Ancestral Reconstruction: Theory and Practice

Mathieu Groussin, Vincent Daubin, Eric Tannier, Manolo Gouy

Ancestral Reconstruction: theory and practice

Mathieu Groussin, Vincent Daubin, Eric Tannier, Manolo Gouy

October 1, 2015

MGroussi@mit.edu, Biological engineering, Massachussets Institute of Technology, USA
Vincent.Daubin@univ-lyon1.fr, Laboratoire de Biométrie et Biologie Evolutive, Université de Lyon, 43 boulevard du 11 novembre 1918, 69622 Villeurbanne, France
Eric.Tannier@inria.fr, Institut National de Recherche en Informatique et Automatique (INRIA), Laboratoire de Biométrie et Biologie Evolutive, Université de Lyon, 43 boulevard du 11 novembre 1918, 69622 Villeurbanne, France
Manolo.Gouy@univ-lyon1.fr, Laboratoire de Biométrie et Biologie Evolutive, Université de Lyon, 43 boulevard du 11 novembre 1918, 69622 Villeurbanne, France

Glossary

Markov process Here, this term indicates a probabilistic process in which there exists a finite set of possible states (e.g., the 4 nucleotides, or the 20 amino acids) and a matrix whose terms are the rates at which a state can change per unit time. This conveniently represents the evolutionary process at the molecular level. Such a process is said to be homogeneous when the matrix is constant through time and through evolutionary lineages.

Dynamic programming is a computer science method for solving a complex problem by breaking it down into a collection of simpler subproblems. Here, it is used to calculate a result for a whole tree by considering progressively larger parts of this tree, starting from the tree leaves for which the desired result can be easily computed.

Phylogenetic tree A tree is an object made of nodes connected by edges such that no closed circuit of edges exists, and there exists a path between any pair of nodes. Tree leaves are those nodes connected to a single other node. In a rooted tree, one node has a special status by being the tree root. A phylogenetic tree represents the evolutionary divergence of biological entities, usually species or genes, the leaves, from their last common ancestor, the tree root. Its nodes figure ancestral entities. Its edges figure evolutionary lineages. In a species tree,

biological entities are species. In a gene tree, they are homologous genes, that is, genes sharing a common ancestor.

Keywords

ancestral chromosome; ancestral sequence reconstruction; continuous characters; discrete characters; gene adjacencies; gene order; gene content; gene tree/species tree reconciliation; maximum likelihood; parsimony;

Synopsis

Biological organisms are the result of a long evolutionary history, and knowledge of the path they have followed through time is very beneficial to the understanding of their extant shapes and functions. Interestingly, each lineage keeps a record of its own evolutionary history. Comparing lineages thus allows to restore the information of ancestral traits (*e.g.* morphologies, molecules, etc) and gives insights on organisms living in the past, as well as on the history of their diversification.

We survey mathematical models and computational methods used to reconstruct ancestral states for different levels of organization: phenotypic traits, DNA and protein sequences, gene repertoires, and genome architecture. We discuss the possibility of reconstructing ancestral genomes in their entirety, integrating all these levels of complexity.

Introduction

The theoretical possibility of reconstructing ancestral molecules from the sequence of these molecules in extant organisms has been formalized by Pauling and Zuckerkandl (1963), and has become practical and widely applied in the recent years. Various types of traits have been considered for ancestral reconstruction: sequence data, either of the nucleotide or protein types, have been extensively studied. Methods for processing sequence data can often be seen in a more general fashion as methods handling characters with a discrete set of values. Therefore, these methods also allow to consider non-sequence data such as binary characters, presence/absence. Other sets of methods consider continuous characters (e.g., body mass) and allow the reconstruction of their ancestral values.

Ancestral reconstruction of gene content has also been studied with the objective of understanding the evolutionary history of gene repertoires over time. Methods for discrete characters can be applied in this case, but a model incorporating the processes of gene evolution is more accurate. Indeed, the processes of gene evolution in genomes include duplications and horizontal gene transfer, which are improperly rendered when genes are coded as discrete character

data. It has triggered the development of new methods specific for the reconstruction of gene history and hence ancestral genic repertoires. Similarly, the order of genes along chromosomes has been considered as a target for ancestral reconstruction. Because this problem becomes very rapidly computationally intractable, the less ambitious target of reconstructing ancestral adjacencies between genes along chromosomes has been studied.

For very shallow divergences, recent attempts have also been performed to reconstruct full ancestral chromosomal sequences, with both genic and non coding nucleotides.

This chapter describes major methods for all these problems, which concern the reconstruction of ancestral traits or sequences along a pre-established phylogenetic tree, and starting from a set of observed trait values (*e.g.,* multiple alignment of extant homologous sequences) (see Figure 1). Notably, the procedures described here are those that have been widely used recently to reconstruct ancestral enzymes in order to study the path followed by a protein along evolutionary time.

1 Ancestral character reconstruction

Inferring the ancestral states of a given trait requires a set of values measured in extant organisms, a binary rooted or unrooted phylogeny relating these organisms and a model assuming a particular process of trait evolution. This model may contain parameters, such as rates of change between states or the branch lengths of the phylogeny, which are usually estimated from the data in the case of probabilistic reconstructions. Three methodological frameworks are commonly considered to infer ancestral characters: Parsimony (Sankoff, 1975), Maximum Likelihood (ML) (Schluter *et al.*, 1997) and the Bayesian framework (Pagel *et al.*, 2004).

Models used to infer ancestral characters are usually homogeneous continuous-time Markov processes (Pagel, 1994; Lewis, 2001) defined over state spaces encompassing extant measures. Examples of such state spaces are finite sets of binary characters (*e.g.* presence/absence of wings, expression/inhibition of a gene, etc), finite unordered spaces (*e.g.* nucleotides, amino-acids, diets, morphology, gene orders), ordered discrete sets (*e.g.* number of genes, genome size), or continuous sets allowing continuous variation along a numerical range of values (*e.g.* body size, mass, optimal growth temperature).

1.1 Discrete characters

Markov models for a finite state space Ω may contain several parameters, depending on the *a priori* assumptions that are made about transition rates between states (Pagel, 1994; Yang, 2006). When applied on binary or unordered (excepted nucleotides or amino acids, see Section 2) sets of characters of size k, the models are called M_k. Extensions of M_k are for example the Binary and Multi-State Speciation and Extinction models (BiSSE and MuSSE, respec-

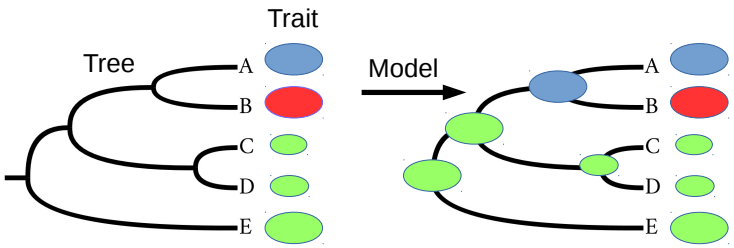

Figure 1: **Reconstruction of ancestral characters along a phylogenetic tree**

tively) developed to simultaneously infer the process of state change with the estimation of speciation and extinction rates along the phylogeny (Maddison *et al.*, 2007; FitzJohn, 2012). M_k model makes the assumption that transition rates are constant across lineages. The Hidden Rates Model (HRM) relax this homogeneity hypothesis and intends to capture the variation of transition rates between states across the phylogeny, yielding more accurate estimations of ancestral states (Beaulieu *et al.*, 2013).

Maximum parsimony reconstructions are approximations of solutions when the rates of events are supposed to be low, so that the probability of multiple events on one branch can be considered negligible. In this case, consider a cost d_{ij} of transition between states on any branch, with $i, j \in \Omega$, and ancestral states are chosen in order to minimise the sum of the costs of transitions on all branches. Calculations typically follow a dynamic programming procedure along the phylogeny (Sankoff, 1975). If the state space is small, a cost $C_u(i)$ is computed for every node u of the phylogeny, and every possible state i, starting from the leaves and following a postorder traversal of the tree. For a leaf l, $C_l(i) = 0$ if the state i is present at leaf l, and $C_l(i) = \infty$ otherwise. Then, assuming that u_1 and u_2 are the children of u, apply

$$C_u(i) = \min_{j \in \Omega}(C_{u_1}(j) + d_{ij}) + \min_{j \in \Omega}(C_{u_2}(j) + d_{ij}).$$

(For u_1, the cost $C_{u_1}(j) + d_{ij}$ is computed for all possible values $j \in \Omega$, and the minimum is retained. Similarly for u_2).

One can thus obtain the set F_{root} of minimal cost states at the tree root, those for which $C_{root}(i)$ is minimal. The set F_p of minimal cost states at internal node p, child of node a, is recursively computed, from root to leaves, with:

$F_p \leftarrow \emptyset$ **#Initially, F_p is the empty set**
#for each minimal cost state assigned to parent node a
for each $i \in F_a$ do
  $m \leftarrow min_j(d_{ij} + C_p(j))$ **#The minimum cost m is computed**
#All states j associated with a cost equal to m are added
#to the set of possible values
  for each j do
    if $d_{ij} + C_p(j) = m$
      $F_p \leftarrow F_p \cup \{j\}$

The set F_p gives all most parsimonious ancestral states at node p, and the result does not depend on the position of the root. There are two primary issues with the use of parsimony to reconstruct ancestral states. First, parsimony ignores branch lengths, which are indicative of variation of evolutionary rates between lineages. Second, parsimony does not account for variation of rates between the different states of the character, excepted with weighted parsimony, which assign weights to state changes. But, in this case it is always difficult to attribute appropriate weights *a priori*.

Likelihood calculations follow the same computational framework, with the advantage of probabilistically accounting for the concerns raised above about parsimony. Let $L_u(i)$ be the probability of observing the state i at node u, p_{ij}^l being the probability for state i to turn into j along a branch of length l, computed according to the evolutionary model. Conditional probabilities are computed recursively along the tree from the leaves to the root. For a leaf l, $L_l(i) = 1$ if i is the observed value at leaf l, and $L_l(i) = 0$ otherwise. For other nodes,

$$L_u(i) = (\sum_{j_1 \in \Omega} p_{ij_1}^{l_1} \times L_{u_1}(j_1))(\sum_{j_2 \in \Omega} p_{ij_2}^{l_2} \times L_{u_2}(j_2))$$

where u_1 and u_2 are the children of node u and l_k is the length of the branch between u and u_k, which has to be a parameter of the model.

The probabilistic approach considers two types of ancestral character reconstruction, the *marginal* and the *joint* reconstructions. The marginal reconstruction gives the probabilities $P(u = i|S)$ of each state i of a given ancestral node u, where S represents the observed values at all leaves of the tree. Using Bayes formula, we obtain:

$$P(u = i|S) = \frac{P(u = i)P(S|u = i)}{P(S)} = \frac{\pi_i P(S|u = i)}{P(S)}.$$

where π_i is the equilibrium frequency of character i. Provided it is computed after rooting the tree at its node r, the vector of conditional probabilities at the tree root defined above corresponds to the term $P(S|r = i)$. This procedure is called "empirical Bayes" by Yang (2006).

The joint reconstruction gives the most probable set of residues across all ancestral nodes, and necessitates a backward traversal of the tree like in the parsimony case (Pupko *et al.*, 2002).

The computations have a quadratic complexity with respect to the state space size k, which makes dynamic programming a weak tool for large or infinite state spaces. If the state space is the integer set, it is possible to see the cost function $C_u(i)$ as an affine function of i and to propagate only the coefficients of the function along the nodes, instead of every value, to reach a complexity that does not depend on the size of the state space (Csűrös, 2014), making parsimony calculations feasible. When the state space is large but without a good structure to order it, as with gene orders (Section 4, there can be $n!$ possible gene orders), then the ancestral reconstruction becomes intractable.

1.2 Continuous characters

When the state space Ω is a continuous set of numbers, the parsimony cost function d_{ij} is usually defined as the absolute value or the square of the difference between i and j. The cost C cannot be computed for all values $i \in \Omega$, and the coefficients of a linear or quadratic function of i is computed at each node, following the same dynamic programming principle (for an exhaustive review on parsimony methods see Csűrös (2014)). If $d_{ij} = (i-j)^2$, then the parsimony solution is also the ML one under a Brownian motion (BM) with a constant rate, which is the most commonly used Markovian process under continuous characters. BM assumes that the trait value changes as a random walk, with multiple and independent steps drawn from a Normal distribution of mean 0 and variance σ^2. The consequence of the independence between steps is that the net change of the trait value along a branch of length t is drawn from a Normal distribution of mean 0 and variance $t \times \sigma^2$. With a BM process, the only parameter is the variance σ^2, the trait changes with a constant rate along the phylogeny and drifts neutrally without directionality or evolves towards an optimum that drifts neutrally (Felsenstein, 2004). BM is a special case of the Ornstein-Uhlenbeck (OU) process, which contains an additional attraction parameter toward a central value, increased with the distance from this central value. OU might be of interest when stabilising selection acts on the trait under consideration (Martins 1994). Extensions of these models were proposed to account for the heterogeneity in rate of change over time (Blomberg *et al.*, 2003; Harmon *et al.*, 2010; Eastman *et al.*, 2011). For instance, the Early Burst Model (Harmon *et al.*, 2010) allows the rate of the BM process to exponentially change in time.

The estimation of ancestral continuous trait values can be performed in ML (Schluter *et al.*, 1997). Likelihoods of transitions in state between adjacent nodes are easily derived from the BM process formulation, and the joint likelihood of the Brownian rate (σ^2) and all internal states is maximized over the entire tree. Restricted ML (REML) is also often used (Felsenstein, 1985; Paradis *et al.*, 2004). REML does not analyze the raw data directly, but instead realises the ML estimations using the contrasts among the observations. REML has been

shown to produce unbiased estimates of variance and covariance parameters. Ancestral values may be estimated in the Bayesian framework (Huelsenbeck *et al.*, 2002; Pagel *et al.*, 2004), allowing to integrate over uncertainty in the tree topology, branch lengths, and rate parameters.

Two alternative methods are commonly considered: the Phylogenetic Independent Contrast (PIC) (Felsenstein, 1985) and the Generalized Least Squares (GLS) methods (Martins and Hansen, 1997). However, they are not usually employed to infer ancestral character values and were rather designed to control for the influence of tree topology on the estimation of correlations between evolving traits. PIC assumes a BM-like model to recursively transform the tip values into statistically independent and identically distributed values, called contrasts, over the internal nodes up to the root. Thus, the ancestral values can be estimated with the data of descendant nodes only, contrarily to ML estimations. GLS estimates the unknown parameters in a linear regression model. It assumes the model $Y = DX$, where D is a vector of observed values at the tips, X is a matrix describing both the phylogeny unifying the tips along which the trait evolves and the process of trait evolution (usually a BM process). Y is the vector of ancestral trait estimates. In many situations, ML, PIC and GLS produce similar values, especially when a simple BM process is assumed to model the evolution of traits (Martins, 1999).

2 Ancestral sequence reconstruction

Ancestral sequence reconstruction (ASR) from extant molecular sequences (DNA or proteins) consists in computing ancestral residues, given extant residues at the leaves of a phylogenetic tree. As such it requires a multiple alignment of sequences, a phylogenetic tree and a substitution model. Computations are usually done independently at each site of the aligned sequences.

ASR inherits from all models and methods for ancestral character reconstruction for small state spaces (DNA or amino-acids). But models of sequence evolution have their specificities. First, depending on the number of parameters defining the transition rates between states, the M_k model takes different names (*e.g.* Jukes and Cantor (JC) when all rates are equal or General Time Reversible (GTR) when all rates are different (Yang, 2006)). Second, the variation of evolutionary rates across sequence sites is usually accounted for, modelled by a gamma distribution discretized in K classes of $\frac{1}{K}$ weight. Then, the likelihood equation of section 1 becomes:

$$L_u^s(i) = \sum_{c=1}^{K} \frac{1}{K} [(\sum_{j_1} p_{ij_1}^{l_{uu_i}}) L_{u_1}^s(j_1))(\sum_{j_2} p_{ij_2}^{l_{uu_i}}) L_{u_2}^s(j_2))].$$

With sequences, the *marginal* reconstruction is the most frequently used algorithm to compute ancestral characters. However, an efficient algorithm developed by Pupko *et al.* (2002) can be employed to perform *joint* reconstruction. It deals with all ancestral nodes together, and the set of sequence residues at all

nodes of largest probability is considered to be the best ancestral reconstruction at a given site. While the marginal reconstruction requires a computation time proportional to the number of sequences, the joint reconstruction is exponential in the sequence number when across-site rate variation is modeled with a discretized gamma distribution (Liberles, 2007). Joint reconstruction is therefore difficult for more than a few tens of sequences.

Yang (2006) indicates that joint and marginal reconstructions usually produce consistent results where the most probable joint reconstruction for a site consists of the best marginal reconstruction at each node. Furthermore, when conflicting results arise, neither reconstruction is very reliable.

In practice, the marginal approach of ancestral sequence reconstruction in ML, retaining the most probable residue for each node at each site is most often employed in ancestral protein resurrection experiments. Several software packages implement this procedure: PAML (Yang, 2007), MEGA (Tamura *et al.*, 2011), DAMBE (Xia, 2013), and bppAncestor, a part of the Bio++ library (Guéguen *et al.*, 2013).

When the goal is the resurrection of ancestral proteins, the residue with largest probability at each node is often retained. In this case, it is interesting to consider the magnitude of this largest probability: values above .9 indicate high confidence in the reconstructed ancestral residue, whereas residues with moderate probabilities are those where reconstruction is uncertain. This strategy, though, is biased towards the most frequent amino acid at each protein site (Yang, 2006). An alternative strategy that avoids this bias is to generate an ancestral sequence by randomly drawing at each site one residue according to the probability vector computed for this site. The second strategy also allows to generate a small number of putative ancestral sequences in order to measure the sensitivity of inferences to the uncertainty about ancestral sequences.

Several studies have compared ASRs by the parsimony and probabilistic approaches (Yang *et al.*, 1995; Zhang and Nei, 1997). The general outcome is that probabilistic methods are more accurate than the parsimony approach, except when sequence divergences are weak where the two approaches perform similarly. The probabilistic approach has also the advantage of estimating the uncertainty of the reconstruction at each ancestral site. It was also shown that complex probabilistic models which aim at capturing the compositional heterogeneity of the substitution process provide more accurate estimates of ancestral sequences (Groussin *et al.*, 2013). However, the downside of parameter-rich models is that they may require large datasets to accurately estimate all parameters, and may also increase the time and memory requirements of the algorithm.

Models of insertions and deletions on a multiple alignment are necessary to correctly infer the presence or absence of a residue at some site in an ancestral sequence. For this, alignments have to be computed together with phylogenies, so that indels are scored according to an evolutionary model. It is done in parsimony by Löytynoja and Goldman (2008), and statistical models of insertion/deletion are included in alignments algorithms of Diallo *et al.* (2007). In that case the alignment, phylogeny and presence or absence of ancestral residues are simultaneously given as an input to ASR methods. Diallo *et al.*

(2010) present a software for computing indels (and thus, an alignment) given a phylogeny, while Suchard and Redelings (2006) and Lunter *et al.* (2005) propose a Bayesian co-estimation of alignment and phylogeny.

Finally, Groussin *et al.* (2015) highlighted that an erroneous tree negatively impact ASR accuracy and that the use of species tree-aware gene trees reconstructed with models of duplication, transfer and loss events (see Section 3) strongly increase ASR accuracy. This calls for an additional integration of duplication, transfer and loss in alignment and phylogeny algorithms.

3 Gene content

Several approaches exist for the reconstruction of ancestral gene content onto a species tree. A first one is the analysis of phylogenetic profiles, i.e. a matrix of binary characters coded as presence/absence of genes at the leaves of a given phylogeny of species (see Figure 2). A slightly more sophisticated version of phylogenetic profiles considers counts of homologous genes as discrete characters to allow the reconstruction of ancestral copy numbers for the genes under study. But before reconstructing ancestral gene content, it is important to first consider the evolutionary events that may affect the history of genes. For instance, in all organisms, genes can be duplicated or lost. In addition, many organisms have the ability to integrate genes from distantly related donors via lateral gene transfer (LGT). LGT is believed to be frequent in Bacteria and Archaea, and is beginning to be recognised as important in unicellular eukaryotes. It is still considered to be much rarer in multicellular eukaryotes, although several cases have been described. Hence, the application of methods that ignore LGT as a process for the evolution of gene repertoires should be restricted to very particular cases. Dollo parsimony (Farris, 1977), which allows loss of characters, but forbids gains after an initial origination of the gene has been very popular for genome reconstruction in the absence of LGT. But more realistic approaches attribute asymmetrical costs for gain and loss to account for the possibility of gene transfer with a relatively high cost. The equivalent likelihood methods use a birth-and-death model in which probabilities of gene loss and gain are different, and can be estimated from the dataset (see e.g. (Szöllősi and Daubin, 2012) for review). When the phylogenetic profile represents the number of copies for each gene, and not only their presence/absence, it is possible to estimate the branch-wise rates of duplication, transfer and loss along a phylogeny of species (Csurös, 2010). There is considerable flexibility in the definition of the birth-and-death model for gene evolution, the most simple models being linear, i.e with rates of gain and loss that are independent of gene family size.

However, the analysis of phylogenetic profiles imperfectly renders the evolution of gene families along the phylogeny of species (see Figure 2). Numerous events of duplication, transfer and loss remain invisible to the examination of gene presence/absence or counts. This is the case in particular for gene replacements by LGT. The possibility of reconstructing gene phylogenies based on sequence alignments reveal such hidden evolutionary events. The mapping

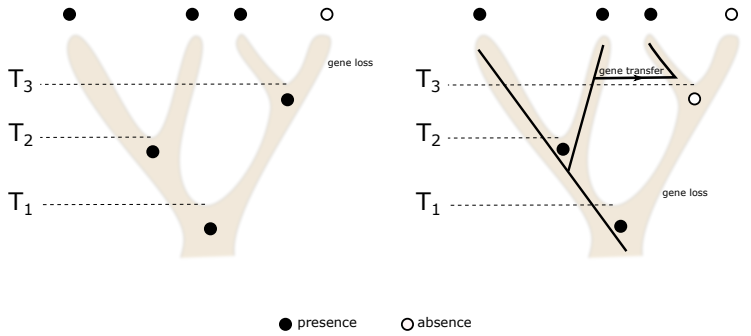

Figure 2: **Using phylogenetic profiles or gene phylogenies may reconstruct different ancestral gene contents.** The solid lines represent the evolution of the gene within the species tree. Horizontal dotted lines represent the dates of species divergence, restricting LGTs between species that co-exist in time.

of a gene family tree onto a species tree, invoking events of duplication, transfer and loss is called reconciliation (Maddison, 1997). A number of algorithms have been proposed, which are more or less complete in the set of events that are modelled (Doyon *et al.*, 2011). An efficient solution to the problem of reconciliation considering duplication and loss is the LCA (for Last Common Ancestor) algorithm (Górecki and Tiuryn, 2006), which can run in linear time with the number of nodes in the gene tree. The modelling of LGT is more complex. LGT reconciliation has been proposed for the problem of gene replacement specifically. The problem here is to define a minimal number of editing operations to transform the gene tree in the species tree. The Subtree Pruning and Regrafting (SPR) approach reproduces the effect of transfers on phylogeny and has hence been the subject of active research (Nakhleh *et al.*, 2005; Beiko and Hamilton, 2006; Than and Nakhleh, 2008). However, the problem is computationally difficult and only applies to gene families without multi-copies. More complete models, that account for duplication, transfer and loss (DTL) and hence apply to the modelling of the evolution of multicopy gene families have been developed, both in a parsimony and probabilistic framework (Doyon *et al.*, 2011; Tofigh, 2009; David and Alm, 2011; Szöllősi *et al.*, 2012, 2013; Sjöstrand *et al.*, 2014). The problem is still hard, but can become tractable through the addition of biologically relevant restrictions. For instance, an important constraint to include into DTL reconciliation algorithms is the time consistency of gene transfers: LGT can only occur among lineages that are contemporary in the history of species (see Figure 2). A promising approach for handling such time constraints is to fully specify the sequence of speciation events in the species tree in which the gene tree is reconciled (Tofigh, 2009), and to allow genes to evolve in lineages that are not explicitly represented in the tree representing the relationships among extent species (Szöllősi *et al.*, 2013). A full reconciliation

between a species tree and all gene trees yields the best possible reconstruction for the gene repertoire at nodes of the species tree.

4 Gene order

Molecular ancestral character reconstruction methods actually began with the organisation of genes along chromosomes. Dobzhansky and Sturtevant (1938) constructed the phylogeny and ancestral chromosome conformations of 17 drosophila species, based on the observation of inversions. Already at that time the computational complexity arising from a model of evolution of gene arrangements subjected to inversions was visible (Sturtevant and Tan, 1937): after only a few events, scenarios are difficult to reconstruct even under a parsimony principle. This partly explains that ancestral gene order reconstruction methods seem underage: models are simple but computationally costly, hard to apply to a complex reality and to integrate with other kind of evolutionary signal.

Nevertheless the organisation of genes along chromosomes contains valuable information on adaptation and modes of evolution of organisms, which has often been overlooked. And the methodology slowly comes of age.

Dobzhansky-Sturtevant-like methods, modelling chromosomes as permutations subject to inversions, often generalized into double cuts-and-joins to capture more possible events (Yancopoulos *et al.*, 2005), have been used to reconstruct gene orders of some mammalian (Alekseyev and Pevzner, 2009) or angiosperm (Sankoff *et al.*, 2009) ancestors with a parsimony principle. Such techniques are limited to few (typically less than 10) closely related genomes, all with equal gene content, with the notable exception of some including whole genome duplications when ohnologous genes are still present in two copies (Zheng and Sankoff, 2013) or when ohnologous segments can be detected (Gavranović and Tannier, 2010). A probabilistic model of evolution of permutations by inversions has been implemented in Badger (Larget *et al.*, 2005), and applied to reconstruct animal mitochondria or *Yersinia pestis* ancestral gene orders (Darling *et al.*, 2008). These cannot benefit from the computational facilities of dynamic programming because of the large state space, and use Monte Carlo techniques for space explorations.

A way to bypass the algorithmic complexity and to scale up to hundreds of genomes with unequal gene content is to model the evolution of independent local characters instead of the evolution of a whole genome, just as it is the case for nucleotide sequences and substitutions. This method was also —and is still— applied independently from bioinformatics with cytogenetics data (Svartman *et al.*, 2006). Gene orders may be seen as sequences of *adjacencies*, which are the links between two consecutive genes (Gallut *et al.*, 2000). Adjacencies can be summed up by a binary character, either two genes are consecutive, or not. The immediate advantage of this view is that it can benefit from the standard methodological arsenal of ancestral character reconstruction on binary characters (see Section 1). Yet there are several drawbacks to such an hypothesis of independent evolution of adjacencies. One is that the information connecting

the adjacencies, used to estimate inversion distances, is lost. Another is that genomes considered as sets of independently evolving adjacencies are not anymore constrained to be linear arrangement of genes (yet the single cut-or-join model of genome evolution has this linearity constraint (Feijão and Meidanis, 2011; Miklós *et al.*, 2014)). Linearization techniques, often related to Travelling Salesman approaches, have to be applied to sets of reconstructed ancestral adjacencies (Maňuch *et al.*, 2012) in order to present bona fide gene orders.

Parsimony approaches (Sankoff or Dollo) on adjacencies have been used, together with linearization procedures to reconstruct ancestors of mammalian (Ma *et al.*, 2006), yeasts (Chauve *et al.*, 2010), monocotyledons (Sankoff *et al.*, 2009), or bacterial (Wang *et al.*, 2006) clades.

Evolution of adjacencies is so simple that it paves the way to integrating gene order and gene content, via gene phylogenies, in a single framework (Sankoff and El-Mabrouk, 2000). Thus, it is now possible to model the evolution of adjacencies along reconciled phylogenies (Ma *et al.*, 2008; Bérard *et al.*, 2012; Louis *et al.*, 2013; Patterson *et al.*, 2013).

Prospective methods have attempted the reconstruction of more ancient animal proto-karyotypes: amniotes (Kohn *et al.*, 2006; Nakatani *et al.*, 2007), bony fishes (Jaillon *et al.*, 2004; Woods *et al.*, 2005; Catchen *et al.*, 2008), vertebrates (Naruse *et al.*, 2004; Kohn *et al.*, 2006; Nakatani *et al.*, 2007), chordates (Putnam *et al.*, 2008), or even eumetazoa (Putnam *et al.*, 2007). However, the accuracy of these ad-hoc methodologies has not been studied as thoroughly as for more generic methods, which probably also miss a satisfactory validation process due to the lack of simulators with a diversified enough model of whole genome evolution.

The complete reconstruction of ancestral genomes

Imagine now a theoretical pipeline that should reconstruct the full sequences of genomes from the past. First give its gene content (Section 3), then order the genes (Section 4), align genic and intergenic regions and reconstruct ancestral sequences (Section 2).

Whereas such integration has been attempted and applied to a small fragment of eutherian genomes (Blanchette *et al.*, 2008), or the chromosome of a medieval bacteria (Rajaraman *et al.*, 2013), the problem is still very challenging for several reasons.

One is that there are specific problems to whole genome reconstruction, like borders between genic and intergenic which can be fuzzy: sometimes homologous border sequences are genic in one organism, and intergenic in another, due to the variation of the start and stop positions of genes. Overlapping alignments, possibly with different phylogenetic histories, have to be handled.

Another is that each step highly depends on the others: the construction of phylogenetic trees depends on sequence alignments, while sequence alignment (especially when modeling indels) depends on evolutionary history along a tree (Section 2); gene order can be informative for gene content (Lafond *et al.*, 2013);

every step depends on the accuracy of phylogenetic trees, which in turn are informed by sequence and gene content evolution (Szöllősi *et al.*, 2012), and could also benefit from information about the evolution of gene order. The sequencial pipeline adds up the errors of each step without the possibility of any backward correction, while real integrative models are missing. Even small error rates can lead to many errors at the genomic scale.

The validity of all the methods described above is a big issue. True ancestral molecules are unknown and the accuracy of computational estimations can be approached via simulations (but no simulator realistically handles all possible events), via extremely rare cases where an ancient sequence is known (Rajaraman *et al.*, 2013), via expectations about ancestral genomic features (e.g., stability of gene content, or linearity of ancestral chromosomes (Boussau *et al.*, 2013)) or via the viability of resurrections (Groussin *et al.*, 2015).

References

Alekseyev, M. and Pevzner, P. (2009). Breakpoint graphs and ancestral genome reconstruction. *Genome Res*, **19**, 943–957.

Beaulieu, J. M., O'Meara, B., and Donoghue, M. (2013). Identifying hidden rate changes in the evolution of a binary morphological character: the evolution of plant habit in campanulid angiosperms. *Systematic Biology*, **62**(5), 725–737.

Beiko, R. G. and Hamilton, N. (2006). Phylogenetic identification of lateral genetic transfer events. *BMC Evol Biol*, **6**, 15.

Bérard, S., Gallien, C., Boussau, B., Szöllősi, G. J., Daubin, V., and Tannier, E. (2012). Evolution of gene neighborhoods within reconciled phylogenies. *Bioinformatics*, **28**(18), i382–i388.

Blanchette, M., Diallo, A. B., Green, E. D., Miller, W., and Haussler, D. (2008). Computational reconstruction of ancestral dna sequences. *Methods Mol Biol*, **422**, 171–184.

Blomberg, S. P., Jr., T. G., and Ives, A. R. (2003). Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution*, **57**, 717–745.

Boussau, B., Szöllosi, G. J., Duret, L., Gouy, M., Tannier, E., and Daubin, V. (2013). Genome-scale coestimation of species and gene trees. *Genome Res*, **23**(2), 323–330.

Catchen, J., Conery, J., and Postlethwait, J. (2008). Inferring ancestral gene order. In J. Keith, editor, *Bioinformatics, Volume I: Data, analysis, and Evolution*, volume 452, pages 365–383. Humana Press, Springer.

Chauve, C., Gavranovic, H., Ouangraoua, A., and Tannier, E. (2010). Yeast ancestral genome reconstructions: the possibilities of computational methods ii. *J Comput Biol*, **17**(9), 1097–1112.

Csurös, M. (2010). Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics*, **26**(15), 1910–1912.

Csűrös, M. (2014). How to infer ancestral genome features by parsimony: dynamic programming over an evolutionary tree. In *Models and Algorithms for Genome Evolution*. Springer.

Darling, A. E., Miklós, I., and Ragan, M. A. (2008). Dynamics of genome rearrangement in bacterial populations. *PLoS Genet*, **4**(7), e1000128.

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

Diallo, A. B., Makarenkov, V., and Blanchette, M. (2007). Exact and heuristic algorithms for the indel maximum likelihood problem. *J Comput Biol*, **14**(4), 446–461.

Diallo, A. B., Makarenkov, V., and Blanchette, M. (2010). Ancestors 1.0: a web server for ancestral sequence reconstruction. *Bioinformatics*, **26**(1), 130–131.

Dobzhansky, T. and Sturtevant, A. H. (1938). Inversions in the chromosomes of drosophila pseudoobscura. *Genetics*, **23**(1), 28–64.

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.

Eastman, J. M., Alfaro, M. E., Joyce, P., Hipp, A. L., and Harmon, L. J. (2011). A novel comparative method for identifying shifts in the rate of character evolution on trees. *Evolution*, **65**, 3578–3589.

Farris, J. S. (1977). Phylogenetic analysis under dollo's law. *Systematic Zoology*, **26**, 77–88.

Feijão, P. and Meidanis, J. (2011). Scj: a breakpoint-like distance that simplifies several rearrangement problems. *IEEE/ACM Trans Comput Biol Bioinform*, **8**(5), 1318–1329.

Felsenstein, J. (1985). Phylogenies and the comparative method. *The American Naturalist*, **125**, 1–15.

Felsenstein, J. (2004). *Inferring phylogenies*. Sunderland (MA): Sinauer Associates.

FitzJohn, R. G. (2012). Diversitree: Comparative phylogenetic analyses of diversification in r. *Methods in Ecology and Evolution*, **3**(6), 1084–1092.

Gallut, C., Barriel, V., and Vignes, R. (2000). Gene order and phylogenetic information. In *Comparative Genomics*. Springer.

Gavranović, H. and Tannier, E. (2010). Guided genome halving: provably optimal solutions provide good insights into the preduplication ancestral genome of saccharomyces cerevisiae. *Pac Symp Biocomput*, pages 21–30.

Groussin, M., Boussau, B., and Gouy, M. (2013). A branch-heterogeneous model of protein evolution for efficient inference of ancestral sequences. *Systematic Biology*, **62**, 523–538.

Groussin, M., Hobbs, J. K., Szöllősi, G. J., Gribaldo, S., Arcus, V. L., and Gouy, M. (2015). Toward more accurate ancestral protein genotype–phenotype reconstructions with the use of species tree-aware gene trees. *Molecular biology and evolution*, **32**(1), 13–22.

Guéguen, L., Gaillard, S., Boussau, B., Gouy, M., Groussin, M., Rochette, N. C., Bigot, T., Fournier, D., Pouyet, F., Cahais, V., *et al.* (2013). Bio++: efficient extensible libraries and tools for computational molecular evolution. *Molecular biology and evolution*, **30**, 1745–1750.

Górecki, P. and Tiuryn, J. (2006). Dls-trees: a model of evolutionary scenarios. *Theor Comput Sci*, **359**, 378–399.

Harmon, L. J., Losos, J. B., Davies, J., Gillespie, R. G., Gittleman, J. L., Jennings, W. B., Kozak, K. H., McPeek, M. A., Moreno-Roark, F., Near, T. J., Purvis, A., Ricklefs, R. E., Schluter, D., II, J. A. S., Seehausen, O., Sidlauskas, B. L., Torres-Carvajal, O., Weir, J. T., and Mooers, A. O. (2010). Early bursts of body size and shape evolution are rare in comparative data. *Evolution*, **64**, 2385–2396.

Huelsenbeck, J. P., Larget, B., Miller, R. E., and Ronquist, F. (2002). Potential applications and pitfalls of bayesian inference of phylogeny. *Syst Biol*, **51**(5), 673–688.

Jaillon, O., Aury, J.-M., Brunet, F., Petit, J.-L., n, N. S.-T., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., Nicaud, S., Jaffe, D., Fisher, S., Lutfalla, G., Dossat, C., Segurens, B., Dasilva, C., Salanoubat, M., Levy, M., Boudet, N., Castellano, S., Anthouard, V., Jubin, C., Castelli, V., Katinka, M., Vacherie, B., Biémont, C., Skalli, Z., Cattolico, L., Poulain, J., de Berardinis, V., Cruaud, C., Duprat, S., Brottier, P., e Coutanceau, J.-P., Gouzy, J., Parra, G., Lardier, G., Chapple, C., McKernan, K. J., McEwan, P., Bosak, S., Kellis, M., Volff, J.-N., Guigó, R., Zody, M. C., irov, J. M., Lindblad-Toh, K., Birren, B., Nusbaum, C., Kahn, D., Robinson-Rechavi, M., et, V. L., Schachter, V., Quétier, F., Saurin, W., Scarpelli, C., Wincker, P., der, E. S. L., Weissenbach, J., and Crollius, H. R. (2004). Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature*, **431**, 946–957.

Kohn, M., Högel, J., Vogel, W., Minich, P., Kehrer-Sawatzki, H., Graves, J., and Hameister, H. (2006). Reconstruction of a 450My-old ancestral vertebrate protokaryotype. *Trends Genet*, **22**, 203–210.

Lafond, M., Semeria, M., Swenson, K. M., Tannier, E., and El-Mabrouk, N. (2013). Gene tree correction guided by orthology. *BMC Bioinformatics*, **14 Suppl 15**, S5.

Larget, B., Kadane, J. B., and Simon, D. L. (2005). A bayesian approach to the estimation of ancestral genome arrangements. *Mol Phylogenet Evol*, **36**(2), 214–223.

Lewis, P. O. (2001). A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology*, **50**(6), 913–925.

Liberles, D. A. (2007). *Ancestral sequence reconstruction*. Oxford University Press USA:.

Louis, A., Muffato, M., and Roest Crollius, H. (2013). Genomicus: five genome browsers for comparative genomics in eukaryota. *Nucleic Acids Res*, **41**(Database issue), D700–D705.

Löytynoja, A. and Goldman, N. (2008). Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science*, **320**(5883), 1632–1635.

Lunter, G., Miklós, I., Drummond, A., Jensen, J. L., and Hein, J. (2005). Bayesian coestimation of phylogeny and sequence alignment. *BMC Bioinformatics*, **6**, 83.

Ma, J., Zhang, L., Suh, B., Rany, B., Burhans, R., Kent, W., Blanchette, M., Haussler, D., and Miller, W. (2006). Reconstructing contiguous regions of an ancestral genome. *Genome Res*, **16**, 1557–1565.

Ma, J., Ratan, A., Raney, B., Suh, B., Zhang, L., Miller, W., and Haussler, D. (2008). Dupcar: Reconstructing contiguous ancestral regions with duplications. *J Comput Biol*, **15**, 1007–1027.

Maddison, W. P. (1997). Gene trees in species trees. *Syst Biol*, **46**(3), 523–536.

Maddison, W. P., Midford, P. E., and Otto, S. P. (2007). Estimating a binary character's effect on speciation and extinction. *Systematic Biology*, **56**, 701–710.

Martins, E. P. (1999). Estimation of ancestral states of continuous characters: a computer simulation study. *Systematic Biology*, **48**, 642–650.

Martins, E. P. and Hansen, T. F. (1997). Phylogenies and the comparative method: A general approach to incorporating phylogenetic information into the analysis of interspecific data. *American Naturalist*, **149**, 646–667.

Maňuch, J., Patterson, M., Wittler, R., Chauve, C., and Tannier, E. (2012). Linearization of ancestral multichromosomal genomes. *BMC Bioinformatics*, **13 Suppl 19**, S11.

Miklós, I., Kiss, S. Z., and Tannier, E. (2014). Counting and sampling SCJ small parsimony solutions. *Theor. Comput. Sci.*, **552**, 83–98.

Nakatani, Y., Takeda, H., and Morishita, S. (2007). Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res*, **17**, 1254–1265.

Nakhleh, L., Ruths, D., and Wang, L.-S. (2005). *RIATA-HGT: a fast and accurate heuristic for reconstructing horizontal gene transfer*. Springer.

Naruse, K., Tanaka, M., Mita, K., Shima, A., Postlethwait, J., and Mitani, H. (2004). A medaka gene map: The trace of ancestral vertebrate proto-chromosomes revealed by comparative gene mapping. *Genome Res*, **14**, 820–828.

Pagel, M. (1994). Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proc. R. Soc. London B*, **255**, 37–45.

Pagel, M., Meade, A., and Barker, D. (2004). Bayesian estimation of ancestral character states on phylogenies. *Syst Biol*, **53**(5), 673–684.

Paradis, E., Claude, J., and Strimmer, K. (2004). Ape: analyses of phylogenetics and evolution in r language. *Bioinformatics*, **20**, 289–290.

Patterson, M., Szöllősi, G., Daubin, V., and Tannier, E. (2013). Lateral gene transfer, rearrangement, reconciliation. *BMC Bioinformatics*, **14 Suppl 15**, S4.

Pauling, L. and Zuckerkandl, E. (1963). Chemical paleogenetics, molecular restoration studies of extinct forms of life. *Acta chemica Scandinavica*.

Pupko, T., Pe'er, I., Hasegawa, M., Graur, D., and Friedman, N. (2002). A branch-and-bound algorithm for the inference of ancestral amino-acid sequences when the replacement rate varies among sites: Application to the evolution of five gene families. *Bioinformatics*, **18**(8), 1116–1123.

Putnam, N., Butts, T., Ferrier, D., Furlong, R., Hellsten, U., Kawashima, T., Robinson-Rechavi, M., Shoguchi, E., Terry, A., Yu, J., Benito-Gutiérrez, E., Dubchak, I., Garcia-Fernàndez, J., Gibson-Brown, J., Grigoriev, I., AC, A. H., de Jong, P., Jurka, J., Kapitonov, V., Kohara, Y., Kuroki, Y., Lindquist, R., Lucas, S., Osoegawa, K., Pennacchio, L., Salamov, A., Satou, Y., Sauka-Spengler, T., Schmutz, J., Shin-I, T., Toyoda, A., Bronner-Fraser, M., Fujiyama, A., Holland, L., Holland, P., Satoh, N., and D.S.Rokhsar (2008). The amphioxus genome and the evolution of the chordate karyotype. *Nature*, **453**, 1064–1071.

Putnam, N. H., Srivastava, M., Hellsten, U., Dirks, B., Chapman, J., Salamov, A., Terry, A., Shapiro, H., Lindquist, E., Kapitonov, V. V., Jurka, J., Genikhovich, G., Grigoriev, I. V., Lucas, S. M., Steele, R. E., Finnerty, J. R.,

Technau, U., Martindale, M. Q., and Rokhsar, D. S. (2007). Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science*, **317**, 86–94.

Rajaraman, A., Tannier, E., and Chauve, C. (2013). Fpsac: fast phylogenetic scaffolding of ancient contigs. *Bioinformatics*, **29**(23), 2987–2994.

Sankoff, D. (1975). Minimal mutation trees of sequences. *SIAM Journal on Applied Mathematics*, **28**(1), 35–42.

Sankoff, D. and El-Mabrouk, N. (2000). *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map alignment and the Evolution of Gene Families*, volume 1 of *Computational Biology Series*, chapter Duplication, rearrangement and reconciliation, pages 537–550. Kluwer Academic Publishers.

Sankoff, D., Zheng, C., Wall, P. K., dePamphilis, C., Leebens-Mack, J., and Albert, V. A. (2009). Towards improved reconstruction of ancestral gene order in angiosperm phylogeny. *J Comput Biol*, **16**(10), 1353–1367.

Schluter, D., Price, T., Mooers, A. O., and Ludwig, D. (1997). Likelihood of ancestor states in adaptive radiation. *Evolution*, **51**, 1699–1711.

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.

Sturtevant, A. and Tan, C. (1937). The comparative genetics of drosophila pseudoobscura and d. melanogaster. *Journal of genetics*, **34**(3), 415–432.

Suchard, M. A. and Redelings, B. D. (2006). Bali-phy: simultaneous bayesian inference of alignment and phylogeny. *Bioinformatics*, **22**(16), 2047–2048.

Svartman, M., Stone, G., and Stanyon, R. (2006). The ancestral eutherian karyotype is present in xenarthra. *PLoS Genet*, **2**, e109.

Szöllősi, G. J. and Daubin, V. (2012). Modeling gene family evolution and reconciling phylogenetic discord. In M. Anisimova, editor, *Evolutionary Genomics*, volume 856 of *Methods in Molecular Biology*, pages 29–51. Humana Press.

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.

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

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). Mega5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular biology and evolution*, **28**(10), 2731–2739.

Than, C. and Nakhleh, L. (2008). Spr-based tree reconciliation: Non-binary trees and multiple solutions. In *APBC; Kyoto 2008*.

Tofigh, A. (2009). Using trees to capture reticulate evolution. *PhD Thesis, KTH School of Computer Science and Communication.*

Wang, Y., Li, W., Zhang, T., Ding, C., Lu, Z., Long, N., Rose, J. P., Wang, B.-C., and Lin, D. (2006). Reconstruction of ancient genome and gene order from complete microbial genome sequences. *J Theor Biol*, **239**(4), 494–498.

Woods, I., Wilson, C., Friedlander, B., Chang, P., Reyes, D., Nix, R., Kelly, P., Chu, F., Postlethwait, J., and Talbot, W. (2005). The zebrafish gene map defines ancestral vertebrates chromosomes. *Genome Res*, **15**, 1307–1314.

Xia, X. (2013). Dambe5: a comprehensive software package for data analysis in molecular biology and evolution. *Molecular biology and evolution*, **30**(7), 1720–1728.

Yancopoulos, S., Attie, O., and Friedberg, R. (2005). Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics*, **21**(16), 3340–3346.

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

Yang, Z. (2007). Paml 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, **24**(8), 1586–1591.

Yang, Z., Kumar, S., and Nei, M. (1995). A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics*, **141**(4), 1641–1650.

Zhang, J. and Nei, M. (1997). Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. *Journal of molecular evolution*, **44**(1), S139–S146.

Zheng, C. and Sankoff, D. (2013). Practical aliquoting of flowering plant genomes. *BMC Bioinformatics*, **14 Suppl 15**, S8.

Selected further reading

(Felsenstein, 2004)
(Liberles, 2007)
(Maddison, 1997)

(Yang, 2006)
(Yang *et al.*, 1995)
(Sankoff, 1975)