# Absolute Convergence:<br>True Trees From Short Sequences

Tandy Warnow* Bernard M.E. Moret† Katherine St. John‡

## Abstract

Fast-converging methods for reconstructing phylogenetic trees require that the sequences characterizing the taxa be of only polynomial length, a major asset in practice, since real-life sequences are of bounded length. However, of the half-dozen such methods proposed over the last few years, only two fulfill this condition without requiring knowledge of typically unknown parameters, such as the evolutionary rate(s) used in the model; this additional requirement severely limits the applicability of the methods. We say that methods that need such knowledge demonstrate *relative fast convergence*, since they rely upon an oracle. We focus on the class of methods that do not require such knowledge and thus demonstrate *absolute fast convergence*. We give a very general construction scheme that not only turns any relative fast-converging method into an absolute fast-converging one, but also turns any statistically consistent method that converges from sequences of length $O(e^{O(diam(T))})$ into an absolute fast-converging method.

## 1 Introduction

Phylogenetic reconstruction methods build an evolutionary tree from a collection of taxa given, for example, by molecular sequences. These methods are designed to recover the "true" evolutionary tree as often as possible. Not all are guaranteed to do so with high probability under reasonable conditions; even those that offer this guarantee vary considerably in their requirements. Under some models of evolution, no method can be guaranteed to recover the true tree with high probability, due to unidentifiability. Under other models, many methods will be able to recover the tree if given long enough sequences. The latter methods are said to be *statistically consistent* under the model of evolution. Formally, a method is statistically consistent for a specific model of evolution if, for every model tree (i.e., rooted tree and the associated random variables) and every $\varepsilon > 0$, there is a sequence length $k$ (which depends on the method, the model tree, and $\varepsilon$) such that the method recovers the topology (the edges) of the model tree with probability at least $1-\varepsilon$, if it is given sequences of length at least $k$. For many models (such as the Jukes-Cantor model [13], the simplest four-state model, as well as more complex models, such as the General Markov (GM) [19] model), even simple distance methods are easily established to be statistically consistent.

The sequence length required by a method is a significant aspect of its performance since real data sets are of limited length. (Computational requirements are also important, but it is possible to wait longer or get more powerful machines, while it is not possible to get longer sequences than exist in nature.) Consequently, experimental and analytical studies have attempted to bound the sequence lengths required by different phylogenetic methods. Methods that perform well (with respect to topology estimation) from sequences of realistic lengths (bounded by at most a few thousand nucleotides) are very desirable, especially if the topological accuracy remains good when the rate of evolution and number of taxa increase.

In an earlier paper [12], we defined the notion of *fast convergence* under the Jukes-Cantor model of evolution. In this paper we introduce the concept of *absolute fast convergence* and extend this definition to more general models, such as the General Time Reversible Markov model. A method is *absolute fast-converging* if it is fast-converging and does not need to know any of the parameters of the model in order to achieve fast convergence. In the Jukes-Cantor model, such parameters might be the minimum ($f$) or maximum ($g$) expected number of times a site changes on any edge in the tree. A method can be fast converging only if it is given knowledge of one or both of these values: such a method is not absolute fast-converging and we say instead that it is *relative fast-converging*.

Only a few methods have been proved to be absolute fast-converging even under the simple Jukes-Cantor model: two "DCM-boosted" quartet methods [12] and

[*] Dept. of Computer Science, U. of Texas at Austin, tandy@cs.utexas.edu; supported by NSF grant 94-57800 and by the David and Lucile Packard Foundation.

[†] Dept. of Computer Science, U. of New Mexico, moret@cs.unm.edu; supported in part by NSF grant 00-81404.

[‡] Graduate Center and Dept. of Math. and Computer Science, Lehman College, City U. of New York, stjohn@lehman.cuny.edu; supported in part by NSF grant 99-73874.

the Short-Quartet methods [8, 9]. Methods that have been proved relative fast-converging under the Jukes-Cantor model include DCM-boosted neighbor-joining (DCM-NJ) [12], the *Harmonic Greedy Triplets* (HGT) method of Csűrös and Kao [4], and a method of Cryan, Goldberg, and Goldberg (CGG) [3]. These methods are only relative fast-converging under the Jukes-Cantor model, rather than absolute fast-converging, because they require knowledge of $f$ or $g$; without such knowledge, they have to guess the parameter (or, more precisely, they have to guess which of the trees they have constructed is the correct tree). Their guessing strategies are not provably correct. Consequently, in absence of knowledge about $f$ and $g$, these methods are not even statistically consistent.

In this paper, we describe a very simple algorithm, which we call *Short Quartet Support (SQS)*. This algorithm selects the true tree (under the GM model) with high probability from a collection of trees, given sequences of only polynomial length. Since SQS does not require knowledge of the model parameters, it can be used to turn a relative fast-converging method into an absolute fast-converging method. Consequently, a straightforward use of SQS produces absolute fast-converging versions of the relative fast-converging methods DCM-Neighbor-Joining [12], HGT [4], and CGG [3].

SQS can also be combined with the first phase of the *Disk-Covering Method (DCM)* [12] to produce a technique we call *DCM** for reducing the dependency of methods on sequence lengths. In particular, DCM* turns methods that converge from sequence lengths that grow *exponentially* in the diameter (the longest path length in the tree) into methods that are absolute fast-converging. Since the diameter of an $n$-leaf tree can be as large as $n-1$ and is typically $\Omega(\sqrt{n})$, this technique turns methods that are statistically consistent, but not even relative fast-converging, into absolute fast-converging methods. Finally, SQS provides a very general framework within which absolute fast-converging methods can be developed.

We state our results in terms of the General Time Reversible Markov model of evolution, which contains, as a special case, the Jukes-Cantor model.

## 2 Basics

### 2.1 Stochastic models of DNA sequence evolution.

A model of DNA sequence evolution must describe the probability distribution of the four states, $A, C, T, G$, at the root, the evolution of a random site (i.e., position within the DNA sequence) and how the evolution differs across the sites. Typically the probability distribution at the root is uniform (so that all sequences of a fixed length are equally likely). The evolution of a single site is modeled through the use of "stochastic substitution matrices," $4 \times 4$ matrices (one for each tree edge) in which every row sums to 1. A stochastic model of how a single site evolves can thus have up to 12 free parameters. The simplest such model is the Jukes-Cantor model, with one free parameter, and the most complex is the General Markov model, with all 12 parameters [19].

Definition 1. *The GM model of single-site evolution is defined as follows.*

1. *The nucleotide in a random site at the root is drawn from a known distribution, in which each nucleotide has positive probability.*
2. *The probability of each site substitution on an edge $e$ of the tree is given by a $4 \times 4$ stochastic substitution matrix $M(e)$ in which $det(M(e))$ is not $0$, $1$, or $-1$.*

This model is generally used in a context where all sites evolve identically and independently (the *iid* assumption), although sometimes a distribution of rates across sites is also given. In this paper, we use the GM model with *iid* site evolution.

We denote a model tree in the GM model as a pair, $(T, \{M_e \colon e \in E(T)\})$, or more simply as $(T, M)$. We assume that the number of changes obeys a Poisson distribution. For each edge $e \in E(T)$, we define the weight of the edge $\lambda(e)$ to be $-\log|det(M_e)|$. This allows us to define the matrix of leaf-to-leaf distances, $\{\lambda_{ij}\}$, with $\lambda_{ij} = \sum_{e \in P_{ij}} \lambda(e)$ and where $P_{ij}$ is the path in $T$ between leaves $i$ and $j$. Note that $\{\lambda_{ij}\}$ is a symmetric matrix. It is a well-known fact that, given the distance matrix $\{\lambda_{ij}\}$, it is easy to recover the underlying leaf-labelled tree $T$ in polynomial time.

This general model of site evolution subsumes the great majority of other models examined in the phylogenetic literature, including the Hasegawa-Kishino-Yano (HKY) model, the Kimura 2-parameter model (K2P), the Kimura 3-ST model (K3ST), the Jukes-Cantor model (JC), etc. These models are all special cases of the General Markov model, because they place restrictions on the form of the stochastic substitution matrices (see [14] for more information about stochastic models of evolution). Most distance-based methods are statistically consistent under the General Markov model, because statistically consistent methods exist for estimating the matrix $\{\lambda_{ij}\}$ above. (A method for estimating the matrix $\{\lambda_{ij}\}$ is statistically consistent if each of the distance estimates $d_{ij}$ converges to the true value, $\lambda_{ij}$, as the sequence length increases, with probability 1.) The "log-det" distances provide such a statistically consistent estimation of the matrix $\{\lambda_{ij}\}$

[19]. In our theorems, we will call these distances simply the *GM corrected distances.*

In [9], the sequence length needed to obtain an arbitrarily good estimate of these distances was estimated, as follows:

THEOREM 2.1. *Let $(T, M)$ be a model tree in the GM model. Set $\lambda(e) = -\log|det(M_e)|$ and $\lambda_{ij} = \sum_{e \in P_{ij}} \lambda(e)$. Assume that $f, g$ are fixed with $0 < f \leq \lambda(e) \leq g$ for all edges $e \in T$. Let $\varepsilon > 0, \delta > 0$ be given. Then, there is a constant $C$ such that, if the sequence length exceeds*

$$C \log n e^{O(g \cdot diam(T))}$$

*then with probability at least $1 - \delta$, we have $L_\infty(d, \lambda) = \max_{ij} |d_{ij} - \lambda_{ij}| < \varepsilon$, where $d$ is the matrix of GM corrected distances, $\lambda$ is the matrix of true distances, $n$ is the number of leaves, and $diam(T)$ is the topological diameter of $T$.*

The significance of Theorem 2.1 is seen in the light of the following theorem:

THEOREM 2.2. *Let $\{d_{ij}\}$ be an $n \times n$ dissimilarity matrix, $\{\lambda_{ij}\}$ an additive matrix defined by a tree $T$ with positive edge-weighting $\lambda(e)$, and $f$ the smallest edge weight in $T$, $f = \min_e w(e)$.*

- *(From [1] and [8]) If $L_\infty(d, \lambda) < f/2$, then the $Q^*$ and Neighbor-Joining methods both return tree $T$ on input $d$.*
- *(From [8]) If $L_\infty(d, \lambda) < f/8$, then the Agarwala* et al. *method returns tree $T$ on input $d$.*

These two theorems imply the following result.

THEOREM 2.3. *Let $T$, $M$, $f$, $g$, $n$, $\delta$, $\lambda$, $d$, and $diam(T)$ be as in Theorem 2.1. Then there is a constant $C > 0$ such that, if the sequence length exceeds*

$$C \log n e^{O(g \cdot diam(T))}$$

*then, with probability at least $1-\delta$, the Neighbor-Joining and $Q^*$ methods recover the true tree. There is also a constant $C' > 0$ such that, if the sequence length exceeds*

$$C' \log n e^{O(g \cdot diam(T))}$$

*then, with probability at least $1-\delta$, the Agarwala method recovers the true tree.*

Since $diam(T)$ can be as large as $n - 1$, the sequence length requirement of each of these methods is bounded from above by a function that grows *exponentially* in $n$, even when $g$ is fixed.

### 2.2 Fast-converging methods.

Since letting $f$ be arbitrarily small or $g$ be arbitrarily large affects the sequence length requirement, we are interested in developing methods for which polynomially long sequences ensure accuracy under the General Markov model, when both $f$ and $g$ are fixed, but arbitrary. In order to define this concept precisely, we first parameterize the General Markov model.

DEFINITION 2. *$GM_{f,g}$ contains those $(T, M) \in GM$ for which $f \leq \lambda(e) \leq g$ holds for all edges $e \in E(T)$.*

We now define two types of "fast convergence."

DEFINITION 3. *A phylogenetic reconstruction method $\Phi$ is* absolute fast-converging (afc) *for the GM model if, for all positive $f, g, \varepsilon$, there is a polynomial $p$ such that, for all $(T, M)$ in the GM model, on set $S$ of $n$ sequences of length at least $p(n)$ generated on $T$, we have $Pr[\Phi(S) = T] > 1 - \varepsilon$.*

Note that method $M$ operates without any knowledge of parameters $f$ or $g$—or indeed any function of $f$ and $g$. Thus, although the polynomial $p$ depends upon both $f$ and $g$, the method itself does not.

DEFINITION 4. *A phylogenetic reconstruction method $\Phi$ is* relative fast-converging (rfc) *for the model $GM_{f,g}$ if, for all positive $f, g, \varepsilon$, there is a polynomial $p$ such that, for all $(T, M) \in GM_{f,g}$ on set $S$ of $n$ sequences of length at least $p(n)$ generated on $T$, we have $Pr[\Phi(S, A(f, g)) = T] > 1-\varepsilon$, where $A(f, g)$ denotes an oracle that provides information about $f$ or $g$.*

Not requiring information about $f$ or $g$ is clearly desirable, since in practice there is no *a priori* way to bound $f$ or $g$. Hence, absolute fast convergence is a more desirable property than relative fast convergence.

However, almost all known fast-converging methods are only rfc, not afc. The only known afc methods are:

- The *Short-Quartet* methods (the *dyadic closure method* [8] and the *witness-antiwitness method* [9]). These methods operate by attempting to produce, for each setting of the parameter $g$, a binary tree that meets certain constraints. They also can explore efficiently (e.g., through binary search) the parameter space of $g$.
- Certain variations of the *Disk-Covering Method (DCM)* [12]. DCM is a two-phase technique used in conjunction with existing phylogenetic methods. The first phase produces a collection of trees, where each tree is obtained by producing a division of the dataset into overlapping subproblems ("disks") of low diameter. Trees on these subproblems are computed using some existing phylogenetic method

and are then merged into a supertree on the entire set of taxa. The second phase selects a tree on the entire set of taxa, based upon a criterion designed specifically with respect to the base phylogenetic method. Using DCM with the Buneman Method (also known as the $Q^*$ method) or with the Naïve Quartet Method (NQM) produces afc methods, because the second phase has provable performance guarantees.

All afc methods operate by producing a set of trees, then selecting the best tree from the set. By designing a suitable selection criterion, one can prove performance guarantees and thus establish absolute fast convergence.

By comparison, rfc methods need to know (the value of or a tight bound on) one or both of the parameters ($f$ and $g$). Absent such knowledge, rfc methods cannot reconstruct the true tree with high probability from sequences of polynomial length. Indeed, rfc methods also operate by producing a set of trees, one for each setting of the unknown but necessary parameter, and then select a tree from the set. Because their selection procedure has no performance guarantees, these methods are not statistically consistent, much less afc. For example, the HGT method [4] or its heuristic modification [5], the CGG method [3], and DCM with neighbor-joining or with the Agarwala method (see [12]), all use heuristics (with no performance guarantees) to select the best tree. The existence of a selection criterion with performance guarantees is thus a crucial ingredient in producing afc methods, prompting us to define the following problem.

True Tree Selection Problem:

- **Input:** A set $S$ of sequences over the nucleotide alphabet $\{A, C, T, G\}$ generated by an unknown GM model tree $(T, M)$ and a collection $\mathcal{T} = \{T_1, T_2, \ldots, T_p\}$ of phylogenies on $S$.
- **Output:** The true tree $T$ if $T$ is in $\mathcal{T}$

In order for an algorithm for True Tree Selection to be useful in producing afc methods, it must itself demonstrate a form of absolute fast convergence.

Definition 5. *An algorithm $\Phi$ for* True Tree Selection *is* absolute fast-converging *under the GM model if, for all $f, g, \varepsilon > 0$, there is a polynomial $p$ such that, for all model trees $(T, M) \in GM_{f,g}$ and for all sets $S$ a set of sequences generated on $(T, M)$ of length at least $p(|S|)$, $T \in \mathcal{T}$ implies $Pr[\Phi(S, \mathcal{T}) = T] > 1 - \varepsilon$.*

## 3 Short Quartet Support

We now present the *Short Quartet Support (SQS)* method and prove that SQS is absolute fast-converging for True Tree Selection under the GM model. We begin with some notation and terminology. Let $T$ be a phylogeny on $S$. We denote by $Q(T)$ the set of quartets on the leaves of $T$ defined by $T$—i.e., quartet $t$ is in $Q(T)$ if and only if the subtree of $T$ induced by the four taxa of $t$ equals $t$. $T$ can be reconstructed from $Q(T)$ in polynomial time [22].

Definition 6. *For a given quartet $q$ on taxa from $S$, define $diam_D(q)$ as*

$$diam_D(q) = max\{D_{i,j} \mid \{i, j\} \subset q\}$$

*In other words, $diam_D(q)$ is the maximum GM distance between the taxa of $q$. For $Q$, a fixed set of quartets, given $D$ we can define the set $Q_w = \{q \in Q : diam_D(q) \leq w\}$.*

Definition 7. *Let $T$ be a fixed tree leaf-labelled by the set $S$, $Q$ a fixed set of quartets on $S$, and $D$ the GM distance matrix on $S$. The* support *of $T$ with respect to $Q$, denoted $s(T, Q)$, is*

$$\max\{l \mid (q \in Q \text{ and } diam_D(q) \leq l) \Longrightarrow q \in Q(T)\}$$

Although we define the support of $T$ with respect to arbitrary sets $Q$ and arbitrary ways of defining the dissimilarity matrices, we will focus our attention on matrices $D$ and sets $Q$ defined in particular ways:

Definition 8. *Let $D$ denote the GM corrected distances. We define $\mathcal{Q}$ to be the set of neighbor-joining quartets on the fourtuples of leaves in $S$ based upon $D$. (Since neighbor-joining on quartets is identical to the Four-Point Method (see [8]) on quartets, this is the same as the set of quartet trees computed in [8, 9] and other papers.)*

We are now ready to present a high-level version of the SQS method. The main theorem of this paper is that SQS is absolute fast-converging under the General Markov (GM) model.

Procedure SQS($\mathcal{T}, S$)

- For each set of four taxa from $S$, compute the neighbor-joining quartet $q$; let $\mathcal{Q}$ be the set of all such quartets.
- Return $T_i \in \mathcal{T}$ such that $s(T_i, \mathcal{Q})$ is maximum; if more than one such tree exists, return the one with the smallest index $i$.

THEOREM 3.1. *SQS is absolute fast-converging under the General Markov model. That is, for all $f, g, \varepsilon > 0$, there is a polynomial $p$ such that, for all $(T, M) \in GM_{f,g}$ on set $S$ of $n$ sequences generated at random by $T$ with length at least $p(n)$, we have*

$$Pr[SQS(\mathcal{T}, S) = T] > 1 - \varepsilon$$

*for all $\mathcal{T}$ such that $T \in \mathcal{T}$.*

We postpone the proof of this theorem to Section 6.

## 4 Applications of SQS

**4.1 Turning rfc Methods into afc Methods:** An alternative way of viewing rfc methods is that they construct a set $\mathcal{T}$ of trees, one for each potential setting of the parameter, then face the problem of selecting the true tree from $\mathcal{T}$. This is a relatively easy task if the value for the parameter is known, but no general solution has been found to the problem when the parameter is unknown. SQS is designed to solve this problem. In particular, SQS can be used with any rfc method as a second phase, turning that method into an afc method.

DEFINITION 9. *Let $\Phi$ be an rfc method. Define $\Phi^*$ to be the method obtained by using $\Phi$ to produce a set $\mathcal{T}$ of trees and using SQS to choose a tree from $\mathcal{T}$.*

THEOREM 4.1. *If $\Phi$ is rfc, then $\Phi^*$ is afc. The additional running time needed by $\Phi^*$ is $O(n^4|\mathcal{T}|)$.*

**4.2 DCM*: Dramatically Reducing the Required Length:** We describe a general-purpose technique for reducing the sequence length requirement of statistically consistent phylogenetic methods. This technique has two phases. The first phase is obtained from the Disk-Covering Method of Huson *et al.* [12]. This phase takes as input a phylogenetic method $\Phi$ and a set $S$ of sequences generated under a model of evolution, and produces a collection $\mathcal{T}$ of trees. The second phase of the technique uses the SQS selection procedure to select the best tree from the set $\mathcal{T}$.

This two-phase procedure, DCM*, has very strong theoretical performance guarantees. In particular, it can be used to create afc methods. For example, if the convergence rate of $\Phi$ is $O(e^{O(\max\{\lambda_{ij}\})})$, then the convergence rate of DCM*-$\Phi$ is $O(n^{O(g)})$, where $g = \max_e \lambda(e)$, so that DCM*-$\Phi$ is afc.

We call the first phase of the DCM method the *MaxClique* method, since the dataset decomposition technique is based upon computing maximal cliques in a graph that we now define:

DEFINITION 10. *Let $D$ be a dissimilarity matrix on the set $S$, and let $w \in R^+$ be given. The* threshold graph *$TG(D, w)$ has vertex set $S$ and edges between those taxa $i, j$ such that $D_{ij} \leq w$. Formally: $TG(D, w) = (S, E_w)$, where $E_w = \{(i, j) \mid D_{ij} \leq w\}$.*

In [12], it was proved that if $D$ is additive, then the threshold graph $TG(D, w)$ is triangulated.

DEFINITION 11. *A graph $G$ is triangulated if every cycle in $G$ of length at least four contains a chord (i.e. a pair of non-sequential adjacent vertices).*

Although our matrices will not in general be additive, they will often be sufficiently close to additive (since they are based upon statistically consistent estimators of GM distances). The significance of this is that finding maximal cliques (i.e. maximal subsets of vertices, every pair of which are adjacent) in triangulated graph is solvable in polynomial time, although hard for the general case.

We describe the MaxClique procedure below.

PROCEDURE MAXCLIQUE

- **Input**: The GM distance matrix $D$, a set $S$ of sequences generated on an unknown GM tree $(T, M)$, and a phylogenetic base method $\Phi$.
- **Output:** A set $\mathcal{T}$ of trees on $S$ obeying $Pr[T \in \mathcal{T}] > 1 - \varepsilon$ if the sequence length exceeds $p(|S|)$ (where $p$ is a polynomial).
- **Algorithm:**
  For each $w \in D_{ij}$:
  - Let $E_w = \{(i, j) \mid D_{ij} \leq w\}$. Construct the threshold graph, $TG(D, w) = (S, E_w)$.
  - If $TG(D, w)$ is not connected, then let $T_w$ be the star tree (i.e., the tree with one internal node attached to each of the $n$ leaves) and exit.
  - Triangulate $TG(D, w)$, minimizing $\max\{D_{ij} \mid (i, j) \text{ added to } (S, E_w)\}$, thus producing the triangulated graph $TG^*(D, w)$.
  - Compute the maximal cliques $C_1, C_2, \ldots, C_l$ of $TG^*(D, w)$ (note $l \leq n$). For each $i$, $1 \leq i \leq l$, let $t_i = \Phi(C_i)$.
  - Merge the subtrees $t_i$ using the *Strict Consensus Merger* [12]. Let $T_w$ be the resulting tree and exit.

  Return $\mathcal{T} = \{T_w \mid w \in \{D_{ij}\}\}$.

THEOREM 4.2. *Assume $(T, M) \in GM_{f,g}$, $S$ a set of sequences generated on $T$, $D$ the GM distance matrix for $S$, and $\varepsilon > 0$. If $\Phi$ is a phylogenetic method with a convergence rate on tree $T$ of $O(e^{O(\max\{\lambda_{ij}\})})$, then MaxClique-$\Phi$ is* rfc.

A stronger version of this theorem, which shows an improvement in the convergence rate for *all* phylogenetic methods whose sequence length requirements depend upon the maximum evolutionary distance $\max \lambda_{ij}$ in the tree, is provided in Section 6. Lemma 6.2 on the conditions under which the *Strict Consensus Merger* is guaranteed to yield the true tree appears in Section 6 as well, since it depends on terminology and theory developed in that section.

For any phylogenetic reconstruction method $\Phi$, we denote by DCM*-$\Phi$ the method obtained by applying this two-phase process:

- Phase 1: Given phylogenetic method $\Phi$, set $S$ of sequences generated on an unknown GM tree, and the GM distance matrix $D$, compute the set $\mathcal{T}$ of trees on $S$ using MaxClique-$\Phi$.
- Phase 2: Compute $\mathcal{Q}$, the set of neighbor-joining quartets on $S$, based upon $D$, then return $SQS(\mathcal{T}, \mathcal{Q})$.

THEOREM 4.3. *Let $(T, M) \in GM_{f,g}$, $S$ a set of sequences generated on $T$, $D$ the GM distance matrix for $S$, and let $\varepsilon > 0$ be given. If $\Phi$ is a phylogenetic method with convergence rate on tree $T$ bounded from above by $O(e^{O(\max\{\lambda_{ij}\})})$, then DCM*-$\Phi$ is* afc.

*Proof.* Follows from Theorem 4.2 and Theorem 3.1.

**4.3 WIGWAM: New rfc and afc methods:** This general two-phase structure for afc methods can be used to design new afc methods. We present such a method, which we call the *WeIGhted Witness-Antiwitness Method (WIGWAM)*. WIGWAM runs in polynomial time, is afc, and by design produces a binary tree on every input; hence, it improves upon some afc methods that tend to return unresolved trees. WIGWAM considers all quartets, but weighs them according to their statistical support (as indicated by $diam_D$); consequently, it should have better performance than quartet methods that do not distinguish between quartets [20].

WIGWAM is a modification of an earlier afc method called the *Witness-Antiwitness Method (WAM)* [9]. We therefore present the WAM method first, and then show how WIGWAM differs from WAM. The proof that WIGWAM is afc is slightly more complicated than the proof for WAM but is quite similar. Due to space limitations, we will only suggest how the proof goes; see [9] for the proof that WAM is afc.

WAM has two phases. In the first phase, it uses an algorithm called WATC (Witness-Antiwitness Tree Construction) to compute a set $\mathcal{T}$ of trees, one for each setting of the parameter $w$ (quartet width). In the second phase, it selects a tree from $\mathcal{T}$ as the true tree. The technique used to construct $\mathcal{T}$ is provably rfc and that used to select the tree from $\mathcal{T}$, while different from SQS, is provably afc, so that WAM is afc.

WIGWAM differs from WAM in the implementation of the first phase and then uses SQS to select the best tree from $\mathcal{T}$. We will focus therefore on describing WATC and show how our version (WIGWATC) operates.

Recall that $S$ is the set of sequences generated on an unknown GM model tree, $D$ is the GM distance matrix, and $\mathcal{Q}$ is the set of NJ quartet trees computed using the GM matrix. Finally, recall that for $q \in \mathcal{Q}$, $diam_D(q)$ is the maximum GM distance between the leaves of $q$. We take the following definition from [9]:

DEFINITION 12. *Let $e$ be an internal edge of $T$ and assume that the removal of $e$ (and its endpoints) from $T$ decomposes $T$ into four subtrees, two of which are $T_1$ and $T_2$. A quartet $\{ab|cd\} \in \mathcal{Q}$ is a* witness *for the siblinghood of $T_1$ and $T_2$ if we have $a \in T_1$, $b \in T_2$ and $c, d \in T - (T_1 \cup T_2)$. A quartet $\{ef|gh\} \in \mathcal{Q}$ is an* antiwitness *for the siblinghood of $T_1$ and $T_2$ if we have $e \in T_1$, $g \in T_2$ and $f, h \in T - (T_1 \cup T_2)$. A* $w$-witness *($w$-*antiwitness*) is a witness (anti-witness) $q$ obeying $diam_D(q) \leq w$.*

WATC takes as input the GM dissimilarity matrix $D$ and a parameter $w$ and proceeds like neighbor-joining, in that it constructs the tree from the "outside-in." Initially every leaf is its own rooted subtree. WATC then repeatedly pairs two rooted subtrees until there are only three subtrees left, at which point the three rooted subtrees are joined into a star. The resulting unrooted tree is the output of WATC. WATC chooses two subtrees to pair from a list of candidates—any candidate can be chosen; two rooted subtrees $t$ and $t'$ are candidates for pairing if and only if there is a $w$-witness and no $w$-antiwitness to their siblinghood. If no candidate pair can be found at some stage, WATC returns the star-tree and exits; otherwise WATC will return an unrooted binary tree. This simple technique is provably rfc [9].

Our modified technique, WIGWATC, uses two separate diameter bounds, one ($W$) for witnesses and one ($A$) for antiwitnesses. Two subtrees are then candidates for pairing only when there is at least a $W$-witness and no $A$-antiwitness. Unlike WATC, which picks an arbitrary candidate, WIGWATC chooses the "best" candidate pair—that which has the least evidence against its pairing. Formally, WIGWAM ranks antiwitnesses according to their diameter, with the antiwitness of least diameter (the "shortest" antiwitness) having the most importance; WIGWAM then selects the candidate pair (i.e. pair of subtrees which are candidates) to maximize the diameter of its shortest antiwitness.

We describe WIGWATC and WIGWAM below.

PROCEDURE WIGWATC:

- **Input:** Two diameter bounds $W$ and $A$, and the input given to WIGWAM, i.e. a set $S$ of taxa, a set $Q$ of quartet trees (that includes a quartet on every four taxa in the set $S$), and a dissimilarity matrix $D$ on $S$.
- Start with each taxon defining a subtree.
- While at least four subtrees remain do:
    - Form the graph $G$ on the set of subtrees (one vertex per subtree), where two vertices are joined by an edge if the corresponding pair of subtrees has at least one $W$-witness and no $A$-antiwitness; define the cost of each edge to be the reciprocal of the diameter of the shortest antiwitness for that edge, 0 otherwise.
    - If no edge is created, return failure.
    - Choose the edge of least cost, merge its two endpoints and the two corresponding subtrees, and update edge costs as necessary (the merging may eliminate antiwitnesses for other edges).
- Merge the remaining trees in the obvious way and return the resulting binary tree.

The difficult part of WIGWAM is to identify suitable values for $W$ and $A$ such that the tree returned for these values is the true the true tree. In its simplest form, WIGWAM searches (nearly) linearly through all $O(n^2)$ possible values of quartet diameters for each of the parameters. Roughly speaking, it starts with a value $w$ for $W$ that ensures that the threshold graph $TG(D, w)$ is connected, then steps through values either one at a time (when the support of the tree returned is not as good as the current diameter) or by skipping over an entire range of values (when the support is larger than the current diameter). The key to the procedure is the updating of the threshold for antiwitnesses in terms of the level of support for the last tree generated. Once a certain level of support $l$ has been reached, no antiwitness of diameter less than or equal to $l$ is tolerated—the new tree under construction is forced to agree with all quartets of diameter less than or equal to $l$. This constraint may prevent the building of a tree, but note that the next iteration, with an unchanged value for support, but a larger threshold for witnesses, may find new candidates that were unavailable at the current iteration and thus may succeed in a building a tree that obeys the constraints. Since a tree that obeys all short quartets is the true tree, WIGWAM eventually produces a true tree if given all of its short quartets. WIGWAM is typical of rfc methods in that it returns a collection of trees, up to one per value of the parameter (here the threshold on the diameter of quartets), but our support-based mechanism incorporates the SQS principle directly into the procedure, so that WIGWAM is afc.

PROCEDURE WIGWAM

- **Input:** A set $S$ of taxa, a set $Q$ of quartet trees (that includes a quartet on every four taxa in the set $S$), and a dissimilarity matrix $D$ on $S$.
- Compute the smallest diameter $d$ such that the initial threshold graph $TG$ (with one node for each taxon) for $(D, d)$ is connected. Set $W = d$ and $A = 0$.
- Repeat until $W$ equals the largest value in $D$:
    - Call WIGWATC with parameters $W$ and $A$.
    - If WIGWATC returns failure, increase $W$ to the next larger value in $D_{ij}$.
    - Otherwise compute the support $s(T, Q_W)$ of the returned tree $T$ with respect to the set $Q_W$ of quartets and set $A = s(T, Q_W)$. If the support is larger than $W$, set $W = s(T, Q_W)$; otherwise increase $W$ to the next larger value in $D_{ij}$.

Correctness follows from the fact that if $s(T, Q_W) = s_0 \geq W$, then a tree with support at least $s_0$ can be built by WIGWATC for each threshold value from $W$ to $s_0$.

## 5 Conclusions

We have introduced the concept of relative *vs.* absolute fast convergence and its associated problem TRUE TREE SELECTION. We have given a simple, polynomial-time technique, SQS, for solving this problem, a technique that turns rfc methods into afc methods and that can be used, particularly in conjunction with the DCM method, for designing new afc methods. As an example, we have introduced a new afc method, WIGWAM, which uses SQS to select its tree.

There is significant experimental evidence that the technique used to select the best tree from the set $\mathcal{T}$ has a dramatic impact on the topological accuracy of the tree returned in phase II. For example, the studies by Csűrös and Kao on their rfc method HGT [5] and by Huson *et al.* on their rfc method DCM-NJ [12] both showed that, under some conditions, the set $\mathcal{T}$ could contain the true tree or a very close approximation thereof, but that $\mathcal{T}$ also contained trees with topological error rates higher than 50%. Thus, designing methods with provable performance guarantees is extremely important, both from a theoretical and a practical point of view.

We note that the set $\mathcal{T}$ does not need to contain a tree for each setting of its unknown parameter. In [9], Erdős *et al.* presented a technique called *sparse-high* for use with their fast-converging WAM, a technique which only requires computing a small number of trees. Hence rfc methods need not be computationally intensive.

Our results prompt new questions: are there other afc methods for TRUE TREE SELECTION? In particular, is maximum likelihood (MLE) such a method (see, e.g., [18] for overview)? Are there generic techniques for speeding up SQS? (Existing methods such the sparse high search mentioned above are very much tailored to the base phylogenetic method.)

## 6 Proofs

We prove Theorem 3.1, and provide a stronger version of Theorem 4.2.

THEOREM 3.1. *The SQS Procedure is absolute fast-converging under the GM model.*

Our proof proceeds by first establishing a condition under which we have $s(T, \mathcal{Q}) > s(T', \mathcal{Q})$ for all $T' \neq T$. We then show that this condition holds with probability $1-\varepsilon$ from sequences of length bounded by a polynomial in $n$, for fixed $f, g, \varepsilon > 0$. Since the SQS procedure does not consider the values of $f$ or $g$, this will prove that it is afc.

Let $(T, M) \in \mathrm{GM}_{f,g}$ be a given GM tree. Recall that $\{\lambda_{ij}\}$ denotes the matrix of true evolutionary distances leaves the taxa in $S$. Let $S$ be a set of taxa generated by $T$ at the leaves of $T$, and let $D_{ij}$ be the GM corrected distance between $i$ and $j$. As mentioned earlier, for all $i, j$, $D_{ij}$ converges to $\lambda_{ij}$ as the sequence length $k$ increases. Recall that $\mathcal{Q}$ denotes the set of neighbor-joining quartets computed on all fourtuples of leaves from $S$ and that $diam_\lambda(q)$ denotes $\max\{\lambda_{ij} \mid \{i,j\} \subset q\}$.

We define various quantities with respect to the two matrices, $\{\lambda_{ij}\}$ and $\{D_{ij}\}$. For each internal edge $e \in E(T)$, the deletion of $e$ from $T$ breaks $T$ into four subtrees, $T_1, T_2, T_3, T_4$. Set

$$w_\lambda(e) = \min\{diam_\lambda(a_1, a_2, a_3, a_4)\},$$

where the minimum is taken over all fourtuples $(a_1, a_2, a_3, a_4)$ with $a_i$ a leaf in $T_i, i = 1,2,3,4$. The *short quartets* around $e$ are formed from taxa $a_1, a_2, a_3, a_4$ with $a_i \in T_i$, so as to minimize $diam_\lambda(a_1, a_2, a_3, a_4)$. Set $w_\lambda(T) = \max_e w_\lambda(e)$.

DEFINITION 13. *The set of short quartet trees of the tree $T$, denoted $Q^*_{short}(T)$, is the set of true trees (i.e., subtrees of the true tree $T$) on those fourtuples of $S$ obeying $diam_\lambda(q) \leq w_\lambda(T)$.*

Our interest in short quartets is based upon the following theorem that was established in [6]:

LEMMA 6.1. *Let $(T, w)$ be a fixed edge-weighted tree leaf-labelled by $S$ and let $T'$ be another tree leaf-labelled by $S$. If we have $Q^*_{short}(T) \subseteq Q(T')$, then $T$ equals $T'$.*

In the definition of the DCM* method, we stated that subtrees on maximal cliques are computed and then merged using a technique called the *Strict Consensus Merger*. This technique is roughly as follows (see [12] for details): merge the trees computed in a pairwise fashion, until all the subtrees are merged into a single tree on the full set of taxa. During each merger, contract, if necessary, a minimum number of edges in order to make the pair of trees agree on the subtrees induced by their common leaf sets, before merging the trees. (The order in which subtrees are merged matters, and follows the perfect elimination ordering given for the triangulation of the threshold graph; see [12] for details). We now characterize the conditions under which the strict consensus merger is guaranteed to be correct.

LEMMA 6.2. *Let $(T, w)$ be an edge-weighted tree on leaf set $S$ and let $G$ be a triangulated graph on vertex set $S$ such that each short quartet of $T$ induces a four-clique in $G$. Assume that, for each maximal clique $C$ of $G$, we have the true tree $t_C$ induced in $T$ by $C$. Then the Strict Consensus Merger of the trees $t_C$ produces $T$.*

*Proof.* The proof of Theorem 4 in [12] proves this assertion when the trees on maximal cliques are obtained by using the Buneman Tree method; however, the proof of that theorem does not depend upon the use of the Buneman Tree method, but only upon the trees being correct. Consequently, this lemma can be proved using the same proof.

Recall that for a fixed set $Q$ of quartets, $Q_w$ is the set of quartets $q \in Q$ obeying $diam_D(q) \leq w$.

DEFINITION 14. *The* width *of $T$ with respect to $D$ is* $w_D(T) = \max\{diam_D(q) \mid q \in Q^*_{short}(T)\}$.

(In words, $w_D(T)$ is the maximum GM corrected distance between any two leaves $i, j$ that appear in a short quartet of $T$.) Note that membership in a short quartet depends upon the true distance $\lambda$, but that $w_D(q)$ is based upon the GM corrected distance matrix $D$. Let $(T, M)$ be a GM tree, with $D$ the GM corrected distance matrix, and let $\lambda$ denote the true distance matrix. We define

$$L^{(q)}_\infty(D, \lambda) = \max\{|D_{ij} - \lambda_{ij}| \colon \min\{D_{ij}, \lambda_{ij}\} \leq q\}$$

We continue with a lemma established in [8].

LEMMA 6.3. *If we have $L^{(w)}_\infty(D,\lambda) < f/2$, then neighbor-joining will return the correct tree for every quartet $q$ with $diam_D(q) \leq w$.*

Now the following lemma follows easily.

LEMMA 6.4. *If $L^{(w)}_\infty(D,\lambda) < f/2$, then $s(T, \mathcal{Q}) \geq w$.*

*Proof.* If $L^{(w)}_\infty(D,\lambda) < f/2$, then $\mathcal{Q}_w \subseteq Q(T)$, and $s(T, \mathcal{Q}) \geq w$.

We can now prove:

COROLLARY 6.1. *If $L^{(w_D(T))}_\infty(D,\lambda) < f/2$, then we have $s(T', \mathcal{Q}) \geq w_D(T)$ if and only if $T'$ equals $T$.*

*Proof.* Follows from Lemma 6.4 and Lemma 6.1.

We are now ready to state the condition under which the SQS procedure is guaranteed to be correct.

THEOREM 6.1. *Set $(T, M) \in GM_{f,g}$, $S$ the set of sequences generated on $T$, and $D$ the GM distance matrix defined on $S$. Let $\mathcal{Q}$ denote the set of neighbor-joining quartets on $S$ based on $D$. If we have $L^{(w_D(T))}_\infty(D,\lambda) < f/2$, then also we have $s(T, \mathcal{Q}) > s(T', \mathcal{Q})$ for all $T \neq T'$.*

*Proof.* Follows from Corollary 6.1.

We now analyze the sequence length that suffices for the above condition to hold with high probability.

LEMMA 6.5. *Let $(T, M) \in GM_{f,g}$ and let $\varepsilon > 0$ be an arbitrary constant. Let $D$ be the GM corrected distance matrix defined on a set of sequences generated on $T$. Then there is a constant $c > 0$ such that, if the sequence length $k$ exceeds $c \cdot n^{O(g)}$, then we have*

$$Pr[L^{(w_D(T))}_\infty(D,\lambda) < f/2] > 1 - \varepsilon$$

*Proof.* Proof omitted due to space constraints, but follows from results in [12].

This concludes the proof that SQS is fast-converging under the GM model (proof of Theorem 3.1).

We now provide the stronger version of Theorem 4.2. from which Theorem 4.2 follows as a special case.

THEOREM 6.2. *Let $(T, M) \in GM_{f,g}$, $S$ a set of sequences generated on $T$, and $D$ the GM distance matrix for $S$. Let $\varepsilon > 0$ be given and assume that $\Phi$ is a phylogenetic method with convergence rate on tree $T$ bounded from above by $F(\max\{\lambda_{ij}\})$ for some function $F$. Then DCM*-$\Phi$ converges from sequences of length bounded by $F(O(w_\lambda(T)))$. Consequently, if $F(x)$ is $O(e^{O(x)})$, then DCM*-$M$ is afc.*

*Proof.* (Sketch:) The tree $T_{w_D}$ computed by the Max-Clique method is based upon the base method applied to subproblems of maximum GM distance $w_D$. When the sequences are of polynomial length, these subproblems have maximum true distance bounded by $O(w_\lambda)$. Thus, the convergence rate bound for the method $M$ guarantees that $M$ is accurate on each subproblem with high probability if the sequence length is at least $F(O(w_\lambda))$, the maximum evolutionary distance in each subproblem.

The threshold graph contains $TG(D, w_D(T))$, so that every short quartet induces a four-clique in the triangulation of this graph. Thus, since each subtree is correctly reconstructed, Lemma 6.2 implies that the strict consensus merger produces the true tree, proving the first assertion.

Finally, if $F(x)$ is $O(e^{O(x)})$, then DCM*-$M$ converges from sequences of length at most $O(e^{O(w_\lambda(T))})$. In [9], $w_\lambda(T)$ was established to be bounded by $O(g \log n)$, from which the result follows.

## 7 Acknowledgements

The authors wish to thank the University of New Mexico's Department of Computer Science for hosting TW and KS in Summer 2000, during which time these results were obtained.

## References

[1] Atteson, K., "The performance of the neighbor-joining methods of phylogenetic reconstruction," *Algorithmica* **25**, 2/3 (1999), 251–278.

[2] Berry, V., Bryant, D., Jiang, T., Kearney, P., Li, M., Wareham, T., and Zhang, H., "A practical algorithm for recovering the best supported edges of an evolutionary tree," preprint, 2000.

[3] Cryan, M., Goldberg, L., and Goldberg, P., "Evolutionary trees can be learned in polynomial time in the two-state general Markov model," *Proceedings FOCS* (1998), 436–445.

[4] Csűrös, M., and Kao, M.Y., "Recovering evolutionary trees through harmonic greedy triplets," In *Proc. 10th Annual ACM-SIAM Symp. on Discrete Algorithms* (1999), 261–270.

[5] Csűrös, M., and Kao, M.Y., "$O(n^2 logL)$-time accurate recovery of evolutionary trees with more than one thousand leaves: an experimental combination of harmonic greedy triplets and the minimum evolution principle," preprint, Yale U., 1999.

[6] Erdős, P.L., Steel, M., Székély, L., and Warnow, T., "Local quartet splits of a binary tree infer all quartet splits via one dyadic inference rule," *Computers and Artificial Intelligence* **16**, 2 (1997), 217–227.

[7] Erdős, P.L., Steel, M., Székély, L., and Warnow, T., "Inferring big trees from short sequences," *Proc. Int'l Congress on Automata, Languages, and Programming ICALP97*, Bologna, Italy (1997).

[8] Erdős, P.L., Steel, M., Székély, L., and Warnow, T., "A few logs suffice to build almost all trees—I," *Random Structures and Algorithms* **14** (1997), 153–184.

[9] Erdős, P.L., Steel, M., Székély, L., and Warnow, T., "A few logs suffice to build almost all trees—II," *Theor. Comput. Sci.* **221** (1999), 77–118.

[10] Huson, D., Nettles, S., Parida, L., Warnow, T., and Yooseph, S., "A divide-and-conquer approach to tree reconstruction," Workshop on Algorithms and Experiments (ALEX98), Trento, Italy, 1998.

[11] Huson, D., Nettles, S., Rice, K., Warnow, T., and Yooseph, S., "The hybrid tree reconstruction method," *ACM J. Experimental Algorithmics* **4**, 5 (1999), `www.jea.acm.org/1999/HusonHybrid/`.

[12] Huson, D., Nettles, S., and Warnow, T., "Disk-covering, a fast-converging method for phylogenetic tree reconstruction," *J. Comput. Biol.* **6**, 3 (1999), 369–386.

[13] Jukes, T.H., and Cantor, C. *Mammalian Protein Metabolism.* Academic Press, New York, 1969; In chapter "Evolution of protein molecules," 21–132.

[14] Li, W.-H. *Molecular Evolution.* Sinauer, Massachusetts, 1997.

[15] Penny, D., and Steel, M.A, "Distributions of tree comparison metrics—some new results," *Syst. Biol.* **42**, 2 (1993), 126–141.

[16] Rice, K., and Warnow, T., "Problems with big trees," SSE/SSB Joint meetings, Vancouver, British Columbia, 1998.

[17] Saitou, N., and Nei, M., "The neighbor-joining method: A new method for reconstructing phylogenetic trees." *Mol. Biol. & Evol.* **4** (1987), 406–425.

[18] Setubal, J., and Meidanis, J. *Introduction to Computational Molecular Biology.* PWS Publishing Co, 1997.

[19] Steel, M.A., "Recovering a tree from the leaf colourations it generates under a Markov model," *Appl. Math. Lett.* **7** (1994), 19–24.

[20] St. John, K., Warnow, T., Moret, B.M.E., and Vawter, L., "Performance study of phylogenetic methods: (unweighted) quartet methods and neighbor-joining," to appear in SODA 2001.

[21] Strimmer, K., and von Haeseler, A., "Quartet puzzling: A quartet maximum likelihood method for reconstructing tree topologies," *Mol. Biol. & Evol.* **13** (1996), 964–969.

[22] Warnow, T., "Some combinatorial problems in phylogenetics," Invited paper, *Proc. Int'l Colloq. Combinatorics and Graph Theory*, Balatonlelle, Hungary (1996), in Vol. 7 of *Bolyai Society Mathematical Studies*, A. Gyárfás, L. Lovász, and L.A. Székély, eds., Budapest, 363–413 (1999).