We report that for population data, where sequences are very similar to one another, it is often possible to use a two-pronged (MinMax Squeeze) approach to prove that a tree is the shortest possible under the parsimony criterion. Such population data can be in a range where parsimony is a maximum likelihood estimator. This is in sharp contrast to the case with species data, where sequences are much further apart and the problem of guaranteeing an optimal phylogenetic tree is known to be computationally prohibitive for realistic numbers of species, irrespective of whether likelihood or parsimony is the optimality criterion. The Squeeze uses both an upper bound (the length of the shortest tree known) and a lower bound derived from partitions of the columns (the length of the shortest tree possible). If the two bounds meet, the shortest known tree is thus proven to be a shortest possible tree. The implementation is first tested on simulated data sets and then applied to 53 complete human mitochondrial genomes. The shortest possible trees for those data have several significant improvements from the published tree. Namely, a pair of Australian lineages comes deeper in the tree (in agreement with archaeological data), and the non-African part of the tree shows greater agreement with the geographical distribution of lineages.

Introduction

In the classical phylogeny problem with sequence data the goal is to reconstruct binary leaf-labeled trees, where the terminal nodes (leaves) of the tree represent species; this is a well-studied problem (Felsenstein 2003; Semple and Steel 2003). For this classical interspecies case, information about the character states at the internal nodes of a tree is lacking (fig. 1A) and deriving an optimal tree under the maximum likelihood criterion requires considering all possible states at these internal nodes. This maximum average likelihood estimator, which averages over all states at nodes with missing information, is appropriate in this situation where the data are sparse (Steel and Penny 2000). Computing an optimal tree is not an easy task; it is known to be NP-hard for the parsimony criterion (Graham and Foulds 1982), where it can be reformulated as the Steiner problem in graphs. Deriving the maximum likelihood (ML) estimator is also NP-hard under the maximum ancestral likelihood criterion (Addario-Berry et al. 2003), and is expected to be hard for maximum average likelihood, although this is still an open problem.

With many types of population data it is also appropriate to seek a phylogenetic tree; example data sources are mitochondrial sequences or portions of the nuclear DNA that have not undergone recombination. However, for population data the trees being sought differ from the interspecies case, in that they are not necessarily bifurcating, and that some present-day sequences are found as internal nodes of the tree (fig. 1B). For ample sets of sequences—i.e., sets in which every pair of sequences can be connected via a chain of sequences that are each one mutation apart (as in fig. 1B)—parsimony has been shown to be equivalent to ancestral maximum likelihood (also known as most-parsimonious likelihood) (Steel and Penny 2004a, 2004b). Note that we are not considering here the case where sequences are collected at different times, either—for example, from ancient DNA or from serial samples from RNA viruses (see for example, Drummond et al. 2003). It is also important to note that figure 1A refers to molecular data where it is appropriate to use a Markov model to approximate the processes of molecular evolution. With morphological data, a Markov model has not been justified, and it is currently necessary to consider missing data along each edge (branch) of the tree (fig. 1C). In this case, maximum evolutionary path likelihood is a natural measure and is the same as parsimony (Steel and Penny 2000).

In Hendy, Foulds, and Penny (1980) an approach is presented for demonstrating, in some cases at least, that a particular tree is guaranteed to be of minimal length under the parsimony criteria. The procedure employs an upper and lower bound, and it aims at reducing the upper bound and increasing the lower bound, until they are equal (The MinMax Squeeze).

An example of this approach is illustrated in figure 2 with a small set of six taxa and seven columns (each with only two character-states). Partitioning the data into separate columns gives a lower bound of seven changes, as each column will require at least one mutation on any tree (fig. 2A). A search through tree space finds a tree with 11 changes, and this is taken as an upper bound—four changes more than the lower bound. Partitioning the columns into pairs leads to an increase in the lower bound to 10 (fig. 2B). The increase arises from detecting pairs of columns that are incompatible: they will not fit onto any tree with only one change for each column. For example, considering columns 1 and 2 together gives four dinucleotide pairs, GA, TA, GC, and TC. The minimum possible number of mutations joining the four dinucleotide pairs on any tree is 3. Thus placing columns 1 and 2 into a single part of the partition increases the lower bound by 1. However, the partition in figure 2B still gives a lower bound one less than the upper bound. In this case, the partition of columns in figure 2C increases the lower bound until it equals the upper bound, as the 3-tuple of columns 1, 2 and 7 gives six unique trinucleotide sequences which will require at least five mutations on any tree. Therefore, by the Partition Theorem of Hendy,
Foulds, and Penny (1980) (which we restate as Theorem 1 in the next section), there can never be a shorter tree for these data, though the analysis in figure 2 does not demonstrate whether the minimal length tree is unique.

The Partition Theorem was, for two reasons, not easily applicable when first published. It required relatively similar sequences in order to prove minimality, and it could not be readily extended to maximum likelihood. However, now that many more population data sets require analysis, and theoretical results show that parsimony is in fact a maximum likelihood estimator for ample sets of sequences (Steel and Penny 2004a), it is appropriate to reconsider uses of this earlier theorem. We now introduce and explore some new algorithms and results for finding tight lower bounds based on this theorem.

In the last 15 years intraspecies phylogenies constructed from mitochondrial data sets have been used to test hypotheses about the origins of modern humans and patterns of human migration (Cann, Stoneking, and Wilson 1987; Vigilant et al. 1991; Murray-McIntosh et al. 1998). The first study to use complete mitochondrial genomes was by Ingman et al. (2000) who used phylogenetic evidence from 53 human mitochondrial genomes to test hypotheses about the origin and dispersal of modern humans. We apply the new lower bounds we have developed to this data set and find that a set of 40 trees (2 steps shorter than the published tree) can be proved minimal.

Methods

Suppose that A is a multiple alignment, which we will consider as an \( n_A \times c_A \) matrix, so that A is an alignment of \( n_A \) sequences with \( c_A \) columns. Note that we will not necessarily assume that all sequences in an alignment are distinct.

Given an alignment A and a phylogenetic tree on A, that is a tree with leaves labeled by the sequences of A, the Small Parsimony Problem is to compute the cost or minimum number of mutations required along the tree to give the sequences (Fitch 1971; Rinnsma, Hendy, and Penny 1990). The Large Parsimony Problem is to find, over all phylogenetic trees, the trees on A that have minimal cost (Felsenstein 2003; Semple and Steel 2003). We denote this minimal cost, or parsimony score, by \( P(A) \). The Large Parsimony Problem is intractable for typical sized alignments (Graham and Foulds 1982), although for data that are nearly homoplasy-free, it is possible to find a tree in polynomial time (Fernandez-Baca and Lagergren 2003). The computational complexity of finding optimal trees means that for many data sets heuristic approaches

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>t₁ G A C T T C G</td>
<td>t₁ G A</td>
<td>1 2 3 4 5 6 7</td>
<td>t₁ G A C T T C G</td>
</tr>
<tr>
<td>t₂ T A C T T A G</td>
<td>t₂ TA</td>
<td>12 34 56 7</td>
<td>t₂ G A C T T C G</td>
</tr>
<tr>
<td>t₃ G C C C T A G</td>
<td>t₃ GC</td>
<td>12 34 56</td>
<td>t₃ G A C T T C G</td>
</tr>
<tr>
<td>t₄ T C A C T A G</td>
<td>t₄ TC</td>
<td>12 34 56</td>
<td>t₄ G A C T T C G</td>
</tr>
<tr>
<td>t₅ T G C A T G A A</td>
<td>t₅ GC</td>
<td>12 34 56</td>
<td>t₅ G A C T T C G</td>
</tr>
<tr>
<td>t₆ T C A T G C A</td>
<td>t₆ TC</td>
<td>12 34 56</td>
<td>t₆ G A C T T C G</td>
</tr>
<tr>
<td>p 1111111111</td>
<td>p 3 3 3 3 3</td>
<td>p 1111111111</td>
<td>t 3 1 2 1 1 2 1</td>
</tr>
<tr>
<td>( \Sigma p &lt; \Sigma x )</td>
<td>( \Sigma p &lt; \Sigma x )</td>
<td>( \Sigma p &lt; \Sigma x )</td>
<td>( \Sigma p &lt; \Sigma x )</td>
</tr>
</tbody>
</table>

Fig. 1.—Trees for different classes of biological data. Gray nodes indicate sampled sequences; white nodes indicate unknown sequences. (A) is the classic phylogeny problem for species from sequence data with leaf-labeled binary trees; usually shown as a rooted tree. Sequences at internal nodes are unknown and maximum likelihood averages over all possible states at these internal nodes. (B) is a tree for population data where the tree is no longer binary, and at least some sequences in the present population represent internal nodes of the tree (the tree is shown unrooted). A set of sequences is said to be ample when each sequence in the set is just a single change from others in the population. (C) is again the classic phylogeny problem on species, but for morphological data where (unlike in A, where Markov models are used) it is not sufficient just to consider internal nodes. In this case standard parsimony is the evolutionary path maximum likelihood (MLep) estimator.

Fig. 2.—Application of the Partition Theorem to prove that a tree has minimal length. (A) is a set of six short sequences (columns 1–7) for taxa \( t₁–t₆ \). \( p \) is the minimum number of changes by column for any tree (all are 2-state characters, therefore \( p = 1 \)). \( t \) (the upper bound) is the parsimony score on the test tree (D), the sum of the observed number of changes on each column (this value can change with the tree). In (A) \( \sum p \) (the lower bound) is less than the upper bound, \( \sum t \). 2B has sites partitioned into subsets \{1,2\}, \{3,4\}, \{5,6\}, \{7\} and the values for \( p \) (independent of any tree) are now increased, but still sum to less than the upper bound. (C) is a revised partition with the new subset \{1,2,7\}. The upper and lower bounds are now identical, implying that no tree shorter than the tree in (D) can exist for this data. (D) shows the test tree, it requires 11 changes (the upper bound) for the data in (A).
must be used. Although a heuristic method cannot guarantee to have found the optimal tree, the score of the tree found does provide an upper bound on the score of the optimal tree.

We now turn to the problem of finding lower bounds for the parsimony score of an alignment. Given an alignment \( A \) with columns indexed by the set \( C_A = \{1, 2, \ldots, c_A\} \), and a subset \( \pi \) of \( C_A \), we let \( A_\pi \) be the alignment consisting of the columns indexed by \( \pi \) (so that in particular \( A_\pi \) is an \( n_A \times |\pi| \) matrix). For example, in figure 2C the subset \( \{1, 2, 7\} \) of \( C_A = \{1, 2, \ldots, 7\} \) gives a \( 7 \times 3 \) subalignment. The approaches that we present below for finding lower bounds for \( P(A) \) depend on the following theorem, which originally appeared in Hendy, Foulds, and Penny (1980).

**Theorem 1** Given an alignment \( A \), and \( \Pi \) a partition of \( C_A \) (i.e., a division of \( C_A \) into non-empty, non-intersecting subsets or parts whose union is \( C_A \)). Then

\[
P(A) \geq \sum_{\pi \in \Pi} P(A_\pi).
\]

Informally, this theorem states that the score of any tree on the alignment \( A \) can never be less than the sum of the parsimony scores of the subalignments of \( A \) given by the partition \( \Pi \).

**Distinct Sequence Bounds**

The first lower bound that we present is based on the following result, whose straightforward proof we omit. For an alignment \( A \), we let \( d_A \) denote the number of distinct sequences aligned in \( A \) minus one in case the set of sequences being aligned is ample, and otherwise define \( d_A \) to be just the number of distinct sequences.

**Theorem 2** Suppose that \( A \) is an alignment. Then \( P(A) \geq d_A \).

Now, given a partition \( \Pi \) of \( C_A \), we let

\[
D_\Pi(A) = \sum_{\pi \in \Pi} d_{A_\pi}.
\]

It follows by the last two theorems that \( P(A) \geq D_\Pi(A) \), so that \( D_\Pi(A) \) is a lower bound for \( P(A) \). We call this the distinct sequence bound for \( A \) or \( D \)-bound for short.

Clearly, for any alignment \( A \) the bound \( D_\Pi(A) \) depends on the choice of partition \( \Pi \). In case \( \Pi \) is the trivial partition of \( C_A \), i.e., the partition \( \Pi \) in which every part has cardinality or size 1 (as in fig. 2A), then, denoting by \( r_i \) the number of distinct character states in column \( i \) of \( A \), the inequality \( P(A) \geq D_\Pi(A) \) can be re-expressed as

\[
P(A) \geq \sum_{i=1}^{c_A}(r_i - 1).
\]

This inequality was stated in Hendy, Foulds, and Penny (1980) and later reported in Fernandez-Baca and Lagergren (2003), and is only a tight bound if \( A \) is homoplasy free. To achieve tight bounds for alignments with homoplasy we require partitions where each part has many distinct sequences.

**Incompatibility Bounds**

For an alignment \( A \) and \( i, j \) distinct, we denote \( A_{ij} \) also by \( A_{ij} \), and call a pair of columns in \( A \) indexed by \( i \) and \( j \) incompatible if \( P(A_{ij}) > (r_i - 1) + (r_j - 1) \). In (Semple and Steel 2003, pg. 97) a lower bound for \( P(A) \) is presented that is based on the number of pairs of incompatible columns in \( A \). We now introduce an alternative lower bound.

**Theorem 3** Suppose that \( A \) is an alignment. Let \( M(A) \) be the minimal value of \( \sum_{i,j} m_{ij} \) taken over all integers \( m_{ij} \geq (r_i - 1) \) satisfying

\[
m_i + m_j \geq P(A_{ij}), \quad \text{for all } 1 \leq i < j \leq c_A.
\]

Then \( P(A) \geq M(A) \).

Proof: For a phylogenetic tree \( T \) on \( A \), let \( m_T \) denote the minimum number of mutations along \( T \) required for column \( i \) of \( A \). Then \( m_T^2 \geq (r_i - 1) \) and \( m_T^2 + m_T^2 \geq P(A_{ij}) \) for all \( 1 \leq i < j \leq c_A \). Moreover, if \( T' \) is a phylogenetic tree on \( A \) with cost \( P(A) \), then \( \sum_{i,j} m_{ij} = P(A) \). The inequality \( P(A) \geq M(A) \) follows directly from these observations.

Informally, equation (3) states that the combined score of any pair of columns on any tree is at least the parsimony score for the subalignment of just those two columns. Theorem 3 states that if we could solve the integer program described it would provide a lower bound on the parsimony score. The next result follows by summing the inequalities in equation (3).

**Corollary 4** Suppose that \( A \) is an alignment with \( c_A \geq 2 \). Then

\[
P(A) \geq \frac{1}{(c_A - 1)} \sum_{1 \leq i < j \leq c_A} P(A_{ij}).
\]

By Theorem 1 and Corollary 4, it follows that

\[
I_\Pi(A) = \sum_{\pi \in \Pi - \{\pi\}} \left[ \frac{1}{(|\pi| - 1)} \sum_{i,j \in \pi, i < j} P(A_{ij}) \right]
+ \sum_{\{i,i\} \in \Pi} \left( r_i - 1 \right)
\]

is a lower bound for \( P(A) \), which we call the incompatibility or \( I \)-bound for \( A \). (Note that \( I_\Pi(A) = I(T_\Pi(A)) \). Since \( P(A_{ij}) \geq (r_i - 1) + (r_j - 1) \) for all \( 1 \leq i < j \leq c_A \), the \( I \)-bound will be of most interest when we have as many pairs \( i, j \) as possible in each part \( \pi \) of \( \Pi \) that correspond to incompatible columns of \( A \). To speed up computation of the \( I \)-bound in practice, all possible values of \( P(A_{ij}) \) were stored in a look-up table that was precomputed using the Fitch algorithm.

Using an approach based on incompatibility suggests a natural greedy algorithm for finding a partition of \( C_A \), this can be obtained as follows. (We will abuse notation slightly and say that a pair of elements of \( C_A \) is incompatible if they index incompatible columns of \( A \)). First compute all pairs of incompatible elements of \( C_A \), then, starting with 1, form a subset \( \pi \) of \( C_A \) by greedily adding in the next smallest element of \( C_A \) that is incompatible with 1,
and then repeatedly adding in the next smallest remaining element of \( C_A \) that is incompatible with all of those elements already in \( \pi \). This is continued until no further element can be added to \( \pi \). Then, starting with the smallest element of \( C_A - \pi \), create a new subset in the same way. This is repeated until a partition \( \Pi^g \) of \( C_A \) is created, which will consist of sets of pairwise incompatible elements together with sets of size one. Note that having created such a partition we can evaluate both \( D_- \) and \( D^g \)-bounds, and the optimization methods described in the Methods have been implemented in the Python scripting language and are available on request. The current implementation supports alignments in Phylip sequential format, which is described at http://www.med.nyu.edu/rcr/rcr/phylip/seqdoc.html.

Simulations

DNA alignments were simulated using Treevolve (Grassly and Rambaut, available from http://evolve.zoo.ox.ac.uk/software.html), which simulates the evolution of DNA sequences according to a coalescent model. As many mtDNA data sets used in population studies are effectively two-state because of the strong transition/transversion bias in mitochondrial data, we simulated such data sets by taking alignments generated by Treevolve, and amalgamating the A and G states to R, and the C and T states to Y. 100 data sets of length 100 base pairs were created under a Jukes-Cantor model of nucleotide substitution (Jukes and Cantor 1969), with either 20 or 40 taxa, and with either a high (\( 1*10^{-5} \)) or low (\( 5*10^{-10} \)) mutation rate. The other Treevolve parameter settings we employed were population size, \( n \), of \( 1*10^8 \) and generation time, \( b \), of 1. In alignments simulated using the lower rate there were fewer parallel substitutions and back substitutions and hence less homoplasy. Although only two different mutation rate settings were used, the stochastic nature of the simulations resulted in alignments with a range of homoplasy levels.

Upper bounds on the parsimony score for the alignments were generated using PAUP* by taking the minimum of the parsimony score of the Neighbor-Joining tree (Saitou and Nei 1987), and the parsimony score of the best tree found using a heuristic search. In practice the Neighbor-Joining tree never gave an upper bound that was smaller than the upper bound derived from the tree found by heuristic search.

Both the \( D_- \) and \( I_- \)-bounds for the parsimony score were computed employing the trivial partition \( \Pi^g \) and the greedily constructed partition \( \Pi^f \) as described in the Methods. We denote these lower bounds by \( D_-^p \), \( D_-^g \), and \( I_-^p \).

When necessary, we also performed local optimization to try to improve bounds, which resulted in bounds which we denote, for example, by \( D_-' \) as opposed to \( D_- \). For alignments involving only 2-state characters, we also calculated the \( I_- ' \)-bound which we denote simply by \( I_- ' \).

Results

To explore the effectiveness of the \( D_- \), \( I_- \), and \( I_- ' \)-bounds, we conducted both simulations and a study of a biological data set (53 human mitochondrial genomes from Ingman et al. 2000). All alignments were pre-processed to remove constant columns. The biological data set had been pre-processed to remove parsimony uninformative sites (the parsimony score resulting from these columns is independent of the tree on which the score is evaluated). All phylogenetic analyses were carried out using PAUP version 4.0b10 with default settings (Swofford 1998). The

\[
I_9^p(A) = 2c_A - |\Pi^g|,
\]

which, again by Theorem 1, is a lower bound for \( P(A) \). We will call this the \( I_- ' \)-bound for \( A \). Note that \( I_9^p(A) \geq I(A) \), so that for binary data, the \( I_- ' \)-bound is never worse than the \( I_- \)-bound.

Optimizing Partitions

For many alignments the lower bounds described above may be less than the upper bound (the parsimony score of a tree found by heuristic search), and hence cannot be used to prove the trees for such alignments minimal. However, in such cases the \( D_- \) and \( I_- \)-bounds (but not the \( I_- ' \)-bound) may be improved by choosing new partitions so as to (locally) optimize the bound. To do this, we consider two types of moves in partition space: The first takes two parts of a partition and merges them to form a single part. The second moves a single element from one part of a partition into a different part. These moves give a neighbour-hood of partitions around any given partition whose lower bounds can be evaluated. A simple hill-climbing procedure can then be used to move through partition space until a local optimum is found.

Results

To explore the effectiveness of the \( D_- \), \( I_- \), and \( I_- ' \)-bounds, we conducted both simulations and a study of a biological data set (53 human mitochondrial genomes from Ingman et al. 2000). All alignments were pre-processed to remove constant columns. The biological data set had been pre-processed to remove parsimony uninformative sites (the parsimony score resulting from these columns is independent of the tree on which the score is evaluated). All phylogenetic analyses were carried out using PAUP version 4.0b10 with default settings (Swofford 1998). The
bounds tended to be larger for the data sets with 40 taxa. However, this difference increased more sharply for the $I$-bounds than for the $D$-bounds.

The overall success rate was higher for alignments generated with the lower mutation rate. This is expected as the lower mutation rate gives rise to less homoplasy and so the set of sequences is closer to being ample (since more sequences are just one mutation apart). To see if the lower bound equalled the best upper bound (bottom entry), and the mean difference between the best lower bound and the upper bound (top entry). The slope of this linear relationship is the expected negative linear relationship between the consistency index for each of the simulated data sets. Figure 3 indicates the expected negative linear relationship between the difference (i.e., how tight the best lower bound is) and the consistency index. The slope of this linear relationship is steeper for 4-state than 2-state character data. As can be seen in figure 3, none of the locally optimized bounds were tight for alignments with a consistency index lower than 0.77. However, for many data sets with lower levels of homoplasy we can prove that an optimal tree has been found. This is a major advance, particularly for the 40 taxon trees, where there are no exact search methods that guarantee an optimal tree.

**Table 1**

<table>
<thead>
<tr>
<th></th>
<th>20 Taxa</th>
<th>40 Taxa</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Low Rate</strong></td>
<td><strong>High Rate</strong></td>
<td><strong>Low Rate</strong></td>
</tr>
<tr>
<td>$I = D_t^h$</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>$I_t$</td>
<td>3.80d</td>
<td>12.90</td>
</tr>
<tr>
<td>$D_g$</td>
<td>68</td>
<td>10</td>
</tr>
<tr>
<td>$D_t^*$</td>
<td>0.44</td>
<td>3.13</td>
</tr>
<tr>
<td>$D_t^*$</td>
<td>75</td>
<td>22</td>
</tr>
<tr>
<td>$I_t$</td>
<td>0.30</td>
<td>2.58</td>
</tr>
<tr>
<td>$I_t$</td>
<td>77</td>
<td>31</td>
</tr>
<tr>
<td>$I_t$</td>
<td>0.26</td>
<td>1.57</td>
</tr>
<tr>
<td>$I_t$</td>
<td>70</td>
<td>10</td>
</tr>
<tr>
<td>$I_t$</td>
<td>0.42</td>
<td>3.35</td>
</tr>
<tr>
<td>$I_t$</td>
<td>74</td>
<td>15</td>
</tr>
<tr>
<td>$I_t$</td>
<td>0.36</td>
<td>2.80</td>
</tr>
<tr>
<td>$I_t$</td>
<td>67</td>
<td>15</td>
</tr>
<tr>
<td>$I_t$</td>
<td>0.45</td>
<td>2.74</td>
</tr>
<tr>
<td>$I_t$</td>
<td>78</td>
<td>17</td>
</tr>
<tr>
<td>$I_t$</td>
<td>0.30</td>
<td>2.05</td>
</tr>
<tr>
<td>Overall$^f$</td>
<td>0.12</td>
<td>1.27</td>
</tr>
</tbody>
</table>

$^a$ For each of the four situations given at the top of the table, 100 alignments were generated as described in the text.

$^b$ For each one of the 400 alignments, the lower bounds indicated on the side of the table were computed.

$^c$ The top entry in each cell records the number of times out of 100 that the lower bound equalled the best upper bound.

$^d$ The bottom entry in each cell records the mean difference between the lower bound and the best upper bound.

$^e$ The highest entry in each column (excluding the Overall row) is shown in bold.

$^f$ The row labeled Overall records the number of times out of 100 that the best lower bound equaled the upper bound (top entry), and the mean difference between the best lower bound and the upper bound (bottom entry).

**Table 2**

<table>
<thead>
<tr>
<th></th>
<th>20 Taxa</th>
<th>40 Taxa</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Low Rate</strong></td>
<td><strong>High Rate</strong></td>
<td><strong>Low Rate</strong></td>
</tr>
<tr>
<td>$I = D_t$</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>$I_t$</td>
<td>3.44</td>
<td>12.38</td>
</tr>
<tr>
<td>$D_g$</td>
<td>69</td>
<td>13</td>
</tr>
<tr>
<td>$D_t$</td>
<td>0.38</td>
<td>3.14</td>
</tr>
<tr>
<td>$D_t$</td>
<td>80</td>
<td>27</td>
</tr>
<tr>
<td>$I_t$</td>
<td>0.22</td>
<td>1.48</td>
</tr>
<tr>
<td>$I_t$</td>
<td>79</td>
<td>16</td>
</tr>
<tr>
<td>$I_t$</td>
<td>0.25</td>
<td>2.74</td>
</tr>
<tr>
<td>$I_t$</td>
<td>74</td>
<td>18</td>
</tr>
<tr>
<td>$I_t$</td>
<td>0.36</td>
<td>2.45</td>
</tr>
<tr>
<td>$I_t$</td>
<td>75</td>
<td>17</td>
</tr>
<tr>
<td>$I_t$</td>
<td>0.31</td>
<td>1.96</td>
</tr>
<tr>
<td>$I_t$</td>
<td>75</td>
<td>28</td>
</tr>
<tr>
<td>$I_t$</td>
<td>0.32</td>
<td>1.93</td>
</tr>
<tr>
<td>Overall$^f$</td>
<td>0.11</td>
<td>1.17</td>
</tr>
</tbody>
</table>

$^a$ See table 1 for an explanation of entries.

**Biological Data**

To investigate how our bounds performed on biological data, we used an alignment from Ingman et al. (2000), containing 202 sites from the full coding-region sequences of 53 human mitochondrial genomes. This alignment does not contain any sites from the D-loop region of the mitochondrial genome, and the parsimony uninformative sites have also been removed. The alignment we used is exactly that shown on the left of figure 4 in Ingman et al. (2000).

We began by constructing a smaller data set containing only the non-African sequences because these are known to be less diverse and are thus expected to have lower levels of homoplasy. In view of the simulations, we expected the lower bounds for this reduced data set to be tighter than for the full data. The reduced alignment consisted of 32 taxa and 75 parsimony informative columns. We used the heuristic search implemented in PAUP* to calculate an upper bound of 95 (20 equally scoring trees were found). The consistency index for this data set was 0.79—near the upper range for obtaining good lower bounds in the simulations.

Table 3 shows the performance of the $D$- and $I$-bounds. For this reduced data set it was possible to prove that the 20 equally scoring trees are optimal using the fact that the upper and lower bounds were the same. Note that this doesn’t rule out the possibility that there are still other most parsimonious trees.

We then considered the complete Ingman et al. (2000) data set. The Neighbor-Joining tree shown in figure 2 of Ingman et al. (2000) has a parsimony score of 271. As discussed in Richards et al. (1996) Neighbor-Joining is not an appropriate tool for analyzing population data. The combined effect of having an ample set of sequences along with homoplasy caused by back substitutions and parallel substitutions means that there is unlikely to be a single tree that best describes the data. Therefore, a method like Neighbor-Joining that arbitrarily resolves ties may not perform well. More recent analyses of complete human mt
Coding sequences have used reduced median networks (see for example Kivisild et al. [2002] and Herrnstadt et al. [2002]). Using the heuristic search implemented in PAUP*, we calculated an upper bound of 269 (40 equally scoring trees were found). The consistency index for these data is 0.75, just outside the range for which tight bounds were found in the simulations. Table 3 shows the performance of the \( D^* \) and \( I^* \)-bounds before and after local optimisation. The \( D^* \) bound was only one less than the upper bound.

For the bound \( D^* \) a further optimization was performed that made use of information given by one of the 40 equally scoring trees (chosen arbitrarily). For each part \( \pi \) in the locally optimal partition used to calculate \( D^* \), we checked whether the lower bound \( D(A_\pi) \) equaled the cost of the columns in \( \pi \) for the selected tree. Those parts in the partition where the cost and the lower bound were equal were retained, and those where they differed were divided into trivial partitions (i.e., into a collection of size 1 subsets), and the procedure of moving through partition space using merge moves was started afresh. This procedure was repeated three times until a lower bound of 269 was attained, indicating that the 40 trees found by heuristic search are optimal (although they may not comprise the set of all most parsimonious trees). The partition that gave the lower bound of 269 (parts of size 1 not shown) was \{0, 11, 82, 85, 161\}, \{1, 8, 46, 47, 49, 72, 84, 181\}, \{3, 35, 69, 70, 106, 143, 171\}, and so on.

**Fig. 3.**—Performance of the best lower bound plotted against the consistency index for 200 2-state alignments on 20 taxa (unfilled circles), 200 2-state alignments on 40 taxa (unfilled triangles), 200 4-state alignments on 20 taxa (filled circles), and 200 4-state alignments on 40 taxa (filled triangles). The consistency index is calculated as \( D_t \) divided by the upper bound.

**Fig. 4.**—The consensus network of the 40 most parsimonious trees found by heuristic search. The network displays all splits (edges) that appear in any of the 40 most parsimonious trees. The edge lengths shown in the network are the average lengths of the corresponding edges in the trees containing that edge. African sequences are labeled in boldface.
Table 3  Lower Bounds for the Human Mitochondrial Genome
Data Sets

<table>
<thead>
<tr>
<th></th>
<th>32 Taxa</th>
<th>53 Taxa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upper Bound</td>
<td>95</td>
<td>269</td>
</tr>
<tr>
<td>$I_e = D_e$</td>
<td>75</td>
<td>202</td>
</tr>
<tr>
<td>$D_g^*$</td>
<td>90</td>
<td>250</td>
</tr>
<tr>
<td>$D_g^{**}$</td>
<td>95</td>
<td>268</td>
</tr>
<tr>
<td>$I_r$</td>
<td>90</td>
<td>238</td>
</tr>
<tr>
<td>$I_r^*$</td>
<td>90</td>
<td>246</td>
</tr>
<tr>
<td>$I_r^{**}$</td>
<td>90</td>
<td>246</td>
</tr>
</tbody>
</table>

199\}, \{4, 9, 14, 27, 68, 114, 180, 191\}, \{5, 12, 13, 21, 28, 37, 77, 117, 177, 188\}, \{6, 25, 44, 45, 88, 104, 110, 190\}, \{7, 17, 53, 60, 108, 192\}, \{16, 23, 83\}, \{20, 69, 89, 189\}, \{26, 143, 193\}, \{31, 109\}, \{32, 194\}, \{42, 52, 87, 186\}, \{43, 197\}, \{50, 116\}, \{61, 187\}, \{74, 179\}, \{75, 201\}}, \{79, 112, 115\}, \{81, 111\}, \{113, 162\}, \{130, 198\}, \{182, 200\}, \{2, 195, 196\}.

In figure 4 we present a consensus network (Holland and Moulton 2003) of the optimal trees with a threshold of 0. In this extreme form a consensus network is simply the median network of all splits (edges) that appear in any of the 40 optimal trees. The network indicates some areas of the phylogeny that are not yet fully resolved. The 40 trees fall into five groups of 8, where within each group there are no hard conflicts but there are different levels of resolution. Denoting the group of Guraani, Siberian Inuit, and Japanese as clade A, and the group of Warao, Evenki, Khirgiz, and Burini as clade B: one group of 8 has the Guraani basal within clade A, and clade A and B as sister groups; another 8 have a Japanese sequence basal within clade A, and clade A and a PNG Coast sequence form a sister group; another 8 have a Japanese sequence basal within clade A, and clade A and Asian Indian form a sister group; another 8 have no resolution within clade A, and clade A and clade B form a sister group; the final 8 trees have no resolution within clade A and no resolved sister group for clade A.

The 40 equally scoring trees proved optimal here had several differences to the Neighbor-Joining tree published in Ingman et al. (2000): In particular a pair of Australian lineages comes deeper within the non-African clade in accordance with the archaeological evidence that modern humans have been in Australia for more than 50,000 years (Roberts et al. 2001; Kirch 2000). The European sequences Saami, English, French, Dutch, and Crimean form a clade in all of the 40 optimal trees, as do the Samoan, Korean, and Chinese sequences. This brings more agreement between the tree and the known geographic distribution. One of the clades reported in the Ingman et al. (2000) Neighbor-Joining tree appears in 8/40 of the optimal trees, but is contradicted in 16/40 (in the other 16 it appears in an unresolved polytomy). Note that although these 40 trees have differences from the tree of Ingman et al. (2000) they do not alter the main conclusion of the Ingman et al. article regarding the African origin of modern humans.

**Discussion**

We have presented a range of methods for computing lower bounds for maximum parsimony, these methods are based on the Partition Theorem (Hendy, Foulds, and Penny 1980). The lower bounds may be used in combination with upper bounds provided by heuristic search to prove that a tree is optimal.

The performance of the lower bounds was tested on both simulated and biological data. As expected, the bounds are better for data with low levels of homoplasy. Our results also indicate that a range of lower bounds should be employed to obtain best results. For data that is almost homoplasy free the bounds that are simple to compute (e.g., $D_g$, $I_r$, or $I_r^{**}$) should suffice. However, for high levels of homoplasy, lower bounds usually require some optimization in partition space to prove that trees are minimal.

For one of the real data sets we also used extra information, from one of the 40 trees having a parsimony score equal to the best upper bound, to improve on our lower bounds. This method might be further developed to cope with data with higher levels of homoplasy by making more use of information about the costs of each column on a tree that provides the upper bound. This could also indicate more effective ways of both constructing an initial partition and moving through partition space. Another approach that could improve the lower bounds would be to derive bounds by solving the integer program that can be formulated using the inequalities in Theorem 3 (and their counterparts for parts in partitions of size greater than 2). In addition, optimizing in partition space might be improved by approaches such as simulated annealing (Dowsland 1993) or tabu search (Glover and Laguna 1993), as these methods may allow the search to escape from local optima.

Here we have been concerned with the problem of proving trees minimal; however, tight lower bounds will also have application to the exact Branch and Bound search methodology (Hendy and Penny 1982; Felsenstein 2003). Better bounds will lead to larger portions of tree space being eliminated in the search. Ultimately we expect that this approach could lead to faster and more accurate phylogeny estimation.

At present, parsimony is only shown to be a maximum likelihood estimator when the set of sequences being analyzed is ample. For the complete mitochondrial genomes here, we are at a boundary between being ample and being sparse, and more theoretical work is required. It may be that a combination of ample subsets and maximum average likelihood between these subsets is the appropriate ML estimator. In reality there is a multi-scale continuum between generations, populations, genera, and orders etc., and it is necessary to use the most appropriate method for any application.

It is expected, with the increase in DNA sequencing potential, that many more population data sets will require analysis. Thus the MinMax Squeeze will have a large number of applications. We have applied the approach to over 100 limpet sequences (N. Gemmel, Christchurch) and in that case it was trivial to prove the tree minimal as the set of sequences is ample (unpublished data).
application here is to 53 complete mitochondrial genomes. This is a large increase in the size of a non-trivial data set for which a tree has been proved optimal, but we have little idea yet on the upper limits for the technique. For example, adding additional genomes in regions where the tree is still unresolved (fig. 4) should reduce the number of optimal trees from the 40 found at present. However, additional genomes may increase the extent of homoplasy, perhaps requiring more powerful methods for determining lower bounds.

Acknowledgments

K.T.H. and V.M. thank The Linnaeus Center for Bioinformatics, where part of this research was undertaken. K.T.H. thanks the Swedish Research Council for its support. All authors thank The Swedish Foundation for International Cooperation in Research and Higher Education (STINT). D.P. thanks the New Zealand Marsden Fund.

Literature Cited


Lisa Matisoo-Smith, Associate Editor

Accepted October 4, 2004