Approximate Maximum Parsimony
and Ancestral Maximum Likelihood

Noga Alon * Benny Chor† Fabio Pardi‡ Anat Rapoport§

Abstract

We explore the maximum parsimony (MP) and ancestral maximum likelihood (AML) criteria in phylogenetic tree reconstruction. Both problems are NP hard, so we seek approximate solutions. We formulate the two problems as Steiner tree problems under appropriate distances. The gist of our approach is the succinct characterization of Steiner trees for a small number of leaves for the two distances. This enables the use of known Steiner tree approximation algorithms. The approach leads to \( \frac{11}{6} \) approximation ratio for AML, and asymptotically to 1.55 approximation ratio for MP.

1 Introduction

The ancestral maximum likelihood (AML) problem, also called *most parsimonious likelihood* [2], is a maximum likelihood variant of phylogenetic tree reconstruction. Given a set of \( m \) sequences, the goal in AML is to find a tree topology \( T \) with the \( m \) sequences at the leaves, an assignment of sequences to internal (ancestral) nodes, and an assignment of substitution parameters for every edge, such that the overall likelihood (the probability of the resulting configuration) is maximized. AML “lies between” maximum parsimony (MP) [6] and maximum likelihood (ML) [5], in that it is a likelihood method (like ML), but sequences for internal tree vertices are also reconstructed (like MP). When the tree topology and its edge lengths are given, it is known how to efficiently find an optimal assignment of internal sequences [13]. When the tree topology is given, but edge lengths are not, it is still unknown if there is an efficient solution. Neither is much known about approximations and heuristics to the general AML problem (where the tree topology is not given), which is NP-hard.

*Schools of Mathematics and Computer Science, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel. Research supported in part by the Israel Science Foundation, by a USA-Israel BSF grant, and by the Hermann Minkowski Minerva Center for Geometry at Tel Aviv University. nogaa@post.tau.ac.il
†School of Computer Science, Tel Aviv University. Part of this work was done while visiting the European Bioinformatics Institute, Hinxton, Cambridge, UK. benny@cs.tau.ac.il
‡European Bioinformatics Institute, Hinxton, Cambridge, UK. pardi@ebi.ac.uk
§School of Computer Science, Tel Aviv University. anat.rapoport@gmail.com
MP can be seen as a special case of AML, which constrains the tree to be fully resolved, all edge lengths to be equal, and where a symmetric substitution model is assumed [7, 15]. Under these constraints, any tree that maximizes the ancestral likelihood is an MP tree.

In this paper we present an approximation algorithm for MP and for AML under the Neyman 2-state substitution model [12]. We remark that this simpler model is biologically significant, for example when DNA sequences are expressed in terms of Purines (Adenine and Guanine) and Pyrimidines (Thymine and Cytosine). In Neyman’s model, for each edge \( e \) of a tree \( T \) there is a corresponding probability \( p_e \) that the character states at the two endpoint vertices of \( e \) differ. Given leaves’ labels, any assignment of substitution probabilities to edge lengths, and of labels to internal nodes, determine the probability of generating this configuration. This probability is termed the ancestral likelihood, and yields the following version of the AML optimization problem:

**Ancestral Maximum Likelihood (Version I)**

**Input:** A set \( S \) of \( m \) binary sequences, each of length \( n \).

**Goal:** Find a tree \( T \) with \( m \) leaves, an assignment \( e \mapsto p_e \in [0,1] \) of edge probabilities, and a labelling \( \lambda : V(T) \to \{0,1\}^n \) of the vertices such that

1. The \( m \) labels of the leaves are exactly the sequences from \( S \), and
2. \( \prod_{e \in E(T)} p_e^{d_e/n}(1-p_e)^{n-d_e} \) (where \( d_e \) is the Hamming distance of the two labels across the edge \( e \)) is maximized.

We remark that in most phylogenetic contexts, evolution is viewed as a “conservative” process. Subsequently, in realistic instances, the edge substitution probabilities are in the range \( 0 \leq p_e \leq 1/2 \). The AML problem may, at first glance, seem like a continuous optimization problem due to the edge probabilities. The following observation, due to [1], shows that this is not the case. Given \( d_e \), the value of \( p_e \) that maximizes the individual contribution of \( e \) to the likelihood, \( p_e^{d_e}(1-p_e)^{n-d_e} \), is \( p_e = d_e/n \). This implies that the optimal \( p_e \) is one of \( n+1 \) possible values. Upon substituting this value and taking the \( n \)-th root, the contribution of the edge to the “normalized likelihood” is

\[
\left( \frac{d_e}{n} \right)^{d_e/n} \left( 1 - \frac{d_e}{n} \right)^{1-d_e/n}.
\]

Taking logarithms, the overall normalized log likelihood becomes

\[
\sum_{e \in E(T)} \frac{d_e}{n} \log \left( \frac{d_e}{n} \right) + \left( 1 - \frac{d_e}{n} \right) \log \left( 1 - \frac{d_e}{n} \right) = \sum_{e \in E(T)} -H_2 \left( \frac{d_e}{n} \right),
\]

where \( H_2 \) is the binary entropy function, \( H_2(p) = -p \log_2(p) - (1-p) \log_2(1-p) \) [4]. This leads to our second AML formulation (we drop the subscript 2 from logarithms and entropies):

\[\sum_{e \in E(T)} \frac{d_e}{n} \log_2 \left( \frac{d_e}{n} \right) + \left( 1 - \frac{d_e}{n} \right) \log_2 \left( 1 - \frac{d_e}{n} \right) \]
Ancestral Maximum Likelihood (Version II)

**Input:** A set $S$ of $m$ binary sequences, each of length $n$.

**Goal:** Find a tree $T$ with $m$ leaves and a labelling $\lambda : V(T) \rightarrow \{0, 1\}^n$ of the vertices such that

1. The $m$ labels of the leaves are exactly the sequences from $S$, and
2. $\sum_{e \in E(T)} H \left(\frac{d_e}{n}\right)$ is minimized.

The last formulation is fairly close to the following formulation of the maximum parsimony problem:

**Maximum Parsimony**

**Input:** A set $S$ of $m$ binary sequences, each of length $n$.

**Goal:** Find a tree $T$ with $m$ leaves and a labelling $\lambda : V(T) \rightarrow \{0, 1\}^n$ of the vertices such that

1. The $m$ labels of the leaves are exactly the sequences from $S$, and
2. $\sum_{e \in E(T)} (d_e/n)$ is minimized.

Finally, we recall the definition of the Steiner tree problem, which plays a central role in our algorithm.

The input is a connected graph $G = (V, E)$ with positive edge weights, and a subset $S \subseteq V$ of vertices, called terminals. A Steiner tree is a minimum weight connected subgraph of $G$, containing all vertices of $S$. It is known that the Steiner tree problem is NP-hard [10]. This motivates the search for efficient algorithms that produce approximate solutions.

If the graph $G$ is complete, and the weights satisfy the triangle inequality, then a minimum spanning tree on $S$ achieves an approximation ratio at most 2 [16]. Consequently, in a series of papers, a number of authors found improved approximation algorithms for the Steiner tree problem, with approximation ratio smaller than 2 [17, 3, 8, 14]. The first such improvement, due to Zelikovski, achieves a $11/6$ approximation ratio [17]. Further improvements applied to the running time of the algorithms, to the achieved approximation ratios, or to both. The best scheme to date approaches the approximation ratio $1.55 = 1 + \frac{\ln 3}{2}$, meaning that the tree produced by the algorithm has weight that is no more than 1.55 the weight of the Steiner tree [14].

2 Results

Both the AML and the MP problems can be thought of as Steiner tree problems, where the underlying graph $G$ is the complete graph over $\{0, 1\}^n$. For any pair of vertices $u, v \in \{0, 1\}^n$ with Hamming distance $d$ between them, the distance is $H(d/n)$ for the AML, and $d/n$ (or, equivalently, $d$) for MP. (We change notation from weight to distance as we will now deal with questions like the triangle inequality.) In the appendix, we show that the entropy measure $H(d/n)$ is indeed a distance, by
showing that it satisfies the triangle inequality. The triangle inequality is needed for applying the Steiner approximation algorithms. We remark that the ML problem cannot be formulated as a Steiner tree problem, at least not directly.

A central idea, due to [17] and then [3], is shared by these (and other) works. Given the graph $G$, the set of terminals (“leaves”) $S$, and an integer $k \geq 3$, find the Steiner trees (in $G$) for all subsets $A \subset S$ of up to $k$ terminals. Then, cleverly combine some of these $\binom{|S|}{k}$ “$k$ trees” to produce an approximate solution to the Steiner tree problem. In terms of running time, this approach is polynomial in $|V| + \binom{|S|}{k}$, which is polynomial in $|V| + |S|$ for any fixed $k$. However, in our AML/MP application, $G$ is not given as part of the input. Furthermore, since the number of $G$’s vertices, $|V| = 2^n$, is exponential in $n$, we cannot exhaustively go over all possible sets of internal nodes from $V$. However, going over all of $V$ is not necessary, provided we can generate, in time polynomial in $|S| + n$, a Steiner tree of $k$ (or fewer) given points. This rules out a direct application of the approximation algorithms mentioned above, as they are applicable to any graph $G$ and any metric on it.

We may, however, be able to take advantage of specific properties of AML/MP in order to identify a Steiner tree for each $A \subseteq S$ of size at most $k$ without exhaustively trying all internal nodes of $G$.

For MP, this is straightforward, as for each subset $A \subseteq S$ of $k$ input sequences, a most parsimonious tree can be found in time polynomial in $n$ (and super exponential in $k$), e.g. by trying exhaustively all tree topologies with the $k$ sequences at their leaves.

What about AML? As pointed out in the previous section, there is no known polynomial solution (polynomial in $n \cdot m$) to the “small AML” problem (that is, when the tree is given but edge lengths are not). We characterize optimal assignments of the internal node for the case of $k = 3$ leaves. We show that this internal assignment can always be taken as one of the three given sequences or as their point-wise majority (their MP solution). We begin with a simpler case, where edge lengths are given (see Figure 1).

**Problem 1:** We are given three sequences $v_1, v_2, v_3 \in \{0, 1\}^n$, and the three “edge lengths” $p_1, p_2, p_3$ ($0 \leq p_i \leq 1/2$). We wish to find a sequence $w \in \{0, 1\}^n$ that maximizes the ancestral likelihood of the sequences, given the tree and its internal node, $w$, namely the expression

$$p_1^{d(v_1,w)}(1-p_1)^{(n-d(v_1,w))} \cdot p_2^{d(v_2,w)}(1-p_2)^{(n-d(v_2,w))} \cdot p_3^{d(v_3,w)}(1-p_3)^{(n-d(v_3,w))}.$$ 

Taking logarithms, the expression becomes

$$\log L(v_1, v_2, v_3 | p_1, p_2, p_3, w) =$$

$$= \log \left(p_1/(1-p_1)\right) d(v_1, w) + \log \left(p_2/(1-p_2)\right) d(v_2, w) + \log \left(p_3/(1-p_3)\right) d(v_3, w)$$

$$+ n \log \left((1-p_1)(1-p_2)(1-p_3)\right)$$
Let \( C_i = \log(p_i/(1-p_i)) \) \((i = 1, 2, 3)\). As the \( p_i \)'s are smaller or equal to 1/2, \( p_i \leq 1 - p_i \), so all the \( C_i \) are non-positive. The last term in \( \log L \) does not depend on \( w \), so it suffices to maximize \( C_1d(v_1, w) + C_2d(v_2, w) + C_3d(v_3, w) \). Expressing the distances coordinate-wise, this equals

\[
\sum_{i=1}^{n} C_1 \delta(v_{1,i}, w_i) + \sum_{i=1}^{n} C_2 \delta(v_{2,i}, w_i) + \sum_{i=1}^{n} C_3 \delta(v_{3,i}, w_i) = \sum_{i=1}^{n} (C_1 \delta(v_{1,i}, w_i) + C_2 \delta(v_{2,i}, w_i) + C_3 \delta(v_{3,i}, w_i)),
\]

where \( \delta(u_i, v_i) = 0 \) if \( u_i = v_i \), and 1 if \( u_i \neq v_i \).

For any coordinate, \( i \), where \( v_{1,i} = v_{2,i} = v_{3,i} \), we should take \( w_i \) to equal this shared value. This makes \( \delta(v_{1,i}, w_i) = \delta(v_{2,i}, w_i) = \delta(v_{3,i}, w_i) = 0 \), and maximizes the contribution of such \( i \)-th term in the sum.

For any coordinates, \( i \), where the three entries are not equal, the optimal value of \( w_i \) depends on the coefficients \( C_1, C_2, C_3 \). Assume, without loss of generality, that \( C_1 \leq C_2 \leq C_3 \). We claim that If \( C_2 + C_3 - C_1 > 0 \) then for all coordinates, the optimal setting for \( w_i \) is \( w_i = v_{1,i} \). If \( C_2 + C_3 - C_1 < 0 \) then the optimal setting for \( w_i \) is the majority value out of \( v_{1,i}, v_{2,i}, v_{3,i} \). (In the case where \( C_2 + C_3 - C_1 = 0 \), any of these two options is optimal.) Now suppose that \( C_2 + C_3 - C_1 > 0 \). If we take \( w_i \neq v_{1,i} \), the contribution of the term \( C_1 \delta(v_{1,i}, w_i) \) to the overall sum is \( C_1 \). If, instead, we take \( w_i = v_{1,i} \), the worst contribution (minimum) of \( C_2 \delta(v_{2,i}, w_i) + C_3 \delta(v_{3,i}, w_i) \) to the sum is \( C_2 + C_3 \). Since \( C_1 < C_2 + c_3 \), we maximize our objective function by making \( \delta(v_{1,i}, w_i) = 0 \). On the other hand, if \( C_2 + C_3 - C_1 < 0 \), then since \( C_1 \leq C_2 \leq C_3 \), the sum of any two of the coefficients is smaller than the third. Therefore setting the entry \( w_i \) to equal the majority of the three bits \( v_{1,i}, v_{2,i}, v_{3,i} \) contributes a single coefficient to the sum, which is a larger contribution than the other two. Finally, it is clear that if \( C_2 + C_3 - C_1 = 0 \) then both options are optimal. We have just shown:

**Lemma 2.1** Let \( v_1, v_2, v_3 \in \{0, 1\}^n \) be three sequences that are the leaves in a tree with corresponding edge lengths \( p_1, p_2, p_3 \) \((0 \leq p_i \leq 1/2)\). Then an internal node that maximizes the ancestral likelihood is among \( v_1, v_2, v_3 \) or the maximum parsimony point (coordinate wise majority) of the three.

![Figure 1: The triplet tree, T](image)
The lemma can easily be generalized to the less realistic cases where some (or all) edge lengths \( p_i \) are greater than 1/2, with corresponding changes like replacing a sequence by its complement. We can also extend the characterization of AML assignment from the tree with three leaves to a star tree with \( k \) leaves \((k \geq 3)\). Let \( C_j = \log \left( \frac{p_j}{1-p_j} \right) \). At every coordinate \( i \) we look at the sequences \( v_j \) whose \( i \)-th coordinate is 0, and those where it is 1. We compute the sum of \( C_j \)'s for both sets. The optimal setting of \( w_i \) is to the value whose corresponding sum of \( C_j \)'s is smaller.

Finally, we come back to our motivating problem: Characterize AML solutions for the optimum tree with \( k = 3 \) leaves, when the edge lengths are not specified in advance.

**Problem 2:** Given three sequences \( v_1, v_2, v_3 \in \{0,1\}^n \), find a sequence \( w \) that minimizes the sum

\[
H(w,v_1) + H(w,v_2) + H(w,v_3)
\]

(1)

The same characterization proved for Problem 1 holds here as well, despite the fact that edge lengths are not specified. To see this, take an optimal assignment for the internal node and its induced edge lengths. For these lengths, Lemma 2.1 implies the optimality of one of the four assignments.

Using the terminology of Zelikovsky, what we showed is that for each triple of terminals (input sequences, or vertices in \( S \)), we can efficiently find the center. Then by the Steiner tree approximation algorithm of [17], which was discussed earlier, we get:

**Corollary 2.2** There is an efficient AML approximation algorithm, using \( k = 3 \) subsets, and achieving an approximation ratio \( \leq 11/6 \).

### 3 Concluding Remarks

By finding solutions to maximum parsimony and to ancestral maximum likelihood on \( k \) sequences of length \( n \), in time that is a fast growing function of \( k \) but polynomial in \( n \) (for any fixed \( k \)), we can employ known Steiner trees approximation algorithms in order to get approximate solutions to MP and AML. For MP, we can do this for every fixed \( k \), leading asymptotically to an approximation ratio of 1.55. For AML, our characterization applies to \( k = 3 \), leading to an 11/6 approximation ratio. It seems that the same approach can be extended to small values beyond \( k = 3 \), even though this becomes substantially more tedious for larger values of \( k \). We demonstrate the case \( k = 4 \) in Appendix 2.

It will also be of interest to extend the AML approximations to “real DNA” (4 states characters) under symmetric substitutions models such as Jukes-Cantor [9] and Kimura 2 and 3 parameter models [11]. Further extension to non-symmetric models of substitution, and to larger alphabets (e.g. proteins) are also of interest. Bounds on inapproximability of MP or AML are of (mostly theoretical) interest.

**Acknowledgments** Thanks to Alexander Zelikovski for his help in clarifying some subtle points in existing Steiner approximation algorithms.
References


Appendix 1: The Triangle Inequality

In this appendix, we establish the triangle inequality with respect to the entropy measure, i.e., we prove that for every $v_1, v_2, v_3 \in \{0,1\}^n$, $h(v_1, v_2) \leq h(v_1, v_3) + h(v_2, v_3)$, where $h(u,v) = H(d(u,v)/n)$.

Consider a process where we start at $v$, and switch each of its $n$ bits independently, each with probability $p$. The probability of reaching $u$ as a result of this process is $p^{d(v,u)} (1-p)^{n-d(v,u)}$. This probability is maximized for $p = d(v,u)/n$, and then the logarithm of this maximum probability is $-nh(v,u)$.

For all $1 \leq i < j \leq 3$, let $p_{ij} = d(v_i, v_j)/n$. Consider the following two phase process: We start with the sequence $v_1$, and switch each of its bits, randomly and independently, with probability $p_{13}$. Then, in the second phase, switch each bit of the resulting sequence, randomly and independently, with probability $p_{23}$. The probability that in this process $v_1$ is converted in the first phase to $v_3$, and then in the second phase to $v_2$, is precisely $2^{-nh(v_1,v_2,v_3)}$. On the other hand, the two phases combined are equivalent to flipping each bit of $v_1$, randomly and independently, with probability $p = p_{13}(1-p_{23}) + (1-p_{13})p_{23}$. Let $P$ denote the probability that starting with $v_1$, we end with $v_2$ in this combined process (not necessarily going through $v_3$). Clearly, $P$ is at least as large as the probability that this happens while passing through $v_3$ in the end of the first phase. On the other hand, $P$ is at most $2^{-nh(v_1,v_2)}$, as this is the probability of starting with $v_1$ and ending with $v_2$ while flipping every bit with the optimal probability $p_{12}$. This can only give larger (or equal) probability than the one we get using $p$, and therefore

$$2^{-nh(v_1,v_2,v_3)} \leq P \leq 2^{-nh(v_1,v_2)} ,$$

and the desired result follows.
Appendix 2: The $k = 4$ Case

In this appendix, we demonstrate how to find Steiner trees on $k = 4$ leaves under the AML (entropy) measure. We first assume that edge lengths are given, and characterize the two internal points in the tree. This enables us to provide a short list of possible Steiner trees for the entropy measure.

Consider the tree $T$ in Figure 2, where the given sequences at its leaves are $a,b,c,d$, and the five substitution probabilities are $p_1,p_2,p_3,p_4,p_5$. Applying Lemma 2.1 to the sequences at the vertices of the two triplet subtrees of $T$, we conclude that any AML solution must satisfy one of

$x = a, x = b, x = y, x = MP(a, b, y)$

as well as one of

$y = c, y = d, y = x, y = MP(c, d, x)$.

The case of equality – $x = a$ or $x = b$, or $y = c$ or $y = d$ brings us back the triplet case. The case of equality $x = y$ brings us back to the star case. We will explicitly demonstrate the solution to this case shortly. The only remaining case is $x = MP(a, b, y)$ and $y = MP(c, d, x)$, and furthermore $x \neq y$. We first argue that such “local maximum parsimony” (on the two subtrees) implies global maximum parsimony (on the whole tree).

Since parsimonious settings are bit-wise independent, it suffices to consider each of the $n$ bits separately. It is not hard to see that by symmetry, it suffices to consider only four patterns, appearing in Figure 3:

In the first two cases (the upper ones), $x_i$ must be 0, forcing $y_i = 0$ as well. In the third (bottom left) case, $x_i$ must be 0 and $y_i$ must be 1, agreeing with the global parsimony. In the remaining (bottom right) case, $x_i$ must equal $y_i$, or otherwise this will not be a locally parsimonious assignment. Either two 0s or two 1s are an acceptable solution, and both yield a global maximum parsimony assignment. Thus, there could be up to $2^n$ different MP assignments. However, depending
Figure 3: Case analysis, local parsimony.

On the specific values of $p_1, p_2, p_3, p_4$, typically one of them (either two 0s or two 1s) will induce higher likelihood. In the borderline case where the two likelihoods are identical, we can take each assignment and maximize the likelihood. (See the next paragraph for exact details.) Therefore, overall, it suffices to consider two internal $n$-bit assignments for this last case.

Finally, we give an explicit solution for the case $x = y$. Let $C_i = \log(p_i/(1-p_i))$ $(i = 1, 2, 3, 4)$. Under the usual “conservativeness” assumption, all $p_i$’s are smaller or equal to $1/2$, $p_i \leq 1 - p_i$, so all the $C_i$ are non-positive. Assume, without loss of generality, that $0 \leq -C_1 \leq -C_2 \leq -C_3 \leq -C_4$. We view this as a “weighted voting” case. There are essentially three instances (which overlap only for borderline cases):

1. One of the “weights”, $-C_4$, is greater or equal than the sum of all the others. Then we have a ”dictatorship” and assigning the internal sequence $x = d$ is optimal.

2. $-(C_2 + C_3) \geq -(C_1 + C_4)$ In this case, examination of all 8 possible patterns shows that the majority vote among $b, c$ and $d$ is optimal. In other words, $x = MP(b, c, d)$.

3. $-(C_2 + C_3) \leq -(C_1 + C_4)$ (and $-C_4 \leq -(C_1 + C_2 + C_3)$, to avoid case 1). In this case, the optimal assignment to the internal sequence is a majority vote among all sequences, with ties decided by $d$. We denote this by $x = MP^*(a, b, c; d)$.

Overall, we get a fixed, small number of candidate solutions (independent of $n$). When edge lengths are not given, we go over all candidates, optimize edges’ lengths for each, and determine the best one, which is the Steiner tree on the 4 given leaves.