How to infer ancestral genome features by parsimony: dynamic programming over an evolutionary tree

Miklós Csűrös

Abstract

We review mathematical and algorithmic problems of reconstructing evolutionary features at ancestors in a known phylogeny. In particular, we revisit a generic framework for the problem that was introduced by Sankoff and Rousseau ["Locating the vertices of a Steiner tree in an arbitrary metric space," *Mathematical Programming*, **9**:240–246, 1975].

1 Introduction

If extant organisms descended from a common ancestor through modifications [13], then their genomes carry clues about the extinct ancestors' genomes. This simple observation can lead to impressive insights when the descendants are sufficiently diverse. For example, by Blanchette et al.'s estimate [4], more than 95% of an early mammalian genome can be inferred soundly from whole-genome sequences.

Linus Pauling proposed the reconstruction of ancestral molecular sequences as early as 1962 (as recounted in [41]). Pauling presented the idea by the example of 70 homologous sites in four human hemoglobin peptide chains ($\alpha, \beta, \gamma, \delta$) and three other related sequences available at the time. The alignment and the ancestral inference were done manually, using only amino acid identities. It took another decade to work out general computational procedures to do alignment and reconstruction. Dynamic programming, the key algorithmic technique for the task, was introduced into molecular biology by Needleman and Wunsch [40] with cursory mathematical exposition. Subsequent work in the early 70s, including notable foundational contributions

Department of Computer Science and Operations Research, University of Montréal; e-mail: csuros@iro.umontreal.ca.

from David Sankoff, rigorously established the utility of the dynamic programming approach in problems related to sequence alignment, phylogeny and RNA secondary structure [44].

Phylogenetic reconstruction methods matured concomitantly with sequence alignment. Edwards and Cavalli-Sforza [17] proposed the idea of "minimal evolution" or *parsimony* [7]: the phylogeny should imply the least evolutionary change leading to the features observed in extant organisms. The principle, recast into probabilistic terms, leads to likelihood methods in phylogenetic inference [22]. Alternative ways of quantifying "evolutionary change" give rise to a number of parsimony varieties [23]. Efficient algorithms have been developed for many special cases of parsimony [7, 19, 24, 20, 49, 28, 35]. Sankoff and Rousseau [47] proposed an elegant method for parsimony inference that is general enough to apply to many specific variants and has been adapted often in contemporary approaches to ancestral reconstruction.

Here, I aim to revisit the power of the Sankoff-Rousseau algorithm and explore some modern applications.

2 Ancestral reconstruction by parsimony

We are interested in the problem of using homologous traits in extant organisms to reconstruct the states of the corresponding phylogenetic character at their ancestors. The corresponding mathematical problem, called here *parsimony labeling*, is introduced in §2.1. The discussed mathematical abstraction searches for an assignment of states to the nodes of a known evolutionary tree which minimizes a penalty imposed on state changes between parents and children. The penalization represents the "surprise value" associated with an assumed state change in the evolution of the studied phylogenetic character; searching for the least surprising evolutionary history is intended to ensure the reconstruction's biological plausibility. The most popular mathematical varieties of parsimony, involving different types of characters and penalties, are reviewed in §2.2. Sankoff and Rousseau's generic algorithm is presented in §2.3, along with its applications in many specific parsimony variants.

2.1 Parsimony labeling

Consider a given *phylogeny* $\Psi = \big(\mathcal{L}, \mathcal{V}, \mathcal{E}\big)$ over the terminal taxa $\mathcal{L}$, which is a tree with node set $\mathcal{V}$, leaves $\mathcal{L} \subseteq \mathcal{V}$, and edge set $\mathcal{E}$. The tree is rooted at a designated *root node* $\rho \in \mathcal{V}$: for every node $u \in \mathcal{V}$, exactly one path leads from ρ to u. A node $u \in \mathcal{V}$ and all its descendants form the *subtree* rooted at u, denoted by Ψ_u.

Every node $u \in \mathcal{V}$ is associated with a *label* $\xi[u] \in \mathcal{F}$ over some feature *alphabet* $\mathcal{F}$. The labels $\xi[x]$ represent the states of a homologous character at different nodes of the phylogeny. Labels are observed at the terminal nodes, but not at the other nodes, which represent hypothetical ancestors; see Figure 1.

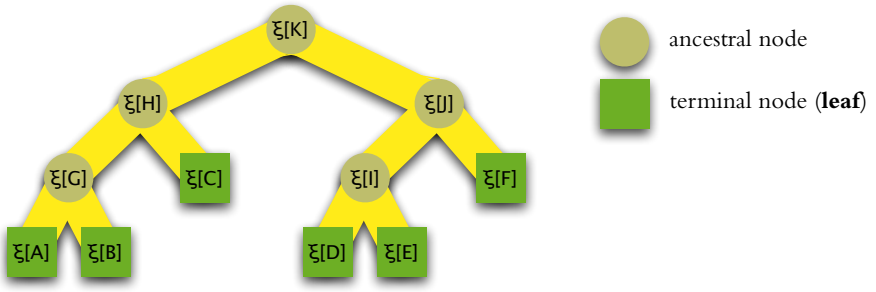

Fig. 1 Ancestral inference. Node labels $\xi[u]$ are observed at the leaves and need to be inferred at ancestral nodes.

We state the problem of ancestral reconstruction in an optimization setting, where the label space is equipped with a *cost function* $d\colon \mathcal{F}\times\mathcal{F} \mapsto [0,\infty]$. Suppose we are given a fixed leaf labeling $\Phi\colon \mathcal{L} \to \mathcal{F}$. The goal is to extend it to a joint labeling $\xi\colon \mathcal{V} \mapsto \mathcal{F}$ with minimum total cost between parent-edge labels on the edges. In systematics, the labels are states of a phylogenetic character at the nodes. The cost function reflects the evolutionary processes at play. Formally, we are interested in the following optimization problem.

General parsimony labeling problem

Consider a phylogeny $\Psi = \big(\mathcal{L}, \mathcal{V}, \mathcal{E}\big)$*, and a label space* $\mathcal{F}$ *equipped with a cost function* $d\colon \mathcal{F} \times \mathcal{F} \mapsto [0,\infty]$*. Find a joint labeling* $\xi\colon \mathcal{V} \mapsto \mathcal{F}$ *that extends a given leaf labeling* $\Phi\colon \mathcal{L} \to \mathcal{F}$ *and minimizes the total change*

$$f^* = \min_{\xi \supset \Phi} f(\xi) = \min_{\xi \supset \Phi} \sum_{uv \in \mathcal{E}} d\big(\xi[u], \xi[v]\big). \tag{1}$$

Parsimony labeling belongs to a large class of optimization problems related to Steiner trees [29]. In a Steiner-tree problem, the pair $(\mathcal{F}, d)$ forms a metric space, and the labels describe the placement of tree nodes in that space. Leaves have a fixed placement and need to be connected by a minimal tree through additional inner nodes, or so-called Steiner vertices. Classically, the placement is considered in k-dimensional Euclidean space with $\mathcal{F} = \mathbb{R}^k$,

and d is the ordinary Euclidean distance. General parsimony labeling gives the optimal placement of Steiner vertices for a fixed topology.

The algorithmic difficulty of parsimony labeling depends primarily on the assumed cost function d. Finding the most parsimonious tree is NP-hard under all traditional parsimony variants [14, 16, 15], but computing the score of a phylogeny is not always difficult.

2.2 A quick tour of parsimony variants

The minimum total change of Equation (1) measures the economy of the assumed phylogeny, or its *parsimony score* $\min_\xi f(\xi)$. Systematicists have been routinely constructing hypothetical phylogenies minimizing the parsimony score over some chosen phylogenetic characters [23]. Provided that the cost function *truly* reflects the economy of the implied evolutionary histories, parsimony is the phylogenetic equivalent of Occam's razor. Common parsimony variants use fairly simple abstractions about evolutionary processes. For instance, Dollo parsimony (§2.2.1) and Fitch parsimony (§2.2.1), use cost functions that penalize every state change the same way. In the following tour, we briefly discuss classic parsimony variants for directed evolution (§2.2.1), numerical characters (§2.2.2) and molecular sequences (§2.2.3), along with some historical notes.

2.2.1 Directed evolution

The earliest formalizations of parsimony [7, 34] framed phylogenetic inference for situations where evolutionary changes have a known (or assumed) directionality. In *Dollo-* and *Camin-Sokal* parsimony, the directionality constraints yield simple solutions to ancestral labeling.

Camin and Sokal [7] examine phylogenetic characters with some fixed ordering across possible states. The ordering is ensured by cost asymmetry. If $x \prec y$, then $0 \le d(x,y) < \infty$ and $d(y,x) = \infty$. The cost function is additive over successive states: for any three successive labels $x \prec y \prec z$, $d(x,z) = d(x,y) + d(y,z)$. In [7], this type of scoring is introduced for morphological characters — nine characters such as foot anatomy encoded by small integers — to establish a phylogeny for ten horse fossils. Ancestral labeling is straightforward: set $\xi[u]$ to the minimum label seen at leaves in u's subtree.

Dollo's law [34] applies to rare evolutionary gains (e.g., complex morphological structures) that subsequent lineages may lose. The principle translates into a binary labeling problem where only one $0 \to 1$ transition is allowed in the entire phylogeny, and the task is to minimize the number of $1 \to 0$ transitions. As first explained by Farris [20], the optimal labeling is directly

determined by the principle. The lowest common ancestor w of leaves u with label $\xi[u] = 1$ is labeled as $\xi[w] = 1$. It is either the root node, or the point of the single gain $0 \to 1$ in the entire phylogeny. Outside w's subtree, all nodes are labeled with 0. Within Ψ_w, every ancestral node v is labeled by the maximum of the labels under it: if all its descendants are labeled with 0, so is $\xi[v] = 0$; otherwise, $\xi[v] = 1$.

2.2.2 Numerical labels

Cavalli-Sforza and Edward [9] pose the problem of inferring a phylogeny of populations from gene allele frequencies. Allele frequencies and many other interesting characters are best captured by numerical labels as real-valued continuous variables. When $\mathcal{F} = \mathbb{R}$, the absolute and the squared distances are common choices for parsimony labeling.

Wagner parsimony [32] applies to a numerical label space (discrete or continuous) and uses the distance $d(x, y) = |x - y|$. Farris [19] describes a linear-time algorithm for labeling a binary tree, which was proven to be correct and adaptable to non-binary trees by Sankoff and Rousseau [47].

Squared parsimony [35] employs the squared distance $d(x, y) = (x - y)^2$ over a continuous label space and gives a linear-time algorithm to compute the ancestral labeling in linear time. Squared parsimony has an attractive probabilistic interpretation involving a Brownian motion model [35]. Suppose that changes along each edge follow a Brownian motion, so child labels have a normal distribution centered around the parent's label with variance proportional to the edge length. If edge lengths are the same, then the ancestral labeling that maximizes the likelihood also minimizes the parsimony score with the cost function $d'(x, y) = -\log\Big(\frac{1}{\sigma\sqrt{2\pi}} \exp\big(-\frac{(x-y)^2}{2\sigma^2}\big)\Big)$, where σ is the common standard deviation of the child labels. After stripping away the constant term and common scaling factors, only the remaining squared distance $d(x, y) = (x - y)^2$ determines the labeling's optimality.

2.2.3 Molecular sequences

For the purposes of phylogenetic inference and ancestral reconstruction from molecular sequences, parsimony scoring has to capture the peculiarities of homologous sequence evolution. One possibility is to consider a fixed multiple alignment and use parsimony with residues at aligned sites. Fitch parsimony [24], which applies to a finite label space such as the four-letter DNA alphabet, simply minimizes the number of different labels on tree edges. Much more challenging is parsimony with edit distance, first addressed by David Sankoff [46, 43], when the label space encompasses all possible sequences, and the scoring includes insertions and deletions. Parsimony labeling with

edit distance is NP-hard, since it is equivalent to multiple alignment with a fixed guide tree, which is known to be NP-hard for any alignment scoring [18].

Originally, Kluge and Farris [32] employed Wagner parsimony to six binary characters derived from distinguishing body features in twelve frog families. When labels are binary ($\mathcal{F} = \{0, 1\}$) in Wagner parsimony, there are only two possible distances: $d(x, y) = 0$ if $x = y$ and $d(x, y) = 1$ if not. *Fitch parsimony* [24] generalizes the same scoring principle to any discrete alphabet:

$$d(x, y) = \{x \neq y\} = \begin{cases} 0 & \text{if } x = y; \\ 1 & \text{if } x \neq y \end{cases}$$

Computing the optimal labeling under this distance takes linear time in the tree size [24, 28]. Fitch parsimony, in contrast to Wagner parsimony, accommodates non-numerical phylogenetic characters, including amino acids and nucleotides. In an early application, Fitch and Farris [25] apply the scoring method to nucleotides in homologous positions in a multiple alignment in order to infer the most parsimonious RNA coding sequences from protein data.

2.3 The Sankoff-Rousseau technique

The crucial insight of [47] is that general parsimony labeling has a recursive structure that calls for solutions by dynamic programming.

For every node $u \in \mathcal{V}$, define the *subtree cost* $f_u \colon \mathcal{F} \mapsto [0, \infty)$ as the minimum total change implied by u's label within its subtree:

$$f_u(x) = \min \sum_{vw \in \mathcal{E}_u; \xi[u]=x} d(\xi[v], \xi[w]),$$

where $\mathcal{E}_u$ are the edges within the subtree Ψ_u. The minimum is taken across all ancestral node labelings with the same assignment x at u. By the principle of optimality, the following recursions hold:

$$f_u(x) = \begin{cases} 0 & \text{if } x = \Phi[u]; \\ \infty & \text{if } x \neq \Phi[u]; \end{cases} \qquad \{u \in \mathcal{L}\} \tag{2a}$$

$$f_u(x) = \sum_{uv \in \mathcal{E}} \min_{y \in \mathcal{F}} \big(d(x, y) + f_v(y)\big) \qquad \{u \in \mathcal{V} \setminus \mathcal{L}\} \tag{2b}$$

In particular, the parsimony score is retrieved by considering the minimum total penalty at different root labelings:

$$f^* = \min_{x \in \mathcal{F}} f_\rho(x).$$

The recursions of Eq. (2) suggest a general outline for computing the best ancestral labeling together with its score. First, compute all f_u in a postfix traversal, visiting parents after child nodes, as shown by the procedure Ancestral below. Second, retrieve the best labeling in in a prefix traversal that realizes the computed minimum f^* by standard backtracking, as shown by the procedure Labeling here.

```
   Ancestral(u)                                   // (computes f_u for a node u)
A1 if u ∈ L then f_u(x) ← {x = Φ[u]}?0 : ∞         // (leaf)
A2 else
A3    for uv ∈ E do                               // (for all children v)
A4       f_v ← Ancestral(v)
A5       compute h_uv(x) ← min_{y∈F} (d(x, y) + f_v(y))
A6    set f_u(x) ← Σ_{uv∈E} h_uv(x)
A7 return f_u
```

(Line A1 uses $\{x = \Phi[u]\}?0 : \infty$ to denote the function for a forced labeling at a terminal node u, from Eq. (2a).) Line A5 computes the *stem cost* $h_{uv}(x)$ for a child of u, which is the minimal change along the edge and within the subtree of v, given that u is labeled with x. Line A6 sums the stem cost functions to construct f_u.

```
   Labeling(v)                                    // (computes the best labeling)
L1 if v is the root then ξ[v] = arg min_x f_v(x)
L2 else
L3    u ← parent of v; x ← ξ[u]
L4    ξ[v] ← arg min_{y∈F} (d(x, y) + f_v(y))
L5 for vw ∈ E do Labeling(w)                      // (for all children of w, if any)
```

For a finite label space, the minimum in Line A5 is found by examining all possible values. At the same time, the best labeling y for each x is saved for backtracking in Lines L1 and L4. For an infinite space or very large label space, it is not immediately clear how the minimization can be done in practice. Luckily, it is possible to track f and h by other means than tabulation in many important cases.

2.3.1 Few possible labels

The general Sankoff-Rousseau outline immediately yields an algorithm when labels have only a few possible values. Figure 2 shows an example with absence-presence labels ($\mathcal{F} = \{0, 1\}$). The distance metric is defined by the gain penalty g and loss penalty 1, which apply on every edge uv to labelings $\xi[u] = 0, \xi[v] = 1$, and $\xi[u] = 1, \xi[v] = 0$, respectively.

A computer implementation can tabulate $f_u(x)$ for all nodes u and possible values x. With a table-based implementation, the running time grows linearly with tree size, but quadratically with possible labels. Note that the tables accommodate arbitrary cost functions, not only true distance metrics.

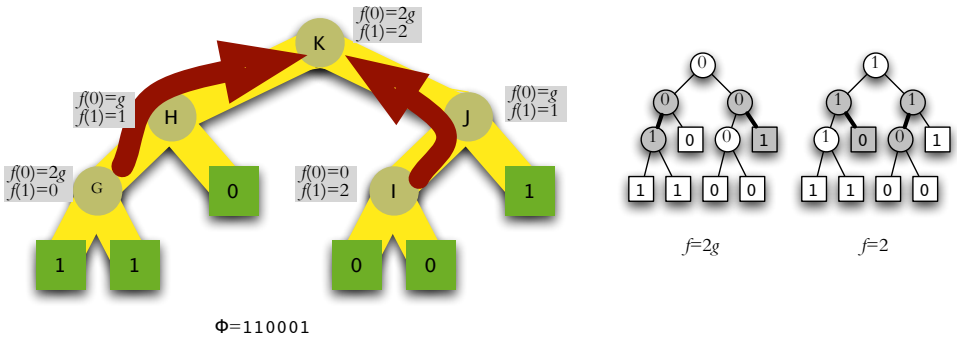

Fig. 2 Inference of ancestral labels by parsimony. This example uses a binary character (presence-absence) with profile Φ, a loss penalty 1 and gain penalty $g \geq 0$; i.e., the distance function over the feature space $\mathcal{F} = \{0, 1\}$ is defined as $d(x, x) = 0$ and $d(0, 1) = g$, $d(1, 0) = 1$. Depending on the gain penalty g, the optimal solution may imply two losses (score $f = 2$) or two gains ($f = 2g$). The dynamic programming proceeds from the leaves towards the root, computing the score $f_u(x)$ of the optimal reconstruction within each subtree Ψ_u, in the order indicated by the arrows.

Theorem 1. *For a finite label space of of size $r = |\mathcal{F}|$, and an evolutionary tree with m edges, algorithms* Ancestral *and* Labeling *compute an optimal labeling in $O(mr^2)$ time and $O(mr)$ space.*

Proof. A table stores $f_u(x)$ for $m + 1$ nodes and r possible labels. Line A1 is executed once for every terminal node. In Line A5, $\min_{y \in \mathcal{F}}$ is found in a loop over possible child labelings y in $O(r)$ time. Lines A5 and A6 are executed for each edge and label x, in $O(mr^2)$ total time.

In order to use with backtracking, the minimal y found in Line A5 is saved for every edge uv and label x, using $m \times r$ entries in a table. Line L4 takes $O(1)$ to retrieve the optimal labels on each edge, and Algorithm Labeling completes in $O(m)$ time.

2.3.2 Molecular sequences

Sankoff and Rousseau [47] generalize Sankoff's previously developed method for the ancestral reconstruction of RNA sequences [46, 43], which is by far the most challenging case of ancestral parsimony. The recursions for multiple alignment and ancestral labeling can be combined to find an optimal solution, but the computation takes an exponentially long time in the number of nodes [43].

2.3.3 Squared parsimony

Squared parsimony was first proposed [9, 42] as an appropriate cost function $d(x, y) = (x - y)^2$ for inference from allele frequencies $\xi \in [0, 1]$ observed in populations. Maddison [35] solved the parsimony labeling problem by directly employing the method of Sankoff and Rousseau [47]. Theorem 2 below restates the key result: subtree and stem costs are quadratic functions, for which the parameters can be computed recursively.

Theorem 2. *In the general parsimony problem with $\mathcal{F} = \mathbb{R}$ and $\dot{(x, y)} = (y - x)^2$, the subtree weight functions are quadratic. In other words, one can write the subtree cost function at each non-leaf node u as*

$$f_u(x) = \alpha_u(x - \mu_u)^2 + \phi_u, \tag{3}$$

with some parameters $\alpha_u, \phi_u, \mu_u \in \mathbb{R}$. The parameters α, μ satisfy the following recursions

$$\alpha_u = \begin{cases} \textit{undefined} & \textit{if } u \textit{ is a leaf;} \\ \sum_{uv \in \mathcal{E}} \beta_v & \textit{otherwise;} \end{cases} \tag{4a}$$

$$\mu_u = \begin{cases} \xi[u] & \textit{if } u \textit{ is a leaf;} \\ \frac{\sum_{uv \in \mathcal{E}} \beta_v \mu_v}{\sum_{uv \in \mathcal{E}} \beta_v} & \textit{otherwise;} \end{cases} \tag{4b}$$

where β_v is defined for all $v \in \mathcal{V}$ by

$$\beta_v = \begin{cases} 1 & \textit{if } v \textit{ is a leaf;} \\ \frac{\alpha_v}{\alpha_v + 1} & \textit{otherwise.} \end{cases} \tag{4c}$$

By Theorem 2, Ancestral can proceed by storing α and μ at every node.

Theorem 3. *For squared parsimony, algorithms* Ancestral *and* Labeling *compute an optimal labeling in $O(m)$ time and $O(m)$ space.*

Proof. Lines A5–A6 compute the subtree costs by the recursions of (4) in $O(m)$ total time. By Eq. (3), Labeling needs to set the root label in Line L1 as

$$\xi[\rho] = \arg\min_x \alpha_\rho(x - \mu_\rho)^2 + \phi_\rho = \mu_\rho \tag{5}$$

since $\alpha_u > 0$ at all nodes. The stem weight function $h_{uv}(x)$ for an inner node v is

$$h_{uv}(x) = \min_y \big((x - y)^2 + f_v(y)\big) = \frac{\alpha_v}{\alpha_v + 1}(x - \mu_v)^2 + \phi_v,$$

with minimum at

$$y^* = \arg\min_y \big((x - y)^2 + f_v(y)\big) = \frac{x + \alpha_v \mu_v}{\alpha_v + 1}. \tag{6}$$

Therefore, Line L4 sets the child labeling as

$$\xi[v] = \frac{\xi[u] + \alpha_v \mu_v}{\alpha_v + 1}.$$

LABELING thus spends $O(1)$ time on each node and finishes in $O(m)$ time.

Squared parsimony can be generalized to k-dimensional features $\mathcal{F} = \mathbb{R}^k$. The optimal parsimony labeling with the squared Euclidean distance between labels $d(x, y) = \sum_{i=1}^{k} = (x_i - y_i)^2$ can be computed component-wise, since plugging it into (1) gives

$$\min_{\xi \supset \Phi} f(\xi) = \sum_{i=1}^{k} \underbrace{\min_{\xi_i \supset \Phi_i} \sum_{uv \in \mathcal{E}} \big(\xi_i[u] - \xi_i[v]\big)^2}_{\text{best labeling in coordinate } i}. \tag{7}$$

It suffices to apply the recursions of (4) in each coordinate separately. In an application where each label represents a distribution over k elements, it is not immediately clear that the coordinate-wise reconstruction yields valid ancestral distributions, i.e., that $\sum_{i=1}^{k} \xi_i[u] = 1$ is assured everywhere. Fortunately, the optimal labeling formulas of (5) and (6) automatically ensure that the separate reconstructions always add up to proper distributions [12].

2.3.4 Wagner (linear) parsimony

Wagner parsimony considers the linear cost function $d(x, y) = |x - y|$ over a numerical label space like $\mathcal{F} = \mathbb{R}$. A consequence of the linear cost is that the subtree costs have a simple recursive structure [19, 47, 49], and the Sankoff-Rousseau method can be carried out by tracking simple intervals. An asymmetric linear cost function, examined by [12], leads to a similarly recursive structure. Namely, *asymmetric Wagner parsimony* uses a linear cost function of the form

$$d(x, y) = \begin{cases} \lambda(y - x) & \{y \geq x\} \\ \nu(x - y) & \{x > y\}, \end{cases}$$

with gain and loss penalties λ, ν. In asymmetric Wagner parsimony, the subtree cost functions are continuous, convex and piecewise linear. Consequently, they can be manipulated symbolically as vectors of slopes and breakpoints, see Figure 3.

Theorem 4. *For every non-leaf node $u \in \mathcal{V} \setminus \mathcal{L}$, there exist $k \geq 1$, $\alpha_0 < \alpha_1 < \cdots < \alpha_k$ (slopes), $x_1 < x_2 \cdots < x_k$ (breakpoints), and $\phi_0, \ldots, \phi_k \in \mathbb{R}$ that define f_u in the following manner.*

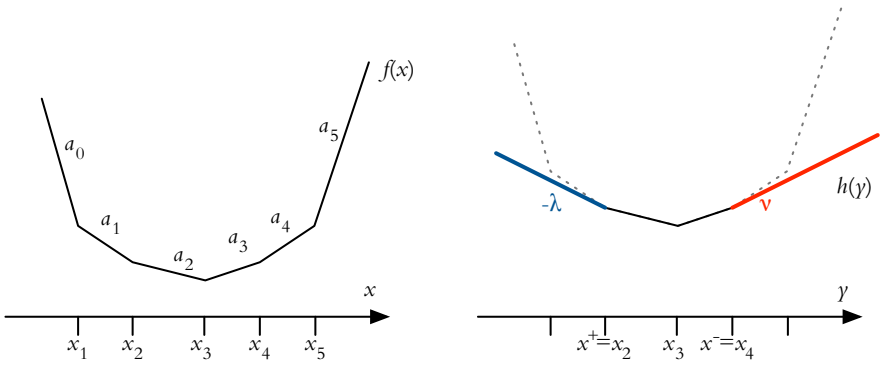

Fig. 3 Illustration of Theorem 4 about the shape of the cost functions. **Left**: for asymmetric Wagner parsimony, the subtree cost function f is always piecewise linear with slopes $a_0, \ldots, a_k$ ($k = 5$ here). **Right**: the stem cost function $h(y) = \min_x(d(y,x) + f(x))$ is obtained by "shaving off" the steep extremities of f and replacing them with slopes of $(-\lambda)$ and ν, respectively.

$$f_u(x) = \begin{cases} \phi_0 + \alpha_0 x & \text{if } x \le x_1; \\ \phi_1 + \alpha_1(x - x_1) & \text{if } x_1 < x \le x_2; \\ \ldots & \\ \phi_{k-1} + \alpha_{k-1}(x - x_{k-1}) & \text{if } x_{k-1} < x \le x_k; \\ \phi_k + \alpha_k(x - x_k) & \text{if } x_k < x, \end{cases} \tag{8}$$

where $\phi_1 = \phi_0 + \alpha_0 x_1$ *and* $\phi_{i+1} = \phi_i + \alpha_i(x_{i+1} - x_i)$ *for all* $0 < i < k$. *Moreover, if* u *has* d *children, then* $a_0 = -d\lambda$ *and* $a_k = d\nu$.

Figure 3 illustrates the proof of Theorem 4 from [12]. The Sankoff-Rousseau algorithm can be implemented by storing the breakpoints for the slopes between $(-\lambda)$ and ν. In classical Wagner parsimony, $\lambda = \nu = 1$, and the two stored breakpoints define an interval of equivalently optimal labelings at every node, used in the original algorithm of [19].

Theorem 5. *Algorithms* Ancestral *and* Labeling *find the optimal labeling by asymmetric Wagner parsimony in* $O(nh \log d_{\max})$ *time for a phylogeny of height* h *with* n *nodes and maximum node degree* $d_{\max}$. *For integer-valued penalties* λ, ν *with* $B = \lambda + \nu$, *the algorithms label the tree in* $O(nB)$ *time.*

Proof. For general real-valued penalties λ and ν, the breakpoints and the slopes defining the subtree cost functions are stored by ordered lists. The symbolic summation of stem costs in Line A6 involves merging ordered lists, leading to a running-time bound of $O(nh \log d_{\max})$ [12].

For integer-valued penalties, it is enough to store the $B = (\lambda + \nu)$ possible breakpoints associated with slopes between $-\lambda$ and ν that can play a role in an optimal labeling. By storing the breakpoints for f_u in an array of length B,

Line A6 sums the stem costs in $O(nB)$ time across all nodes, and an optimal labeling is found in $O(1)$ time per node.

Wagner parsimony easily generalizes to the k-dimensional labels $\mathcal{F} = \mathbb{R}^k$ with the Manhattan distance $d(x, y) = \sum_{i=1}^{k} |x_i - y_i|$. Just like with squared parsimony, the optimal labeling can be decomposed by coordinates.

2.3.5 Multiple reconstructions

The optimal ancestral labeling for Camin-Sokal and Dollo is always unique, due to the directionality of the parsimony cost function. Squared parsimony, as well, has a unique optimal solution by Theorem 2. Otherwise, the most parsimonious labeling of Equation (1) is not necessarily unique. For example, the two ancestral labelings depicted in Figure 2 are both minimal when gains and losses are penalized equally ($g = 1$): they entail either two loss events or two gain events. Even if multiple solutions are possible, the ancestral labeling algorithm can resolve the ties at will between ancestral labels that yield the same minimum subtree costs either at the root (Line L1 in LABELING) or on the edges (Line L4), following the normal order of the algorithm.

Theorem 4 shows an important property of Wagner parsimony, first recognized by Farris [19]. Namely, the minimum subtree score for an ancestral node is attained either at a single label, or within a closed interval where the cost function has slope 0. The ambiguity of optimal ancestral labelings can be characterized by computing the set of possible labels (a closed interval, possibly containing a single point) at each ancestral node in linear time [49]. When multiple ancestral labels are equally optimal, one of two heuristics are traditionally chosen to resolve ties. The first one, proposed in [19] and named ACCTRAN (for "accelerated transformation") by [49], chooses a reconstruction where label changes are placed closer to the root. Mathematically, ACCTRAN is the unique optimal reconstruction in which all subtree scores are minimized; i.e., $\sum_{vw \in T_u} d\big(\xi[v], \xi[w]\big)$ takes its minimum value in each subtree T_u among all reconstructions ξ with minimum parsimony score [39]. The other heuristic, called DELTRAN ("delayed transformation"), defers changes away from the root [49]. ACCTRAN is believed to give biologically more plausible reconstructions by minimizing parallel gains in different lineages, although closer scrutiny shows that ACCTRAN and DELTRAN do not always behave as expected [1].

3 Applications and extensions

Many authors built on the Sankoff-Rousseau technique to develop related efficient algorithms. We sample a few algorithmic extensions in §3.1. A few

biological applications in §3.2 illustrate the pertinence of parsimony-based reconstructions in contemporary studies of genome evolution.

3.1 Algorithmic extensions

Tree-additive cost functions

A somewhat inconvenient property of the generic Sankoff-Rousseau algorithm is that it entails a quadratic dependence on the alphabet size (Theorem 1). Its simplicity, however, lends itself to efficient parallel implementation [31]. The quadratic factor can be avoided for certain cost functions with an additive structure, e.g., if $d(x, y)$ is an ultrametric distance [10].

Parsimony on a phylogenetic network

Kannan and Wheeler [30] address the generalization of the Sankoff-Rousseau algorithm to a phylogenetic network. Specifically, they consider networks with some reticulate nodes having two incoming and one outgoing edges. The parsimony score sums the costs occurred along all the edges, including those connecting reticulate vertices. The optimal labeling can be computed by enumerating all possible joint labelings at reticulate nodes, in $O(mr^{k+2})$ time for m edges and k reticulate nodes over a r-letter alphabet.

Gain and loss edges

After constructing an optimal ancestral labeling, it is trivial to collect the lineages with similar state transitions such as the edges on which some feature was lost (transition $1 \to 0$). In a binary labeling problem for absence-presence data, it is practicable to track such sets of edges by the single traversal of Ancestor [37]. Specifically, define the sets $L_{uv}(x)$ and $G_{uv}(x)$ for all edges uv as sets of loss and gain edges, respectively, affecting v's subtree when $\xi[u] = x$ in an optimal labeling ξ. For a terminal edge uv, $L_{uv}(0) = G_{uv}(1) = \emptyset$, $L_{uv}(1) = \{uv\}$ if $\xi[v] = 0$, and $G_{uv}(0) = \{uv\}$ if $\xi[v] = 1$. For all non-leaf nodes u, let $L_{u*}(x) = \bigcup_{uv \in \mathcal{E}} L_{uv}(x)$ and $G_{u*}(x) = \bigcup_{uv \in \mathcal{E}} G_{uv}(x)$. The following recursions hold, depending on the event on the edge uv

$$\Big\langle L_{uv}(1), G_{uv}(1) \Big\rangle = \begin{cases} \big\langle \{uv\} \cup L_{v*}(0), G_{v*}(0) \big\rangle & \textit{(loss on } uv\textit{)} \\ \big\langle L_{v*}(1), G_{v*}(1) \big\rangle & \textit{(no loss on } uv\textit{)} \end{cases}$$

$$\Big\langle L_{uv}(0), G_{uv}(0) \Big\rangle = \begin{cases} \big\langle L_{v*}(1), \{uv\} \cup G_{v*}(1) \big\rangle & \textit{(gain on } uv\textit{)} \\ \big\langle L_{v*}(0), G_{v*}(0) \big\rangle & \textit{(no gain on } uv\textit{)} \end{cases}$$

The stem cost for the edge h_{uv} counts the edges within the L_{uv} and G_{uv} sets. Using asymmetric gain-loss costs as in §2.3.4, $h_{uv}(x) = \lambda\big|G_{uv}(x)\big|+\nu\big|L_{uv}(x)\big|$. The choices between loss vs. no-loss and gain vs. no-gain are made to minimize the associated costs.

3.2 Applications

Parsimony's simple assumptions are appreciated even in contemporary studies of complex genome features. A case in point is Wagner parsimony that was recently used to study genome size evolution [6] and short sequence length polymorphisms [51]. Genome size and tandem repeat copy numbers as well as the other examples to follow are common in that they are difficult to address in probabilistic models, either for technical reasons or simply because the relevant evolutionary processes are still not understood well enough.

Phylogenetic footprinting

In phylogenetic footprinting, conserved short sequence motifs are discovered in a sample of unaligned sequences associated with the terminal nodes of a known phylogeny [5]. It is assumed that the sequences contain common regulatory signals with some level of conservation. The corresponding ancestral reconstruction problem (Substring Parsimony) labels the nodes with k-letter sequences $\mathcal{F} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}^k$ for a fixed k. Leaves must be labeled by a word that appears somewhere in the input sequence, and edge costs measure distance between parent and child labels. The algorithm of Sankoff and Rousseau can be readily modified to initialize a set of possible labels at the leaves with 0 cost. [5] propose practical algorithmic improvements for small k, but Substring Parsimony is NP-hard in general [18].

Gene family evolution

A number of software packages implement asymmetric Wagner parsimony for the inference of ancestral gene family sizes [11, 8, 2]. Weighted parsimony has been used to study the frequence of gene loss and gain, and to estimate ancestral genome sizes [37, 36]. Pioneering large-scale studies on the phylogenetic distribution of gene families revealed a surprisingly gene-rich last universal common ancestor by ancestral reconstruction [33]. Genes tend to be lost more often than gained in the course of evolution, and asymmetric gain-loss penalties can capture known discrepancies between the intensities of the two processes. Selecting the relative costs between loss and gain en-

tails additional considerations such as the plausibility of the reconstructed ancestral genome size [33].

Han and Hahn [27] use k-dimensional linear parsimony to study gene duplications and losses concomitantly with transpositions between chromosomes. Homolog genes for a family are encoded in a k-dimensional integer vector by the copy numbers on k chromosomes. Homolog families over ten complete Drosophila reveal patterns of sex-specific gene movement to and from the X chromosome, overly active functional categories, and other idiosyncrasies that characterize fly genome evolution.

Splice sites and intron length

Gelfman and coauthors [26] resort to squared parsimony to infer ancestral intron length from homologous introns in 17 vertebrate genomes. The study links length constraints to splicing signal strength (the stronger the signal, the longer the intron can be) and shows that the correlations specifically pertain to vertebrates. In a related study, Schwartz et al. [48] infer ancestral splice site signals. Starting with aligned 5' splice sites and branch sites in different introns, the nucleotide frequencies in each motif position are compiled into probabilistic sequence motifs that label the leaf genomes. Ancestral nucleotide frequencies are reconstructed separately in each motif position by squared parsimony. The reconstruction implies that sites were degenerate in the earliest eukaryotes, hinting at the prevalence of alternative splicing in deep eukaryotic ancestors.

Gene order

Ancestral reconstruction of gene order was pioneered as a parsimony problem by Sankoff et al. [45]. In this context, nodes are labeled by gene orders, which are the permutations defined by the physical order of genes (or other genetic markers) along the chromosomes. Appropriate cost functions for such permutations can be defined using an edit distance penalizing various rearrangement events, or by counting conserved segments and breakpoints. Other contributions in this volume address the mathematically rich field of gene order comparisons in more details (in particular, [38] discusses distances between gene orders): here, we mention only some recent connections to classic parsimony variants.

Wang and Tang [50] explored an encoding of adjacencies that are suitable to submit as phylogenetic characters to parsimony labeling. The reconstructions need to be corrected to yield valid gene orders (the inferred ancestral adjacencies may imply a circular chromosome) — the correction is shown to be NP-hard. Feijão and Meidanis [21] give an edit distance function for which parsimony labeling is feasible in polynomial time. They show in particular

that that by simply using adjacencies as binary phylogenetic characters, and applying Fitch parsimony (with some small algorithmic adjustments), one recovers a most parsimonious gene order history under the so-called single-cut-and-join distance. Bérard et al. [3] also use parsimony labeling for inferring ancestral adjacencies; the novelty of their approach is that it incorporates gene duplications and losses by carrying out Sankoff's dynamic programming method simultaneously along two gene phylogenies reconciled with a known species tree.

4 Conclusion

Parsimony is not as popular as it once was, mostly because today's large data sets contain enough statistical signal to employ sophisticated probabilistic models. Nevertheless, parsimony remains a viable choice in many contemporary applications, where parameter-rich stochastic models are not available or are impractical. Indeed, different scoring policies are available for the evolutionary analysis of "unconventional" genome features, including sequence motifs, copy numbers, genomic lengths and distributions. Surprisingly diverse policies are amenable to exact optimization by dynamic programming following the same basic recipe. Along with Felsenstein's seminal work on likelihood calculations [22], Sankoff's parsimony minimization [47] established fundamental algorithmic techniques for modeling evolutionary changes, which proved to be versatile enough to tackle new computational biology challenges for the past forty years.

References

1. Agnarsson, I., Miller, J.A.: Is ACCTRAN better than DELTRAN? Cladistics **24**, 1–7 (2008)
2. Ames, R.M., Money, D., Ghatge, V.P., Whelan, S., Lovell, S.C.: Determining the evolutionary history of gene families. Bioinformatics **28**(1), 48–55 (2012). DOI 10.1093/bioinformatics/btr592
3. Bérard, S., Gallien, C., Boussau, B., Szöllősi, G.J., Daubin, V., Tannier, E.: Evolution of gene neighborhoods within reconciled phylogenies. Bioinformatics **28**, i382–i388 (2012)
4. Blanchette, M., Green, E.D., Miller, W., Haussler, D.: Reconstructing large regions of an ancestral mammalian genome in silico. Genome Res. **12**, 2412–2423 (2004)
5. Blanchette, M., Schwikowski, B., Tompa, M.: Algorithms for phylogenetic footprinting. J. Comput. Biol. **9**(2), 211–223 (2002)
6. Caetano-Anollés, G.: Evolution of genome size in the grasses. Crop Sci. **45**, 1809–1816 (2005)
7. Camin, J.H., Sokal, R.R.: A method for deducing branching sequences in phylogeny. Evolution **19**, 311–326 (1965)
8. Capra, J.A., Williams, A.G., Pollard, K.S.: Proteinhistorian: tools for the comparative analysis of eukaryote protein origin. PLoS Comput. Biol. **8**(6), e1002,567 (2011)

9. Cavalli-Sforza, L.L., Edwards, A.W.H.: Phylogenetic analysis models and estimation procedures. Am. J. Hum. Genet. **19**(3), 233–267 (1967)
10. Clemente, J.C., Ikeo, K., Valiente, G., Gojobori, T.: Optimized ancestral state reconstruction using sankoff parsimony. BMC Bioinformatics **10**, 51 (2009)
11. Csűrös, M.: Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinformatics **26**(15), 1910–1912 (2010)
12. Csűrös, M.: Ancestral reconstruction by asymmetric Wagner parsimony over continuous characters and squared parsimony over distributions. Springer Lecture Notes in Bioinformatics **5267**, 72–86 (2008). Proc. Sixth RECOMB Comparative Genomics Satellite Workshop
13. Darwin, C.: On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. John Murray, London (1859)
14. Day, W.H.E.: Computationally difficult parsimony problems in phylogenetic systematics. J. Theor. Biol. **103**, 429–438 (1983)
15. Day, W.H.E., Johnson, D.S., Sankoff, D.: The computational complexity of inferring rooted phylogenies by parsimony. Math. Biosci. **81**, 33–42 (1986)
16. Day, W.H.E., Sankoff, D.: Computational complexity of inferring phylogenies by compatibility. Syst. Zool. **35**, 224–229 (1986)
17. Edwards, A.W.F., Cavalli-Sforza, L.L.: Reconstructing evolutionary trees. In: V.H. Heywood, J. McNeill (eds.) Phenetic and phylogenetic classification, 6, pp. 67–76. Systematics Association, London (1963)
18. Elias, I.: Settling the intractability of multiple alignment. J. Comput. Biol. **13**(7), 1323–1339 (2006)
19. Farris, J.S.: Methods for computing Wagner trees. Syst. Zool. **19**(1), 83–92 (1970)
20. Farris, J.S.: Phylogenetic analysis under Dollo's law. Syst. Zool. **26**(1), 77–88 (1977)
21. Feijão, P., ao Meidanis, J.: SCJ: A breakpoint-like distance that simplifies several rearrangement problems. IEEE/ACM Trans. Comput. Biol. Bioinf. **8**(5), 1318–1329 (2011)
22. Felsenstein, J.: Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst. Zool. **22**(3), 240–249 (1973)
23. Felsenstein, J.: Parsomony in systematics: Biological and statistical issues. Annu. Rev. Ecol. Syst. **14**, 313–333 (1983)
24. Fitch, W.M.: Toward defining the course of evolution: minimum changes for a specific tree topology. Syst. Zool. **20**, 406–416 (1971)
25. Fitch, W.M., Farris, J.S.: Evolutionary trees with minimum nucleotide replacements from amino acid sequences. J. Mol. Evol. **3**(4), 263–278 (1974)
26. Gelfman, S., Burstein, D., Penn, O., Savchenko, A., Amit, M., Schwartz, S., Pupko, T., Ast, G.: Changes in exonintron structure during vertebrate evolution affect the splicing pattern of exons. Genome Res. **22**(1), 35–50 (2012). DOI 10.1101/gr.119834.110
27. Han, M.V., Hahn, M.W.: Inferring the history of interchromosomal gene transposition in Drosophila using n-dimensional parsimony. Genetics **190**, 813–825 (2012)
28. Hartigan, J.: Minimum mutation fits to a given tree. Biometrics **29**, 53–65 (1973)
29. Hwang, F.K., Richards, D.S.: Steiner tree problems. Networks **22**, 55–89 (1992)
30. Kannan, L., Wheeler, W.C.: Maximum parsimony on phylogenetic networks. Algorithms Mol. Biol. **7**, 9 (2012)
31. Kasap, S., Benkrid, K.: High performance phylogenetic analysis with maximum parsimony on reconfigurable hardware. IEEE Trans. VLSI Syst. **19**(5) (2011)
32. Kluge, A.R., Farris, J.S.: Quantitative phyletics and the evolution of anurans. Syst. Zool. **18**, 1–32 (1969)
33. Koonin, E.V.: Comparative genomics, minimal gene sets and the last universal common ancestor. Nat. Rev. Microbiol. **1**, 127–136 (2003)
34. Le Quesne, W.J.: The uniquely evolved character concept and its cladistic application. Syst. Zool. **23**, 513–517 (1974)
35. Maddison, W.P.: Squared-change parsimony reconstructions of ancestral states for continuous-valued characters on a phylogenetic tree. Syst. Zool. **40**(3), 304–314 (1991)

36. Makarova, K.S., Sorokin, A.V., Novichkov, P.S., Wolf, Y.I., Koonin, E.V.: Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. Biology Direct **2**, 33 (2007)
37. Mirkin, B.G., Fenner, T.I., Galperin, M.Y., Koonin, E.V.: Algorithms for computing evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. BMC Evol. Biol. **3**, 2 (2003)
38. Moret, B.M.E., Lin, Y., Tang, J.: Rearrangements in phylogenetic inference: Compare or encode? (2013). In this volume
39. Narushima, H., Misheva, N.: On characteristics of ancestral-state reconstructions under the accelerated transformation optimization. Discrete Appl. Math. **122**, 195–209 (2002)
40. Needleman, S.B., Wunsch, C.B.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. **48**, 443–453 (1970)
41. Pauling, L., Zuckerkandl, E.: Chemical paleogenetics: Molecular "restoration studies" of extinct forms of life. Acta Chemica Scandinavica **17**, S9–S16 (1963)
42. Rogers, J.S.: Deriving phylogenetic trees from allele frequencies. Syst. Zool. **52–63** (1984)
43. Sankoff, D.: Minimal mutation trees of sequences. SIAM J. Appl. Math. **28**(1) (1975)
44. Sankoff, D.: The early introduction of dynamic programming into computational biology. Bioinformatics **16**(1), 41–47 (2000)
45. Sankoff, D., Leduc, G., Antoine, N., Paquin, B., Lang, B.F., Cedergren, R.: Gene order comparisons for phylogenetic inference: Evolution of the mitochondrial genome. Proc. Natl. Acad. Sci. USA **89**, 6575–6579 (1992)
46. Sankoff, D., Morel, C., Cedergren, R.J.: Evolution of 5S RNA and the non-randomness of base replacement. Nature New Biology **245**, 232–234 (1973)
47. Sankoff, D., Rousseau, P.: Locating the vertices of a Steiner tree in arbitrary metric space. Math. Program. **9**, 240–246 (1975)
48. Schwartz, S., Silva, J., Burstein, D., Pupko, T., Eyras, E., Ast, G.: Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. Genome Res. **18**, 88–103 (2008)
49. Swofford, D.L., Maddison, W.P.: Reconstructing ancestral states using Wagner parsimony. Math. Biosci. **87**, 199–229 (1987)
50. Tang, J., Wang, L.S.: Improving genome rearrangement phylogeny using sequence-style parsimony. In: Proceedings of the IEEE Fifth Symposium on Bioinformatics and Bioengineering (BIBE'05), pp. 137–144 (2005)
51. Witmer, P.D., Doheny, K.F., Adams, M.K., Boehm, C.D., Dizon, J.S., Goldstein, J.L., Templeton, T.M., Wheaton, A.M., Dong, P.N., Pugh, E.W., Nussbaum, R.L., Hunter, K., Kelmenson, J.A., Rowe, L.B., Brownstein, M.J.: The development of a highly informative mouse simple sequence length polymorphism (SSLP) marker set and construction of a mouse family tree using parsimony analysis. Genome Res. **13**, 485–491 (2003)