mixes happen with no assumption of important human migrations between North-East Africa and Europe in the last 5000 years [3].

3.2.3 Using phylogenetic trees and the gene level

What makes the approach of Alexia and Vincent in our team different from previous studies is the use of phylogenies and consideration of the gene level instead of mainly using clustering techniques and data at a SNPs level.

The literature contains few phylogenetic trees, an exception being [163] with phylogenies from MLST concatenates, from whole genome, and for genes associated with pathogenicity. Nevertheless, there is no approach explicitly taking into account multiple gene trees.

Gene tree is a practical intermediate level, a way to sum up, with biological meaning, the nucleotide level. We get a usable number of phylogenies for further analysis. Amalgamation makes it even possible to keep some variation in the trees to consider opposite information from different parts of the gene.

Binary trees are not a straightforward answer to the representation of the diversification of populations. The same goes for gene trees with important recombination (see, for instance, in [76] different SNPs inside the same MLST gene can be assigned to different ancestral populations). A more adapted model is phylogenetic networks, and it was used for *pylori* in [246] and [163]. However, networks are not as easy to use as trees, and multiple methods need trees instead of networks. Interestingly enough, it is not the case for phylogenetic reconciliation: the upper tree can be a network [203], though it has not been used much. However, using binary trees, reconciliation can use horizontal transfers to model introgression, and reconciliation scenarios are a way to account for discordant signal between the genes.

Nevertheless, trees have been used to study human populations. For instance, in [121], a maximum likelihood tree is constructed from 150000 SNPs and is coherent with what we know of human migrations. It is also congruent with ancestral population structure constructed with methods similar to the one used for *pylori* that display similar mixed individual genetics.

Our approach here is to assume gene diversification can be represented by trees and that the representation of the strains by a phylogenetic tree can be informative.

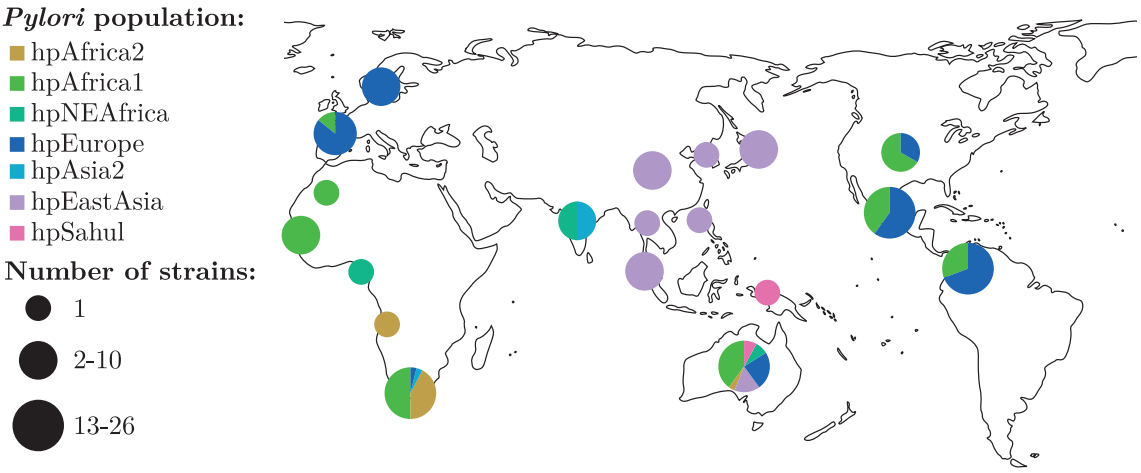


Figure 3.3: Distribution of *Helicobacter pylori* populations depending on the geographical sequencing area. Figure similar to figure 3.a presented in [76] but with our dataset of 119 strains.

3.2.4 Dataset presentation

Alexia Nguyen Trung collected available current strains of *H. pylori* from the NCBI and used pubMLST to find supplementary populations assignments by MLST allelic profiles [2, 103].

From the starting 1136 strains, 483 were kept under the following conditions:

- At least one population assignment
- If two assignments, they must be the same
- Whole genome available

While they are important to our study of European introgression, few hpAsia2 and hpNEAfrica strains were available with whole genomes. A phylogenetic tree was built based on the concatenation of the universal-unicopy genes (322 genes) and a sample of 113 strains representing the diversity of *H. pylori* in the old world was obtained using Treemmer, a Python tool to reduce the size and redundancy of phylogenetic datasets [149]. HspAmerind strains, the subpopulation of hpEastAsia associated with Amerinds, were discarded to focus on the old world. Then, six non-*pylori* strains were added (*H. hepaticus, H. acinonychis, H. canadensis, H felis, H. bizzozeronii, H. cetorum*) as an external group.

Then, Alexia kept the 1034 gene families, including 322 universal unicopy families, which displayed strains from the external group and population assignments from at least three continents.

All leaves were associated with a *pylori* population (and sometimes subpopulation), as well as the geographical position of sequencing.

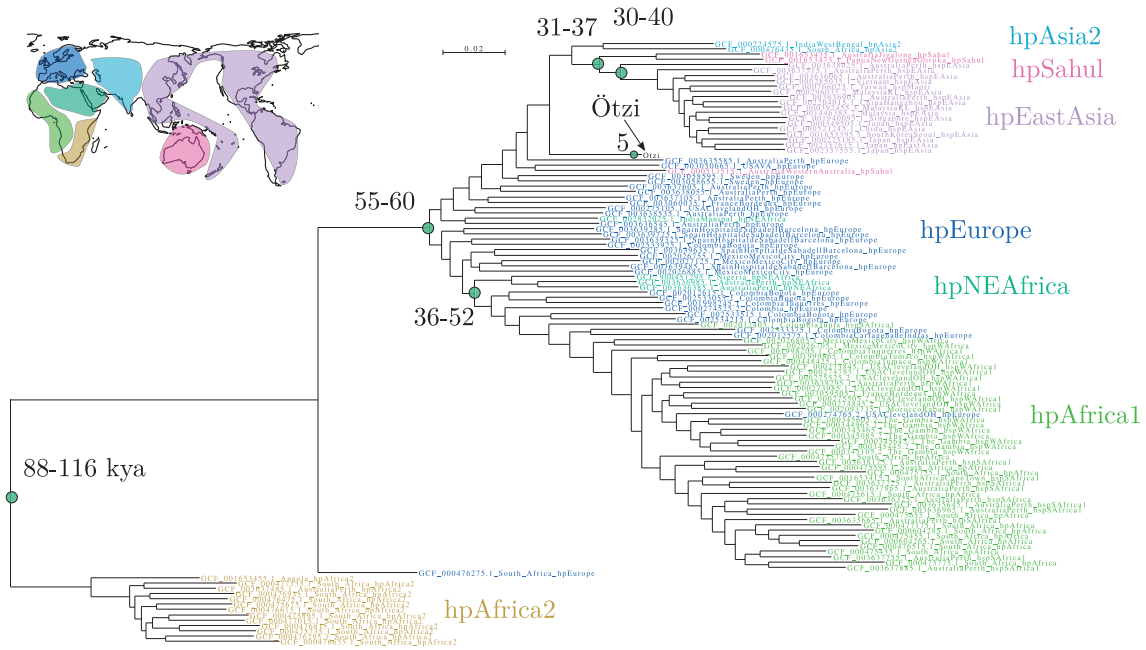


Figure 3.4: Our phylogenetic tree, with the nodes colored depending on the structure group. The time estimates are the one from [154], associated to the corresponding divergences in our phylogeny.
A landscape version is on the next page.

Alexia added Ötzi's *pylori* genes to the dataset after getting a reannotation with Prokka [205], which was validated by Jean Pierre Flandrois using PGAP [252]. From the 322 universal-unicopy genes, 319 are also present in this ancient strain.

3.2.5 Phylogenetic tree of the *pylori* strains

From a concatenate of the 319 universal-unicopy gene trees, Alexia and Vincent constructed a phylogenetic tree of the strains.

Except for Europe, the rest of the populations placement in the tree is similar what is found with the Structure approach. Here, Europe is not a clade and is split between Asia and Africa1. If directly interpreting this phylogeny, the European population would be the ancestor population of Africa1 and Asia. The two clades of Europe could be coherent with the subgroup defined with FineStructure, but we do not have this subpopulation information in our data.

Finally, Ötzi is placed at the root of the Asian clade. Previous studies assigned it to Europe, but as a pure representative of the European ancestral population sister to Asia. Is this position a coincidence? Is Ötzi just a random European? Is Ötzi part of a different population? Or is the tree artefactual, biased by the hpNEAfrica genetic fragments in the European strains?

We could formulate another hypothesis based on this tree. The two European groups were separated, one migrating towards Africa, the other towards Asia, com-

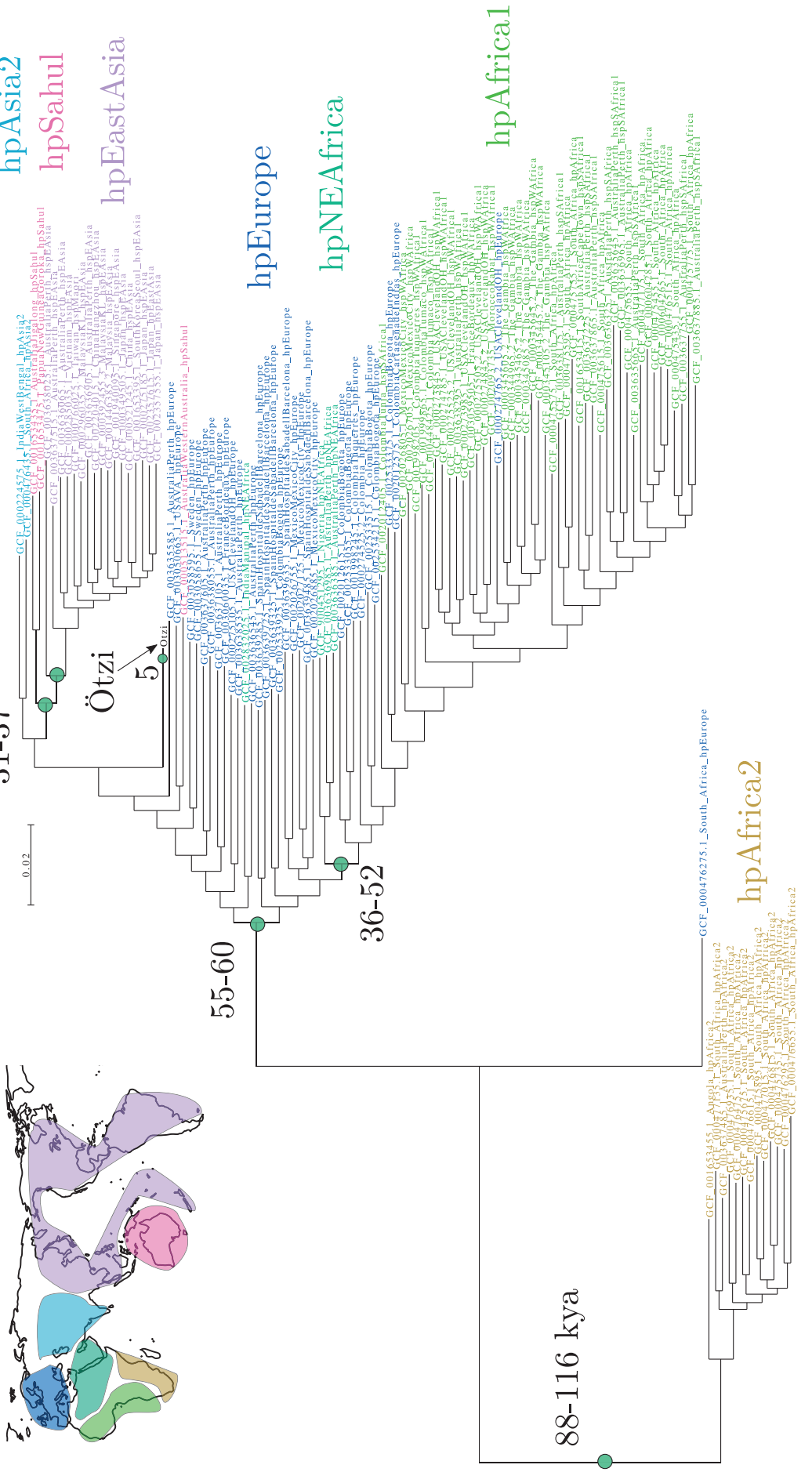


Figure 3.5: *Pylori* phylogeny.

Figure 3.6: An example matching of population and *pylori* species tree to give an idea of the size difference between the two trees.

parable to our phylogenetic tree, then mixing between the two Europeans to gain a global European structure and get Ötzi at the root of Asia.

In the rest of this chapter, we will assume that gene topologies are well supported, which means that introgression at the subgene level is negligible and that we can use the gene trees to separate the different possibilities for the position of Ötzi and the European branches positions. We will investigate the European gene positions in the phylogeny:

Where do the European genes branch? Are they the results of recent introgression and then must branch inside the putative introgression source, Africa or Asia? Or do they branch around these groups? Where do Ötzi genes branch? In the middle of Europeans or not?

SECTION 3.3

Reassignation of the European strains

To investigate the questions posed by the phylogenetic tree, I went back to the gene level and used reconciliation to compare the gene trees and the population trees.

3.3.1 Material and method

Our first question was to look at the position of the hpEurope branches in the different gene families, compared to the other populations. As the European population

is thought to be the results of an introgression, we wanted to see how this would translate at the gene level. For this experiment, we started from a population tree and gene trees. We wanted to challenge the European assignation of the strains, so we developed a simple experiment using reconciliation to see where the branches of the gene trees assigned to Europe would go in a population tree.

We started from a population tree without the European branch, and added potential hpEurope branches at every node in the population tree. We then enabled uniformly the hpEurope strains to match in all the branches of this tree. Usually, the initialization of the induction that defines reconciliation probabilities is to give a probability of 1 to all observed matchings, with a one-to-many matching from species to genes. And then this 1 is processed along the tree with the induction equation. However it is possible to put something else than 1 in this probability, to account for an uncertainty for instance. To the best of our knowledge it is not something that has been used in other reconciliation approaches, and is it also suitable to parsimonious approaches, even though the definition is a bit different then (cost of 0 to start from different matches for instance). As we discussed in the compartment inference section in previous chapter, it can be useful to construct compartments and let the genes choose their compartment based on reconciliation.

With $|H_L|$ the number of population tree leaves, denoting the matching as $\in$, for all gene leaf u and population leaf e:

$$\begin{aligned} P_{e,u} &= \frac{1}{|H_L|} \text{ if } u \text{ is a hpEurope strain and } e \text{ is a leaf} \\ P_{e,u} &= 1 \text{ if } u \text{ is not a hpEurope strain and } u \in e \\ P_{e,u} &= 0 \text{ if } u \text{ is not a hpEurope strain and } u \notin e \end{aligned} \tag{3.1}$$

After reconciliation of the gene trees with this population tree in a DTL model, we looked at posterior probabilities to see where the branch preferentially went. For each couple of gene families and european strains we obtained a posterior distribution of population leaves.

3.3.2 Results

We reconciled the 1034 gene families with the population tree, and sampled 50 reconciliation scenarios. We observed the leaves of the population trees where the European leaves matched in these scenarios. European leaves could match in two kinds of leaves, the leaves of the initial population tree, where the other populations strains match (for instance hpAfrica1), or newly added branches, noted as hpEurope_i, where only the European strains can match.

We see that European strains preferentially match to these newly added branches

instead of the sister branch corresponding to the other populations and where the strains assigned to other populations match (figure 3.7). This could be the sign of ancient introgressions. We also see that Asian population, and even more, African populations attract these branches, while the basal populations, Outgroup and Africa2 are less observed (which is a hint at the validity of this approach). However we do not really see here a preferential attraction toward the exact populations said to be the sisters of ancestral European ones, NEAfrica and Asia2. Though this might be linked to an uneven number of strains in each population (with numerous hpAfrica1 strains notably, and few hpNEAfrica and hpAsia2 strains).

To get a more detailed view of the matching, we looked at it genome by genome 3.8. One of the strains, sequenced in Cleveland USA (number GCF_000274765.2) is strongly assigned to hpAfrica1, and it is coherent with its position inside hpAfrica1 in the phylogenetic tree. Another strain is quite different from the others, sequenced in South Africa (number GCF_000476275.1), with a signal linking it with the hpAfrica2 population. Its position in the phylogenetic tree is also unique. Those two strains might have been wrongly assigned. For the rest of the strains, we find mosaic genomes, with proportions of assignations to both Africa and Asia. As noted with the bar plot in figure 3.7, the European branches are mostly assigned to the added branches, and not to the other populations, which indicates more ancient genetic exchanges, as recent exchanges should match inside the populations from which it originates. As displayed by the heatmap cluster, apart from 5 stains, the 2 strains matching in hpAfrica1 and hpAfrica2 we already mentioned, and three others with a strong matching in hpAfrica1, the other strains are clustered in 2 groups, one with more matchings in Africa, the other in Asia. Though, all those strains have matchings in both. It might be possible to link these two clusters to the two European subpopulations, deemed North and South, identified in [220] using fineStructure, by reapplying fineStructure to our dataset. These cluster are also coherent with the phylogeny, where the European strains are separated after a diversfication, with one side containing the African clade, and the other the Asian one. We regrouped the African assignations and the Asian ones to have a more generic view and differentiate the two origins of European introgression. The resulting heatmap is presented in figure 3.11. This heatmap makes more evident the presence of both Asian and African origins to all European strains.

To check the validity of our approach, we tested it with strains assigned to the other populations, and that are not said to be results of introgression. We deleted all European strains from the gene trees, and selected ten strains representing the different populations. We then applied the same experiment with these strains as we did for the European strains, with the same prior matching. We did this check with 64 genes. If we use the heatmap of genomes and leaves matchings presented in

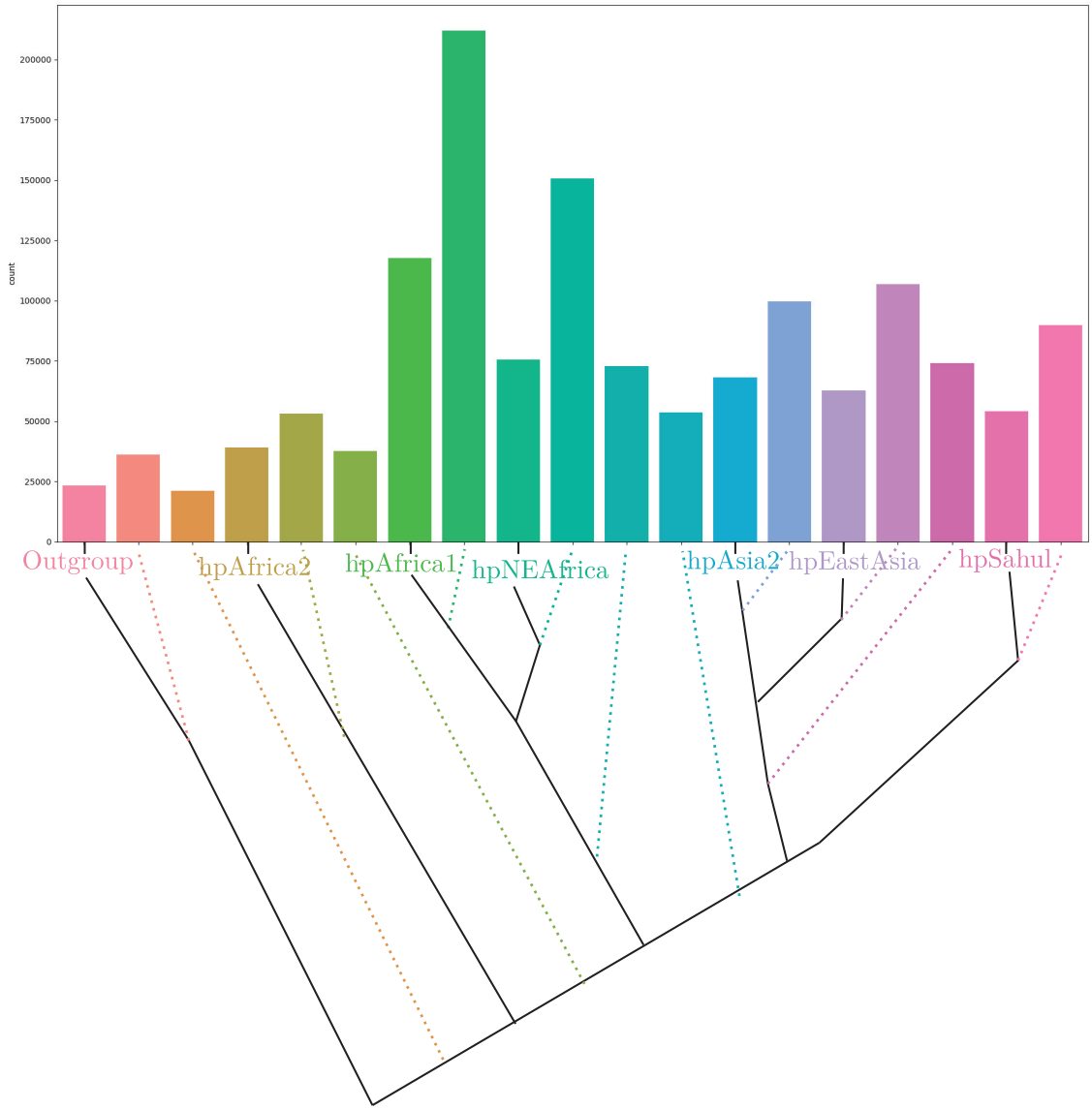


Figure 3.7: Bar plot of the matchings of the European branches. The observed matchings are summed over the 1034 gene families, the 30 hpEurope strains, and 50 sampled scenarios. Under the plot I reproduced the population tree with the added branches in dotted lines.

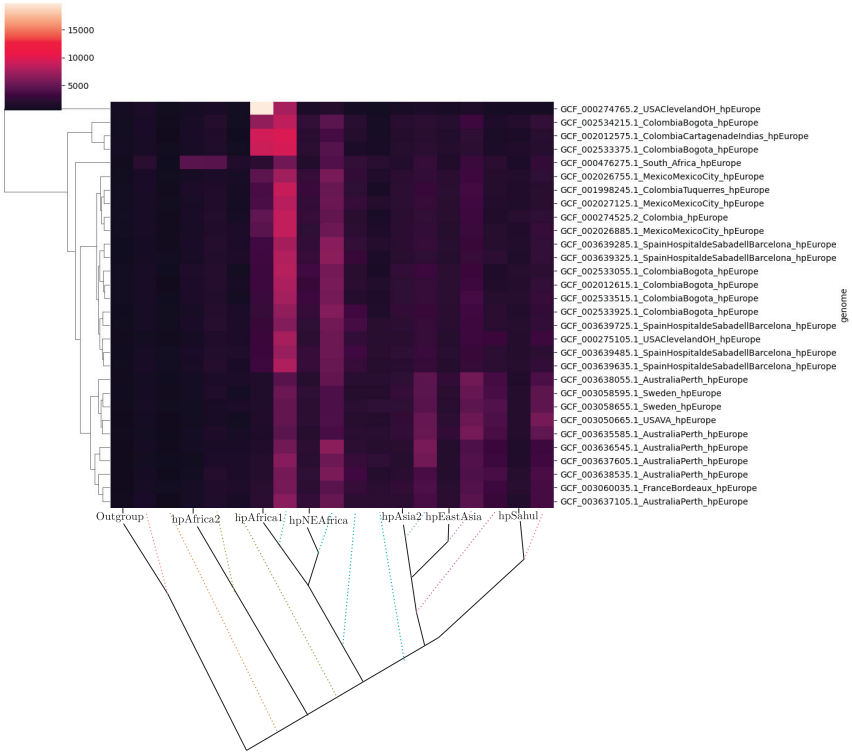


Figure 3.8: Heatmap of the matchings by genome. The light colors correspond to a high number of observation, and dark colors to small one.

figure 3.9 to assign populations to the 10 strains, we would make only one mistake, which is an outgroup strain. We see that for hpAfrica1 strains they mostly match inside the hpAfrica1 population, instead of the sister branches. However it is more balanced for the Asian strains, but still less than the hpEurope strains that mostly match to the sister branch. Apart from the hspMaori strain, the other Asian strains display matchings with branches around Africa.

Finally, we also applied the approach with the gene trees with the Ötzi strains. We used the 319 universal unicopy genes. As expected Ötzi strain is matched with Asian populations, but not African ones 3.10. Its matching is quite reminiscent of the ones of the hpAsia2 strain in figure 3.9, though with a lower match to the sister branch of Asia2.

3.3.3 Discussion

In this section we presented a new approach based on reconciliation to assign lower leaves from a prior uniform matching. The intuition behind the approach is that a leaf is matched depending on its surrounding branches, that have a fixed match, using a known model to get a posterior probability: reconciliation. The approach could be suitable for other datasets, for instance to assign new strains in a set of already assigned strains.

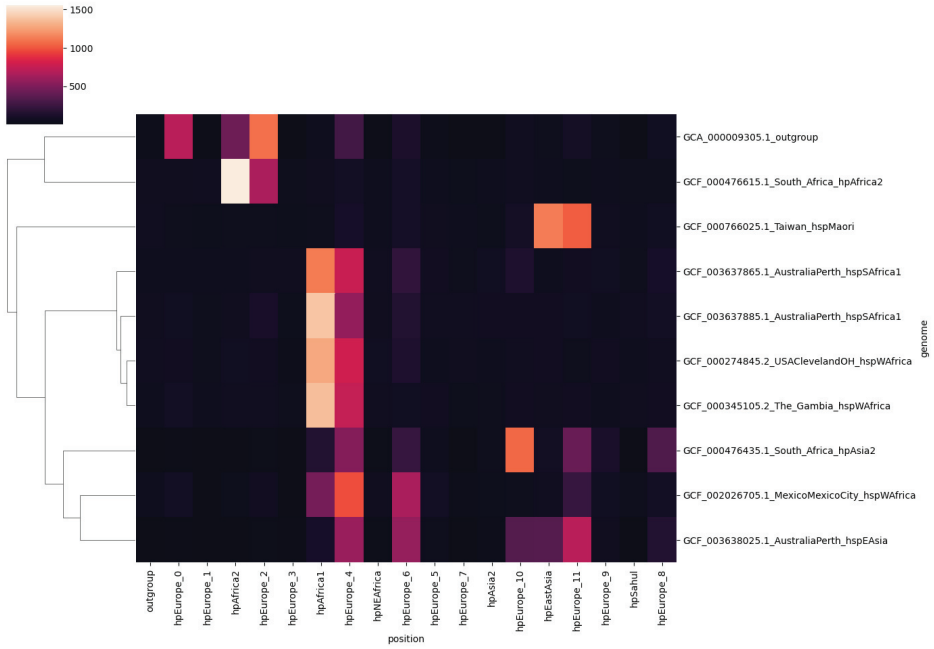


Figure 3.9: Heatmap corresponding to a test with non European branches to check the soundness of the method. The population to which the strains were previously assigned are displayed in their name, with the following correspondences of subpopulations: hspMaori is a hpEastAsia subpopulation, hspWAfrica a hpAfrica1 supbpopulation, hspEAsia a hpEastAsia subpopulation.

The results obtained on *pylori* show an expected mosaic genomes of hpEurope strains. However, there is a distinction between two European clusters, one connected to Asia, the other Africa, not agreeing with the idea of a continuous gradient between all European strains, and more congruent with our phylogenetic tree. The main limits of these results is the number of strains in the different populations, that is not well balanced. It might be useful to redo these analyses with more strains. Nevertheless, the heatmaps and barplot indicate a position of genes not inside the other populations but next to them, which might indicate an ancient origin of the introgression leading to the European strains population, so not really compatible with a Europe populated with hpAsia2 *pylori* strains during Ötzi time, 5kya. The alternative hypothesis is the presence from the divergence of Africa1, NEAfrica and Asian populations of two European populations, on each side of this diversification, leading to the two clusters we see today. And introgression between these two populations, in Europe after 5kya, might have lead to the emergence of an "artifactual" European population structure.

The interest of this new approach also lies in a gene level analysis of the phylogeny, and at a more granular level than just populations. Notably when we go from figure 3.8 to figure 3.12, we see a lot of information has been lost. What seemed like clear clusters looks like a continuous gradient. This is an incentive to add more

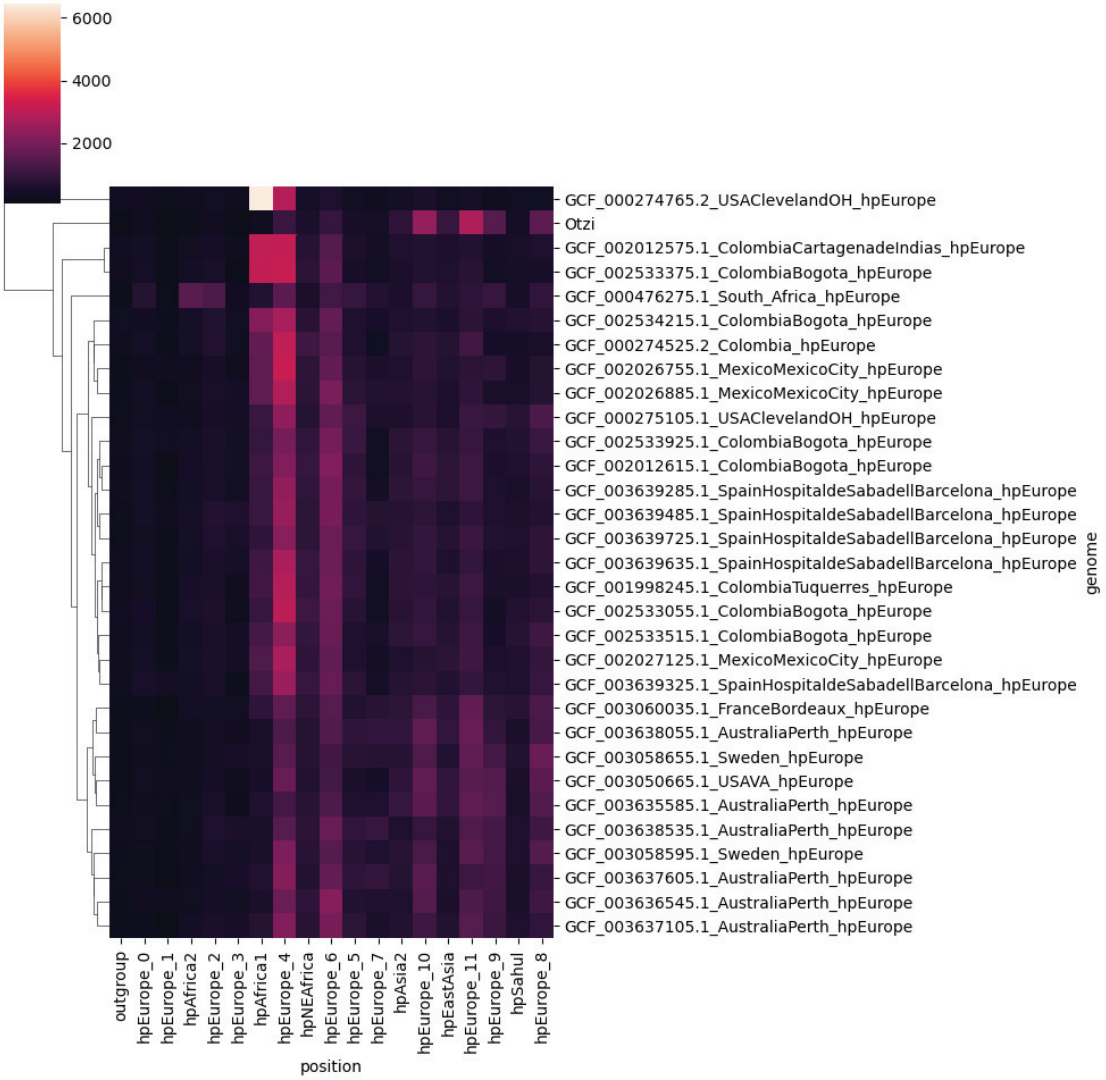


Figure 3.10: Heatmap of the matchings by genome, with the Ötzi strain and 319 universal unicopy gene trees.

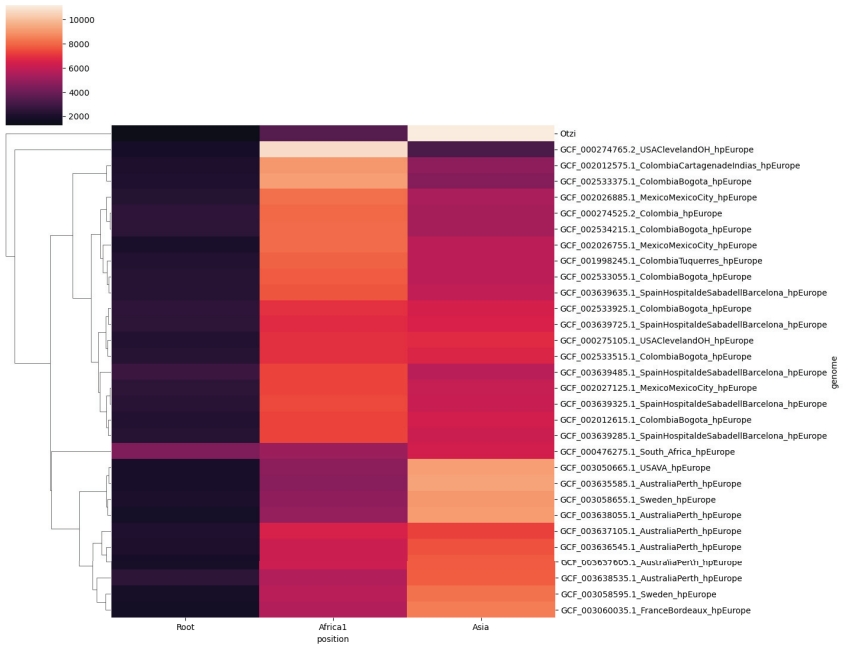


Figure 3.11: Heatmap and cluster with grouped elements, with Ötzi and 319 universal unicopy gene trees.

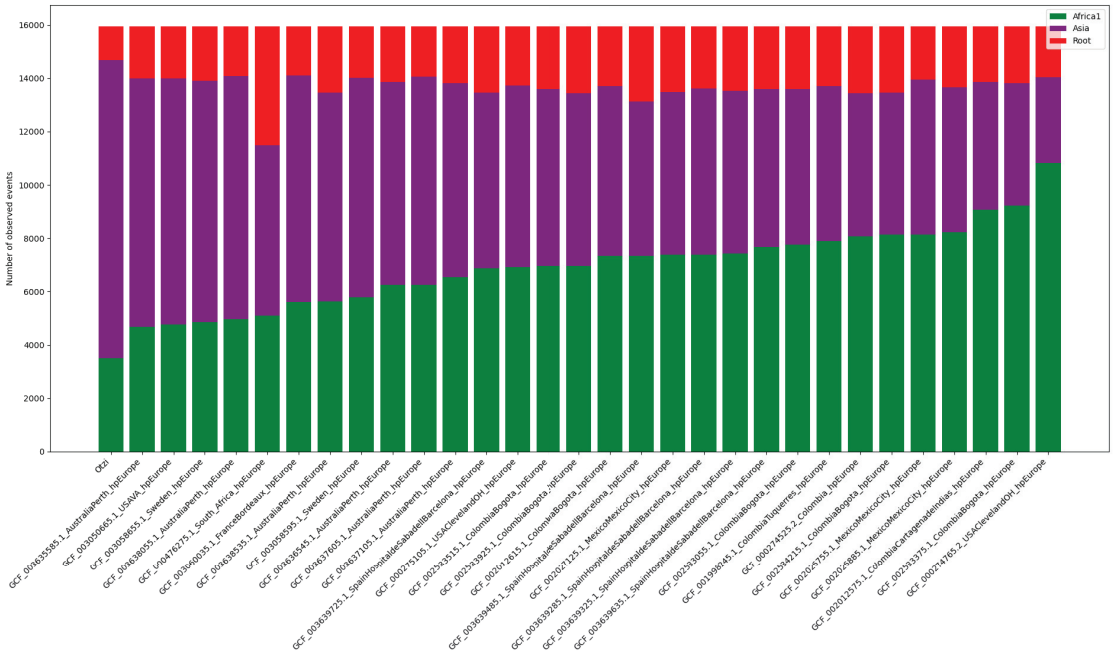


Figure 3.12: We order the strains depending on the number of matching to the African group in their gene families and sampled scenarios. It is a different representation of the same data used in figure 3.11. The structure that appears in figures 3.11 and even more in 3.10 is lost in this representation that looks like a continuous gradient between the different European strains. The figure is similar to figure 1.a in [126].

phylogeny in such populations analyses.

SECTION 3.4

3-level reconciliation with *pylori*

We then applied our 3-level reconciliation to population/strain tree/genes to compare different population trees and reconstruct the strain tree. It was also the occasion of using the dataset as a test case for our method.

We used amalgamation of the universal unicopy genes inside the population tree to reconstruct a strain tree, and then we compared the likelihood of the different population trees.

As we saw on the simulated dataset in our previous chapter, likelihood is a powerful tool for understanding and comparing models. In this case, we wanted to compare different host topologies regarding the question about the introgression and see how it could transpire in the population topology.

3.4.1 The question of random topologies

We wanted to test different topologies for the host tree as, if the European population was a clade, it was not clear where to place it. We constructed four trees that seemed the most probable to us, based on the position of Europe added on the rooted subtree of Asia, Af1, and NEAf. We did not consider the position with Europe closer to Af1 than to NEAf, as it did not seem to be a hypothesis presents in the literature.

To keep a verification step, we added a random population tree: a random tree on the Asia, Europe, Af1, Af2, NEAf and Outgroup leaves. That random tree is presented in figure 3.13 along with our four hypotheses.

However, the random tree does not fare worse than the others when investigating these trees, even better sometimes. We had three hypotheses:

- The random tree is not so bad and could be a potential hypothesis.
- We are not in the hypothesis of our model, so we can not use it.
- The variation we observe are artefactual, and the tree topology does not influence the gene reconciliation.

Other random topologies on the same leaves gave similar results. As we test only one host and strains reconciliation (without the Monte Carlo approach), and because the 2-level is more likely than the 3-level on this dataset, a reconciliation of the host and symbiont with many horizontal transfer events creates as many routes for the gene to do intra transfers, and as such a more likely gene and strains one.

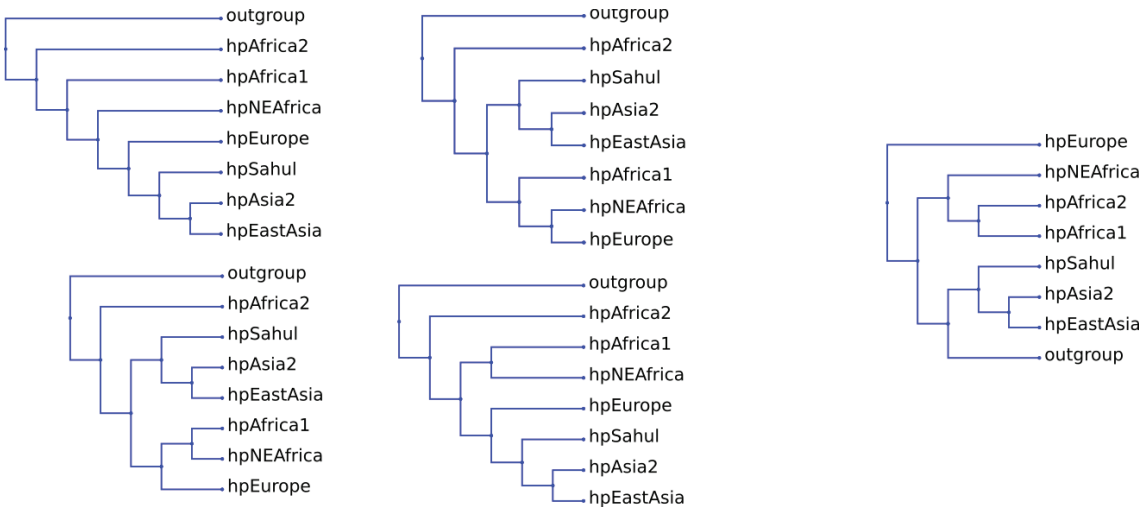


Figure 3.13: The four *pylori* pop trees hypothesis, and the random one on the far right.

We modified the model to make the topology more important, but we did not find a proper distinction between random and hypothetical ones. A description of these modifications is given in the following sections.

Nevertheless, the structure of the leaves is helpful for the reconciliation. If we use a random matching between the strain leaves and the population tree, that will impact the classification of transfers as inter or intra, the 3-level reconciliation likelihood decreases. Similarly, looking at the events in the reconciliation scenarios, there are many intra transfers between strains in the same population.

3.4.2 Modification of the 3-level model

I modified our model to account for the specific case of a population and strain levels and make the host topology more important to the reconciliation.

I event

The I event, which stands for incomplete sorting, combines a D and an SL event for one of the two copies. An example is given in figure 3.14.

It is an event that seemed necessary when looking at the topology of the symbiont tree and the distribution of population in the leaves. All populations are not monophyletic, and we have more of a comb topology.

We discussed this shape, notably when looking at the tree with branch length, and concluded that the shape of the *pylori* phylogeny we used, where only few clades are present, and it could be the result of the manipulation of Treemmer [149], when keeping a representative set of leaves from a tree with more strains[3], thus eliminating

[3]See the dataset presentation in subsection 3.2.4

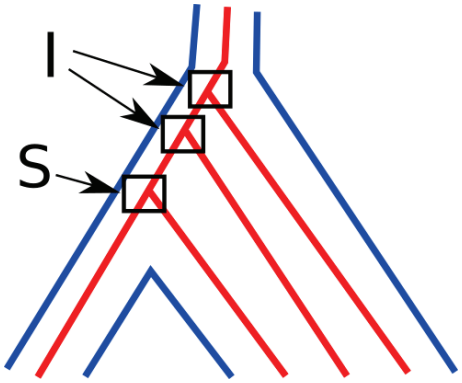


Figure 3.14: The new Incomplete Sorting event

some of the monophyletic terminal clades.

Regardless, it is helpful to add this event when considering the population tree as a geographic tree and with the intuition that some populations might be included in one another.

At the most basic, these adjustments are needed because we do not have a one-to-one matching between population and strains. Thus, we are not precisely in the cases of gene/species or host/symbiont reconciliation, but more in the one of biogeography. In biogeography, that kind of event is sometimes considered one of the default events [191].

The interest in adding a new event that is just a combination of already considered events is that this new event will have its own rate, making it different from a D and an SL.

Rates inference

When estimating the rates for the reconciliation of the population and the strain trees via an expectation-maximization algorithm, the speciation rate came to 0, as numerous duplications were needed. Moreover, there were almost no speciations, as there are only a handful of host tree nodes, while the symbiont one is big. It was more likely to undergo a horizontal transfer than a speciation with those rates. Thus, the strain tree originated in one of the host leaves and transferred to the others without visiting the host's internal nodes.

Adding the I event, the distance-dependent transfers, and considering I, D, and S as events with the same rates as neutral populational ones, made it possible to estimate the rates. But even with that possibility, we fixed the rates to compare the host topologies, as we discussed it in paragraph 2.1.5.

Distance dependent transfers

In order to increase the difference between the reconciliations depending on the host tree, I added distance-dependent transfers for the host and symbiont reconciliation. It is simple to add constraints to transfer in DTL reconciliation, as the possible transfers can be explicitly listed in the same fashion as the common host constraint in 3-level. Distance bounded transfers are discussed in [61], notably seeing how forbidding transfers to happen when the distance is bigger than a certain threshold can help reduce the number of most parsimonious solutions. Distance dependent transfers are discussed in [12], citing [5] that showed a bias in horizontal transfer toward more closely related species. The simple definition of Bansal *et al.* for the cost of transfer between two upper nodes a and b, is, with two parameters for the transfer cost t_1 and t_2, and denoting $d_S(a, b)$ the distance in number of nodes in a path between the two nodes:

$$C_T(a \to b) = t_1 + d_S(a, b)t_2 \tag{3.2}$$

We use a similar implementation, this time in probability. The sum of all possible transfers for one gene is p^T (which sums to 1 with the other events rates), we add a parameter α that gives how much we rely on the distance (0 it has no impact, and then more and more as it increases) :

$$P_T(a \to b) = \frac{P_t^{d_S(a,b)\times\alpha}}{\sum_c P_t^{d_S(a,c)\times\alpha}} \tag{3.3}$$

This modification of transfer rates for host and symbiont reconciliation can then be pushed to the gene symbiont aware of the host one, with our computation of inter transfer from intra and symbiont transfer rate.

3.4.3 2-level vs 3-level likelihood

The 2-level is better than the 3-level in terms of likelihood. However, as we pointed out in the method chapter, our approach is not perfectly ready yet.

The other point relevant here is that in our inter-transfer model, we assume the transfer of a symbiont ghost sister lineage to explain the gene transfer, and we count the probability of this ghost transferring (or doing a chain of events leading it to the gene transfer receiver). However, we count it for each transfer, while one ghost could explain more than one transfer. It could be necessary to correct this for a dataset with many gene families and transfers between the same nodes.

SECTION 3.5

Discussion

3.5.1 A perfect dataset for our model?

The coevolution of *Helicobacter pylori* and *Homo sapiens* is a fascinating dataset. There are multiple levels: geography, human, bacterial symbiont, and genes. There is substantial literature regarding the population structure with scenarios to explain the discrepancies between the host and the symbiont history. An important number of strains have been sequenced, all over the world, and in a variety of human populations, in an effort to understand this system.

However, as it is centered on humans, it is more complicated than with insect datasets to have both the symbiont genome and the corresponding host genome sequenced. In this chapter, we extensively used geography as a proxy for the relationship between the known human phylogeny and the one of *pylori*. Moreover, the evolution of a strain phylogeny inside a population tree might be slightly different from that of a host and symbiont or gene and genome, and our approach might lack essential population-level aspects.

One of the great difficulties of this model is how to disentangle the levels. It is hard to pinpoint the level responsible for an introgression. Have human populations migrated? Have *pylori* switched hosts? Or is it simply the results of multiple gene exchanges? While we need all of that to get to the introgression, one of them can be the main one. We only need one *pylori* donor to have a new gene in a significant part of a population. We only need one host migrating to transport a new *pylori* strain that can colonize another population.

3.5.2 A multidisciplinary, international and local question

It is easy to feel related to the dataset, as it is about our past migrations and a symbiont that has followed us for a long time, and that can be a pathogen.

As Europeans, the question of the peopling of Europe rings something. The as exciting question of the *pylori* population structure in Latin America is at the center of research led from Mexico (with collaborators in other Latin American countries, the USA, and Europe) [163]. While led from Europe, the studies of the African origin of *pylori* were co-authored by researchers from Sudan, Cameroon, and South Africa [167] [126]. The first study to use fineStructure, is from a Japanese team and is used on newly sequenced Japanese *pylori* genomes. A paper presenting a difference between *pylori* strains between stool and saliva in the same host is from Thailand and uses new local sequences [243]. The Austronesian expansion study,

led by European authors, is cosigned by Sénégal, Nouvelle-Calédonie, Taiwan and Papua New Guinea authors [155]. The history of *pylori* in East Asia [29] is studied by a first author affiliated to Cambodia and Senegal, with an American last author, with co-authors from République Centrafricaine, Cambodia, Europe, and the USA. *Pylori* population structure in Senegal and Madagascar is investigated by a similar team (with an exchange of the first and last author), with new authors from Senegal and Madagascar [127].

It is a truly international question, from the question itself about early migrations, that can interest any country but is anchored at a local level, by the necessity of sampling sequences, and moreover, by the medical aspect of *pylori*. Maybe it would even be possible to find common patterns between *pylori* studies authors and the *pylori* they consider (and in that case, we would be sure it is not coevolution!) It is also a really interdisciplinary system interesting both evolutionary biologists, the medical community - often at a more local level - and anthropologists.

What is also very interesting in the study of this model is that it went hand in hand with methodological development from the same team that presented interesting biological results (notably the Structure software, which then went on to be used for other models). It is enlightening to see this kind of cooperation: how methods can be applied and how biological models can be studied.

3.5.3 What happens next?

Maybe we have to take home more new questions than new answers. European genes do not seem to come from inside the other populations, which would correspond to an ancient introgression. It is hardly compatible with the idea that 5kya hpEuropean population was an hpAsia2 population that went through introgression from strains from the hpNEAfrica populations.

Alexia and Vincent are currently working on dating the tree using molecular evolution and calibrating with Ötzi. They might also add some constraints from horizontal gene transfer inferred from reconciliation to this dating. Notably, they try to date the common ancestor of Ötzi *pylori* and extant strains. They seem to think the results might be quite different from what we await, given the supposed coevolution of *pylori* and humans or previous attempts at dating this tree. I am eagerly awaiting those results that could shed new light on the *Homo sapiens* and *Helicobacter pylori* coevolution question.

Chapter 4

Open questions

Contents

In this chapter, I will present open questions more or less related with reconciliation.

These questions are the following:

- Which of parsimony or probability undated reconciliation is the closest to the dated simulation framework?
- Is the maximum likelihood species tree inferred from a reconciliation score the same for multiple unicopy-universal gene trees and for the amalgamation of these unicopy-universal genes?
- Is it possible to construct a "dual" of the reconciliation problem, and is it an interesting biological model?
- In a birth-death model what is the coalescent time to the first species with alive extant descendants?

I detail each of these questions in the following sections, with what is known for the moment.

SECTION 4.1

Parsimony and probability transfer models

4.1.1 Transfer rate and transfer cost

In undated DTL reconciliation, the probability and parsimony approaches are almost directly exchangeable by simple rewriting the induction equations and using the same underlying computational approach. They mostly differ in their interpretation and the use of their output. To rewrite the equations, we replace probabilities sums by taking the max, and products by sums, which can be seen as taking the log of the equations. So we also take the log of the probabilities p to obtain the costs c of events. However, the rewriting, using the same dynamic programming method, with max and addition, instead of addition and multiplication, do not give exactly the same equations, one event is considered differently in the two models: horizontal transfer.

In a probability setting, when a transfer is chosen with transfer probability p^T, we must then choose toward which receiver the transfer is headed, with a uniform probability for all receiver, we divide by $|S|$ the number of branches of the species tree, so we get a probability of one specific transfer to be $p^T/|S|$ while in parsimony the cost is simply c^T, with $c^T = \log p^T$. Transfers are thus dependent on tree size in probability, but not in parsimony. At fixed rate, with an artificially big tree, we can make some transfers more expensive than duplication and losses for instance. I first thought that maybe, via rates estimation, if the tree is big we will simply estimate a higher transfer rate than if it is smaller. So a first question, that we answer in the next paragraph, is: does rates estimation at maximum likelihood (via expectation maximisation for instance) enable us to bypass this problem, and get a solution that does not depend on the size of the tree? Will both models give the same output if we let them estimate rates at maximum likelihood?

4.1.2 Rates inference is not a solution to the problem

It is easy to construct examples where, even with rate inference, the two models, the probability one with $p^T/|S|$ and the parsimony one with c^T do not give the same output. For instance with a simple repetition of separate three leaves trees, with their two possible resolutions as the upper and lower tree. Figure 4.1 illustrates this example.

If we speak in term of probability, the left scenario of figure 4.1 has $1L + 1T + 2S$, the right one $1D + 3L + 3S$. In the first model we have a probability of the transfer scenario of $\frac{(p^S)^2 p^L p^T}{|S|}$, and we lose the $|S|$ to get to the second model, and

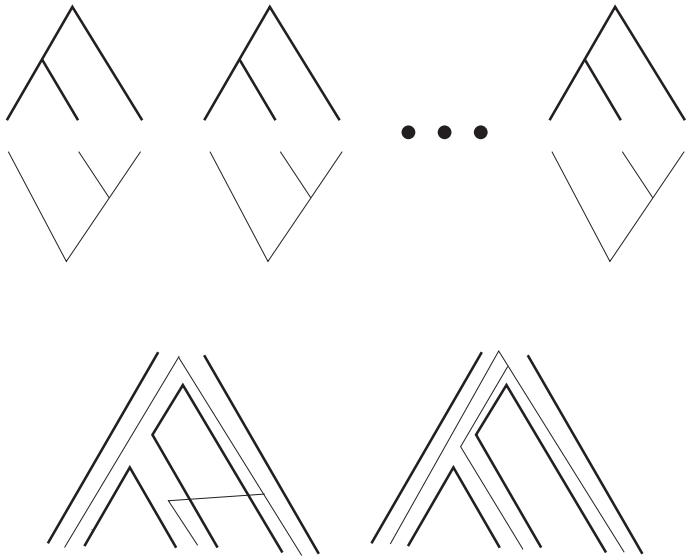


Figure 4.1: A simple reconciliation input for which depending on the number of elements a scenario involving transfers or duplications will be more likely.

the duplication scenario is $(p^S)^3(p^L)^3p^T$ for both models.

Let us replace $p^T, p^L, p^D, |S|$ respectively by t, l, d, h and using $p^S+p^T+p^L+p^D = 1$, and look for the maximum likelihood rates for both scenario, as $|S|$ does not interfere in this choice with this simple example.

$$g(s,t,l) = s^2tl \tag{4.1}$$

The extreme values get 0. We can assume that $d = 0$, as it can only get the probability lower, so $s = t + l$. Let us find the max value using the gradient:

$$\nabla g = = \begin{pmatrix} 2sl - 3s^2l - 2sl^2 \\ s^2 - s^3 - 2s^2l \end{pmatrix} \tag{4.2}$$

Solving $\nabla g = 0$, after factorizing by sl and s^2 we get $s = \frac{1}{2}, l = \frac{1}{4}, t = \frac{1}{4}, d = 0$.
We do the same with $k(s,d,l) = dl^3s^3$ and we get $s = \frac{3}{7}, l = \frac{3}{7}, t = 0, d = \frac{1}{7}$.
The maximum likelihood value is then:

- for the transfer scenario: $\simeq \frac{1.5610^{-2}}{h}$
- for the duplication scenario: $\simeq 8.8510^{-4}$

From $h \geq 17,6$ the duplication scenario is preferred. In its own tree, the transfer has 3 possible receivers, and in the other trees it has 5. So, from 4 three leaves widgets $(3 + 3 \times 5)$ the duplication scenario is preferred in the probabilistic model, while in the parsimony model, the transfer scenario is always preferred.

So rates inference with maximum likelihood does not give a satisfying answer to the question in all cases.

Other toy models can be constructed, for instance using a comb shape where it is more likely to transfer from the last branches of the tree than from the first.

4.1.3 Unconsistency of undated models

Our model is an undated one, and seems as such, though it is a probabilistic one, closer to parsimonious setting than to dated probabilistic models.

Felsenstein showed the unconsistency of parsimony for the inference of phylogeny from an alignment, by showing that as it did not take into account branch length, it was easily fooled. It is presented in [78] pages 107 to 117. Two of the branches are quite long, it is easier to do one event on each, than to do only one event on the short branch.

We can thus intuit that both parsimonious DTL and our probabilistic undated DTL can be fooled by similar premises if we take a simple dated model, as the one we used in this thesis.

4.1.4 The question

Both the parsimony and probability models are inconsistent compared to a dated model.

Is one of the two inference methods closer to the dated model?

SECTION 4.2

Amalgamation and consistency

In the introduction we showed that amalgamation could be a powerful tool to efficiently visit multiple trees and construct chimeric trees from the different clades observed. We also used amalgamation to use universal-unicopy gene trees as a distribution of species tree topology, making it a simple way to avoid using a concatenate. Figure 4.2 and 4.3 show the evolution of the number of clades (in which reconciliation has a linear running time) with the number of trees sampled, in the first figure using a sample of tree obtained from a tree reconstruction bayesian framework, or in the second figure using multiple universal unicopy gene trees seen as a distribution for a species tree.

Reconciliation is a method fast enough to consider thousands of gene trees on hundreds of genomes, but too slow to be used as a measure to guide a MCMC or similar bayesian frameworks. Speed up, like the ones implemented in GeneRax presented in the introduction, are important for this kind of use.

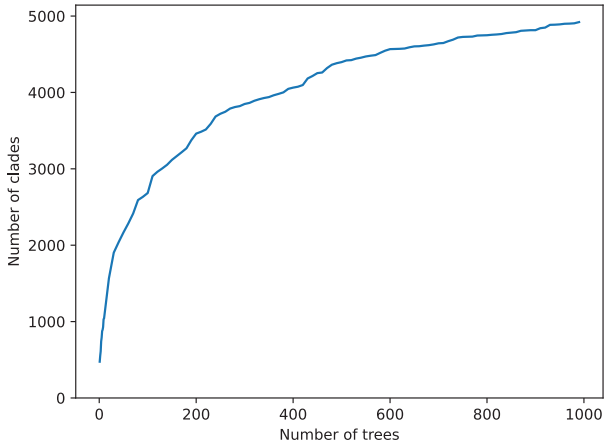


Figure 4.2: The evolution of the number of clades with the number of sampled trees for one of the gene families in our *pylori* dataset. The gene trees contain 119 leaves, which means 471 clades. At first the number of clades grows fast, and then each new tree adds few new information. But even from the first new elements, it is faster than just doing each tree independently.

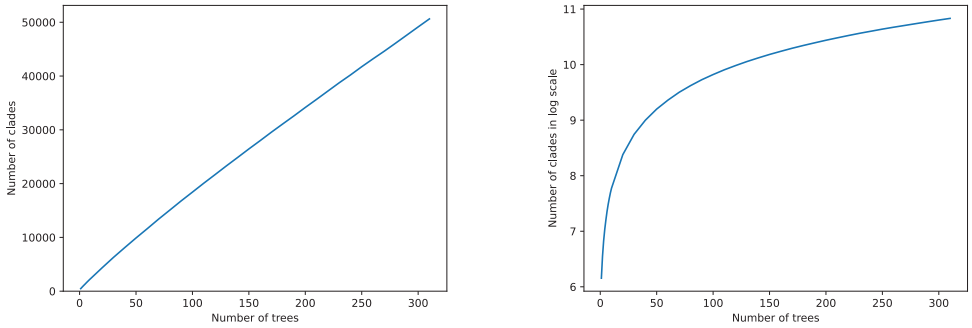


Figure 4.3: The evolution of the number of clades with the number of trees amalgamated from the 322 unicopy universal gene families in the *pylori* dataset. Like in the previous example the number of clades starts at 471, but from then the evolution is quite different. It is still useful to use amalgamation as each gene adds around 160 new clades (instead of 471). The figure in log scale shows that there still is some point where the increase slows down.

To use reconciliation as a measure of likelihood for a bayesian approach, all we need to know is how well a new model gives similar indications than a starting one. One simple way to express this question is:

Is the maximum likelihood species tree inferred from a reconciliation score the same for multiple unicopy-universal gene trees and for the amalgamation of these unicopy-universal genes?

Or, in a simple case with two gene trees G_1 and G_2, and two potential species trees S_1 and S_2, looking at the reconciliation likelihood, and noting AG_1, G_2 the amalgamated tree-like structure obtained from conditional clade frequencies observed in G_1 and G_2, do we have:

$$P(G_1, G_2|S_1) > P(G_1, G_2|S_2) \iff P(A_{G_1,G_2}|S1) > P(A_{G_1,G_2}|S_2) \tag{4.3}$$

The question can be seen as exchanging amalgamation and likelihood computation. In the first case we compute likelihood for each gene family then we "amalgamate" these likelihoods by product. In the second case, we first amalgamate the gene trees and then we compute the likelihood.

It could not be used however for scenario inference though, for instance to see the frequency of a particular transfer between two hosts, whatever the gene family. If one of the topologies is closer to the host, it will be more likely to follow it, while with multiple gene family we always have to keep all trees in the sample.

SECTION 4.3

Exchanging upper and lower

When working on the "Phylogenetic reconciliation" review, we compared gene/species and host/symbiont reconciliation models. Notably we considered two events specific to each of these frameworks. Failure to diverge, on one hand, [225], allows a symbiont species to colonize multiple host species. Incomplete lineage sorting, on the other hand, [35] is a population level effect ending with a different history for gene and species trees. Both are in a way population effects, as failure to diverge could be seen as a symbiont species that we could divide into two populations depending on their host (in case of only one host by individual, which is not always the case), and that inevitably the populations would diverge into two species and the failure to diverge would end. For ILS, what we divide is the species population depending on the version of a gene. These are similar events, in Failure to diverge lower level have multiple upper ones, while in ILS, the upper level has multiple lower ones. Paragraph 1.4.7 in our review [150] is devoted to those two events.

What is interesting here is that we have a similar event, but with an inversion

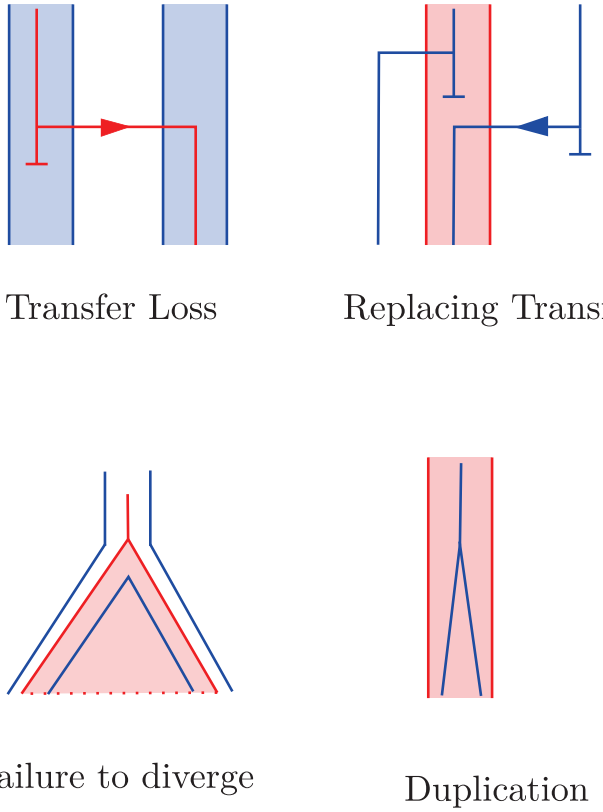


Figure 4.4: We can look at the dual events of the events considered in reconciliation, by inverting the upper and lower levels. For instance the "Transfer Loss" event becomes some sort of Replacing Transfer, while Failure to Diverge is simply a Duplication.

of the role of the upper and the lower levels. In one case the upper level has a polymorphic lower one (species and gene), in the other the lower level has a polymorphic upper one (symbiont and host). So we asked ourselves if inverting the upper and lower levels could give rise to an interesting model, as we could see that with some tinkering failure to diverge could be identified as an inverted ILS. Interestingly, replacing transfer seemed to invert to a classic transfer, while it is hard to reconcile with replacing transfer ([91]). We tried to consider the dual model to DTL reconciliation of a tree A in a tree B, by the reconciliation of the tree B in the tree A with the inverted model.

However the resolution in this dual model of replacing transfers was an illusion, as the dual model presents strong new constraints on the events of the upper level. The new model did not seem to have a biological interest, as it obligates upper elements to always carry a lower one, so we did not went further in this analysis.

A similar model, that tries to keep the symmetry between the upper and lower roles is presented in a really interesting paper [237], where the authors also insist on the similarity between host/symbiont, gene/species and biogeography reconciliation, with a model that can account for any type of events expressed in terms of common

patterns between the two trees.

SECTION 4.4

Coalescent time in birth-death models

At some point last year, Eric added me to a discussion with Damien De Vienne in our research team, concerning birth death models, to add a theoretical contribution for a paper on the impact of extinct lineages on a phylogenetic method from Pittis and Gabaldon [184] to order waves of gene acquisition by transfer. That paper is out now [223], though without this additional analyses.

With a tree generated under a birth death model (and before pruning the branches without descendant at present), the goal was to find the probability distribution of d the time such that a branch at time t has its closest ancestor with non extinct descendants at time distance d (figure 4.5).

The second question, that was the one of interest, is: if we take two branches at given times t_1 and t_2 in the complete tree, what is the probability that the order between t_1 and t_2 is conserved in their first ancestor with alive descendant, so $P(t_1 < t_2 | t_1 + d_1 < t_2 + d_2)$. The goal was to challenge an approach presented in [184] to give relative time order to gene waves of acquisition during eukaryogenesis, with the presence of ghost lineages. In a way, we wanted to see how much can be said about the ordering of transfer givers (that are, at best, traced back to their first alive ancestor).

Even though we thought the question was quite straightforward we could not find existing results. A result like the coalescent time of two nodes dates from 2015 [210]. We stopped working on this when Damien discussed it with Helene Morlon, who seemed to think it was not trivial and so not in the timeframe for that paper.

In the rest of this section I give some leads and some of the ideas we had considered for the first question, the coalescent time to an ancestor with extant descendants.

To better define the question, we simulated a tree with a birth death process with parameters (δ, μ), and we consider the complete tree (before we prune the extinct lineages). Given a time t we choose uniformly one of the branches of the tree present at time t. Two cases, the branch drawn can have descendants at present, and so we say it is at distance 0 of its closest ancestor with descendants at present. Or, the branch has no alive descendants and so we look for the first of its ancestor with alive descendants at present. We call that distance d. We would like to know the distribution of d given δ, μ and t.

Let us start by retrieving the extinction probability $E(t)$ for a branch chosen at time t in a birth death process (δ, μ). In our question it corresponds to $P(d > 0)$.

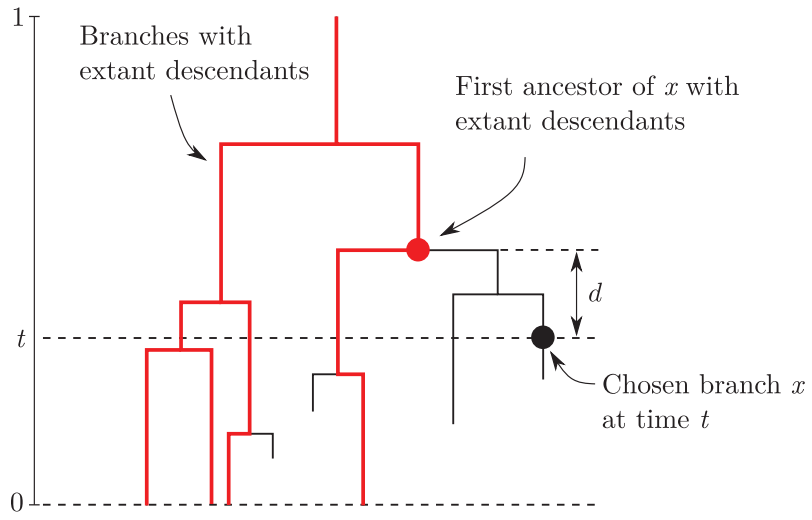


Figure 4.5: Illustration of the problem. Branches with extant descendants are drawn in red. The question is to determine the distribution of d_0, the distance to the first ancestor with extant descendants from a branch drawn uniformly from all branches present at a given time t.

As it is often done, we put present time at 0 and we increase t to go back toward the past. Between $t + dt$ and t, our chosen lineage can go extinct, speciate, or nothing happen. Neglecting the possibility of multiple events in that delta of time, that have a probability of the order at most dt^2, we have:

$$E(t + dt) = \mu dt + \delta dt E(t)E(t) + (1 - (\mu + \delta)dt)E(t) \tag{4.4}$$

And with dt going toward 0:

$$E'(t) = \mu + \delta E(t)E(t) - (\mu + \delta)E(t) \tag{4.5}$$

The probability of going extinct at time 0 (when already at present) is 0 so $E(0) = 0$. We can solve this Ricatti equation with constant coefficient. We get:

$$E(t) = \frac{1 - e^{(\delta-\mu)t}}{1 - \frac{\delta}{\mu}e^{(\delta-\mu)t}} \tag{4.6}$$

We can then look for $P(d > d_0)$ for $d_0 > 0$. To define a bit more formally the question, we name x the branch that we chose. A first approximation of this, is what we called $EE(t + d_0)$ which is the probability for the ancestor of x at time $t + d_0$ to not have any alive descendants at time 0. As we did for E, we can find a differential equation for EE. In a dt time interval, the ancestor of x can speciate, and then there are two possible branch that can be attributed to x, but it cannot go extinct as it would not be x ancestor in that case:

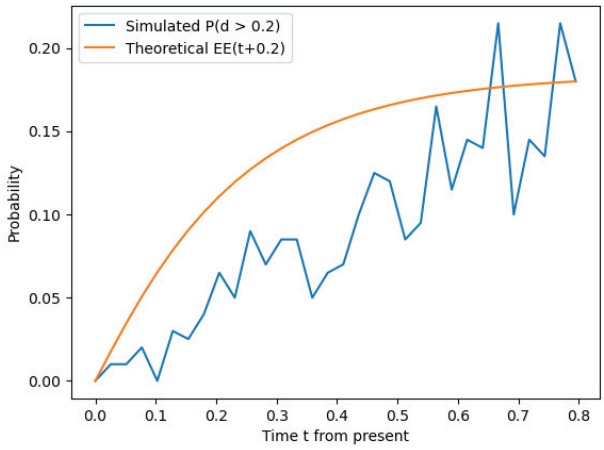


Figure 4.6: Comparison of our theoretical probabilities for EE and observed frequencies on simulated data for the probability $P(d > d_0)$. For $d_0 = 0.2$ and $\mu = 5, \delta = 8$, and for varying time t.

$$EE'(t(x) + d0) = 2\delta EE(t(x) + d0)E(t(x) + d0) - (\delta + \mu)EE(t(x) + d0) \tag{4.7}$$

With initial condition $EE(t + 0) = E(t)$, we find for this first order linear differential equation:

$$EE(t + d0) = \frac{e^{(\delta-\mu)d0} - (1 + \frac{\delta}{\mu})e^{(\delta-\mu)(t+d0)} + \frac{\delta}{\mu}e^{(\delta-\mu)(2t+d0)}}{(1 - \frac{\delta}{\mu}e^{(\delta-\mu)(t+d0)})^2} \tag{4.8}$$

What is the difference between EE and our question? What we want is EE knowing that x was chosen at time t, so a branch with a lot of descendants is more likely to be chosen than one with few.

$P(T > d_0) = P(x$ ancestor at $t + d_0$ has no descendants at time $0\ |x$ was drawn from the branches present at time $t)$

A first lead on the question is to condition on the number of branches at time $t + d_0$ and time t, as the distribution of the number of extant leaves in a given time is known. An equivalent question is the number i of cousins of a branch chosen at time t, from the time $t + d_0$, and then we just have each cousin to go extinct with probability $E(t)^i$.

The following equation might work, obtained from conditioning on the number of leaves at several times in the tree and of different subtrees, if rewrote properly, as it is not tractable as written. We note $p_i(t)$ the probability of having i leaves in time t from one ancestor:

$$\sum_{j=0}^{+\infty} p_j(1-t+d_0) \sum_{n=1}^{+\infty} \sum_{\sum_{i=1}^{j} k_i = n} \frac{n!}{k_1! \ldots k_n!} \prod_{i=1}^{j} p_{k_i}(d_0) \times \frac{k_i}{n} E(t)^{k_i} \tag{4.9}$$

Another path to the question might be to look at the derivation of the coalescent time in [210], where they do so by considering the vector of coalescent events.

Chapter 5

General conclusion

I have already proposed several discussions in the previous chapters of this thesis. At the end of our review of phylogenetic reconciliation 1.4.10, I gave a small insight into what might be interesting future directions in the subject but using a neutral point of view without direct mentions to the models we were developing. At the end of chapter 2 (2.5), I discussed possible biological models for our 3-level reconciliation approach that we considered without having the possibility to pursue them. I closed the matter of *Helicobacter pylori* at the end of chapter 3 (3.5).

What do we still have to discuss? Maybe we can first look back at the introduction and our review of phylogenetic reconciliation, and we can see how this thesis stands in relation to the recent advances in reconciliation. Our review presents the different questions at hand for the two main communities in reconciliation, the host/symbiont one and the gene/species one.

A recurring question, notably in recent works in host/symbiont reconciliation, is how to deal with the uncertainty in the output scenarios. The question presents two sides, how to produce a ready-to-use output for a potential naive user or on the other side how to give interesting information to a user with a good idea of how the methods work. Proposing Thirdkind, our reconciliation viewer, and using a probabilistic framework is a way to answer this call. It is a simple way to aggregate multiple outputs, focusing on the position of horizontal transfers, while methods like Empress[201] or Capybara[232] give more complex outputs based on clustering or well-defined equivalence classes. One downfall of our approach is that we cannot simply differentiate between transfers that cannot happen in the same scenarios. Even though it is not something we presented as such, the 3-level reconciliation, coupled with Thirdkind, is a way to structure the graphical output of reconciliation scenarios. For instance, we gave a visual representation of the reconciliation of 1034 gene trees of 119 leaves with the *Helicobacter pylori* strains tree, with a population structure on top of it. This scenario could not be so easily visually embraced without

the structure of the host.

We also saw in our review that the DTL model of reconciliation resulted from continuous complexifications. In recent developments making reconciliation models more integrative is more about integrating new concepts than adding new patterns in the same framework. For instance, adding replacing transfers is more about adding interdependency between sister lineages than a new pattern for the trees to evolve. The same is true for ILS, which is, for the moment, also made by this consideration of interdependency and a new pattern, but which ultimately is a step toward the integration of population genetics inside phylogenetic reconciliation (or the other way round, the integration of phylogenetic reconciliation to population genetics). Finally, the addition of a third level is also quite present, with application in a gene domain/gene/species framework, and our 3-level approach is a continuation of these advances.

Another important goal of reconciliation, notably for the gene/species community, is how to use reconciliation to correct trees or construct them. That is an interrogation we faced multiple times during this thesis, but we never really fully answered. We implemented amalgamation as a way to take this into account. A discussion and ideas to that regard are given in Chapter 2 (2.3). With the question of more realistic simulations I discuss in the next paragraph, I think this is where we went the closer toward the concept of the holobiont. Notably, we had the idea of grouping genes in compartments of common evolution that they could follow or escape. Multiple biological datasets could have been used to test such a method, as we could have used gene/species ones. I think that is the part I would have the most liked to pursue more if I had the time.

In the 3-level reconciliation article, we wrote that the Monte Carlo and the Sequential heuristics gave similar results on simulated datasets. Our idea to explain this, when the Monte Carlo is a more robust method, in theory, is that uncertainty at the gene symbiont level was not linked to uncertainty at the host and symbiont one. However, I think this independence might not hold in nature. Symbiont host switches or speciation could make possible new horizontal gene transfers or gene losses, with added redundancy making it possible to explore new combinations of symbionts and genes inside a host. It would be interesting to have new simulation frameworks to consider this kind of model. For instance, with a functional view of genes, we could model the necessity that all functions be present in the same host but possibly in different compartments corresponding to symbionts. Doing this would introduce a strong dependence between the gene lineages, requiring new ways of simulating to get an efficient approach, with maybe the need for important methodological developments or the possibility of borrowing existing methods from other fields.

New Zealand Project

I wrote a project on this subject of simulations when I was in my second year. I was thinking about going to SMBE in 2021 (and then 2022) in New Zealand to present my work. If I had to cross half the planet, it seemed reasonable to stay some time around, so I asked a local researcher, Mike Steel, if I could come work with him on a subject parallel to my thesis for a few months. I applied for a fund from the European Society of Evolutionary Biology for such mobility, proposing a project on more realistic simulations with dependence between gene trees, and I was granted the fund. However, the pandemic made border opening uncertain until the last moment, and with the calendar for the end of the thesis, it did not happen.

The development of 3-level reconciliation methods must go hand in hand with studying complex biological models. It is sad to do without a biological dataset, as we did for some time. Looking back, what would have been interesting is to focus on a biological model and use the method for it instead of constructing a method out of our imagination and forcing it onto a biological question. Few test cases can be used for 3-level models, furthermore, there is no classic example like pocket gophers and chewing lice, on which multiple studies could come and try their approaches to propose better models or methods. Maybe *Helicobacter pylori* could be such a model? The geographical/population level might be a barrier, but there should be some way?

The question of dataset availability is also crucial for reconciliation in general. It is a part missing from our review and maybe the only point where we took more the point of view of a user than a methods developer and proposed a list of software but no list of useful datasets. In host and symbiont, multiple datasets can be used, a list of 11 datasets with node numbers varying from 13 to 773 is used in multiple papers [232, 61] (you can look at the list in this thesis page 66 [231]). Nevertheless, as far as I know, these datasets are not openly compiled somewhere. It is important in the development of benchmarks, a database could be very interesting, with the possibility to get the input format used by the different reconciliation software. In gene species, Hogenom[180] looks like such a dataset. However, few datasets are used in each paper, while we could imagine it is easier to get gene trees than host and symbiont systems with known phylogenies. A often used dataset is one on cyanobacteria, that mixes real data and simulations, compiled and generated in [216], it is used in multiple studies [100][158], mostly to test tree inference capabilities. The contributive nature of Wikipedia could make it possible to add such a list of datasets to our review.

The 3-level approach is very interesting in that it enables us to consider the

information at a third level. However, it is slower and, as a reviewer penned it when submitting our paper to ISMB, produces no "slam dunk" cases. Compared to previous 3-level reconciliations, the specificity of our approach is the use of likelihood as a way to differentiate the 3-level and 2-level models. However, we only showed its capacity on simulations, and even there, it is only an intuition given from the figure representing the evolution of the likelihood differences between the two models when the simulation framework changes. On real data, amalgamation could be used as a way to correct for transfers only here to correct uncertainty on the trees, it is something we could test with simulation, with a more complete simulation pipeline generating sequences on the trees, and then reconstructing sample of these trees to use for amalgamation. We would also need other biological datasets to test this possibility. In the next months we want to continue our efforts in that direction to better understand and show the possibilities of this likelihood comparison to improve our article.

In this thesis we took a phylogenetic approach to multiple problems, and more precisely we use trees to represent diversification histories, by splitting discordant signals into smaller parts. Species trees were split into gene trees. We followed [199] use of recombination detection and profile to split coronaviruses genomes into regions on which construct trees. We approached *Helicobacter pylori* with gene phylogenies, supposing that genes would be mostly supported, and used these genes to understand the diverging information of a potential introgression forming the hpEurope population. Reconciliation is a way to combine diverging information, staying at a binary tree level, that makes possible the use of all the models develop for this simplistic model, instead of, for instance using networks, or staying at the SNPs level, that may be too numerous for complex models.

In regard to the implementation, I still have some work to do. I would like to make the input similar to the one of Generax. It could also be interesting to make it a simple option to use the branch matching with no prior on some leaves we used for the *pylori* dataset (it is already easy to use the uniform prior, but the reading of the output is not direct). I would also like to implement some speed up to make it easier to try new features. There is certainly some work to simplify the current options and propose complete documentation.

Computer science

I spent a significant amount of time implementing the method, working almost from scratch to reimplement ALE undated and the function we needed before applying it for multilevel approaches. Reimplementing ALE was motivated mainly by two reasons and a third one. I wanted to understand how reconciliation and amalgamation worked, propose a new implementation, easier to

install than the previous one (which discouraged some potential users), and not have to dive into someone else's code and project and lose time. Retrospectively using the current ALE implementation might have been a gain of time, and I would certainly have exchanged the knowledge I gained on reconciliation for knowledge on software development and proper code techniques. What is done is done, and I think I acquired some new usage through time spent with myself and this project.

Another point where I behaved strangely for a computer scientist, at least, was that I fend off using the cluster and computing facility available at the lab to the furthest point possible. I motivated it by saying it was good not to overuse a common, and money and carbon expensive resource, but I was really scared of falling in over use. So I designed most of the tests without access to other computing resources than my laptop. I started using the cluster at the beginning of my third year to flesh out my simulated dataset experiments and advance on the *pylori* dataset, by not having to let my computer run all night to test the generalization of what I found on 16 genes to the 1034 families of the complete dataset.

In a way, I think the role of computer science, through the definition of NP-completeness, by working on complexity, is to reduce the computation time, not increase it.

Bibliography

Chapter 1 - Introduction

[1] S. S. Abby, E. Tannier, M. Gouy, and V. Daubin. "Detecting lateral gene transfers by statistical reconciliation of phylogenetic forests". In: *BMC Bioinformatics* 11 (2010), p. 324 (*cited on page* 58).

[3] M. Achtman. "How old are bacterial pathogens?" In: *Proceedings of the Royal Society B: Biological Sciences* 283.1836 (2016), p. 20160990 (*cited on pages* 61, 129, 138).

[4] Ö. Åkerborg, B. Sennblad, L. Arvestad, and J. Lagergren. "Simultaneous Bayesian gene tree reconstruction and reconciliation analysis". In: *Proceedings of the National Academy of Sciences* 106.14 (2009). Publisher: National Academy of Sciences Section: Physical Sciences, pp. 5714–5719 (*cited on pages* 56, 57).

[7] L. Arvestad, A.-C. Berglund, J. Lagergren, and B. Sennblad. "Bayesian gene/species tree reconciliation and orthology analysis using MCMC". In: *Bioinformatics* 19 (suppl_1 2003). Publisher: Oxford Academic, pp. i7–i15 (*cited on pages* 48, 57).

[8] L. Arvestad, A.-C. Berglund, J. Lagergren, and B. Sennblad. "Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution". In: *Proceedings of the eighth annual international conference on Resaerch in computational molecular biology*. RECOMB '04. San Diego, California, USA: Association for Computing Machinery, 2004, pp. 326–335 (*cited on page* 48).

[9] C. P. Bagowski, W. Bruins, and A. J. te Velthuis. "The Nature of Protein Domain Evolution: Shaping the Interaction Network". In: *Current Genomics* 11.5 (2010), pp. 368–376 (*cited on page* 40).

[10] M. Bailly-Bechet, P. Martins-Simões, G. J. Szöllősi, G. Mialdea, M.-F. Sagot, and S. Charlat. "How Long Does Wolbachia Remain on Board?" In: *Molecular Biology and Evolution* 34.5 (2017), pp. 1183–1193 (*cited on pages* 68, 73).

[11] M. J. Ballinger, R. M. R. Gawryluk, and S. J. Perlman. "Toxin and Genome Evolution in a Drosophila Defensive Symbiosis". In: *Genome Biology and Evolution* 11.1 (2019), pp. 253–262 (*cited on page* 68).

[12] M. S. Bansal, E. J. Alm, and M. Kellis. "Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss". In: *Bioinformatics* 28.12 (2012). Publisher: Oxford Academic, pp. i283–i291 (*cited on pages* 50, 153).

[13] M. S. Bansal, E. J. Alm, and M. Kellis. "Reconciliation Revisited: Handling Multiple Optima when Reconciling with Duplication, Transfer, and Loss". In: *Journal of Computational Biology* 20.10 (2013). Publisher: Mary Ann Liebert, Inc., publishers, pp. 738–754 (*cited on page* 54).

[14] M. S. Bansal, G. Banay, J. P. Gogarten, and R. Shamir. "Detecting Highways of Horizontal Gene Transfer". In: *Journal of Computational Biology* 18.9 (2011), pp. 1087–1114 (*cited on page* 64).

[15] M. S. Bansal and O. Eulenstein. “The multiple gene duplication problem revisited”. In: *Bioinformatics* 24.13 (2008), pp. i132–i138 (*cited on page* 62).

[16] M. S. Bansal and R. Shamir. “A Note on the Fixed Parameter Tractability of the Gene-Duplication Problem”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 8.3 (2011). Conference Name: IEEE/ACM Transactions on Computational Biology and Bioinformatics, pp. 848–850 (*cited on page* 57).

[17] M. S. Bansal, Y.-C. Wu, E. J. Alm, and M. Kellis. “Improved gene tree error correction in the presence of horizontal gene transfer”. In: *Bioinformatics* 31.8 (2015), pp. 1211–1218 (*cited on page* 56).

[18] C. Baudet, B. Donati, B. Sinaimeri, P. Crescenzi, C. Gautier, C. Matias, and M.-F. Sagot. “Cophylogeny reconstruction via an approximate Bayesian computation”. In: *Systematic Biology* 64.3 (2015), pp. 416–431 (*cited on page* 51).

[19] V. Berry, F. Chevenet, J.-P. Doyon, and E. Jousselin. “A geography-aware reconciliation method to investigate diversification patterns in host/parasite interactions”. In: *Molecular Ecology Resources* 18.5 (2018), pp. 1173–1184 (*cited on pages* 66, 124).

[20] H. C. Betts, M. N. Puttick, J. W. Clark, T. A. Williams, P. C. J. Donoghue, and D. Pisani. “Integrated genomic and fossil evidence illuminates life’s early evolution and eukaryote origin”. In: *Nature Ecology & Evolution* 2.10 (2018). Number: 10 Publisher: Nature Publishing Group, pp. 1556–1562 (*cited on page* 23).

[21] S. R. Bordenstein and K. R. Theis. “Host Biology in Light of the Microbiome: Ten Principles of Holobionts and Hologenomes”. In: *PLoS biology* 13.8 (2015), e1002226 (*cited on pages* 37, 60).

[22] M. Bordewich and C. Semple. “On the Computational Complexity of the Rooted Subtree Prune and Regraft Distance”. In: *Annals of Combinatorics* 8 (2005), pp. 409–423 (*cited on page* 58).

[23] D. Bork, R. Cheng, J. Wang, J. Sung, and R. Libeskind-Hadas. “On the computational complexity of the maximum parsimony reconciliation problem in the duplication-loss-coalescence model”. In: *Algorithms for Molecular Biology* 12.1 (2017), p. 6 (*cited on page* 59).

[24] E. Bortolini, L. Pagani, E. R. Crema, S. Sarno, C. Barbieri, A. Boattini, M. Sazzini, S. G. da Silva, G. Martini, M. Metspalu, D. Pettener, D. Luiselli, and J. J. Tehrani. “Inferring patterns of folktale diffusion using genomic data”. In: *Proceedings of the National Academy of Sciences* 114.34 (2017), pp. 9140–9145 (*cited on page* 62).

[25] B. Boussau and V. Daubin. “Genomes as documents of evolutionary history”. In: *Trends in Ecology & Evolution* 25.4 (2010), pp. 224–232 (*cited on page* 55).

[26] B. Boussau and C. Scornavacca. “Reconciling Gene trees with Species Trees”. In: *Phylogenetics in the Genomic Era*. Ed. by C. Scornavacca, F. Delsuc, and N. Galtier. No commercial publisher — Authors open access book, 2020, 3.2:1–3.2:23 (*cited on page* 41).

[27] B. Boussau, G. J. Szöllosi, L. Duret, M. Gouy, E. Tannier, and V. Daubin. “Genome-scale coestimation of species and gene trees”. In: *Genome Research* 23.2 (2013), pp. 323–330 (*cited on pages* 56, 58).

[30] D. R. Brooks. “Hennig’s Parasitological Method: A Proposed Solution”. In: *Systematic Zoology* 30.3 (1981). Publisher: [Oxford University Press, Society of Systematic Biologists, Taylor & Francis, Ltd.], pp. 229–249 (*cited on pages* 44, 46, 49).

[31] D. Bryant and M. W. Hahn. “The Concatenation Question”. In: (), p. 24 (*cited on page* 34).

[32] J. Burleigh, M. Bansal, A. Wehe, and O. Eulenstein. “Locating Large-Scale Gene Duplication Events through Reconciled Trees: Implications for Identifying Ancient Polyploidy Events in Plants”. In: *Journal of Computational Biology* 16.8 (2009). Publisher: Mary Ann Liebert, Inc., publishers, pp. 1071–1083 (*cited on page* 62).

[35] Y.-b. Chan, V. Ranwez, and C. Scornavacca. “Inferring incomplete lineage sorting, duplications, transfers and losses with reconciliations”. In: *Journal of Theoretical Biology* 432 (2017), pp. 1–13 (*cited on pages* 59, 161).

[37] W.-C. Chang, G. J. Burleigh, D. F. Fernández-Baca, and O. Eulenstein. "An ILP solution for the gene duplication problem". In: *BMC Bioinformatics* 12.1 (2011), S14 (*cited on page* 57).

[38] M. A. Charleston. "Jungles: a new solution to the host/parasite phylogeny reconciliation problem". In: *Mathematical Biosciences* 149.2 (1998), pp. 191–223 (*cited on pages* 44, 45, 49, 50).

[39] M. Charleston and R. Libeskind-Hadas. "Event-Based Cophylogenetic Comparative Analysis". In: *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology*. Springer Berlin Heidelberg, 2014, pp. 465–480 (*cited on page* 41).

[40] M. A. Charleston and S. L. Perkins. "Traversing the tangle: Algorithms and applications for cophylogenetic studies". In: *Journal of Biomedical Informatics*. Phylogenetic Inferencing: Beyond Biology 39.1 (2006), pp. 62–71 (*cited on page* 41).

[41] R. Chaudhary, M. S. Bansal, A. Wehe, D. Fernández-Baca, and O. Eulenstein. "iGTP: A software package for large-scale gene tree parsimony analysis". In: *BMC Bioinformatics* 11.1 (2010), p. 574 (*cited on page* 57).

[42] R. Chaudhary, J. G. Burleigh, and D. Fernández-Baca. "Inferring species trees from incongruent multi-copy gene trees using the Robinson-Foulds distance". In: *Algorithms for Molecular Biology* 8.1 (2013), p. 28 (*cited on page* 57).

[43] C. Chauve and N. El-Mabrouk. "New Perspectives on Gene Family Evolution: Losses in Reconciliation and a Link with Supertrees". In: *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2009, pp. 46–58 (*cited on page* 49).

[44] C. Chauve, A. Rafiey, A. A. Davín, C. Scornavacca, P. Veber, B. Boussau, G. j. Szöllősi, V. Daubin, and E. Tannier. "MaxTiC: Fast ranking of a phylogenetic tree by Maximum Time Consistency with lateral gene transfers". In: *bioRxiv* (2017), p. 127548 (*cited on page* 52).

[45] F. Chevenet, J.-P. Doyon, C. Scornavacca, E. Jacox, E. Jousselin, and V. Berry. "SylvX: a viewer for phylogenetic tree reconciliations". In: *Bioinformatics* 32.4 (2016), pp. 608–610 (*cited on page* 54).

[46] F. D. Ciccarelli, T. Doerks, C. von Mering, C. J. Creevey, B. Snel, and P. Bork. "Toward Automatic Reconstruction of a Highly Resolved Tree of Life". In: *Science* 311.5765 (2006). Publisher: American Association for the Advancement of Science, pp. 1283–1287 (*cited on page* 35).

[47] N. Comte, B. Morel, D. Hasic, L. Guéguen, B. Boussau, V. Daubin, S. Penel, C. Scornavacca, M. Gouy, A. Stamatakis, E. Tannier, and D. P. Parsons. "Treerecs: an integrated phylogenetic tool, from sequences to reconciliations". In: *Bioinformatics* (2020) (*cited on page* 56).

[48] C. Conow, D. Fielder, Y. Ovadia, and R. Libeskind-Hadas. "Jane: a new tool for the cophylogeny reconstruction problem". In: *Algorithms for Molecular Biology* 5.1 (2010), p. 16 (*cited on pages* 52, 54, 55, 59).

[49] T. H. Cormen, ed. *Introduction to algorithms*. 3rd ed. OCLC: ocn311310321. Cambridge, Mass: MIT Press, 2009. 1292 pp. (*cited on page* 25).

[50] M. Csűös. "Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood". In: *Bioinformatics* 26.15 (2010), pp. 1910–1912 (*cited on page* 49).

[51] M. Csűrös and I. Miklós. "Streamlining and Large Ancestral Genomes in Archaea Inferred with a Phylogenetic Birth-and-Death Model". In: *Molecular Biology and Evolution* 26.9 (2009), pp. 2087–2095 (*cited on page* 46).

[52] J. F. H. Cuthill and M. Charleston. "Wing patterning genes and coevolution of Müllerian mimicry in Heliconius butterflies: Support from phylogeography, cophylogeny, and divergence times". In: *Evolution* 69.12 (2015), pp. 3082–3096 (*cited on page* 68).

[53] T. Dagan and W. Martin. "The tree of one percent". In: *Genome Biology* 7.10 (2006), p. 118 (*cited on page* 35).

[54] L. A. David and E. J. Alm. "Rapid evolutionary innovation during an Archaean genetic expansion". In: *Nature* 469.7328 (2011), pp. 93–96 (*cited on pages* 51, 56, 71, 81).

[55] A. A. Davín, T. Tricou, E. Tannier, D. M. de Vienne, and G. J. Szöllősi. "Zombi: a phylogenetic simulator of trees, genomes and sequences that accounts for dead linages". In: *Bioinformatics* 36.4 (2020), pp. 1286–1288 (*cited on page* 52).

[56] A. A. Davín, E. Tannier, T. A. Williams, B. Boussau, V. Daubin, and G. J. Szöllősi. "Gene transfers can date the tree of life". In: *Nature Ecology & Evolution* 2.5 (2018). Number: 5 Publisher: Nature Publishing Group, pp. 904–909 (*cited on page* 52).

[57] J. H. Degnan and L. A. Salter. "Gene Tree Distributions Under the Coalescent Process". In: *Evolution* 59.1 (2005), pp. 24–37 (*cited on page* 59).

[58] R. Denise, S. S. Abby, and E. P. C. Rocha. "Diversification of the type IV filament superfamily into machines for adhesion, protein secretion, DNA uptake, and motility". In: *PLOS Biology* 17.7 (2019), e3000390 (*cited on page* 43).

[60] T. Dobzhansky and A. H. Sturtevant. "Inversions in the Chromosomes of Drosophila Pseudoobscura." eng. In: *Genetics* 23.1 (1938), pp. 28–64 (*cited on page* 40).

[61] B. Donati, C. Baudet, B. Sinaimeri, P. Crescenzi, and M.-F. Sagot. "EUCALYPT: efficient tree reconciliation enumerator". In: *Algorithms for Molecular Biology* 10.1 (2015), p. 3 (*cited on pages* 52, 54, 55, 59, 153, 169).

[62] R. Dondi, M. Lafond, and C. Scornavacca. "Reconciling multiple genes trees via segmental duplications and losses". In: *Algorithms for Molecular Biology* 14.1 (2019), p. 7 (*cited on page* 62).

[63] J.-P. Doyon, V. Ranwez, V. Daubin, and V. Berry. "Models, algorithms and programs for phylogeny reconciliation". In: *Briefings in Bioinformatics* 12.5 (2011), pp. 392–400 (*cited on pages* 41, 50, 77).

[64] J.-P. Doyon, C. Scornavacca, K. Y. Gorbunov, G. J. Szöllősi, V. Ranwez, and V. Berry. "An Efficient Algorithm for Gene/Species Trees Parsimonious Reconciliation with Losses, Duplications and Transfers". In: *Comparative Genomics*. Ed. by E. Tannier. Vol. 6398. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 93–108 (*cited on pages* 50, 55, 74).

[65] B. Drinkwater and M. A. Charleston. "An improved node mapping algorithm for the cophylogeny reconstruction problem". In: *Coevolution* 2.1 (2014). Publisher: Taylor & Francis, pp. 1–17 (*cited on page* 52).

[66] B. Drinkwater and M. A. Charleston. "RASCAL: A Randomized Approach for Coevolutionary Analysis". In: *Journal of Computational Biology* 23.3 (2016), pp. 218–227 (*cited on page* 52).

[67] P. Du, H. A. Ogilvie, and L. Nakhleh. "Unifying Gene Duplication, Loss, and Coalescence on Phylogenetic Networks". In: *Bioinformatics Research and Applications*. Ed. by Z. Cai, P. Skums, and M. Li. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 40–51 (*cited on page* 60).

[69] W. Duchemin, Y. Anselmetti, M. Patterson, Y. Ponty, S. Bérard, C. Chauve, C. Scornavacca, V. Daubin, and E. Tannier. "DeCoSTAR: Reconstructing the Ancestral Organization of Genes or Genomes Using Reconciled Phylogenies". In: *Genome Biology and Evolution* 9.5 (2017), pp. 1312–1319 (*cited on pages* 43, 62).

[70] W. Duchemin, G. Gence, A.-M. Arigon Chifolleau, L. Arvestad, M. S. Bansal, V. Berry, B. Boussau, F. Chevenet, N. Comte, A. A. Davín, C. Dessimoz, D. Dylus, D. Hasic, D. Mallo, R. Planel, D. Posada, C. Scornavacca, G. Szöllősi, L. Zhang, É. Tannier, and V. Daubin. "RecPhyloXML: a format for reconciled gene trees". In: *Bioinformatics* 34.21 (2018), pp. 3646–3652 (*cited on pages* 54, 68, 86, 116).

[71] D. Durand, B. V. Halldórsson, and B. Vernot. "A Hybrid Micro–Macroevolutionary Approach to Gene Tree Reconstruction". In: *Journal of Computational Biology* 13.2 (2006), pp. 320–335 (*cited on pages* 52, 54, 56).

[72] R. Ekblom and J. Galindo. "Applications of next generation sequencing in molecular ecology of non-model organisms". In: *Heredity* 107.1 (2011). Number: 1 Publisher: Nature Publishing Group, pp. 1–15 (*cited on page* 35).

[76] D. Falush, T. Wirth, B. Linz, J. K. Pritchard, M. Stephens, M. Kidd, M. J. Blaser, D. Y. Graham, S. Vacher, G. I. Perez-Perez, Y. Yamaoka, F. Mégraud, K. Otto, U. Reichard, E. Katzowitsch, X. Wang, M. Achtman, and S. Suerbaum. "Traces of human migrations in Helicobacter pylori populations". In: *Science (New York, N.Y.)* 299.5612 (2003), pp. 1582–1585 (*cited on pages* 18, 19, 133, 137–139).

[77] M. R. Fellows, M. T. Hallet, and U. Stege. "On the Multiple Gene Duplication Problem". In: *Proceedings of the 9th International Symposium on Algorithms and Computation*. ISAAC '98. Berlin, Heidelberg: Springer-Verlag, 1998, pp. 347–356 (*cited on page* 62).

[78] J. Felsenstein. *Inferring Phylogenies*. Oxford, New York: Oxford University Press, 2003. 580 pp. (*cited on pages* 43, 46, 48, 71, 75, 108, 159).

[79] E. D. Fountain, J. N. Pauli, J. E. Mendoza, J. Carlson, and M. Z. Peery. "Cophylogenetics and biogeography reveal a coevolved relationship between sloths and their symbiont algae". In: *Molecular Phylogenetics and Evolution* 110 (2017), pp. 73–80 (*cited on pages* 36, 64).

[80] Y. Fu, M. Pistolozzi, X. Yang, and Z. Lin. *A Comprehensive Classification of Coronaviruses and Inferred Cross-Host Transmissions*. preprint. Bioinformatics, 2020 (*cited on pages* 61, 109).

[81] M. Goodman, J. Czelusniak, G. W. Moore, A. E. Romero-Herrera, and G. Matsuda. "Fitting the Gene Lineage into its Species Lineage, a Parsimony Strategy Illustrated by Cladograms Constructed from Globin Sequences". In: *Systematic Zoology* 28.2 (1979), p. 132 (*cited on pages* 41, 48, 56).

[82] P. Górecki, O. Eulenstein, and J. Tiuryn. "Unrooted Tree Reconciliation: A Unified Approach". In: *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM* 10 (2013), pp. 522–36 (*cited on page* 55).

[84] R. D. Gray, D. Bryant, and S. J. Greenhill. "On the shape and fabric of human history". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 365.1559 (2010), pp. 3923–3933 (*cited on page* 40).

[85] M. Groussin, F. Mazel, J. G. Sanders, C. S. Smillie, S. Lavergne, W. Thuiller, and E. J. Alm. "Unraveling the processes shaping mammalian gut microbiomes over evolutionary time". In: *Nature Communications* 8.1 (2017). Number: 1 Publisher: Nature Publishing Group, p. 14319 (*cited on pages* 20, 36, 46, 64).

[86] R. Guigo, I. Muchnik, and T. F. Smith. "Reconstruction of Ancient Molecular Phylogeny". In: *Molecular Phylogenetics and Evolution* 6.2 (1996), pp. 189–213 (*cited on pages* 57, 62).

[87] E. Haeckel. *Systematische Phylogenie*. Verlag von Georg Reimer, 1896 (*cited on page* 40).

[88] M. S. Hafner and S. A. Nadler. "Phylogenetic trees support the coevolution of parasites and their hosts". In: *Nature* 332.6161 (1988), pp. 258–259 (*cited on page* 44).

[89] M. W. Hahn. "Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution". In: *Genome Biology* 8.7 (2007), R141 (*cited on page* 55).

[90] M. T. Hallett and J. Lagergren. "Efficient algorithms for lateral gene transfer problems". In: *Proceedings of the fifth annual international conference on Computational biology - RECOMB '01*. the fifth annual international conference. Montreal, Quebec, Canada: ACM Press, 2001, pp. 149–156 (*cited on pages* 44, 52).

[91] D. Hasic and E. Tannier. "Gene tree reconciliation including transfers with replacement is hard and FPT". In: *Journal of Combinatorial Optimization* 38.2 (2019), pp. 502–544. arXiv: `1709.04459` (*cited on pages* 58, 162).

[92] D. Hasić and E. Tannier. "Gene tree species tree reconciliation with gene conversion". In: *Journal of Mathematical Biology* 78.6 (2019), pp. 1981–2014 (*cited on page* 58).

[93] J. Hein. "A heuristic method to reconstruct the history of sequences subject to recombination". In: (), p. 10 (*cited on page* 44).

[94] J. Hein, T. Jiang, L. Wang, and K. Zhang. "On the complexity of comparing evolutionary trees". In: *Discrete Applied Mathematics* 71.1 (1996), pp. 153–169 (*cited on page* 58).

[96] S. Höhna and A. J. Drummond. "Guided Tree Topology Proposals for Bayesian Phylogenetic Inference". In: *Systematic Biology* 61.1 (2012), pp. 1–11 (*cited on page* 81).

[98] K. T. Huber, V. Moulton, M.-F. Sagot, and B. Sinaimeri. "Geometric medians in reconciliation spaces of phylogenetic trees". In: *Information Processing Letters* 136 (2018), pp. 96–101 (*cited on page* 54).

[99] D. H. Huson, R. Rupp, and C. Scornavacca. *Phylogenetic Networks by Daniel H. Huson*. Cambridge Core. 2010 (*cited on page* 23).

[100] E. Jacox, C. Chauve, G. J. Szöllősi, Y. Ponty, and C. Scornavacca. "ecceTERA: comprehensive gene tree-species tree reconciliation using parsimony". In: *Bioinformatics (Oxford, England)* 32.13 (2016), pp. 2056–2058 (*cited on pages* 50, 53, 71, 77, 78, 81, 169).

[101] E. Jacox, M. Weller, E. Tannier, and C. Scornavacca. "Resolution and reconciliation of non-binary gene trees with transfers, duplications and losses". In: *Bioinformatics* 33.7 (2017), pp. 980–987 (*cited on page* 56).

[102] H. Jeong, B. Arif, G. Caetano-Anollés, K. M. Kim, and A. Nasir. "Horizontal gene transfer in human-associated microorganisms inferred by phylogenetic reconstruction and reconciliation". In: *Scientific Reports* 9.1 (2019), pp. 1–18 (*cited on page* 61).

[105] L. Kloub, S. Gosselin, M. Fullmer, J. Graf, J. P. Gogarten, and M. S. Bansal. "Systematic Detection of Large-Scale Multigene Horizontal Transfer in Prokaryotes". In: *Molecular Biology and Evolution* 38.6 (2021), pp. 2639–2659 (*cited on page* 64).

[106] M. Kordi and M. S. Bansal. "Exact Algorithms for Duplication-Transfer-Loss Reconciliation with Non-Binary Gene Trees". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 16.4 (2019). Conference Name: IEEE/ACM Transactions on Computational Biology and Bioinformatics, pp. 1077–1090 (*cited on page* 56).

[107] M. Kordi and M. S. Bansal. "On the Complexity of Duplication-Transfer-Loss Reconciliation with Non-binary Gene Trees". In: *Bioinformatics Research and Applications*. Ed. by R. Harrison, Y. Li, and I. Măndoiu. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015, pp. 187–198 (*cited on page* 56).

[108] M. Kordi and M. S. Bansal. "TreeSolve: Rapid Error-Correction of Microbial Gene Trees". In: *Algorithms for Computational Biology*. Ed. by C. Martín-Vide, M. A. Vega-Rodríguez, and T. Wheeler. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 125–139 (*cited on page* 56).

[109] M. Kordi, S. Kundu, and M. S. Bansal. "On Inferring Additive and Replacing Horizontal Gene Transfers Through Phylogenetic Reconciliation". In: *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. BCB '19. Niagara Falls, NY, USA: Association for Computing Machinery, 2019, pp. 514–523 (*cited on page* 58).

[111] S. Kundu and M. S. Bansal. "On the impact of uncertain gene tree rooting on duplication-transfer-loss reconciliation". In: *BMC bioinformatics* 19 (Suppl 9 2018), p. 290 (*cited on pages* 54, 55).

[112] S. Kundu and M. S. Bansal. "SaGePhy: an improved phylogenetic simulation framework for gene and subgene evolution". In: *Bioinformatics* 35.18 (2019), pp. 3496–3498 (*cited on pages* 67, 74).

[113] M. Lafond, E. Noutahi, and N. El-Mabrouk. "Efficient Non-Binary Gene Tree Resolution with Weighted Reconciliation Cost". In: (2016), p. 12 (*cited on page* 56).

[114] H. Lai, M. Stolzer, and D. Durand. "Fast Heuristics for Resolving Weakly Supported Branches Using Duplication, Transfers, and Losses". In: *Comparative Genomics*. Ed. by J. Meidanis and L. Nakhleh. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2017, pp. 298–320 (*cited on page* 56).

[115] A. H. Land and A. G. Doig. "An Automatic Method of Solving Discrete Programming Problems". In: *Econometrica* 28.3 (1960). Publisher: [Wiley, Econometric Society], pp. 497–520 (*cited on page* 25).

[116] B. Larget. "The Estimation of Tree Posterior Probabilities Using Conditional Clade Probability Distributions". In: *Systematic Biology* 62.4 (2013), pp. 501–511 (*cited on page* 81).

[117] N. Lartillot and H. Philippe. "A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process". In: *Molecular Biology and Evolution* 21.6 (2004), pp. 1095–1109 (*cited on page* 56).

[119] M. S. Y. Lee and A. Palci. "Morphological Phylogenetics in the Genomic Age". In: *Current Biology* 25.19 (2015), R922–R929 (*cited on page* 32).

[120] B. Legried, E. K. Molloy, T. Warnow, and S. Roch. "Polynomial-Time Statistical Estimation of Species Trees under Gene Duplication and Loss". In: *bioRxiv* (2020). Publisher: Cold Spring Harbor Laboratory Section: New Results, p. 821439 (*cited on page* 57).

[122] L. Li and M. S. Bansal. "An Integrated Reconciliation Framework for Domain, Gene, and Species Level Evolution". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 16.1 (2019), pp. 63–76 (*cited on page* 67).

[123] L. Li and M. S. Bansal. "An Integer Linear Programming Solution for the Domain-Gene-Species Reconciliation Problem". In: *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. BCB '18. event-place: Washington, DC, USA. New York, NY, USA: ACM, 2018, pp. 386–397 (*cited on page* 67).

[124] L. Li and M. S. Bansal. "Simultaneous Multi-Domain-Multi-Gene Reconciliation Under the Domain-Gene-Species Reconciliation Model". In: *Bioinformatics Research and Applications*. Ed. by Z. Cai, P. Skums, and M. Li. Lecture Notes in Computer Science. Springer International Publishing, 2019, pp. 73–86 (*cited on pages* 67, 101).

[125] R. Libeskind-Hadas, Y.-C. Wu, M. S. Bansal, and M. Kellis. "Pareto-optimal phylogenetic tree reconciliation". In: *Bioinformatics* 30.12 (2014), pp. i87–i95 (*cited on page* 51).

[128] L. Liu and D. K. Pearl. "Species Trees from Gene Trees: Reconstructing Bayesian Posterior Distributions of a Species Phylogeny Using Estimated Gene Tree Distributions". In: *Systematic Biology* 56.3 (2007). Publisher: Oxford Academic, pp. 504–514 (*cited on page* 59).

[130] S. López-Madrigal and R. Gil. "Et tu, Brute? Not Even Intracellular Mutualistic Symbionts Escape Horizontal Gene Transfer". In: *Genes* 8.10 (2017) (*cited on page* 61).

[131] E. L. S. Loreto, C. M. A. Carareto, and P. Capy. "Revisiting horizontal transfer of transposable elements in Drosophila". In: *Heredity* 100.6 (2008), pp. 545–554 (*cited on page* 43).

[132] B. Ma, M. Li, and L. Zhang. "From Gene Trees to Species Trees". In: *SIAM Journal on Computing* 30.3 (2000). Publisher: Society for Industrial and Applied Mathematics, pp. 729–752 (*cited on page* 57).

[133] W. Ma, D. Smirnov, J. Forman, A. Schweickart, C. Slocum, S. Srinivasan, and R. Libeskind-Hadas. "DTL-RnB: Algorithms and Tools for Summarizing the Space of DTL Reconciliations". In: *IEEE/ACM transactions on computational biology and bioinformatics* 15.2 (2018), pp. 411–421 (*cited on page* 54).

[134] W. Ma, D. Smirnov, and R. Libeskind-Hadas. "DTL reconciliation repair". In: *BMC Bioinformatics* 18 (Suppl 3 2017) (*cited on page* 52).

[135] W. P. Maddison. "Gene trees in species tree". In: *Systematic Biology* (1997), p. 14 (*cited on page* 41).

[136] W. P. Maddison and L. L. Knowles. "Inferring phylogeny despite incomplete lineage sorting". In: *Systematic Biology* 55.1 (2006), pp. 21–30 (*cited on pages* 57, 59).

[137] M. C. J. Maiden, J. A. Bygraves, E. Feil, G. Morelli, J. E. Russell, R. Urwin, Q. Zhang, J. Zhou, K. Zurth, D. A. Caugant, I. M. Feavers, M. Achtman, and B. G. Spratt. "Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms". In: *Proceedings of the National Academy of Sciences of the United States of America* 95.6 (1998), pp. 3140–3145 (*cited on pages* 35, 131).

[139] A. Manzano-Marín, A. C. D'acier, A.-L. Clamens, C. Orvain, C. Cruaud, V. Barbe, and E. Jousselin. "Serial horizontal transfer of vitamin-biosynthetic genes enables the establishment of new nutritional symbionts in aphids' di-symbiotic systems". In: *The ISME Journal* (2019), pp. 1–15 (*cited on pages* 20, 21, 35, 36, 61, 68).

[140] L. Margulis and R. Fester, eds. *Symbiosis as a Source of Evolutionary Innovation: Speciation and Morphogenesis*. Cambridge, MA, USA: MIT Press, 1991. 470 pp. (*cited on pages* 37, 60).

[141] J. Marin, G. Achaz, A. Crombach, and A. Lambert. "The genomic view of diversification". In: *Journal of Evolutionary Biology* 33.10 (2020). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jeb.1 pp. 1387–1404 (*cited on pages* 59, 73).

[143] A. Martínez-Aquino. "Phylogenetic framework for coevolutionary studies: a compass for exploring jungles of tangled trees". In: *Current Zoology* 62.4 (2016), pp. 393–403 (*cited on page* 41).

[144] A. Martínez-Aquino, F. S. Ceccarelli, L. E. Eguiarte, E. Vázquez-Domínguez, and G. P.-P. d. León. "Do the Historical Biogeography and Evolutionary History of the Digenean Margotrema spp. across Central Mexico Mirror Those of Their Freshwater Fish Hosts (Goodeinae)?" In: *PLOS ONE* 9.7 (2014), e101700 (*cited on page* 64).

[145] N. J. Matzke. "Model Selection in Historical Biogeography Reveals that Founder-Event Speciation Is a Crucial Process in Island Clades". In: *Systematic Biology* 63.6 (2014), pp. 951–970 (*cited on page* 53).

[146] N. J. Matzke. "Probabilistic historical biogeography: new models for founder-event speciation, imperfect detection, and fossils allow improved accuracy and model-testing". In: *Frontiers of Biogeography* 5.4 (2013) (*cited on page* 53).

[147] R. Mawhorter and R. Libeskind-Hadas. "Hierarchical clustering of maximum parsimony reconciliations". In: *BMC Bioinformatics* 20.1 (2019), p. 612 (*cited on page* 55).

[148] F. Mégraud, P. Lehours, and F. F. Vale. "The history of Helicobacter pylori: from phylogeography to paleomicrobiology". In: *Clinical Microbiology and Infection* 22.11 (2016), pp. 922–927 (*cited on pages* 18, 129).

[151] D. Merkle, M. Middendorf, and N. Wieseke. "A parameter-adaptive dynamic programming approach for inferring cophylogenies". In: *BMC Bioinformatics* 11 (Suppl 1 2010), S60 (*cited on pages* 51, 54).

[153] E. K. Molloy and T. Warnow. "FastMulRFS: Fast and accurate species tree estimation under generic gene duplication and loss models". In: *bioRxiv* (2020). Publisher: Cold Spring Harbor Laboratory Section: New Results, p. 835553 (*cited on page* 57).

[154] Y. Moodley, B. Linz, R. P. Bond, M. Nieuwoudt, H. Soodyall, C. M. Schlebusch, S. Bernhöft, J. Hale, S. Suerbaum, L. Mugisha, S. W. v. d. Merwe, and M. Achtman. "Age of the Association between Helicobacter pylori and Man". In: *PLOS Pathogens* 8.5 (2012), e1002693 (*cited on pages* 61, 135–137, 140).

[156] R. D. Mooi and A. C. Gill. "Phylogenies without Synapomorphies—A Crisis in Fish Systematics: Time to Show Some Character". In: *Zootaxa* 2450.1 (2010), p. 26 (*cited on page* 32).

[157] N. A. Moran, J. P. McCutcheon, and A. Nakabachi. "Genomics and evolution of heritable bacterial symbionts". In: *Annual Review of Genetics* 42 (2008), pp. 165–190 (*cited on page* 60).

[158] B. Morel, A. M. Kozlov, A. Stamatakis, and G. J. Szöllősi. *GeneRax: A tool for species tree-aware maximum likelihood based gene family tree inference under gene duplication, transfer, and loss*. preprint. Bioinformatics, 2019 (*cited on pages* 41, 56, 71, 74, 75, 86, 169).

[159] B. Morel, P. Schade, S. Lutteropp, T. A. Williams, G. J. Szöllősi, and A. Stamatakis. *SpeciesRax: A tool for maximum likelihood species tree inference from gene family trees under duplication, transfer, and loss*. Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article. 2021, p. 2021.03.29.437460 (*cited on page* 71).

[161] S. A. Muhammad, B. Sennblad, and J. Lagergren. “Species tree-aware simultaneous reconstruction of gene and domain evolution”. In: *bioRxiv* (2018), p. 336453 (*cited on page* 67).

[164] N. Nair, Y. Lin, A. Manasovska, J. Antic, P. Grnarova, A. Sahu, P. Bucher, and B. M. Moret. “Study of cell differentiation by phylogenetic analysis using histone modification data”. In: *BMC Bioinformatics* 15.1 (2014), p. 269 (*cited on page* 40).

[165] A. Nakabachi, R. Ueoka, K. Oshima, R. Teta, A. Mangoni, M. Gurgui, N. J. Oldham, G. van Echten-Deckert, K. Okamura, K. Yamamoto, H. Inoue, M. Ohkuma, Y. Hongoh, S.-y. Miyagishima, M. Hattori, J. Piel, and T. Fukatsu. “Defensive Bacteriome Symbiont with a Drastically Reduced Genome”. In: *Current Biology* 23.15 (2013), pp. 1478–1484 (*cited on page* 61).

[166] L. Nakhleh. “Computational approaches to species phylogeny inference and gene tree reconciliation”. In: *Trends in ecology and evolution* 28 (2013), pp. 719–728 (*cited on page* 41).

[168] G. Nelson and N. Platnick. *Systematics and Biogeography: Cladistics and Vicariance*. Vol. 31. Journal Abbreviation: Systematic Zoology Publication Title: Systematic Zoology. 1981 (*cited on page* 44).

[169] T. H. Nguyen, V. Ranwez, S. Pointet, A.-M. A. Chifolleau, J.-P. Doyon, and V. Berry. “Reconciliation and local gene tree rearrangement can be of mutual profit”. In: *Algorithms for Molecular Biology* 8.1 (2013), p. 12 (*cited on page* 56).

[170] T.-H. Nguyen, V. Ranwez, V. Berry, and C. Scornavacca. “Support Measures to Estimate the Reliability of Evolutionary Events Predicted by Reconciliation Methods”. In: *PLOS ONE* 8.10 (2013). Publisher: Public Library of Science, e73667 (*cited on page* 54).

[171] C. Nieberding, E. Jousselin, and Y. Desdevises. “The use of co-phylogeographic patterns to predict the nature of host– parasite interactions, and vice versa”. In: *The biogeography of host-parasite interactions*. 2010, pp. 59–69 (*cited on page* 64).

[172] N. Nikoh, T. Hosokawa, M. Moriyama, K. Oshima, M. Hattori, and T. Fukatsu. “Evolutionary origin of insect–Wolbachia nutritional mutualism”. In: *Proceedings of the National Academy of Sciences of the United States of America* 111.28 (2014), pp. 10257–10262 (*cited on page* 61).

[173] Y. Ovadia, D. Fielder, C. Conow, and R. Libeskind-Hadas. “The co phylogeny reconstruction problem is NP-complete”. In: *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 18.1 (2011), pp. 59–65 (*cited on page* 52).

[174] R. D. Page. “GeneTree: comparing gene and species phylogenies using reconciled trees.” In: *Bioinformatics* 14.9 (1998), pp. 819–820 (*cited on page* 57).

[175] R. D. M. Page and J. A. Cotton. “Vertebrate phylogenomics: reconciled trees and gene duplications”. In: *Biocomputing 2002*. World Scientific, 2001, pp. 536–547 (*cited on page* 62).

[176] R. D. M. Page. “Component Analysis: A Valiant Failure?” In: *Cladistics* 6.2 (1990), pp. 119–136 (*cited on page* 44).

[177] R. D. M. Page. “Extracting Species Trees From Complex Gene Trees: Reconciled Trees And Vertebrate Phylogeny”. In: *Molecular Phylogenetics and Evolution* 14.1 (2000), pp. 89–106 (*cited on page* 57).

[178] R. D. M. Page. “Maps Between Trees and Cladistic Analysis of Historical Associations among Genes,Organisms, and Areas”. In: *Systematic Biology* 43.1 (1994), pp. 58–77 (*cited on page* 44).

[179] R. D. M. Page. “Parallel Phylogenies: Reconstructing the History of Host-Parasite Assemblages”. In: *Cladistics* 10.2 (1994), pp. 155–173 (*cited on pages* 45, 49).

[182] T. Penz, S. Schmitz-Esser, S. E. Kelly, B. N. Cass, A. Müller, T. Woyke, S. A. Malfatti, M. S. Hunter, and M. Horn. “Comparative Genomics Suggests an Independent Origin of Cytoplasmic Incompatibility in Cardinium hertigii”. In: *PLOS Genetics* 8.10 (2012), e1003012 (*cited on page* 61).

[183] M. Pinto-Carbó, S. Sieber, S. Dessein, T. Wicker, B. Verstraete, K. Gademann, L. Eberl, and A. Carlier. “Evidence of horizontal gene transfer between obligate leaf nodule symbionts”. In: *The ISME Journal* 10.9 (2016), pp. 2092–2105 (*cited on pages* 61, 68).

[186] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Elsevier, 2014. 313 pp. (*cited on page* 25).

[187] C. Raffalli, R. David, and K. Nour. *Introduction à la Logique, Théorie de la démonstration (2nd édition)*. Sciences Sup. Dunod, 2004. 368 pp. (*cited on page* 25).

[188] B. Rannala, S. V. Edwards, A. Leaché, and Z. Yang. "Chapter 3.3 The Multi-species Coalescent Model and Species Tree Inference". In: (), p. 21 (*cited on page* 59).

[189] B. Rannala and Z. Yang. "Bayes Estimation of Species Divergence Times and Ancestral Population Sizes Using DNA Sequences From Multiple Loci". In: *Genetics* 164.4 (2003). Publisher: Genetics Section: Investigations, pp. 1645–1656 (*cited on page* 59).

[190] M. D. Rasmussen and M. Kellis. "Unified modeling of gene duplication, loss, and coalescence using a locus tree". In: *Genome Research* 22.4 (2012), pp. 755–765 (*cited on page* 60).

[191] R. H. Ree, B. R. Moore, C. O. Webb, and M. J. Donoghue. "A likelihood framework for inferring the evolution of geographic range on phylogenetic trees". In: (2005), p. 13 (*cited on pages* 53, 152).

[192] R. H. Ree and S. A. Smith. "Maximum Likelihood Inference of Geographic Range Evolution by Dispersal, Local Extinction, and Cladogenesis". In: *Systematic Biology* 57.1 (2008), pp. 4–14 (*cited on pages* 50, 53).

[193] E. M. Rodrigues, M.-F. Sagot, and Y. Wakabayashi. "The maximum agreement forest problem: Approximation algorithms and computational experiments". In: *Theoretical Computer Science* 374.1 (2007), pp. 91–110 (*cited on page* 58).

[194] F. Ronquist. "Dispersal-Vicariance Analysis: A New Approach to the Quantification of Historical Biogeography". In: *Systematic Biology* 46.1 (1997), pp. 195–203 (*cited on page* 53).

[195] F. Ronquist. "Reconstructing the history of host-parasite associations using generalised parsimony". In: *Cladistics* 11.1 (1995), pp. 73–89 (*cited on page* 45).

[196] F. Ronquist and S. Nylin. "Process and Pattern in the Evolution of Species Associations". In: *Systematic Biology* 39.4 (1990), pp. 323–344 (*cited on page* 44).

[197] E. Rosenberg, O. Koren, L. Reshef, R. Efrony, and I. Zilber-Rosenberg. "The role of microorganisms in coral health, disease and evolution". In: *Nature Reviews. Microbiology* 5.5 (2007), pp. 355–362 (*cited on pages* 37, 60).

[198] R. M. Ross, S. J. Greenhill, and Q. D. Atkinson. "Population structure and cultural geography of a folktale in Europe". In: *Proceedings of the Royal Society B: Biological Sciences* 280.1756 (2013). Publisher: Royal Society, p. 20123065 (*cited on pages* 61, 68).

[200] S. Santichaivekin, R. Mawhorter, and R. Libeskind-Hadas. "An efficient exact algorithm for computing all pairwise distances between reconciliations in the duplication-transfer-loss model". In: *BMC Bioinformatics* 20 (Suppl 20 2019) (*cited on page* 55).

[201] S. Santichaivekin, Q. Yang, J. Liu, R. Mawhorter, J. Jiang, T. Wesley, Y.-C. Wu, and R. Libeskind-Hadas. "eMPRess: a systematic cophylogeny reconciliation tool". In: *Bioinformatics* (btaa978 2020) (*cited on pages* 54, 55, 167).

[202] C. Scornavacca, E. Jacox, and G. J. Szöllősi. "Joint amalgamation of most parsimonious reconciled gene trees". In: *Bioinformatics (Oxford, England)* 31.6 (2015), pp. 841–848 (*cited on pages* 53, 56).

[203] C. Scornavacca, J. C. P. Mayol, and G. Cardona. "Fast algorithm for the reconciliation of gene trees and LGT networks". In: *Journal of Theoretical Biology* 418 (2017), pp. 129–137 (*cited on pages* 64, 138).

[204] C. Scornavacca, W. Paprotny, V. Berry, and V. Ranwez. "Representing a set of reconciliations in a compact way". In: *Journal of Bioinformatics and Computational Biology* 11.2 (2013), p. 1250025 (*cited on page* 54).

[206] B. Sennblad, E. Schreil, A.-C. Berglund Sonnhammer, J. Lagergren, and L. Arvestad. "primetv: a viewer for reconciled trees". In: *BMC bioinformatics* 8 (2007), p. 148 (*cited on page* 54).

[207] S. G. da Silva and J. J. Tehrani. “Comparative phylogenetic analyses uncover the ancient roots of Indo-European folktales”. In: *Royal Society Open Science* 3.1 (2015). Publisher: Royal Society, p. 150645 (*cited on pages* 61, 68).

[208] P. Simion, F. Delsuc, and H. Philippe. “To What Extent Current Limits of Phylogenomics Can Be Overcome?” In: (), p. 35 (*cited on page* 35).

[209] J. Sjöstrand, A. Tofigh, V. Daubin, L. Arvestad, B. Sennblad, and J. Lagergren. “A Bayesian Method for Analyzing Lateral Gene Transfer”. In: *Systematic Biology* 63.3 (2014), pp. 409–420 (*cited on page* 57).

[211] M. Stolzer, H. Lai, M. Xu, D. Sathaye, B. Vernot, and D. Durand. “Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees”. In: *Bioinformatics* 28.18 (2012), pp. i409–i415 (*cited on pages* 56, 59).

[212] M. Stolzer, K. Siewert, H. Lai, M. Xu, and D. Durand. “Event inference in multidomain families with phylogenetic reconciliation”. In: *BMC Bioinformatics* 16.14 (2015), S8 (*cited on page* 66).

[213] G. J. Szöllosi, E. Tannier, N. Lartillot, and V. Daubin. “Lateral gene transfer from the dead”. In: *Systematic Biology* 62.3 (2013), pp. 386–397 (*cited on pages* 52, 53).

[214] G. J. Szöllősi, B. Boussau, S. S. Abby, E. Tannier, and V. Daubin. “Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations”. In: *Proceedings of the National Academy of Sciences* 109.43 (2012). Publisher: National Academy of Sciences Section: Biological Sciences, pp. 17513–17518 (*cited on pages* 51, 57, 73, 75).

[215] G. J. Szöllősi, A. A. Davín, E. Tannier, V. Daubin, and B. Boussau. “Genome-scale phylogenetic analysis finds extensive gene transfer among fungi”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 370.1678 (2015). Publisher: Royal Society, p. 20140335 (*cited on page* 53).

[216] G. J. Szöllősi, W. Rosikiewicz, B. Boussau, E. Tannier, and V. Daubin. “Efficient Exploration of the Space of Reconciled Gene Trees”. In: *Systematic Biology* 62.6 (2013), pp. 901–912 (*cited on pages* 50, 56, 57, 81, 169).

[217] G. J. Szöllősi, E. Tannier, V. Daubin, and B. Boussau. “The Inference of Gene Trees with Species Trees”. In: *Systematic Biology* 64.1 (2015), e42–e62 (*cited on pages* 41, 59).

[218] J. J. Tehrani. “The Phylogeny of Little Red Riding Hood”. In: *PLoS ONE* 8.11 (2013). Ed. by R. A. Bentley, e78871 (*cited on page* 40).

[219] K. R. Theis, N. M. Dheilly, J. L. Klassen, R. M. Brucker, J. F. Baines, T. C. G. Bosch, J. F. Cryan, S. F. Gilbert, C. J. Goodnight, E. A. Lloyd, J. Sapp, P. Vandenkoornhuyse, I. Zilber-Rosenberg, E. Rosenberg, and S. R. Bordenstein. “Getting the Hologenome Concept Right: an Eco-Evolutionary Framework for Hosts and Their Microbiomes”. In: *mSystems* 1.2 (2016). Ed. by J. A. Gilbert (*cited on page* 60).

[221] A. Tofigh, J. Sjostrand, B. Sennblad, L. Arvestad, and J. Lagergren. “Detecting LGTs using a novel probabilistic model integrating duplications, LGTs, losses, rate variation, and sequence evolution”. In: (), p. 18 (*cited on pages* 71, 75).

[222] A. Tofigh, M. Hallett, and J. Lagergren. “Simultaneous identification of duplications and lateral gene transfers”. In: *IEEE/ACM transactions on computational biology and bioinformatics* 8.2 (2011), pp. 517–535 (*cited on page* 52).

[224] I. Ullah, P. Parviainen, and J. Lagergren. “Species Tree Inference Using a Mixture Model”. In: *Molecular Biology and Evolution* 32.9 (2015). Publisher: Oxford Academic, pp. 2469–2482 (*cited on page* 58).

[225] L. Urbini. “Models and algorithms to study the common evolutionary history of hosts and symbionts”. PhD thesis. Université de Lyon, 2017 (*cited on pages* 59, 161).

[226] L. Urbini, B. Sinaimeri, C. Matias, and M.-F. Sagot. “Exploring the Robustness of the Parsimonious Reconciliation Method in Host-Symbiont Cophylogeny”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 16.3 (2019), pp. 738–748 (*cited on page* 55).

[227] M. H. Van Dam and N. J. Matzke. "Evaluating the influence of connectivity and distance on biogeographical patterns in the south-western deserts of North America". In: *Journal of Biogeography* 43.8 (2016), pp. 1514–1532 (*cited on page* 53).

[228] D. M. d. Vienne, G. Refrégier, M. López-Villavicencio, A. Tellier, M. E. Hood, and T. Giraud. "Cospeciation vs host-shift speciation: methods for testing, evidence from natural associations and relation to coevolution". In: *New Phytologist* 198.2 (2013), pp. 347–385 (*cited on page* 37).

[229] D. M. de Vienne. "Tanglegrams Are Misleading for Visual Evaluation of Tree Congruence". In: *Molecular Biology and Evolution* 36.1 (2019), pp. 174–176 (*cited on page* 44).

[230] D. M. d. Vienne. "Lifemap: Exploring the Entire Tree of Life". In: *PLOS Biology* 14.12 (2016). Publisher: Public Library of Science, e2001624 (*cited on pages* 23, 24).

[232] Y. Wang, A. Mary, M.-F. Sagot, and B. Sinaimeri. "Capybara: equivalence ClAss enumeration of coPhylogenY event-BAsed ReconciliAtions". In: *Bioinformatics* 36.14 (2020). Publisher: Oxford Academic, pp. 4197–4199 (*cited on pages* 55, 167, 169).

[233] T. Warnow. "Supertree Construction: Opportunities and Challenges". In: *arXiv:1805.03530 [q-bio]* (2018). arXiv: 1805.03530 (*cited on pages* 34, 57).

[234] J. D. Weckstein. "Biogeography Explains Cophylogenetic Patterns in Toucan Chewing Lice". In: *Systematic Biology* 53.1 (2004), pp. 154–164 (*cited on pages* 36, 64).

[235] A. Wehe, M. S. Bansal, J. G. Burleigh, and O. Eulenstein. "DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony". In: *Bioinformatics* 24.13 (2008), pp. 1540–1541 (*cited on page* 57).

[236] S. Weiner and M. S. Bansal. "Improved Duplication-Transfer-Loss Reconciliation with Extinct and Unsampled Lineages". In: *Algorithms* 14.8 (2021). Number: 8 Publisher: Multidisciplinary Digital Publishing Institute, p. 231 (*cited on page* 53).

[237] N. Wieseke, M. Bernt, and M. Middendorf. "Unifying Parsimonious Tree Reconciliation". In: *arXiv:1307.7831 [cs, q-bio]* (2013). arXiv: 1307.7831 (*cited on pages* 41, 44, 52, 162).

[238] N. Wieseke, T. Hartmann, M. Bernt, and M. Middendorf. "Cophylogenetic Reconciliation with ILP". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 12.6 (2015), pp. 1227–1235 (*cited on page* 52).

[239] B. K. Wijayawardena, D. J. Minchella, and J. A. DeWoody. "Hosts, parasites, and horizontal gene transfer". In: *Trends in Parasitology* 29.7 (2013), pp. 329–338 (*cited on page* 61).

[240] E. O. Wiley. "Parsimony Analysis and Vicariance Biogeography". In: *Systematic Zoology* 37.3 (1988), pp. 271–290 (*cited on page* 46).

[241] C. R. Woese and G. E. Fox. "Phylogenetic structure of the prokaryotic domain: the primary kingdoms." In: *Proceedings of the National Academy of Sciences of the United States of America* 74.11 (1977), pp. 5088–5090 (*cited on page* 34).

[242] C. R. Woese, O. Kandler, and M. L. Wheelis. "Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya." In: *Proceedings of the National Academy of Sciences* 87.12 (1990), pp. 4576–4579 (*cited on page* 40).

[244] Y.-C. Wu, M. D. Rasmussen, M. S. Bansal, and M. Kellis. "Most parsimonious reconciliation in the presence of gene duplication, loss, and deep coalescence using labeled coalescent trees". In: *Genome Research* (2013), gr.161968.113 (*cited on page* 60).

[245] Y.-C. Wu, M. D. Rasmussen, and M. Kellis. "Evolution at the subgene level: domain rearrangements in the Drosophila phylogeny". In: *Molecular Biology and Evolution* 29.2 (2012), pp. 689–705 (*cited on page* 66).

[249] Y. Yu, R. M. Barnett, and L. Nakhleh. "Parsimonious Inference of Hybridization in the Presence of Incomplete Lineage Sorting". In: *Systematic Biology* 62.5 (2013), pp. 738–751 (*cited on page* 64).

[250] Y. Yu, J. Dong, K. J. Liu, and L. Nakhleh. "Maximum likelihood inference of reticulate evolutionary histories". In: *Proceedings of the National Academy of Sciences* 111.46 (2014), pp. 16448–16453 (*cited on pages* 49, 64).

[251] Y. Yu and L. Nakhleh. “Fast Algorithms for Reconciliation under Hybridization and Incomplete Lineage Sorting”. In: *arXiv:1212.1909 [cs, q-bio]* (2012). arXiv: 1212.1909 (*cited on page* 64).

[253] Y. Zheng, T. Wu, and L. Zhang. “Reconciliation of Gene and Species Trees With Polytomies”. In: *arXiv:1201.3995 [q-bio]* (2012). arXiv: 1201.3995 (*cited on page* 57).

[254] Y. Zheng and L. Zhang. “Reconciliation with Non-binary Gene Trees Revisited”. In: *Research in Computational Molecular Biology*. Ed. by R. Sharan. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2014, pp. 418–432 (*cited on page* 56).

[255] I. Zilber-Rosenberg and E. Rosenberg. “Role of microorganisms in the evolution of animals and plants: the hologenome theory of evolution”. In: *FEMS microbiology reviews* 32.5 (2008), pp. 723–735 (*cited on pages* 37, 60).

[256] E. Zuckerkandl and L. Pauling. “Molecules as documents of evolutionary history.” eng. In: *J Theor Biol* 8.2 (1965), pp. 357–366 (*cited on pages* 40, 48).

[257] A. Zwaenepoel and Y. Van de Peer. *Ancient whole genome duplications and the evolution of the gene duplication and loss rate*. preprint. Evolutionary Biology, 2019 (*cited on pages* 43, 62).

Chapter 2 - 3-level reconciliation

[6] S. J. Anthony, C. K. Johnson, D. J. Greig, S. Kramer, X. Che, H. Wells, A. L. Hicks, D. O. Joly, N. D. Wolfe, P. Daszak, W. Karesh, W. I. Lipkin, S. S. Morse, PREDICT Consortium, J. A. K. Mazet, and T. Goldstein. “Global patterns in coronavirus diversity”. In: *Virus Evolution* 3 (vex012 2017) (*cited on page* 109).

[19] V. Berry, F. Chevenet, J.-P. Doyon, and E. Jousselin. “A geography-aware reconciliation method to investigate diversification patterns in host/parasite interactions”. In: *Molecular Ecology Resources* 18.5 (2018), pp. 1173–1184 (*cited on pages* 66, 124).

[33] K. P. Burnham and D. R. Anderson. “Multimodel Inference: Understanding AIC and BIC in Model Selection”. In: *Sociological Methods & Research* 33.2 (2004), pp. 261–304 (*cited on page* 104).

[36] Y.-b. Chan, V. Ranwez, and C. Scornavacca. “Reconciliation-based detection of co-evolving gene families”. In: *BMC Bioinformatics* 14 (2013), p. 332 (*cited on pages* 114, 126).

[68] L. Duchemin, P. Veber, and B. Boussau. *Bayesian investigation of SARS-CoV-2-related mortality in France*. 2021 (*cited on page* 109).

[70] W. Duchemin, G. Gence, A.-M. Arigon Chifolleau, L. Arvestad, M. S. Bansal, V. Berry, B. Boussau, F. Chevenet, N. Comte, A. A. Davín, C. Dessimoz, D. Dylus, D. Hasic, D. Mallo, R. Planel, D. Posada, C. Scornavacca, G. Szöllősi, L. Zhang, É. Tannier, and V. Daubin. “RecPhyloXML: a format for reconciled gene trees”. In: *Bioinformatics* 34.21 (2018), pp. 3646–3652 (*cited on pages* 54, 68, 86, 116).

[78] J. Felsenstein. *Inferring Phylogenies*. Oxford, New York: Oxford University Press, 2003. 580 pp. (*cited on pages* 43, 46, 48, 71, 75, 108, 159).

[80] Y. Fu, M. Pistolozzi, X. Yang, and Z. Lin. *A Comprehensive Classification of Coronaviruses and Inferred Cross-Host Transmissions*. preprint. Bioinformatics, 2020 (*cited on pages* 61, 109).

[83] K. Gori, T. Suchan, N. Alvarez, N. Goldman, and C. Dessimoz. “Clustering Genes of Common Evolutionary History”. In: *Molecular Biology and Evolution* 33.6 (2016), pp. 1590–1605 (*cited on page* 114).

[110] E. Kuitche, Y. Qi, N. Tahiri, J. Parmer, and A. Ouangraoua. “DoubleRecViz: a web-based tool for visualizing transcript–gene–species tree reconciliation”. In: *Bioinformatics* 37.13 (2021), pp. 1920–1922 (*cited on page* 124).

[124] L. Li and M. S. Bansal. “Simultaneous Multi-Domain-Multi-Gene Reconciliation Under the Domain-Gene-Species Reconciliation Model”. In: *Bioinformatics Research and Applications*. Ed. by Z. Cai, P. Skums, and M. Li. Lecture Notes in Computer Science. Springer International Publishing, 2019, pp. 73–86 (*cited on pages* 67, 101).

[181] S. Penel, H. Menet, T. Tricou, V. Daubin, and E. Tannier. “Thirdkind: displaying phylogenetic encounters beyond 2-level reconciliation”. In: *Bioinformatics (Oxford, England)* (2022), btac062 (*cited on page* 116).

[199] E. Sallard, J. Halloy, D. Casane, E. Decroly, and J. van Helden. “Tracing the origins of SARS-COV-2 in coronavirus phylogenies: a review”. In: *Environmental Chemistry Letters* (2021), pp. 1–17 (*cited on pages* 109, 170).

[247] Z. Yang. *Computational molecular evolution*. Oxford series in ecology and evolution. Oxford: Oxford University press, 2006 (*cited on page* 105).

Chapter 3 - *Helicobacter pylori*

[2] M. Achtman, T. Azuma, D. E. Berg, Y. Ito, G. Morelli, Z. J. Pan, S. Suerbaum, S. A. Thompson, A. van der Ende, and L. J. van Doorn. “Recombination and clonal groupings within Helicobacter pylori from different geographical regions”. In: *Molecular Microbiology* 32.3 (1999), pp. 459–470 (*cited on pages* 131, 139).

[3] M. Achtman. “How old are bacterial pathogens?” In: *Proceedings of the Royal Society B: Biological Sciences* 283.1836 (2016), p. 20160990 (*cited on pages* 61, 129, 138).

[5] C. P. Andam and J. P. Gogarten. “Biased gene transfer in microbial evolution”. In: *Nature Reviews Microbiology* 9.7 (2011). Number: 7 Publisher: Nature Publishing Group, pp. 543–555 (*cited on page* 153).

[12] M. S. Bansal, E. J. Alm, and M. Kellis. “Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss”. In: *Bioinformatics* 28.12 (2012). Publisher: Oxford Academic, pp. i283–i291 (*cited on pages* 50, 153).

[28] D. Bravo, A. Hoare, C. Soto, M. A. Valenzuela, and A. F. Quest. “Helicobacter pylori in human health and disease: Mechanisms for local gastric and systemic effects”. In: *World Journal of Gastroenterology* 24.28 (2018), pp. 3071–3089 (*cited on page* 130).

[29] S. Breurec, B. Guillard, S. Hem, S. Brisse, F. B. Dieye, M. Huerre, C. Oung, J. Raymond, T. Sreng Tan, J.-M. Thiberge, S. Vong, D. Monchy, and B. Linz. “Evolutionary History of Helicobacter pylori Sequences Reflect Past Human Migrations in Southeast Asia”. In: *PLoS ONE* 6.7 (2011), e22058 (*cited on page* 155).

[34] L. L. Cavalli-Sforza, L. Cavalli-Sforza, P. Menozzi, and A. Piazza. *The History and Geography of Human Genes*. Google-Books-ID: FrwNcwKaUKoC. Princeton University Press, 1994. 1144 pp. (*cited on page* 133).

[59] X. Didelot and D. Falush. “Inference of Bacterial Microevolution Using Multilocus Sequence Data”. In: *Genetics* 175.3 (2007), pp. 1251–1266 (*cited on page* 136).

[61] B. Donati, C. Baudet, B. Sinaimeri, P. Crescenzi, and M.-F. Sagot. “EUCALYPT: efficient tree reconciliation enumerator”. In: *Algorithms for Molecular Biology* 10.1 (2015), p. 3 (*cited on pages* 52, 54, 55, 59, 153, 169).

[73] M. Eppinger, C. Baar, B. Linz, G. Raddatz, C. Lanz, H. Keller, G. Morelli, H. Gressmann, M. Achtman, and S. C. Schuster. “Who Ate Whom? Adaptive Helicobacter Genomic Changes That Accompanied a Host Jump from Early Humans to Large Felines”. In: *PLOS Genetics* 2.7 (2006). Publisher: Public Library of Science, e120 (*cited on page* 130).

[74] D. Falush, C. Kraft, N. S. Taylor, P. Correa, J. G. Fox, M. Achtman, and S. Suerbaum. “Recombination and mutation during long-term gastric colonization by *Helicobacter pylori* : Estimates of clock rates, recombination size, and minimal age”. In: *Proceedings of the National Academy of Sciences* 98.26 (2001), pp. 15056–15061 (*cited on page* 130).

[75] D. Falush, M. Stephens, and J. K. Pritchard. "Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies". In: *Genetics* 164.4 (2003), pp. 1567–1587 (*cited on page* 133).

[76] D. Falush, T. Wirth, B. Linz, J. K. Pritchard, M. Stephens, M. Kidd, M. J. Blaser, D. Y. Graham, S. Vacher, G. I. Perez-Perez, Y. Yamaoka, F. Mégraud, K. Otto, U. Reichard, E. Katzowitsch, X. Wang, M. Achtman, and S. Suerbaum. "Traces of human migrations in Helicobacter pylori populations". In: *Science (New York, N.Y.)* 299.5612 (2003), pp. 1582–1585 (*cited on pages* 18, 19, 133, 137–139).

[95] J. Hey and R. Nielsen. "Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics". In: *Proceedings of the National Academy of Sciences* 104.8 (2007). Publisher: Proceedings of the National Academy of Sciences, pp. 2785–2790 (*cited on page* 136).

[97] J. K. Y. Hooi, W. Y. Lai, W. K. Ng, M. M. Y. Suen, F. E. Underwood, D. Tanyingoh, P. Malfertheiner, D. Y. Graham, V. W. S. Wong, J. C. Y. Wu, F. K. L. Chan, J. J. Y. Sung, G. G. Kaplan, and S. C. Ng. "Global Prevalence of Helicobacter pylori Infection: Systematic Review and Meta-Analysis". In: *Gastroenterology* 153.2 (2017), pp. 420–429 (*cited on page* 130).

[103] K. A. Jolley, J. E. Bray, and M. C. J. Maiden. "Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications". In: *Wellcome Open Research* 3 (2018), p. 124 (*cited on page* 139).

[104] E. T. Kabamba, V. P. Tuan, and Y. Yamaoka. "Genetic populations and virulence factors of Helicobacter pylori". In: *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases* 60 (2018), pp. 109–116 (*cited on page* 129).

[118] D. J. Lawson, G. Hellenthal, S. Myers, and D. Falush. "Inference of Population Structure using Dense Haplotype Data". In: *PLOS Genetics* 8.1 (2012). Publisher: Public Library of Science, e1002453 (*cited on page* 134).

[121] J. Z. Li, D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto, S. Ramachandran, H. M. Cann, G. S. Barsh, M. Feldman, L. L. Cavalli-Sforza, and R. M. Myers. "Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation". In: *Science* 319.5866 (2008). Publisher: American Association for the Advancement of Science, pp. 1100–1104 (*cited on page* 138).

[126] B. Linz, F. Balloux, Y. Moodley, A. Manica, H. Liu, P. Roumagnac, D. Falush, C. Stamer, F. Prugnolle, S. W. van der Merwe, Y. Yamaoka, D. Y. Graham, E. Perez-Trallero, T. Wadstrom, S. Suerbaum, and M. Achtman. "An African origin for the intimate association between humans and Helicobacter pylori". In: *Nature* 445.7130 (2007), pp. 915–918 (*cited on pages* 133, 135, 149, 154).

[127] B. Linz, C. R. R. Vololonantenainab, A. Seck, J.-F. Carod, D. Dia, B. Garin, R. M. Ramanampamonjy, J.-M. Thiberge, J. Raymond, and S. Breurec. "Population genetic structure and isolation by distance of Helicobacter pylori in Senegal and Madagascar". In: *PloS One* 9.1 (2014), e87355 (*cited on page* 155).

[129] S. López, L. van Dorp, and G. Hellenthal. "Human Dispersal Out of Africa: A Lasting Debate". In: *Evolutionary Bioinformatics Online* 11 (Suppl 2 2015), pp. 57–68 (*cited on pages* 131, 132).

[137] M. C. J. Maiden, J. A. Bygraves, E. Feil, G. Morelli, J. E. Russell, R. Urwin, Q. Zhang, J. Zhou, K. Zurth, D. A. Caugant, I. M. Feavers, M. Achtman, and B. G. Spratt. "Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms". In: *Proceedings of the National Academy of Sciences of the United States of America* 95.6 (1998), pp. 3140–3145 (*cited on pages* 35, 131).

[138] F. Maixner, B. Krause-Kyora, D. Turaev, A. Herbig, M. R. Hoopmann, J. L. Hallows, U. Kusebauch, E. E. Vigl, P. Malfertheiner, F. Megraud, N. O'Sullivan, G. Cipollini, V. Coia, M. Samadelli, L. Engstrand, B. Linz, R. L. Moritz, R. Grimm, J. Krause, A. Nebel, Y. Moodley, T. Rattei, and A. Zink. "The 5300-year-old Helicobacter pylori genome of the Iceman". In: *Science* 351.6269 (2016). Publisher: American Association for the Advancement of Science, pp. 162–165 (*cited on pages* 135, 136).

[142] B. J. Marshall and J. R. Warren. "Unidentified curved bacilli in the stomach of patients with gastritis and peptic ulceration". In: *Lancet (London, England)* 1.8390 (1984), pp. 1311–1315 (*cited on page* 130).

[148] F. Mégraud, P. Lehours, and F. F. Vale. "The history of Helicobacter pylori: from phylogeography to paleomicrobiology". In: *Clinical Microbiology and Infection* 22.11 (2016), pp. 922–927 (*cited on pages* 18, 129).

[149] F. Menardo, C. Loiseau, D. Brites, M. Coscolla, S. M. Gygli, L. K. Rutaihwa, A. Trauner, C. Beisel, S. Borrell, and S. Gagneux. "Treemmer: a tool to reduce large phylogenetic datasets with minimal loss of diversity". In: *BMC Bioinformatics* 19.1 (2018), p. 164 (*cited on pages* 139, 151).

[152] H. L. Mobley, G. L. Mendz, and S. L. Hazell, eds. *Helicobacter pylori: Physiology and Genetics*. Washington (DC): ASM Press, 2001 (*cited on pages* 129, 130).

[154] Y. Moodley, B. Linz, R. P. Bond, M. Nieuwoudt, H. Soodyall, C. M. Schlebusch, S. Bernhöft, J. Hale, S. Suerbaum, L. Mugisha, S. W. v. d. Merwe, and M. Achtman. "Age of the Association between Helicobacter pylori and Man". In: *PLOS Pathogens* 8.5 (2012), e1002693 (*cited on pages* 61, 135–137, 140).

[155] Y. Moodley, B. Linz, Y. Yamaoka, H. M. Windsor, S. Breurec, J.-Y. Wu, A. Maady, S. Bernhöft, J.-M. Thiberge, S. Phuanukoonnon, G. Jobb, P. Siba, D. Y. Graham, B. J. Marshall, and M. Achtman. "The Peopling of the Pacific from a Bacterial Perspective". In: *Science (New York, N.Y.)* 323.5913 (2009), pp. 527–530 (*cited on pages* 134, 136, 155).

[160] G. Morelli, X. Didelot, B. Kusecek, S. Schwarz, C. Bahlawane, D. Falush, S. Suerbaum, and M. Achtman. "Microevolution of Helicobacter pylori during prolonged infection of single hosts and within families". In: *PLoS genetics* 6.7 (2010), e1001036 (*cited on page* 130).

[162] Z. Y. Muñoz-Ramirez, B. Pascoe, A. Mendez-Tenorio, E. Mourkas, S. Sandoval-Motta, G. Perez-Perez, D. R. Morgan, R. L. Dominguez, D. Ortiz-Princz, M. E. Cavazza, G. Rocha, D. M. M. Queiroz, M. Catalano, G. Z. D. Palma, C. G. Goldman, A. Venegas, T. Alarcon, M. Oleastro, F. F. Vale, K. J. Goodman, R. C. Torres, E. Berthenet, M. D. Hitchings, M. J. Blaser, S. K. Sheppard, K. Thorell, and J. Torres. "A 500-year tale of co-evolution, adaptation, and virulence: Helicobacter pylori in the Americas". In: *The ISME Journal* 15.1 (2021). Number: 1 Publisher: Nature Publishing Group, pp. 78–92 (*cited on page* 134).

[163] Z. Y. Muñoz-Ramírez, A. Mendez-Tenorio, I. Kato, M. M. Bravo, C. Rizzato, K. Thorell, R. Torres, F. Aviles-Jimenez, M. Camorlinga, F. Canzian, and J. Torres. "Whole Genome Sequence and Phylogenetic Analysis Show Helicobacter pylori Strains from Latin America Have Followed a Unique Evolution Pathway". In: *Frontiers in Cellular and Infection Microbiology* 7 (2017), p. 50 (*cited on pages* 134, 138, 154).

[167] S. Nell, D. Eibach, V. Montano, A. Maady, A. Nkwescheu, J. Siri, W. F. Elamin, D. Falush, B. Linz, M. Achtman, Y. Moodley, and S. Suerbaum. "Recent Acquisition of Helicobacter pylori by Baka Pygmies". In: *PLOS Genetics* 9.9 (2013). Publisher: Public Library of Science, e1003775 (*cited on pages* 137, 154).

[185] J. K. Pritchard, M. Stephens, and P. Donnelly. "Inference of Population Structure Using Multilocus Genotype Data". In: *Genetics* 155.2 (2000), pp. 945–959 (*cited on page* 133).

[191] R. H. Ree, B. R. Moore, C. O. Webb, and M. J. Donoghue. "A likelihood framework for inferring the evolution of geographic range on phylogenetic trees". In: (2005), p. 13 (*cited on pages* 53, 152).

[203] C. Scornavacca, J. C. P. Mayol, and G. Cardona. "Fast algorithm for the reconciliation of gene trees and LGT networks". In: *Journal of Theoretical Biology* 418 (2017), pp. 129–137 (*cited on pages* 64, 138).

[205] T. Seemann. “Prokka: rapid prokaryotic genome annotation”. In: *Bioinformatics (Oxford, England)* 30.14 (2014), pp. 2068–2069 (*cited on page* 140).

[220] K. Thorell, K. Yahara, E. Berthenet, D. J. Lawson, J. Mikhail, I. Kato, A. Mendez, C. Rizzato, M. M. Bravo, R. Suzuki, Y. Yamaoka, J. Torres, S. K. Sheppard, and D. Falush. “Rapid evolution of distinct Helicobacter pylori subpopulations in the Americas”. In: *PLoS Genetics* 13.2 (2017), e1006546 (*cited on pages* 134, 144).

[243] P. Wongphutorn, C. Chomvarin, B. Sripa, W. Namwat, and K. Faksri. “Detection and genotyping of Helicobacter pylori in saliva versus stool samples from asymptomatic individuals in Northeastern Thailand reveals intra-host tissue-specific H. pylori subtypes”. In: *BMC Microbiology* 18.1 (2018), p. 10 (*cited on page* 154).

[246] K. Yahara, Y. Furuta, K. Oshima, M. Yoshida, T. Azuma, M. Hattori, I. Uchiyama, and I. Kobayashi. “Chromosome painting in silico in a bacterial species reveals fine population structure”. In: *Molecular Biology and Evolution* 30.6 (2013), pp. 1454–1464 (*cited on pages* 134, 138).

[248] S.-i. Yokota, M. Konno, S.-i. Fujiwara, N. Toita, M. Takahashi, S. Yamamoto, N. Ogasawara, and T. Shiraishi. “Intrafamilial, Preferentially Mother-to-Child and Intraspousal, Helicobacter pylori Infection in Japan Determined by Mutilocus Sequence Typing and Random Amplified Polymorphic DNA Fingerprinting”. In: *Helicobacter* 20.5 (2015), pp. 334–342 (*cited on page* 130).

[252] Y. Zhao, J. Wu, J. Yang, S. Sun, J. Xiao, and J. Yu. “PGAP: pan-genomes analysis pipeline”. In: *Bioinformatics* 28.3 (2012), pp. 416–418 (*cited on page* 140).

Chapter 4 - Open questions

[35] Y.-b. Chan, V. Ranwez, and C. Scornavacca. “Inferring incomplete lineage sorting, duplications, transfers and losses with reconciliations”. In: *Journal of Theoretical Biology* 432 (2017), pp. 1–13 (*cited on pages* 59, 161).

[78] J. Felsenstein. *Inferring Phylogenies*. Oxford, New York: Oxford University Press, 2003. 580 pp. (*cited on pages* 43, 46, 48, 71, 75, 108, 159).

[91] D. Hasic and E. Tannier. “Gene tree reconciliation including transfers with replacement is hard and FPT”. In: *Journal of Combinatorial Optimization* 38.2 (2019), pp. 502–544. arXiv: 1709.04459 (*cited on pages* 58, 162).

[150] H. Menet, V. Daubin, and E. Tannier. “Phylogenetic reconciliation”. 2021 (*cited on page* 161).

[184] A. A. Pittis and T. Gabaldón. “Late acquisition of mitochondria by a host with chimeric prokaryotic ancestry”. In: *Nature* 531.7592 (2016), pp. 101–104 (*cited on page* 163).

[210] T. Stadler, T. G. Vaughan, A. Gavryushkin, S. Guindon, D. Kühnert, G. E. Leventhal, and A. J. Drummond. “How well can the exponential-growth coalescent approximate constant-rate birth–death population dynamics?” In: *Proceedings of the Royal Society B: Biological Sciences* 282.1806 (2015). Publisher: Royal Society, p. 20150420 (*cited on pages* 163, 166).

[223] T. Tricou, E. Tannier, and D. M. de Vienne. *Ghost lineages invalidate or reverse several results on gene flow*. preprint. Evolutionary Biology, 2022 (*cited on page* 163).

[225] L. Urbini. “Models and algorithms to study the common evolutionary history of hosts and symbionts”. PhD thesis. Université de Lyon, 2017 (*cited on pages* 59, 161).

[237] N. Wieseke, M. Bernt, and M. Middendorf. “Unifying Parsimonious Tree Reconciliation”. In: *arXiv:1307.7831 [cs, q-bio]* (2013). arXiv: 1307.7831 (*cited on pages* 41, 44, 52, 162).

Chapter 5 - General conclusion

[61] B. Donati, C. Baudet, B. Sinaimeri, P. Crescenzi, and M.-F. Sagot. "EUCALYPT: efficient tree reconciliation enumerator". In: *Algorithms for Molecular Biology* 10.1 (2015), p. 3 (*cited on pages* 52, 54, 55, 59, 153, 169).

[100] E. Jacox, C. Chauve, G. J. Szöllősi, Y. Ponty, and C. Scornavacca. "ecceTERA: comprehensive gene tree-species tree reconciliation using parsimony". In: *Bioinformatics (Oxford, England)* 32.13 (2016), pp. 2056–2058 (*cited on pages* 50, 53, 71, 77, 78, 81, 169).

[158] B. Morel, A. M. Kozlov, A. Stamatakis, and G. J. Szöllősi. *GeneRax: A tool for species tree-aware maximum likelihood based gene family tree inference under gene duplication, transfer, and loss.* preprint. Bioinformatics, 2019 (*cited on pages* 41, 56, 71, 74, 75, 86, 169).

[180] S. Penel, A.-M. Arigon, J.-F. Dufayard, A.-S. Sertier, V. Daubin, L. Duret, M. Gouy, and G. Perrière. "Databases of homologous gene families for comparative genomics". In: *BMC bioinformatics* 10 Suppl 6 (2009), S3 (*cited on page* 169).

[199] E. Sallard, J. Halloy, D. Casane, E. Decroly, and J. van Helden. "Tracing the origins of SARS-COV-2 in coronavirus phylogenies: a review". In: *Environmental Chemistry Letters* (2021), pp. 1–17 (*cited on pages* 109, 170).

[201] S. Santichaivekin, Q. Yang, J. Liu, R. Mawhorter, J. Jiang, T. Wesley, Y.-C. Wu, and R. Libeskind-Hadas. "eMPRess: a systematic cophylogeny reconciliation tool". In: *Bioinformatics* (btaa978 2020) (*cited on pages* 54, 55, 167).

[216] G. J. Szöllősi, W. Rosikiewicz, B. Boussau, E. Tannier, and V. Daubin. "Efficient Exploration of the Space of Reconciled Gene Trees". In: *Systematic Biology* 62.6 (2013), pp. 901–912 (*cited on pages* 50, 56, 57, 81, 169).

[231] Y. Wang. "Algorithmic investigations of the dynamics of species interactions". In: (), p. 186 (*cited on page* 169).

[232] Y. Wang, A. Mary, M.-F. Sagot, and B. Sinaimeri. "Capybara: equivalence ClAss enumeration of coPhylogenY event-BAsed ReconciliAtions". In: *Bioinformatics* 36.14 (2020). Publisher: Oxford Academic, pp. 4197–4199 (*cited on pages* 55, 167, 169).