Ancestral reconstruction by asymmetric Wagner parsimony over continuous characters and squared parsimony over distributions

Miklós Csűrös

Department of Computer Science and Operations Research
University of Montréal
C.P. 6128, succ. Centre-Ville, Montréal, Québec, H3C 3J7, Canada
csuros@iro.umontreal.ca

Abstract. Contemporary inferences about evolution occasionally involve analyzing infinitely large feature spaces, requiring specific algorithmic techniques. We consider parsimony analysis over numerical characters, where knowing the feature values at terminal taxa allows one to infer ancestral features, namely, by minimizing the total number of changes on the edges using continuous-valued distance measures. In particular, we show that ancestral reconstruction is possible in linear time for both an asymmetric linear distance measure (Wagner parsimony) over continuous-valued characters, and a quadratic distance measure over finite distributions. The former can be used to analyze gene content evolution with asymmetric gain and loss penalties, and the latter to reconstruct ancestral diversity of regulatory sequence motifs and multi-allele loci. As an example of employing asymmetric Wagner parsimony, we examine gene content evolution within Archaea.

1 Introduction

Phylogenetic studies commonly operate with molecular sequence data, where homologous characters take values over a finite space. When working with characters such as numbers of paralogs within homologous gene families, allele frequencies, sequence length polymorphisms, or DNA sequence motif distributions, the analysis of theoretically infinite feature spaces becomes necessary [1]. In such situations, one can resort to parsimony criteria to infer ancestral states, or score candidate phylogenies by minimizing the total change of the feature in question. Change is quantified by using different types of distance measures which are appropriate for the study. A popular parsimony criterion for features that can be ordered linearly is the so-called Wagner parsimony [2, 3] in which change is penalized simply by the absolute value of the numerical difference on an edge. Another criterion used sometimes is the minimization of squared distance between the numerical values [4].

Wagner parsimony has been used to infer the evolution of gene family size. Change in the family size, however, is not always equally likely in both directions,

as losses may be more frequent than gains, or vice versa. We propose a modification of the original Wagner parsimony criterion for such situations, where increases and decreases are penalized linearly, but with different penalty factors. We discuss the resulting optimization problem, and show how to compute the parsimony score, as well as the ancestral states in linear time, regardless of the actual values at the terminal taxa. We also show that squared parsimony over finite distributions can be computed efficiently, by performing the minimization in each coordinate separately, without considering the restriction to the probability simplex.

We demonstrate the utility of asymmetric Wagner parsimony by an analysis of gene content evolution in Archaea.

2 Algorithmic results

2.1 Problem statement

Consider the following general parsimony framework, introduced by Sankoff and Rousseau [5]. Let $\mathcal{T} = (\mathcal{V}, \mathcal{E})$ be a rooted tree that represents a phylogeny, with node set $\mathcal{V}$ and edge set $\mathcal{E}$. The set of tree leaves is denoted by $\mathcal{L}$. It is assumed that every non-leaf node has at least two children. Each node $u \in \mathcal{V}$ is associated with a *label* $\xi[u] \in \mathcal{X}$ where $\mathcal{X}$ is the space of possible labels. The focus of this study is the case when $\mathcal{X}$ is a numerical infinite space such as $\mathcal{X} = \mathbb{R}^d$ or $\mathcal{X} = \{0, 1, 2, \dots\}$. The label space is equipped with a *change weight* function $\Delta\colon \mathcal{X}^2 \mapsto [0, \infty)$. (Classically, Δ is a proper distance metric, but we will consider asymmetric functions, as well.) We are interested in the following problem.

General parsimony labeling problem *Given the tree T, label space $(\mathcal{X}, \Delta)$, and fixed assignments $\xi[u]$ at the leaves $u \in \mathcal{L}$, find $\xi[v]$ for all inner nodes $v \in \mathcal{V} \setminus \mathcal{L}$ that minimize the total change*

$$\sum_{uv \in \mathcal{E}} \Delta(\xi[u] \to \xi[v]).$$

The problem in this form was introduced in [5] as a Steiner tree problem [6] with a distance metric Δ. Some specific cases of the general problem have been extensively studied. The case of nonnegative integers $\mathcal{X} = \mathbb{N}$ and $\Delta(y \to x) = |y - x|$, is known as Wagner parsimony that can be solved in linear time [2, 3]. The case $\mathcal{X} = \mathbb{R}$ and $\Delta(y \to x) = (y - x)^2$ is known as squared parsimony, which also has a linear-time solution [4]

The parsimony labeling problem is encountered in phylogenetic studies when one wants to estimate the ancestral state of some feature that is represented by the labels [7, 8]. An unknown phylogeny can also be inferred by searching for the topology $\mathcal{T}$ over the leaf set $\mathcal{L}$ that minimizes the parsimony score [1].

Features in question may be (continuous-valued) allele frequencies, in which case squared-parsimony is in fact equivalent to likelihood maximization under a Brownian motion model [4, 1]. Wagner parsimony has been used to infer the

evolution of sequence length polymorphisms [9], genome size [10], and gene family size.

2.2 General solution by dynamic programming

The general parsimony problem has a solution by dynamic programming, as elucidated in the pioneering paper of Sankoff and Rousseau [5]. The key idea is to define the *subtree weight functions* $f_u \colon \mathcal{X} \mapsto [0, \infty]$ for each node $u \in \mathcal{V}$, so that $f_u(x)$ gives the minimum weight within the subtree $\mathcal{T}_u$ rooted at u when $\xi[u] = x$. For leaves, $f_u(x) = 0$ if $x = \xi[u]$; otherwise, $f_u(x) = \infty$. For an inner node u, the following recursion holds.

$$f_u(y) = \sum_{v \in \text{children}(u)} \min_{x \in \mathcal{X}} \Big(\Delta(y \to x) + f_v(x) \Big). \tag{1}$$

For every edge $uv \in \mathcal{E}$, define the *stem weight functions*

$$h_v(y) = \min_{x \in \mathcal{X}} \Big(\Delta(y \to x) + f_v(x) \Big), \tag{2}$$

so that

$$f_u(y) = \sum_{v \in \text{children}(u)} h_v(y). \tag{3}$$

The minimum total weight is then $\min_y f_{\text{root}}(y)$, and the optimal labeling can be determined by backtracking. For a finite label space, the general solution takes $O(|\mathcal{X}|^2)$ time on each edge. For an infinite space, it is not immediately clear how the minimization can be done in practice. Luckily, it is possible to compute f and h efficiently in many important cases [2, 4, 5].

2.3 Asymmetric Wagner parsimony

Often, the labels represent features that are more easily lost than gained [11, 7]. Gene content evolution, in particular, is characterized by frequent gene loss, which may be properly captured in parsimony methods by penalizing gains more than losses [12]. We define the *asymmetric Wagner parsimony* problem as that of general parsimony labeling when

$$\mathcal{X} \subseteq \mathbb{R} \quad \text{and} \quad \Delta(y \to x) = \begin{cases} \gamma(x - y) & \text{if } y < x; \\ \lambda(y - x) & \text{if } x < y, \end{cases}$$

where $\gamma, \lambda > 0$ are gain and loss penalty factors, respectively. The pivotal observation for an algorithmic solution is given by the following lemma; the claim is illustrated in Figure 1.

Lemma 1. *For every non-leaf node $u \in \mathcal{V} \setminus \mathcal{L}$, the subtree weight function is a continuous, convex, piecewise linear function. In other words, there exist $k \geq 1$,*

$\alpha_0 < \alpha_1 < \cdots < \alpha_k$, $x_1 < x_2 \cdots < x_k$, *and* $\phi_0, \ldots, \phi_k \in \mathbb{R}$ *that define* f_u *in the following manner.*

$$f_u(x) = \begin{cases} \phi_0 + \alpha_0 x & \text{if } x \le x_1; \\ \phi_1 + \alpha_1(x - x_1) & \text{if } x_1 < x \le x_2; \\ \ldots & \\ \phi_{k-1} + \alpha_{k-1}(x - x_{k-1}) & \text{if } x_{k-1} < x \le x_k; \\ \phi_k + \alpha_k(x - x_k) & \text{if } x_k < x, \end{cases} \tag{4}$$

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* $a_0 = -d\gamma$ *and* $a_k = d\lambda$.

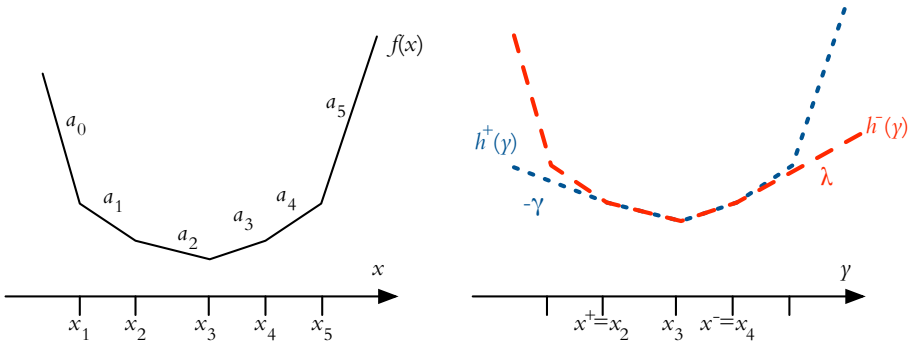

Fig. 1. Illustration of Lemma 1. **Left**: for asymmetric Wagner parsimony, the subtree weight function f is always piecewise linear with slopes $a_0, \ldots, a_k$ ($k = 5$ here). **Right**: the stem weight function h is determined by the two auxiliary functions h^+ and h^-, which are obtained by "shaving off" the steep extremities of f, and replacing them with slopes of $-\gamma$, and λ, respectively.

Proof. The proof proceeds by induction over the tree in a postorder traversal, following the recursion of (1). By the definition of Δ, if v is a leaf, then

$$h_v(x) = \begin{cases} \gamma(\xi[v] - x) & \text{if } x \le \xi[v]; \\ \lambda(x - \xi[v]) & \text{if } \xi[v] < x. \end{cases} \tag{5}$$

Base case. If all d children of u are leaves, then (3) and (5) imply that (4) holds with some $k \le d$, $\alpha_0 = -d\gamma$ and $\alpha_k = d\lambda$. For a more precise characterization, let $\mathcal{C}$ be the set of children of u, and consider the set of leaf labels $\mathcal{S} = \{\xi[v] \colon v \in \mathcal{C}\}$. Then $k = |\mathcal{S}|$, and $\{x_1, \ldots, x_k\} = \mathcal{S}$. Furthermore, for all $i = 1, \ldots, k$, $\alpha_i = t_i\lambda - (d - t_i)\gamma$ with $t_i = \sum_{v \in \mathcal{C}} \{\xi[v] \le x_i\}$, where $\{\cdot\}$ denotes the indicator for the event within the braces; i.e., t_i is the number of children that carry a label that is not larger than x_i. Finally, $\phi_0 = \gamma \sum_{v \in \mathcal{C}} \xi[v]$.

Induction step. Assume that u is an inner node at which (4) holds for every non-leaf descendant. Let v be a non-leaf child of u. By the induction hypothesis, $f_v(x)$ is a piecewise linear function as in (4) with some parameters $(\alpha_i \colon i = 0, \ldots, k)$, and $(x_i \colon i = 1, \ldots, k)$.

In order to compute $h_v(y) = \min_{x \in \mathcal{X}} \Big(\Delta(y \to x) + f_v(x) \Big)$, consider the two minimization problems over $\mathcal{X}$ split into half by y:

$$h_v^+(y) = \min_{x \in \mathcal{X}; x > y} \Big(\gamma(x - y) + f_v(x) \Big)$$
$$h_v^-(y) = \min_{x \in \mathcal{X}; x \le y} \Big(\lambda(y - x) + f_v(x) \Big).$$

Clearly, $h_v(y) = \min\Big\{ h_v^+(y), h_v^-(y) \Big\}$. Figure 1 illustrates the shapes of h^+ and h^-.

Recall that $\alpha_0 < \alpha_1 < \cdots < \alpha_k$ by the induction hypothesis. Since the constant term $(-\gamma y)$ can be ignored in the minimization for h^+, the solution is determined by the point $x^+ = x_j$ with $j = \min\{i \colon \alpha_i + \gamma \ge 0\}$. In particular,

$$h_v^+(y) = \begin{cases} \gamma \cdot \big(x^+ - y\big) + f_v\big(x^+\big) & \text{if } y < x^+; \\ f_v(y) & \text{if } y \ge x^+. \end{cases}$$

In a similar manner, let $x^- = x_j$ with $j = \min\{i \colon \alpha_i - \lambda \ge 0\}$. Then

$$h_v^-(y) = \begin{cases} f_v(y) & \text{if } y < x^-; \\ \lambda \cdot \big(y - x^-\big) + f_v\big(x^-\big) & \text{if } y \ge x^-. \end{cases}$$

Notice that the induction hypothesis implies that $x^+ \le x^-$, since $\alpha_0 + \gamma < 0$ and $\alpha_k - \lambda > 0$ hold. By the definition of x^+ and x^-, it is also true that $h_v^+(y) \le f_v(y)$ if $y < x^+$, and that $h_v^-(y) \le f_v(y)$ if $y \ge x^-$. Hence, h_v is a piecewise linear function in the form

$$h_v(y) = \begin{cases} \gamma \cdot \big(x^+ - y\big) + f_v\big(x^+\big) & \text{if } y < x^+; \\ f_v(y) & \text{if } x^+ \le y < x^-; \\ \lambda \cdot \big(y - x^-\big) + f_v\big(x^-\big) & \text{if } y \ge x^-. \end{cases} \tag{6}$$

The formula also shows that when $\xi[u] = y$, the best labeling for v is either $x = x^+$ for $y < x^+$ (i.e., net gain on edge uv), or $x = x^-$ for $y \ge x^-$ (i.e., net loss), or else $x = y$ (no change).

Equations (5) and (6) show that $h_v(y)$ is always a continuous, convex, piecewise linear function with slopes $(-\gamma)$ on the extreme left and λ on the extreme right. Consequently, $f_u(y) = \sum_{v \in \text{children}(u)} h_v(y)$ is also a continuous, convex, piecewise linear function, with slopes $(-d\gamma)$ on the left and $d\lambda$ on the right. Hence, the induction hypothesis holds for u. □

The proof provides the recipe for implementing the dynamic programming of (1). The algorithm has to work with piecewise linear functions as in (4),

parametrized by the set of slopes $(\alpha_i : i = 0, \ldots, k)$, breakpoints $(x_i : i = 1, \ldots, k)$ and shift ϕ_0. The parameters are naturally sorted as $\alpha_0 < \alpha_1 < \cdots < \alpha_k$ and $x_1 < x_2 < \cdots < x_k$, and can be thus stored as ordered arrays. The algorithm is sketched as follows.

```
W1  Dynamic programming for asymmetric Wagner
W2  initialize h_u(·) and f_u(·) as null at each node u ∈ V
W3  for all nodes u in postorder traversal
W4  if u is a leaf
W5  then set h_u(x) as in (5)
W6  else        ▷ h_v(x) is computed for all children v already
W7      compute f_u(x) = Σ_{v∈children(u)} h_u(x)
W8      if u is not the root then compute h_u(y) by (6)
W9  find the minimum of f_root(x)
W10 backtrack for the optimal labeling if necessary
```

Theorem 1. *For a tree $\mathcal{T}$ of height h and $n = |\mathcal{V}|$ nodes, asymmetric Wagner parsimony can be solved in $O\big(n \min\{h, D\} \log d_{\max}\big)$ time where D is the number of different leaf labels and $d_{\max}$ is the maximum arity.*

Proof. First, notice that the breakpoints at each f_u and h_u are exactly the set of different leaf labels in the subtree rooted at u, with at most D elements. Line W5 takes $O(1)$ time at each leaf. In Line W8, a binary search for x^+ and x^- takes $O(\log k)$ time if there are k breakpoints. In Line W7, piecewise linear functions need to be summed, which can be done by straightforward modification of well-known linear-time merging algorithms for ordered lists [13]. In order to sum the piecewise linear functions, the breakpoints must be processed in their combined order, and the intermediate slopes need to be computed. The procedure takes $O(k \log d)$ time, if the node has d children, and there are a total of k breakpoints at the children's stem weight functions. Now, $k \leq D$, and, thus, every node can be processed in $O(D \log d_{\max})$ time. The $O(nh \log d_{\max})$ bound comes from the fact that k is bounded by the number of leaves in the subtree. The total computing time for nodes that are at the same distance from the root is then $O(n \log d_{\max})$. By summing across all levels, we get $O(nh \log d_{\max})$ computing time. □

Remark. Lemma 1 and its proof show that there is an optimal solution where every non-leaf node carries a label that appears at one of the leaves. Accordingly, it is enough to keep track of $f_u(x)$ only where x takes one of the leaf label values. Adapting Sankoff's general parsimony algorithm over the discrete finite label space defined by the D label values of interest yields an $O(nD^2)$ algorithm.

2.4 Squared parsimony

In certain applications, node labels are distributions such as allele frequencies [14], or probabilistic sequence motifs [15]. Suppose, for example, that we

identified homologous regulatory sequence motifs in some genomes related by a known phylogeny. A particular instance of the motif is a DNA oligomer $s_1 s_2 \cdots s_\ell$ with a fixed length ℓ. From the set, we compile sequence motifs describing each terminal node by the labels $\xi_{is}[u]$, which give the relative frequency of each nucleotide s at motif position $i = 1, \ldots, \ell$. From the node labels, we would like to infer the compositional distribution of the motif at ancestral nodes. In a recent example, Schwartz and coworkers [15] examined the evolution of splicing signals in eukaryotes. The authors deduced that the 5' splice site and the branch site were degenerate in the earliest eukaryotes, in agreement with previous studies by Irimia and coworkers [16]. These findings are intriguing as they hint at the prevalence of alternative splicing in the earliest eukaryotes. Schwartz et al. [15] reconstructed the diversity of ancestral splicing signals by using a squared change penalty $\Delta(\mathbf{y} \to \mathbf{x}) = \sum_{i=1}^{\ell} \sum_{s=\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}} \big(x_{is} - y_{is}\big)^2$. An equivalent sum-of-squares penalty was suggested by Rogers [14] in a different context, where $i = 1, \ldots, \ell$ would stand for genetic loci and s would index possible alleles at each locus. Since the positions can be handled separately, we consider the problem of general parsimony at a given position. Specifically, we assume that the labels are distributions over a finite set $\mathcal{A} = \{1, 2, \ldots, r\}$. The change penalty is defined by

$$\Delta(\mathbf{y} \to \mathbf{x}) = \sum_{s \in \mathcal{A}} \big(x_s - y_s\big)^2.$$

The case of a binary alphabet $r = 2$ was shown to be solvable in linear time by Maddison [4]. The algorithm is stated for the general parsimony problem with $\mathcal{X} = \mathbb{R}$ and $\Delta(y \to x) = (y - x)^2$. While Maddison's algorithm is trivially extended to any dimension with $\mathcal{X} = \mathbb{R}^r$ and $\Delta(y \to x) = \sum_i (y_i - x_i)^2$, the extension to distributions with $r > 2$ is not immediately obvious. In [15], the distributions were discretized to an accuracy of 0.02, and then solved on the corresponding grid by using Sankoff's dynamic programming. Notice that there are 23426 such discretized distributions, and dynamic programming over a finite alphabet takes quadratic time in the alphabet size. Here we show that Maddison's algorithm can be carried out at each coordinate independently, as the computed solution is automatically a distribution.

Squared parsimony for a continuous character

For a discussion, we restate the result of [4].

Lemma 2. *In the general parsimony problem with $\mathcal{X} = \mathbb{R}$ and $\Delta(y \to x) = (y - x)^2$, subtree weight functions are quadratic. In other words, at each non-leaf node u, there exist $\alpha, \mu, \phi \in \mathbb{R}$ such that*

$$f_u(x) = \alpha(x - \mu)^2 + \phi. \tag{7}$$

Proof. We will use the simple arithmetic formula that

$$\sum_{i=1}^{d} \alpha_i (x - \mu_i)^2 = \alpha(x - \bar{\mu})^2 + \alpha\big(\mu^{(2)} - (\bar{\mu})^2\big) \tag{8}$$

with

$$\alpha = \sum_{i=1}^{d} \alpha_i, \quad \bar{\mu} = \frac{\sum_{i=1}^{d} \alpha_i \mu_i}{\sum_{i=1}^{d} \alpha_i}, \quad \mu^{(2)} = \frac{\sum_{i=1}^{d} \alpha_i \mu_i^2}{\sum_{i=1}^{d} \alpha_i}.$$

The proof proceeds by induction over the tree in a postorder traversal, following the recursion structure of Eq. (1).

Base case. Let u be an inner node with d children $\{v_1, \ldots, v_d\}$ that are all leaves. By (1),

$$f_u(y) = \sum_{i=1}^{d} (y - \xi[v_i])^2.$$

Hence (8) applies with $\alpha_i = 1$ and $\mu_i = \xi[v_i]$. Specifically, (7) holds with $\mu = \sum_{i=1}^{d} \xi[v_i]/d$.

Induction step. Suppose that u is an inner node with d children $\{v_1, \ldots, v_d\}$, which are all either leaves, or inner nodes for which (7) holds. Let $v = v_i$ be an arbitrary child node. If v is a leaf, then $h_v(y) = (y - \xi[v])^2$. If v is an inner node with $f_v(x) = \alpha(x - \mu)^2 + \phi$, then

$$\begin{aligned} h_v(y) &= \min_x \big((y - x)^2 + \alpha(x - \mu)^2 + \phi \big) \\ &= \min_x \left\{ (\alpha + 1) \Big(x - \frac{y + \alpha\mu}{\alpha + 1} \Big)^2 \right\} + \frac{\alpha}{\alpha + 1} (y - \mu)^2 + \phi \\ &= \frac{\alpha}{\alpha + 1} (y - \mu)^2 + \phi. \end{aligned}$$

Notice that the best labeling at v is achieved with $x = \frac{y + \alpha\mu}{\alpha + 1}$.

Consequently, the stem weight function can be written as $h_{v_i}(x) = \alpha_i (x - \mu_i)^2 + \phi_i$ for every child v_i with some $\alpha_i, \mu_i, \phi_i \in \mathbb{R}$. By (3),

$$f_u(x) = \sum_{i=1}^{d} \Big(\alpha_i (x - \mu_i)^2 + \phi_i \Big) = \alpha (y - \bar{\mu})^2 + \phi,$$

where $\phi = \alpha \big(\mu^{(2)} - (\bar{\mu})^2 \big) + \sum_{i=1}^{d} \phi_i$, and α, $\bar{\mu}$, $\mu^{(2)}$ are as in (8). Therefore, (7) holds at u. □

The proof of Lemma 2 shows how the parameters α and μ need to be computed in a postorder traversal. Namely, for every node u, the following recursions hold for the parameters $\alpha = \alpha_u$ and $\mu = \mu_u$ of (7).

$$\alpha_u = \begin{cases} \text{undefined} & \text{if } u \text{ is a leaf;} \\ \sum_{v \in \text{children}(u)} \beta_v & \text{otherwise;} \end{cases} \tag{9a}$$

$$\mu_u = \begin{cases} \xi[u] & \text{if } u \text{ is a leaf;} \\ \dfrac{\sum_{v \in \text{children}(u)} \beta_v \mu_v}{\sum_{v \in \text{children}(u)} \beta_v} & \text{otherwise;} \end{cases} \tag{9b}$$

where

$$\beta_v = \begin{cases} 1 & \text{if } v \text{ is a leaf;} \\ \frac{\alpha_v}{\alpha_v + 1} & \text{otherwise.} \end{cases} \tag{9c}$$

Squared parsimony for distributions

Suppose that the nodes are labeled with finite distributions over a set $\mathcal{A} = \{1, 2, \ldots, r\}$. Accordingly, we write $\xi_i[u]$ with $i = 1, \ldots, r$ for the i-th probability value at each node u. Node labelings are scored by the square parsimony penalty: $\Delta(\mathbf{y} \to \mathbf{x}) = \sum_{i=1}^{r}(y_i - x_i)^2$, where $\mathbf{y}$ and $\mathbf{x}$ are distributions over $\mathcal{A}$, i.e., points of the $(r-1)$-dimensional simplex in $\mathbb{R}^r$ defined by $0 \leq \xi_i[u]$ for all i, and $\sum_{i=1}^{r} \xi_i[u] = 1$. Suppose that one carries out the minimization coordinate-wise, for each i separately, without making particular adjustments to ensure that the ancestral labels also define a distribution. By Lemma 2, such an independent ancestral reconstruction finds the subtree weight functions of the form $f_{u,i}(x) = \alpha_u\big(x - \mu_{u,i}\big)^2 + \phi_{u,i}$ in each coordinate i. (Equations (9a) and (9c) show that α_u and β_u are determined by the tree topology alone, and are thus the same in each coordinate.)

Theorem 2. *The coordinate-wise independent ancestral reconstruction produces the optimal solution for distributions.*

Proof. Let $f_{u,i}(x)$ denote the subtree weight function for coordinate i at node u. Clearly, $\sum_{i=1}^{r} f_{u,i}(x_i)$ is a lower bound on the true subtree weight function $f_u(x_1, \ldots, x_r)$ for the distributions. Consequently, it is enough to show that the solution by coordinate-wise reconstruction leads to valid distributions. From Equation (9b), if u is an inner node, then $\sum_{i=1}^{r} \mu_{u,i} = \sum_{v \in \text{children}(u)} \frac{\beta_v}{\alpha_u} \sum_{i=1}^{r} \mu_{v,i}$. As $\sum_{i=1}^{r} \mu_{u,i} = 1$ holds at every leaf u, the equality holds at all nodes by induction. It is also clear that $\mu_{u,i} \geq 0$ is always true, since β_v is never negative. In particular, the optimal labelings at the root define a distribution with $\xi_i[\text{root}] = \mu_{\text{root},i}$.

In the proof of Lemma 2, we showed that if the parent of an inner node v is labeled by $\mathbf{y} = (y_1, \ldots, y_r)$, then the optimal labeling at v is $\xi_i[v] = x_i = \frac{y_i + \alpha_v \mu_{v,i}}{\alpha_v + 1}$. Now, $\sum_{i=1}^{r} x_i = \frac{\sum_i y_i + \alpha_v \sum_i \mu_{v,i}}{\alpha_v + 1} = 1$ if $\sum_i y_i = 1$ holds. Since the independent ancestral reconstructions produce a distribution at the root, the backtracking procedure produces a distribution at every inner node v. □

3 Gene content evolution in Archaea

We applied asymmetric Wagner parsimony to the analysis of gene content evolution in Archaea. We note that parsimony-based analysis has its well-known shortcomings, such as the underestimation of gene loss, and the imposition of uniformity across lineages and genes, which may be avoided with sophisticated probabilistic methods [17, 18]. Nevertheless, parsimony may give important insights by providing a conservative estimate of ancestral gene content, and by underlining some general idiosyncrasies without much procedural difficulty.

Makarova and coauthors [19] delineated homologous gene families across 41 completely sequenced and annotated archaeal genomes. They analyzed some characteristic features of archaeal genome evolution, and extrapolated the gene

composition of the last archaeal common ancestor, or LACA. The analysis relied on so-called phyletic profiles, which are binary patterns of family presence-absence, in conjunction with parsimony-based ancestral reconstruction algorithms [20]. In our analysis, we used the available information on the number of paralogs within different genomes.

3.1 Data and methods

Data was downloaded from `ftp://ftp.ncbi.nih.gov/pub/wolf/COGs/arCOG`. The data set defines 7538 families (so-called archaeal clusters of orthologous genes, or *arCOGs*) in 41 genomes. Figures 2 and 3 show the organisms and their phylogenetic relationships. The abbreviations are those used in [19] and the arCOG database: the Appendix lists the organism names and the abbreviations. The archaeal phylogeny is based on the one used by Makarova et al. (Figure 7 in [19]) for inferring gene content evolution, using additional considerations to partially resolve certain polytomies. Namely, we assume the monophyly of the Pyrococcus genus within Thermococcales [21], and the monophyly of Methanomicrobia excluding Halobacteriales [22], as depicted in Figure 3.

In order to perform the analysis, an adequate gain and loss penalization needed to be chosen. The ratio between the two penalty factors influences how much of the reconstructed history is dominated by gene loss [12]. Since the inference depends only on the ratio of the gain and loss penalties, we set $\lambda = 1$, and performed the reconstruction at different gain penalties γ. We selected a gain penalty of $\gamma = 1.6$, matching the estimate of [19] the closest. The reconstruction results in a LACA genome of 984 families and 1106 genes, which is similar in the corresponding statistics to such extant archaea as *Methanopyrus kandleri* (Metka; 1121 arCOGs with 1336 genes) and *Cenarchaeum symbiosum* (Censy; 918 arCOGs with 1296 genes).

3.2 Results

Gene content at LACA. The reconstructed set of ancient families contains 96 families inferred as present, and 107 as absent in contradiction with [19]. The two reconstructions qualitatively give a very similar picture, pointing to a LACA genome complexity comparable to the simplest free-living prokaryotes such as Mycoplasma. Table 1 shows a summary of the functional categorization for the inferred primordial gene families. Among the gene families present in LACA, 91 (9%) included more than one gene. The majority of these families (77 of 91) have closer homologs among Bacteria than among Eukaryota, which would be expected if Archaea emerged from a bacterial lineage. These multi-gene families are indicative of ancestral adaptations: notable cases include reverse gyrase (2 paralogs), hinting at a hyperthermophilic LACA, and various genes implicated in pyruvate oxidation that has a pivotal importance in archaeal metabolism [21].

Losses and gains of families. Figures 2 and 3 show further details of the ancestral reconstruction. Using asymmetric Wagner parsimony, it was possible to

Cat[a]	Description[b]	Multi[c]	Fam[d]
	Information storage and processing		
J	Translation	4	153
K	Transcription	6	59
L	Replication	7	57
	Cellular processes and signaling		
D	Cell cycle control	3	5
V	Defense mechanisms	3	19
T	Signal transduction mechanisms	2	8
M	Cell wall, membrane and envelope biogenesis	7	23
N	Cell motility	1	5
U	Intracellular trafficking and secretion	1	11
O	Posttranslational modification, protein turnover, chaperones	5	41
	Metabolism		
C	Energy production and conversion	10	77
G	Carbohydrate transport and metabolism	6	37
E	Amino acid transport and metabolism	14	101
F	Nucleotide transport and metabolism	3	46
H	Coenzyme transport and metabolism	3	70
I	Lipid transport and metabolism	2	23
P	Inorganic ion transport and metabolism	1	45
Q	Secondary metabolites biosynthesis, transport and catabolism	1	23
R,**S**	Poorly characterized or unknown	12	197
	Total	91	984

Table 1. Ancestral gene content at LACA. Columns: (a) arCOG functional category code, (b) functional category description, (c) LACA families with more than one member, (d) total number of families at LACA.

postulate expansions and reductions within gene families, in addition to the families' appearance and elimination. Numerous losses, just as in the reconstruction of [19], are associated with symbiotic lifestyles (Censy and Naneq). Our studies also agree on examples of significant losses coupled with major gains in Thermococcales (node 7) and Thermoplasmales (node 9), hinting at unusually dynamic genomes. Our reconstructions of lineage-specific changes, however, often differ numerically, as illustrated in Table 2. Namely, Wagner parsimony tends to postulate fewer genes at inner nodes, and family gains on deep branches also tend to be lower. Our reconstruction seems more conservative, and at times even more plausible. For instance, we posit major gains in Desulfurococcales and Sulfolobales (nodes 4 and 5) lineages, whereas [19] postulates an extremely large genome for their common ancestor (node 3) instead.

Patterns of diversification. Interestingly, large losses are not always associated with compact genomes: Methanosarcina species (cf. Fig. 3) are among the archaea with the largest genomes, but terminal lineages have disposed of many families to end up with their current gene repertoire. The finding points to

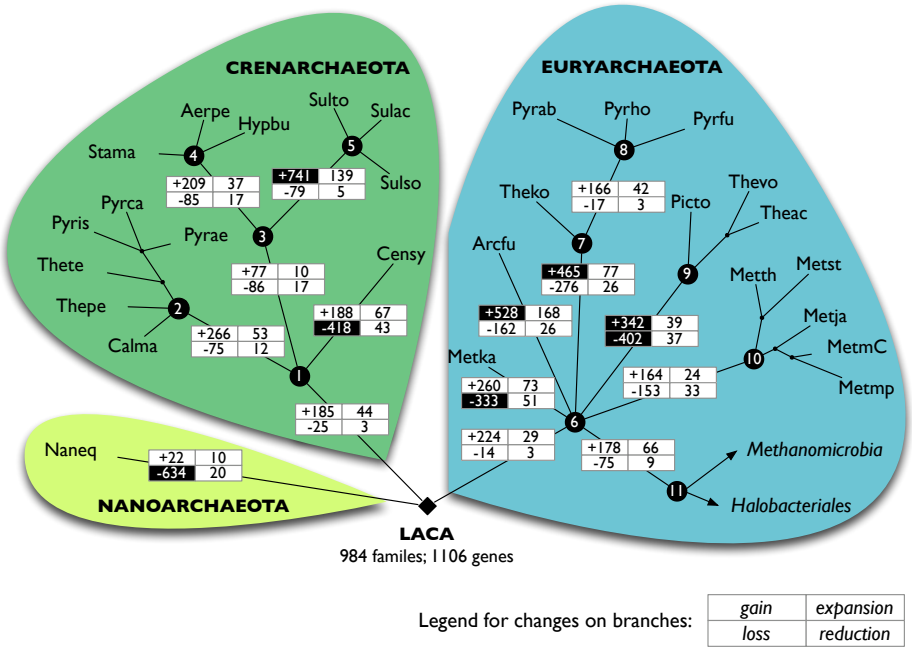

Fig. 2. Changes of gene repertoire in main lineages. On each branch, we computed the number of arCOG families gained and lost, as well as those that were retained but underwent changes in the number of paralogs (i.e., expansions or reductions). The numbers are shown in the small tables, in which darkened cells highlight major losses and gains. Correspondence between numbered nodes and taxonomic groups is given in Table 2. The subtree below node 11 is shown in Figure 3.

different paths of specialization from a versatile ancestor, accompanied by the elimination of redundant functions.

On branches leading to major lineages, newly appearing families typically outnumber expanding families by a factor of two to eight. It is not surprising that gains on those branches would be so frequent: the substantial differences in lifestyles are presumably possible only by acquiring genes with adequate new functionalities through lateral transfer or other means of evolutionary innovations. At the same time, terminal branches often display abundant family expansions: in 29 of the 41 terminal lineages, there are less than twice as many newly acquired genes than expanding families. This point is illustrated in Figure 3, showing a detailed reconstruction within a subtree. The most dramatic expansions are seen in Sulfolobales (below node 5 in Fig. 2), Methanosarcina and Halobacteriales (cf. Fig. 3). The branches leading to the progenitors of the same groups are precisely those with the most gains inferred in this study. The abundance of expansions is not a simple consequence of relatively large genome sizes, since expansions are frequent even in relative terms. Within Halobacteriales, 7.5–18% of families expanded on terminal branches; on the terminal branches

		This study		Makarova et al. (Fig. 7)	
Node number	**Group**	**Presence**	**Gain**	**Presence**	**Gain**
1	Crenarchaeota	1148	185	1245	291
2	Thermoproteales	1339	266	1404	237
3	Thermoprotei	1139	77	2128	928
4	Desulfurococcales	1263	209		*not shown*
5	Sulfulobales	1801	741		*not shown*
6	Euryarchaeota	1194	224	1335	349
7	Thermococcales	1413	465	1715	720
8	Pyrococcus	1562	166		*not shown*
9	Thermoplasmales	1134	342	1474	643
10	"Class I" methanogens	1205	164	1563	415

Table 2. Inferred gene content history in major linages. "Presence" columns give the number of arCOG families inferred at the listed taxonomic groups. "Gain" columns list the number of families that appear on the branch leading to the listed nodes.

of *M. hungatei* (Methu) and *M. acetivorans* (Metac), more than 12% of families did, in contrast with an overall average of 5.7% on terminal branches.

The observed patterns exemplify adaptations to new environments. Such an adaptation may be prompted by the acquisition of new functions, with ensuing series of gene duplications that lead to sub-functionalization, and, thus, specialization. A further scrutiny of such scenarios, is unfortunately difficult, because a substantial number of lineage-specific expansions are within poorly characterized families. In the most extreme case of *H. marismortui* (Halma), for example, 126 (31%) of 396 expanding families are poorly characterized. The top arCOG functional categories represented by the remaining expansions are C (energy: 35 families), E (amino acid metabolism: 33), K (transcription: 26), and T (signal transduction: 25). The functional variety of lineage-specific expansions illustrates the wide-ranging consequences of adapting to extreme environments.

4 Conclusion

When small data sets need to be analyzed, or reasonable assumptions for probabilistic analysis are not available, parsimony is a well-justified method of choice. Even in phylogenetic reconstructions, parsimony may enjoy an advantage over sophisticated likelihood methods, as it enables the faster exploration of the search space by quick scoring of candidate phylogenies [1]. The present work augments the set of parsimony tools available for the analysis of numerical evolutionary characters in a range of applications, including the analysis of gene content, regulatory motifs, and allele frequencies.

References

1. Felsenstein, J.: Inferring Phylogenies. Sinauer Associates, Sunderland, Mass. (2004)

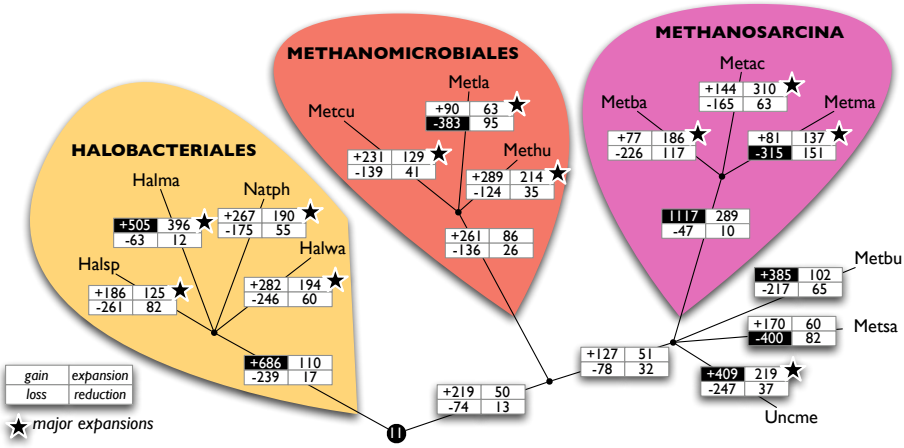

Fig. 3. Gene content evolution within Halobacteriales and Methanomicrobia. Stars highlight substantial expansions (at least half as many as family gains).

2. Farris, J.S.: Methods for computing Wagner trees. Syst. Zool. **19** (1970) 83–92
3. Swofford, D.L., Maddison, W.P.: Reconstructing ancestral states using Wagner parsimony. Math. Biosci. **87** (1987) 199–229
4. Maddison, W.P.: Squared-change parsimony reconstructions of ancestral states for continuous-valued characters on a phylogenetic tree. Syst. Zool. **40** (1991) 304–314
5. Sankoff, D., Rousseau, P.: Locating the vertices of a Steiner tree in arbitrary metric space. Math. Program. **9** (1975) 240–246
6. Hwang, F.K., Richards, D.S.: Steiner tree problems. Networks **22** (1992) 55–89
7. Cunningham, C.W., Omland, K.E., Oakley, T.H.: Reconstructing ancestral character states: a critical reappraisal. Trends Ecol. Evol. **13** (1998) 361–366
8. Pagel, M.: Inferring the historical patterns of biological evolution. Nature **401** (1999) 877–884
9. 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** (2003) 485–491
10. Caetano-Anollés, G.: Evolution of genome size in the grasses. Crop Sci. **45** (2005) 1809–1816
11. Omland, K.E.: Examining two standard assumptions of ancestral reconstructions: repeated loss of dichromatism in dabbling ducks (Anatini). Evolution **51** (1997) 1636–1646
12. Koonin, E.V.: Comparative genomics, minimal gene sets and the last universal common ancestor. Nat. Rev. Microbiol. **1** (2003) 127–136
13. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms. Second edn. MIT Press, Cambridge, Mass. (2001)
14. Rogers, J.S.: Deriving phylogenetic trees from allele frequencies. Syst. Zool. **52–63** (1984)

15. 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** (2008) 88–103
16. Irimia, M., Penny, D., Roy, S.W.: Coevolution of genomic intron number and splice sites. Trends Genet. **23** (2007) 321–325
17. Csűrös, M., Miklós, I.: A probabilistic model for gene content evolution with duplication, loss, and horizontal transfer. Lecture Notes in Computer Science **3909** (2006) 206–220 Proc. Tenth Annual International Conference on Research in Computational Molecular Biology (RECOMB).
18. Iwasaki, W., Takagi, T.: Reconstruction of highly heterogeneous gene-content evolution across the three domains of life. Bioinformatics **23** (2007) i230–i239
19. 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** (2007) 33
20. 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** (2003) 2
21. Fukui, T., Atomi, H., Kanai, T., Matsumi, R., Fujiwara, S., Imanaka, T.: Complete genome sequence of the hyperthermophilic archaeon Thermococcus kodakaraensis KOD1 and comparison with Pyrococcus genomes. Genome Res. **15** (2005) 352–363
22. Brochier, C., Forterre, P., Gribaldo, S.: An emerging phylogenetic core of Archaea: phylogenies of transcription and translation machineries converge following addition of new genome sequences. BMC Evol. Biol. **5** (2005) 36

A Species names and abbreviations

The following organisms are included in the study.
Aerpe *Aeropyrum pernix*, **Arcfu** *Archaeoglobus fulgidus*, **Calma** *Caldivirga maquilingensis* IC-167, **Censy** *Cenarchaeum symbiosum*, **Halma** *Haloarcula marismortui* ATCC 43049, **Halsp** Halobacterium species strain NRC-1, **Halwa** *Haloquadratum walsbyi*, **Hypbu** *Hyperthermus butylicus*, **Metac** *Methanosarcina acetivorans*, **Metba** *Methanosarcina barkeri fusaro*, **Metbu** *Methanococcoides burtonii* DSM 6242, **Metcu** *Methanoculleus marisnigri* JR1, **Methu** *Methanospirillum hungatei* JF-1, **Metja** *Methanocaldococcus jannaschii*, **Metka** *Methanopyrus kandleri*, **Metla** *Methanocorpusculum labreanum* Z, **Metma** *Methanosarcina mazei*, **MetmC** *Methanococcus maripaludis* C5, **Metmp** *Methanococcus maripaludis* S2, **Metsa** *Methanosaeta thermophila* PT, **Metst** *Methanosphaera stadtmanae*, **Metth** *Methanothermobacter thermoautotrophicus*, **Naneq** *Nanoarchaeum equitans*, **Natph** *Natronomonas pharaonis*, **Picto** *Picrophilus torridus* DSM 9790, **Pyrab** *Pyrococcus abyssi*, **Pyrae** *Pyrobaculum aerophilum*, **Pyrca** *Pyrobaculum calidifontis* JCM 11548, **Pyrfu** *Pyrococcus furiosus*, **Pyrho** *Pyrococcus horikoshii*, **Pyris** *Pyrobaculum islandicum* DSM 4184, **Stama** *Staphylothermus marinus* F1, **Sulac** *Sulfolobus acidocaldarius* DSM 639, **Sulso** *Sulfolobus solfataricus*, **Sulto** *Sulfolobus tokodaii*, **Theac** *Thermoplasma acidophilum*, **Theko** *Thermococcus kodakaraensis* KOD1, **Thepe** *Thermofilum pendens* Hrk 5, **Thete** *Thermoproteus tenax*, **Thevo** *Thermoplasma volcanium*, **Uncme** Uncultured methanogenic archaeon.