BIOINFORMATICS

Vol. 24 ISMB 2008, pages i123–i131
doi:10.1093/bioinformatics/btn149

Efficient inference of bacterial strain trees from genome-scale multilocus data

C. Than [1], R. Sugino [2], H. Innan [2] and L. Nakhleh [1,*]

[1]Department of Computer Science, Rice University, Houston, TX 77005, USA and [2]Department of Evolutionary Studies of Biosystems, Graduate University for Advanced Studies, School of Advanced Sciences, Hayama, Kanagawa 240-0193, Japan

ABSTRACT

Motivation: In bacterial evolution, inferring a *strain tree*, which is the evolutionary history of different strains of the same bacterium, plays a major role in analyzing and understanding the evolution of strongly isolated populations, population divergence and various evolutionary events, such as horizontal gene transfer and homologous recombination. Inferring a strain tree from multilocus data of these strains is exceptionally hard since, at this scale of evolution, processes such as homologous recombination result in a very high degree of gene tree incongruence.

Results: In this article we present a novel computational method for inferring the strain tree despite massive gene tree incongruence caused by homologous recombination. Our method operates in three phases, where in phase I a set of candidate strain-tree topologies is computed using the *maximal cliques* concept, in phase II divergence times for each of the topologies are estimated using *mixed integer linear programming* (MILP) and in phase III the optimal tree (or trees) is selected based on an optimality criterion. We have analyzed 1898 genes from nine strains of the *Staphylococcus aureus* bacteria, and identified a fully resolved (binary) strain tree with estimated divergence times, despite the high degrees of sequence identity at the nucleotide level and gene tree incongruence. Our method's efficiency makes it particularly suitable for analysis of genome-scale datasets, including those of strongly isolated populations which are usually very challenging to analyze.

Availability: We have implemented the algorithms in the PhyloNet software package, which is available publicly at http://bioinfo.cs.rice.edu/phylonet/

Contact: nakhleh@cs.rice.edu

1 INTRODUCTION

Genome sequencing technologies are amassing large amounts of data from various organisms that span the Tree of Life, and in the case of bacteria, genomes of several strains of the same bacterium are becoming available (e.g. see the Microbial Genome Project of the US Department of Energy at http://microbialgenomics.energy.gov/). These data are enabling biologists to analyze the relationships among populations and species, as well as understand speciation and population divergence. To elucidate these relationships and understand these processes among different strains of the same bacterium, an accurate reconstruction of the evolutionary history of these strains—the *strain tree*—is essential, since it serves as the backbone against which events such as horizontal gene transfer and homologous recombination can be identified and assessed. In a sequence of papers, Roger Milkman and co-workers pioneered some of the work in this area, mainly focusing on mapping the 'clonal ancestry' in several strains of *Escherichia coli* (e.g. Milkman and Stoltzfus, 1988; Stoltzfus *et al.*, 1988).

In this article, we focus on the problem of inferring the strain tree from a genome-scale set of gene trees whose incongruence is mainly due to *homologous recombination*. In bacteria, homologous recombination through transformation or conjugation allows for the integration of homologous alien DNA into a host genome (Errington *et al.*, 2001). This process plays an important role in DNA repair as well as bacterial genome diversification.

From an evolutionary perspective, and barring any recombination, the evolutionary history of a set of genomes would be depicted by a tree that is the same tree that models the evolution of each gene in these genomes. However, homologous recombination among bacteria decouples the evolution of different genes in their genomes, thus resulting in incongruent (or, discordant) gene trees—a scenario that is illustrated in Figure 1.

For example, in Figure 1c, looking backwards in time, the gene lineage from strain A and the gene lineage from B persist deep enough into the past that they have not coalesced by the time of the ancestral strain to A, B and C. Thus, the lineage from B may coalesce with the lineage from C more recently than with the lineage from A. As the ancestries of different parts of the genome may take different paths through the phylogeny, e.g. due to homologous recombination, gene trees may differ in topology from the strain tree topology, and an individual gene history might not reflect the shape of the strain tree. Even if this gene history is correctly estimated, the strain-tree estimate based on a single locus may be incorrect. As genome-scale sequence data from thousands of loci in different strains of bacteria become available, it is now critical that appropriate methods and tools be developed for understanding and overcoming the problem of gene-tree discordance in strain-tree inference.

A few methods have been introduced recently for analyzing gene trees, reconciling their incongruities and inferring species trees despite these incongruities. To the best of our knowledge, none of these methods have been applied to bacterial genomes, particularly different strains of the same bacterium, with massive gene tree incongruence due to homologous recombination. Generally speaking, each of these methods follows one of two approaches: the *combined analysis* approach or the *separate analysis* approach. In the combined analysis approach, the sequences from multiple loci are concatenated, and the resulting 'supergene' dataset is analyzed using traditional phylogenetic methods, such as maximum parsimony and maximum likelihood (e.g. Rokas *et al.*, 2003.) In the separate analysis approach, the sequence data from each locus is first analyzed individually, and a reconciliation of the gene trees is then

*To whom correspondence should be addressed.

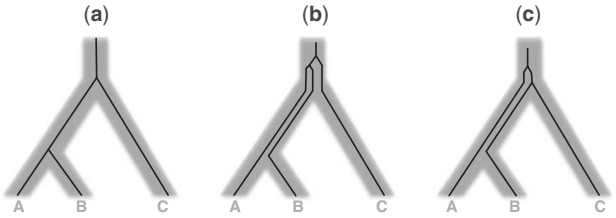

Fig. 1. Three different gene trees within the branches of a strain tree. (**a**) Coalescent times coincide with divergence times, and strain/gene tree topologies are concordant. (**b**) Coalescent times do not coincide with divergence times, and strain/gene-tree topologies are concordant. (**c**) Coalescent times do not coincide with divergence times, and strain/gene-tree topologies are discordant. When topology (tree shape) alone is considered, the gene trees in (a) and (b) are identical, and differ from the one in (c). However, when times are also taken into consideration, all three gene trees are different.

sought so as to optimize certain criterion (e.g. Edwards *et al.*, 2007). Shortcomings of both approaches have been recently reported by various researchers (e.g. Degnan and Rosenberg, 2006; Kubatko and Degnan, 2007). A particular challenge that was not addressed in these recent studies concerns the analysis of very closely related groups of genomes (strains of the same bacterium, for example). In this case, sequence identity at the nucleotide level is very high, which gives rise to gene trees with low resolution, a fact that further complicates the task of inferring the strain tree. Last but not least, the genomic scale of the available data necessitates the development of efficient tools for tackling the task of strain tree inference.

In this article, we address the problem of strain-tree inference from genome-scale multilocus data, where gene-tree incongruence is due to homologous recombination. Our proposed model of the optimal strain tree (topology and divergence times) is one that minimizes the amount of deep coalescent events, which is similar to that used in Maddison and Knowles (2006), and our proposed method to infer the optimal tree under this model is based on two widely encountered optimization problems: maximal cliques and mixed integer linear programming. Our method operates in three phases, where in phase I a set of candidate tree topologies is computed using the maximal cliques concept, in phase II divergence times for each of the topologies are estimated using mixed integer linear programming (MILP) and in phase III one tree, or a set of trees, is selected based on an optimality criterion. To assess our method's performance, we have analyzed 1898 genes from nine strains of the *Staphylococcus aureus* bacteria. A compatibility graph of all different clusters in 1898 corresponding gene trees was built, whose maximal cliques were then computed to reconstruct candidate tree topologies. The compatibility graph has 36 vertices and 304 edges, which correspond to 304 pairs of compatible clusters, and all its maximal cliques were identified in about 0.046 s. For the 24 trees that corresponded to the maximal cliques, we computed divergence times using a novel MILP formulation, which we solved using the CPLEX tool from ILOG. It took CPLEX approximately 1 h to compute the optimal divergence time assignment of a strain-tree topology, given a set of 1898 gene trees, on a 3.2 GHz Intel Pentium 4 machine, running Linux, with 1GB of RAM. The optimal strain tree that our method identified is fully resolved (binary) despite the high degree of sequence identity at the nucleotide level, which further affirms the suitability of the method to analysis of very closely related organisms.

2 METHODS

2.1 Definitions and notations

Let $T=(V,E)$ be a tree, where $V(T)$ and $E(T)$ are the *tree nodes* and *tree edges* (or, *tree branches*), respectively, and let $\mathscr{L}(T)$ denote its leaf set. Further, let $\mathscr{X}$ be a set of taxa. Then, T is a phylogenetic tree over $\mathscr{X}$ if there is a bijection between $\mathscr{X}$ and $\mathscr{L}(T)$. A tree T is said to be *rooted* if the edges in E are directed and there is a single internal node x with in-degree 0. In this article, we assume only rooted trees, unless stated otherwise. Let $T=(V,E)$ be a rooted tree, and u be a node in V. Given a tree $T=(V,E)$ leaf-labeled by set $\mathscr{X}$ of taxa, a node $v\in V$, an edge $e=(u,v)$ and a set $X\subseteq\mathscr{X}$, we use the following notations: $p^T(v)=u$; $T[v]$ is the *clade*, or subtree, rooted at node v; c_e^T is the cluster, i.e. the set of leaves of $T[v]$; and, $MRCA_T(X)$ is the *most recent common ancestor* of X—i.e. the node $v\in V(T)$ such that $X=c_e^T$ where $e=(u,v)\in E(T)$. Tree T induces the set $C_T=\{c_e^T : e\in E(T)\}$ of clusters. The topology of the tree T naturally defines a partial order $\subseteq_T$ on C. In this article, we assume that any strain tree T always has a special node r with a special edge $re=(r,x)$, where x is the MRCA of all leaves in the tree (e.g. see the strain tree in Fig. 4a).

Let $\tau : V(T)\rightarrow(\mathbb{R}^+\cup\{0\})$ be a function assigning each node a time such that (1) $\tau(u)=\tau(v)$ for $u,v\in\mathscr{L}(T)$ and (2) $\tau(u)>\tau(v)$ for $(u,v)\in E(V)$.

2.2 Strain-tree inference and gene-tree reconciliation

As indicated above, our proposed model of the optimal strain tree (topology and divergence times) is one that minimizes the amount of deep coalescent events. The input to our problem is a set of gene trees (topologies and coalescent times), and the output is a strain tree (topology and divergence times) that minimizes the amount of deep coalescent events and incongruence of the gene trees when reconciled within the branches of the inferred strain tree. The strain tree is built in three phases. First, a set of topology candidates is computed based on the set of clades in the input gene trees. Second, the times for nodes in each of the candidate trees are inferred based on the coalescent times of the input gene trees. Third, the gene trees are reconciled within the branches of each of the tree candidates, and the tree (topology and times) that optimizes a certain criterion (a weighted sum of deep coalescent events, gene/strain-tree incongruence and shallow coalescent events) is selected as the strain tree.

2.2.1 Phase I: inferring strain-tree topology candidates Given a set of gene-tree topologies $\{T_1,\ldots,T_k\}$, it may be that the tree topology that represents the most frequent coalescent history does not reflect the true divergence patterns (Degnan and Rosenberg, 2006). Further, the tree built from the concatenated 'supergene' may also not reflect the true speciation patterns (Kubatko and Degnan, 2007). Our working hypothesis is that the strain-tree topology is most probably formed from a set of clusters, each of which appears in at least one of the gene trees. For a set of clusters to define a (rooted) tree, they have to be *pairwise compatible*. Two clusters (sets of taxa) $c_1, c_2\subseteq\mathscr{X}$ are compatible if at least one of the three intersections $c_1\cap c_2$, $c_1\cap\overline{c_2}$ and $\overline{c_1}\cap c_2$ is empty ($\overline{c}$ denotes the set $\mathscr{X}-c$). A classical result in phylogenetics states that a set of pairwise compatible clusters defines a unique tree (Semple and Steel, 2003). Based on our working hypothesis and the relationship between clusters and trees, we formulate our heuristic algorithm for finding candidate strain-tree topologies from the set of gene-tree topologies, as outlined in Figure 2. The algorithm first computes C, the set of all clusters appearing in any of the gene trees. It then builds the compatibility graph $H=(V_H,E_H)$, where $V_H=C$, and $E_H\subseteq V_H\times V_H$ where $E_H=\{(c_i,c_j) : c_i \text{ is compatible with } c_j\}$. Based on the aforementioned relationship between clusters and trees, our next task entails computing all maximal sets of pairwise compatible clusters, which amounts to computing the set K of all maximal cliques in the compatibility graph H. Finally, strain tree topology candidates are constructed in a straightforward manner from the set K, where each maximal clique corresponds to a unique tree. Figure 3 illustrates the algorithm on three input gene trees. The set C contains seven distinct clusters, and the compatibility graph H is shown. There are six

EstimateSTTopology($\mathscr{G}$)

1. $C \leftarrow$ the set of all clusters in $\mathscr{G}$;
2. $H \leftarrow$ the compatibility graph of C;
3. $K \leftarrow$ the set of all maximal cliques of H;
4. $\mathscr{T} \leftarrow \{T_k : T_k$ is the tree of maximal clique $k \in K\}$;
5. **Return** $\mathscr{T}$;

Fig. 2. The algorithm for estimating a set $\mathscr{T}$ of strain-tree topology candidates from an input set $\mathscr{G}$ of gene-tree topologies. In Step 1, the set C of all clusters that appear in any of the gene trees is computed. In Step 2, the compatibility graph of C is built and in Step 3, the set of all maximal cliques is computed. Each maximal clique corresponds to one tree, and the set of all such trees is computed in Step 4.

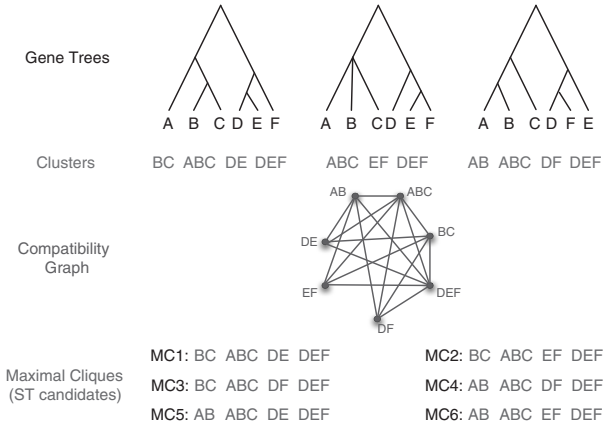

Fig. 3. Example illustrating algorithm EstimateSTTopology. At the top are three gene trees, which are the input to the algorithm. The set of all clusters occurring in these gene trees is then computed, and their compatibility graph is built. Finally, the set of all maximal cliques is computed, and each defines a strain-tree topology candidate.

maximal cliques in H, which implies that the clusters of the input gene trees give rise to six different strain tree topology candidates.

2.2.2 Phase II: estimating strain-tree divergence times Our next task entails estimating the divergence times at internal nodes of each of the strain-tree topology candidates that we computed so as to optimize the weighted sum criterion, as described above. We present a novel optimization based on solving an MILP formulation. The MILP formulation involves a special labeling of the strain-tree topology branches, formulation of temporal constraints based on information from the gene trees, linking coalescence and temporal information and finally putting together all steps into one MILP program. We now describe in details each of these four steps.

(1) *Labeling the strain-tree branches.* In order to model the coalescent of genes on the strain-tree branches, we need to label these branches. As we seek to minimize deep coalescent events (genes that coalesce deeper than their MRCA), we seek a labeling that reflects the 'depth' of the coalescent event, i.e. how far the coalescent event of a set X occurred away from the MRCA of X.

For each internal node x in the strain tree ST, let $P(x)=\langle x_1,x_2,\ldots,x_p\rangle$ be the sequence of nodes where: (1) $x_1=x$, (2) $x_p=r(ST)$ and (3) $(x_i,x_{i-1})\in E(ST)$, for all $2\le i\le p$. Further, $E_{P(x)}$ denotes the *list* of edges defined by $P(x)$; i.e. $E_{P(x)}=\langle(x_i,x_{i-1}):2\le i\le p\rangle$. For example, we have $P(x_2)=\langle x_2,x_3,x_5,r\rangle$ and $E_{P(x_2)}=\langle(x_3,x_2),(x_5,x_3),(r,x_5)\rangle$ in the strain tree in Figure 4a.

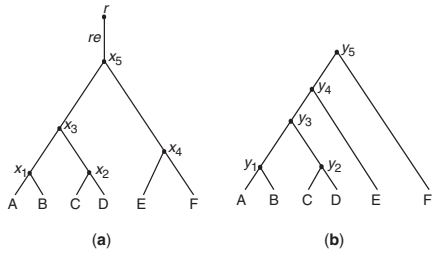

Fig. 4. A strain tree (**a**) and a gene tree (**b**) on six taxa, used for illustrating the strain-tree branch labeling.

Given these sequences, a clade rooted at node y in a gene tree may coalesce only on any edge in $E_{P(x)}$, where $x=MRCA_{ST}(y)$. For example, the clade (C,D) in the gene tree in Figure 4b may coalesce only on one of the edges in $E_{P(x_2)}$, where x_2 is the node in the strain tree in Figure 4a.

Given $E_{P(x)}=\langle(x_2,x_1),(x_3,x_2),\ldots,(x_p,x_{p-1})\rangle$ for some node x in a strain-tree topology, we label the edges in $E_{P(x)}$ by the numbers $1,2,\ldots,p$ such that $\ell((x_s,x_{s-1}))=s-2$, for $2\le s\le p$. For example, for $E_{P(x_2)}$, where x_2 is the node in the strain tree in Figure 4a, we have the labels: $\ell((x_3,x_2))=1$, $\ell((x_5,x_3))=2$ and $\ell((r,x_5))=3$. This labeling is essential for our MILP formulation, since it will be used to reflect the 'depth' of the coalesce events. For example, if clade (C,D) from the gene tree in Figure 4b coalesces on branch (r,x_5) in the strain tree, then the depth of that coalescence event is $\ell((r,x_5))-1$, which is 2 (the reason we choose a label that is larger by 1 than the actual depth value is to accommodate shallow coalescence events, as we discuss below). Indeed, in this scenario, (C,D) coalesced two branches deeper than it could have coalesced [which is branch (x_3,x_2)]. We denote by LabelTree the procedure that computes the lists $P(x)$ and $E_{P(x)}$, as well as the labeling of each edge in $E_{P(x)}$.

(2) *Temporal constraints.* The topology of the strain tree defines a partial order on the times of the internal nodes. This can be represented using linear constraints as $\tau_u>\tau_v$ for every branch (u,v) in the strain tree. For example, in the strain tree in Figure 4a, we have the constraint $\tau_{x_5}>\tau_{x_3}$.

Further, each clade in a gene tree may coalesce on any branch in the strain tree on the path from the MRCA of the clade to the branch *re*. Temporally, this imposes the (linear) constraint $\tau_x\le\tau_y\le\tau_r$, where y is a clade (equivalently in this case, the set of leaves in that clade) in a gene tree, $x=MRCA_{ST}(y)$, and r is the special root of the strain tree. For example, in Figure 4, we have the constraint $\tau_{x_2}\le\tau_{y_2}\le\tau_r$. However, since the coalescence times may be underestimated or gene transfer may have occurred after divergence of the strains, we relax this constraint, by allowing the coalescence time of certain clades to be smaller than the time of their MRCA in the strain tree. Nonetheless, we wish to minimize such events. We achieve via the two constraints

$$[\tau_y<\tau_x]\Rightarrow[g_y=1] \qquad \text{and} \qquad [\tau_y\ge\tau_x]\Rightarrow[g_y=0],$$

for every clade y in a gene tree and its MRCA x in the strain tree. The binary variable g_y here takes the value 1 when the coalescence time of y is lower than that of its MRCA in the strain tree and 0 otherwise. Defining $T^{\max}$ to be the maximum time of the root of any of the gene trees in $\mathscr{G}$, we write these as linear constraints, as follows:

$$\begin{array}{lrcl} \text{(A)} & \tau_y-(1-g_y)T^{\max} & \le & \tau_x-\varepsilon \\ \text{(B)} & (1-g_y)\tau_y+g_yT^{\max} & \ge & \tau_x \\ \text{(C)} & g_y & \in & \{0,1\} \end{array}$$

In this case, we add a small value ε (e.g. $\varepsilon=1\times10^{-8}$) to emulate the $<$ relation.

(3) *Associating times with branches through their labels.* Let y be a node in the gene tree, $x=MRCA_{ST}(y)$, and $(u,v)\in E_{P(x)}$ such that $\ell((u,v))=m$. If node y coalesces on branch (u,v) in the strain tree, this introduces a

constraint of the form $\tau_u \geq \tau_y \geq \tau_v$, which translates into the constraint

$$[f_y = m] \Rightarrow [\tau_u \geq \tau_y \geq \tau_v]. \quad (1)$$

Notice that if $f_y \neq m$, then τ_y is not constrained, which we emulate by constraining τ_y from above by $T^{\max}$ and from below by 0. In other words, we have the constraint

$$[f_y \neq m] \Rightarrow [T^{\max} > \tau_y \geq 0]. \quad (2)$$

Let $M_y = \{1, \ldots, \kappa(y)\}$, where $\kappa(y) = |P(x)| - 1$ for $x = MRCA_{ST}(y)$. For branch $e = (u, v) \in E_{P(x)}$, where $\ell(e) = m$, we denote $s^y(m) = u$ and $t^y(m) = v$.

For each clade y in a gene tree, we convert the conjunction of constraints (1) and (2) into linear constraints by introducing $\kappa(y)$ binary variables α_i for $1 \leq i \leq \kappa(y)$, and then writing the following constraints:

$$\begin{array}{lrcll}
\text{(A)} & \tau_y - (1-\alpha_i)T^{\max} & \leq & \tau_{s^y(i)} & \forall 1 \leq i \leq \kappa(y) \\
\text{(B)} & \tau_y + (1-\alpha_i)T^{\max} & \geq & \tau_{t^y(i)} & \forall 1 \leq i \leq \kappa(y) \\
\text{(C)} & g_y + \sum_{i=1}^{\kappa(y)} \alpha_i & = & 1 & \\
\text{(D)} & f_y - \sum_{i=1}^{\kappa(y)} (\alpha_i \cdot i) & = & 0 & \\
\text{(E)} & \alpha_i & \in & \{0,1\} & \forall 1 \leq i \leq \kappa(y)
\end{array}$$

Constraints (A) and (B) connect the branch assignment with the times of that branch, as they guarantee that $\alpha_i = 1$ if $\tau_{s^y(i)} \geq \tau_y \geq \tau_{t^y(i)}$ and $\alpha_i = 0$ otherwise. Constraint (C) guarantees that either $g_y = 1$ and all the α values are 0, thus resulting in $f_y = 0$ based on constraint (D), which corresponds to the case where the coalescence times of clade y in the gene tree is lower than that of its MRCA in the strain tree, or $g_y = 0$ and exactly one of the α values is 1, which corresponds to the case where y coalesces, under the time assignment to the strain tree, on a unique branch on the path from the MRCA of y to the root. Constraint (D) guarantees that the unique value is chosen from the set M_y. Constraint (E) states that all the α variables are binary.

(4) *Putting it all together: the MILP formulation.* Now that we have described the constraints and how to write them as linear constraints for CPLEX, we are in a position to introduce the complete MILP formulation for solving the problem of estimating divergence times in a strain tree ST, given a set $\mathscr{G}$ of gene trees with coalescence times at internal nodes. We denote $I(T)$ by the set of all internal nodes of tree T, and by $\mathscr{I}$ the set $\cup_{GT \in \mathscr{G}} I(GT)$.

We seek τ_x, for every internal node x in the strain tree, and f_y for every internal node y in all gene trees so as to minimize the amount of deep coalescence events and the amount of shallow coalescence. A MILP formulation of this problem, which we refer to as EstimateSTTimes, is given in Figure 5.

Notice that since $g_y = 1$ (which indicates 'shallow coalescence') if and only if $f_y = 0$, the objective function correctly captures the amount of deep coalescence events, and chooses the solution that minimizes it.

2.2.3 Phase III: strain/gene-tree reconciliation and optimality. Given a strain tree (ST, τ_{ST}) and a gene tree (GT, τ_{GT}), we seek the coalescence history of the gene, given its tree, on the branches of the strain tree. Because both the strain tree and gene trees have times at internal nodes at this stage of the method, this problem is trivial: the coalescence event of a set c of taxa at time t in gene tree GT must occur at time t on the path between the root of ST and the MRCA of c in ST. There is exactly one such point in ST, so this mapping is unique for each cluster in a gene tree.

Considering trees with times at internal nodes is very important since temporal constrains implied by divergence and coalescent times render certain coalescent histories invalid. Therefore, whenever such temporal information is available, it must be used, not only for accuracy reasons, but also to achieve further reductions in the size of the space of strain/gene-tree reconciliations, which in turn affects the computational efficiency of existing and newly developed reconciliation methods.

Let $\mathscr{C}(\mathscr{G})$ be the set of all internal nodes in the gene trees in $\mathscr{G}$ (we use an internal node and the subtree rooted at it interchangeably here) and denote by $c \in ST$ that c is a clade in ST. Our optimality criterion, $\eta(ST, \mathscr{G})$, is defined as the sum of (1) weighted number of missing clades $w_{il}(\sum_{\{c \in \mathscr{C}(\mathscr{G}): c \notin ST\}} 1)$, (2) weighted number of deep coalescence events $w_{dc}(\sum_{\{c \in \mathscr{C}(\mathscr{G}): c \in ST, f_c > 0\}} (f_c - 1))$ and (3) weighted number of shallow coalescence events $w_{sc}(\sum_{\{c \in \mathscr{C}(\mathscr{G}): c \in ST, f_c = 0\}} g_c)$. The first term is the number of clades in the gene trees that do not occur in the strain tree. For a clade c in gene tree GT and which also appears in the strain tree ST, the quantity f_c captures how far (in terms of the number of branches) c coalesced away from its MRCA in ST and g_c captures if it may not coalesce, given the time assignment in the strain tree. Note that if all gene trees have the same topology as the strain tree, and each cluster coalesces on the branch immediately above its MRCA, then $\eta(ST, \mathscr{G}) = 0$. The weights w_{il}, w_{dc} and w_{sc} can be set in a way to reflect the significance given to each of the three terms in the criterion. For example, if only topological difference among the gene trees and strain tree matters, w_{dc} and w_{sc} can be set to 0.

2.3 The algorithm

Now that we have defined our optimality criterion, the complete algorithm for inferring an optimal strain tree (topology and times) is described in Figure 6.

EstimateSTTimes$(ST, \mathscr{G})$

minimize: $w_{dc} \sum_{y \in \mathscr{I}} (f_y + g_y - 1) + w_{sc} \sum_{y \in \mathscr{I}} g_y + \sum_{(u,v) \in S(ST)} (\tau_u - \tau_v)$

subject to

$$\begin{array}{rcll}
\tau_u & > & \tau_v & \forall (u,v) \in E(ST) \\
[\tau_y < \tau_x] & \Rightarrow & [g_y = 1] & \forall y \in \mathscr{I} \text{ and } x = MRCA(y) \\
[\tau_y \geq \tau_x] & \Rightarrow & [g_y = 0] & \forall y \in \mathscr{I} \text{ and } x = MRCA(y) \\
[f_y = m] & \Rightarrow & [\tau_{s^y(m)} \geq \tau_y \geq \tau_{t^y(m)}] & \forall y \in \mathscr{I}, m \in M_y \\
[f_y \neq m] & \Rightarrow & [T^{\max} > \tau_y \geq 0] & \forall y \in \mathscr{I}, m \in M_y
\end{array}$$

Fig. 5. Algorithm EstimateSTTimes, which is an MILP formulation for estimating the divergence times of a strain-tree topology ST given a set $\mathscr{G}$ of gene trees with times at internal nodes. Solving this MILP yields the divergence time τ_u, for every node u in the strain tree, and for each clade y in any of the gene trees in $\mathscr{G}$, the strain-tree branch f_y on which clade y coalesces under the τ time assignment, and g_y if y cannot coalesce, so as to minimize the number of deep coalescence (and shallow coalescence) events and the branch lengths. For k gene trees, each on n leaves, this formulation generates an MILP program with $O(kn^2)$ variables (including binary ones) and $O(kn^2)$ constraints. In our analysis of 1898 genes from nine strains of *S.aureus* bacteria, the MILP program contained over 30 000 constraints, which CPLEX solved in about 1 h.

ComputeStrainTree($\mathscr{G}$)

1. $\mathscr{T} \leftarrow$ EstimateSTTopology($\mathscr{G}$);
2. $ST \leftarrow NIL$;
3. $M \leftarrow \infty$;
4. **For each** $T \in \mathscr{T}$

   a. LabelTree(T);
   b. EstimateSTTimes(T,$\mathscr{G}$);
   c. $Q \leftarrow \eta(T, \mathscr{G})$;
   d. **If** $Q < M$

      (1) $M \leftarrow Q$;
      (2) $ST \leftarrow T$;

5. **End For**;
6. **Return** ST;

Fig. 6. The algorithm for computing the strain-tree topology and divergence times (ST) from an input set of gene trees with coalescence times at internal nodes ($\mathscr{G}$). Step 1 computes a set $\mathscr{T}$ of candidate strain-tree topologies, based on the set $\mathscr{G}$ of gene trees. The loop in Step 4 goes through each tree T in the set $\mathscr{T}$, labels it (Step 4a) estimates the divergence times of T (Step 4b), and keeps track of the optimal candidate strain tree (Step 4d), which is the tree returned at the end. Steps 4c and 4d compute the optimality score, based on the formula described in Section 2.2.3, of the estimated strain tree (topology and times), given the set $\mathscr{G}$ of gene trees with coalescence times at internal nodes.

3 MATERIALS AND ANALYSIS

3.1 Sequence data

In our experimental study, we used the *S.aureus* bacteria, which infect humans in the community and hospitals and cause a variety of diseases. We obtained all the sequence data from the site ftp://ftp.ncbi.nih.gov/genomes/. Table 1 summarizes the nine strains we used. NC_002745 is *S.aureus* subsp. *aureus* N315, which is a prototype of methicillin-resitant *S.aureus* (MRSA; Kuroda *et al.*, 2001). NC_002758 is *S.aureus* subsp. *aureus* Mu50, which has a moderate resistance to vancomycin by the thickened cell wall. NC_002951 is *S.aureus* subsp. *aureus* COL,which is an early methicillin-resistant isolate. The first isolation was found in a British hospital in 1961 (Gill *et al.*, 2005). NC_002952 is *S.aureus* subsp. *aureus* MRSA252. NC_002953 is *S.aureus* subsp. *aureus* MSSA476. These strains were isolated from hospital and community (Holden *et al.*, 2004). MRSA252 belongs to the clinically important EMRSA-16 clone that is responsible for half of the MRSA infections in the United Kingdom and is one of the major MRSA clones found in the USA (USA200). MSSA476 causes severe invasive diseases in immunocompetent children in the community and belongs to a major clone associated with community-acquired disease. NC_003923 is *S.aureus* subsp. *aureus* MW2 (Baba *et al.*, 2002). This strain was isolated from the community, and caused fatal septicaemia. This strain was reported in mid-west USA. NC_007622 is *S.aureus* subsp. *aureus* RF122 (Herron-Olson *et al.*, 2007). NC_007793 is *S.aureus* subsp. *aureus* USA300 (Diep *et al.*, 2006). USA300 is one of the major strains in the USA, Canada and Europe. NC_007795 is *S.aureus* subsp. *aureus* NCTC 8325 (Gillaspy *et al.*, 2006).

3.2 Identifying orthologous genes

To identify orthologous genes, we used the information of both DNA sequence identity and synteny (gene order) as follows. All-against-all BLASTN search with default parameters (Altschul *et al.*, 1997) was performed for the genes in NC_002745 versus all others. Then, we produced a list of BLASTN hits of the 2669 genes in NC_002745 for each of the other strains. The lists include genes that have at least 90% sequence identity to the reference gene in NC_002745 and the length of the BLASTN hit region covers >50% of the entire gene. We excluded BLASTN hits when there are more than one hit for each reference gene. As there were not many such cases, this restriction did not result in much loss of data.

In order to identify orthologous genes conservatively, we considered that orthologous genes should be in a large block of a region in which the gene order is well conserved for all investigated strains. A block is defined such that genes from all strains are continuously located on their genomes with less than three gene skips, which could be created by small indels and annotation errors. To detect such blocks, we performed a synteny survey from the first gene in NC_002745 (NC_002745_1) to downstream genes. Then, we identified 222 such blocks, which covered in total 1898 × 9 genes.

3.3 Gene- and strain-tree analysis

For each gene, we built a maximum parsimony (MP) tree from its DNA sequences by using PAUP* 4.0 (Swofford, 2003), and rooted the tree using the midpoint method. When the MP heuristic identified more than one tree for a given gene, we used the strict consensus of these trees. We inferred coalescence times at internal nodes in the gene trees using the formula

$$\tau_y = \left(\sum_{(a,b)\in B(y)} \frac{d_s(a,b)}{2r_s} \right) / |B(y)| \qquad (3.3)$$

for coalescence time of node y in a gene tree, where $B(y)=\{(a,b): MRCA(a,b)=y\}$, d_s is the number of synonymous substitutions per synonymous sites and r_s is the rate of synonymous substitutions. In other words, τ_y is the average of all coalescence times of every pair of genes whose MRCA is node y. Given that the rate of synonymous substitutions is similar across genes (Nei and Kumar, 2000), this allowed us to compare the coalescence times across gene trees and use them to infer divergence times in the strain tree. We used $r_s = 10^{-8}$, following the findings of Ochman and Wilson (1987).

It has been suggested that d_s may not be constant across the genome due to different codon bias among genes (Retchless and Lawrence, 2007). We found that d_s and the codon adaptation index (CAI) are in a negative correlation, therefore, we used a linear regression method to correct d_s for bias caused by non-random usage of codons. The correction is made such that a corrected d_s corresponds to that with the mean CAI. However, the corrected d_s measure did not change the relative times we obtained for the strain trees and results are not shown.

To get the strain-tree candidates, we used the algorithm EstimateSTTopology described in Figure 2. The compatibility graph H contained 36 nodes and 304 edges. We used the MaxClique

Table 1. Strain information

Refseq	subsp. *aureus*~	Genome size (nt)	Annotated gene number	Reference
NC_002745	N315	2 814 816	2669	Kuroda *et al.* (2001)
NC_002758	Mu50	2 878 529	2775	Kuroda *et al.* (2001)
NC_002951	COL	2 809 422	2724	Gill *et al.* (2005)
NC_002952	MRSA252	2 902 619	2845	Holden *et al.* (2004)
NC_002953	MSSA476	2 799 802	2723	Holden *et al.* (2004)
NC_003923	NW2	2 820 462	2712	Baba *et al.* (2002)
NC_007622	RF122	2 742 531	2665	Herron-Olson *et al.* (2007)
NC_007793	USA300	2 872 769	2648	Diep *et al.* (2006)
NC_007795	NCTC 8325	2 821 361	2969	None

tool of Kevin O'Neill to compute the maximal cliques. The tool identified all maximal cliques in 0.046 s. Additionally, we considered five other candidate tree topologies: (1) T_{conc}: the tree topology obtained by the maximum parsimony heuristic, as implemented in PAUP*, on the concatenation of all 1898 gene data sets; (2) T_{hf}: the topology of the gene tree that is compatible with the largest number of other gene trees (this tree, shown in Figure 11, is compatible with 1645 of the gene trees); (3) T_{avgds}: a tree topology built using the neighbor joining (Saitou and Nei, 1987) method from the average d_s distances among nine strains; (4) T_{avghd}: a tree topology built using the neighbor joining method from the average Hamming distances among nine strains and (5) $T_{majcons}$: the topology of the majority consensus tree of all 1898 gene trees. In total, we have 29 candidate strain-tree topologies.

We then estimated the divergence times of each of the strain tree topology candidates, using the CPLEX tool (from ILOG) to solve the algorithm (MILP program) described in Figure 5. We have implemented a software tool for generating the MILP program from a set of gene trees with coalescence times, following the formulation in Figure 5, in the PhyloNet software package, which is available publicly at http://bioinfo.cs.rice.edu/phylonet/. In the nine-genome dataset that we considered in this study, each MILP program had ~4000 variables and 30000 constraints. Nonetheless, CPLEX solved each program in about 1h.

4 RESULTS AND DISCUSSION

Our first task was to measure the 'heterogeneity' in the data, which consisted of the 9×1898 gene sequences and 1898 gene trees. In this task, we considered two measures of heterogeneity: topological differences among the gene trees, and distributions of coalescence times of each cluster of genes across all gene trees. Figure 7 shows the topological differences between every pair of the 1898 gene trees, as computed by the Robinson–Foulds (RF; Robinson and Foulds, 1981) distance measure. The RF measure quantifies, for a given pair of trees, the average number of clades that appears in one, but not both, of the trees. Hence, if two trees are identical, the RF distance between them is 0; if they do not share any clades, then the RF distance is 1; and, trees with varying degrees of shared clades have RF distance values between 0 and 1.

As shown in Figure 7, while blue (low RF values) is the dominating color, there are many pairs of trees that have RF distance of at least 0.3. In fact, among the 1898 gene trees, there were over

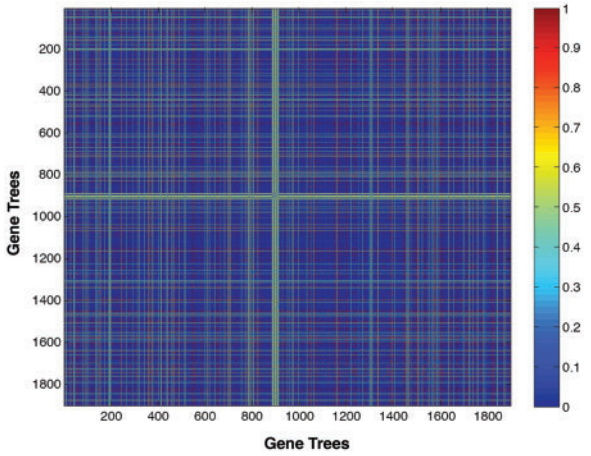

Fig. 7. The RF distances between every pair of the 1898 gene trees. RF distance of 0 indicates the two trees are identical, and RF distance of 1 indicates that the two trees do not share any clades in common.

400 different topologies. Given our conservative selection of the orthology groups, which almost eliminates the possibility of gene tree discordance due to events such as horizontal gene transfer and gene duplication/loss, this result indicates massive gene-tree discordance due to stochastic effects of the coalescent (incomplete lineage sorting).

Furthermore, it is important to point out that the majority of the gene trees were not binary, since the percent identity among the orthologous sequences was very high. This lack of resolution of the gene tree topologies may give a false indication of high concordance (low RF values) among the gene trees, even though this may not be the case in reality. Alternatively, one may quantify the 'compatibility', rather than 'similarity' (as measured by the RF distance), among gene trees. However, this suffers from the fact that compatibility measures are not true metrics, and in particular do not satisfy the triangle inequality property, which may distort the picture emerging from such an analysis.

As indicated in the Section 1 and illustrated in Figure 1, it may be the case the gene trees have the same topology, yet they disagree in their coalescence times (times at their internal nodes). Therefore, what we studied next was the distribution of coalescence times of each cluster of taxa across all gene trees in which the cluster occurs

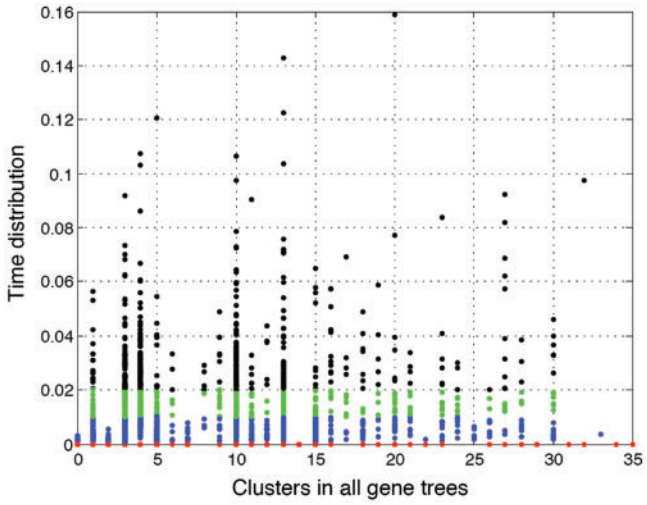

Fig. 8. The distributions of coalescence times of all 36 clusters of taxa in the 1898 gene trees, as calculated by Formula (3.3), yet without division by $r_s \approx 10^{-8}$.

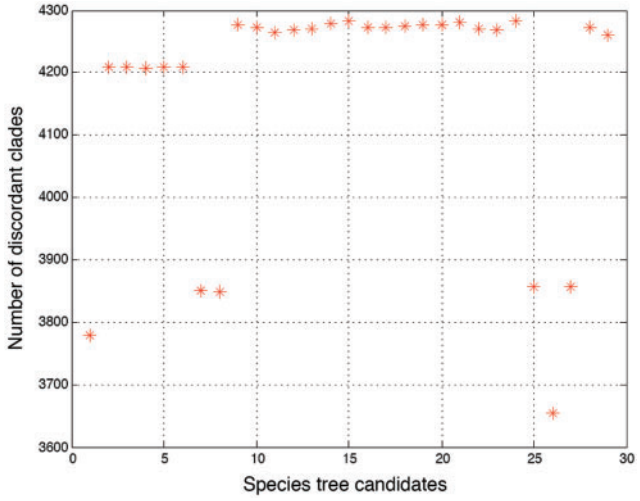

Fig. 9. The number of gene tree clades that do not appear in the strain tree. Trees 1 to 24 are built from maximal cliques. Trees 25, 26, 27, 28 and 29 are T_{conc}, T_{hf}, T_{avgds}, T_{avghd} and $T_{majcons}$, respectively.

(recall that a cluster occurs in a tree if the tree contains a clade whose leaves are the only members of that cluster); the results are shown in Figure 8.

The figure shows that, even with the exclusion of possible outliers, each cluster of taxa has a wide distribution of coalescence times across all gene trees in which it occurs. Further, what makes the computational analysis of such a dataset particularly challenging is that large extent of overlap of distributions of the different clusters. Dealing with this overalp is where most of the computational time of solving our MILP formulation is spent.

After we characterized the heterogeneity in the data, we turned to the main issue, namely estimating the strain-tree topology and divergence times from the set of 1898 gene trees. As described in the previous section, we considered 29 strain-tree topology candidates. For each of these 29 topology candidates, we solved the MILP formulation as outlined in Figure 5, once with $w_{dc} = w_{sc} = 1$, and another with $w_{sc} = 5w_{dc}$. In both cases, the same tree topology candidate of all 24 maximal cliques emerged as the optimal one, yet with differing times. Therefore, we report the results of only the optimal solution under $w_{dc} = w_{sc} = 1$.

For a clearer presentation, we show each of the three terms in the optimality criterion described in Section 2.2.3 individually, with Figure 9 showing the number of missing (or, discordant) clades, and the stacked bars in Figure 10 showing the sum of the depths of deep coalescence events (the blue bars) and the number of shallow coalescence events (the red bars).

Figure 9 shows that the first tree out of the 24 maximal clique trees has the least disagreements with the set of 1898 gene trees, with trees 8 and 9 differing from it by about 70 clades. The other 21 maximal clique trees are much less optimal in this context, with the best of them disagreeing with the gene trees in at least 400 more clades. We denote by T_{mc} the first tree, which is the best in this context among all 24 maximal clique trees. Out of the additional five trees, T_{hf} is clearly the best in this context, and the only one that is better than T_{mc}. Both trees, T_{mc} and T_{hf} are shown in Figure 11. The tree

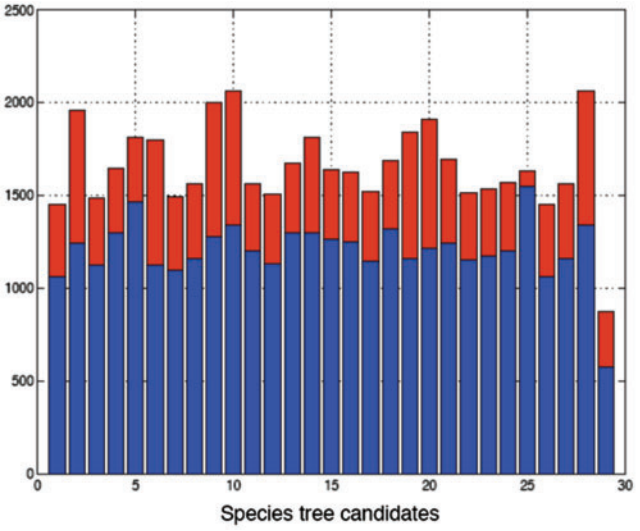

Fig. 10. The number of deep coalescences, $sf = \sum_{c \in \mathscr{C}(\mathscr{G}), f_c > 0} [c \in ST] \times (f_c - 1)$, and the number of shallow coalescences, $sg = \sum_{c \in \mathscr{C}(\mathscr{G}), f_c = 0} [c \in ST] \times g_c$, for all 29 strain-tree candidates. Trees 25, 26, 27, 28 and 29 are T_{conc}, T_{hf}, T_{avgds}, T_{avghd}, and $T_{majcons}$, respectively.

T_{mc} is a *refinement* of the tree T_{hf}; that is, T_{mc} contains all the clades in T_{hf}, plus additional ones. In this case, T_{hf} has the clade (*USA*300, *NCTC*8325, *COL*) unresolved, while T_{mc} has it resolved as (*NCTC*8325, (*USA*300, *COL*)).

When considering the optimality of both trees, T_{mc} and T_{hf}, as measured by the amount of deep coalescence and shallow coalescence events, as shown in Figure 10, they are identical. The significance of this result comes from the fact that, while the unresolved clade (*USA*300, *NCTC*8325, *COL*) has three possible refinements (1) (*NCTC*8325, (*USA*300, *COL*)), (2) ((*NCTC*8325,

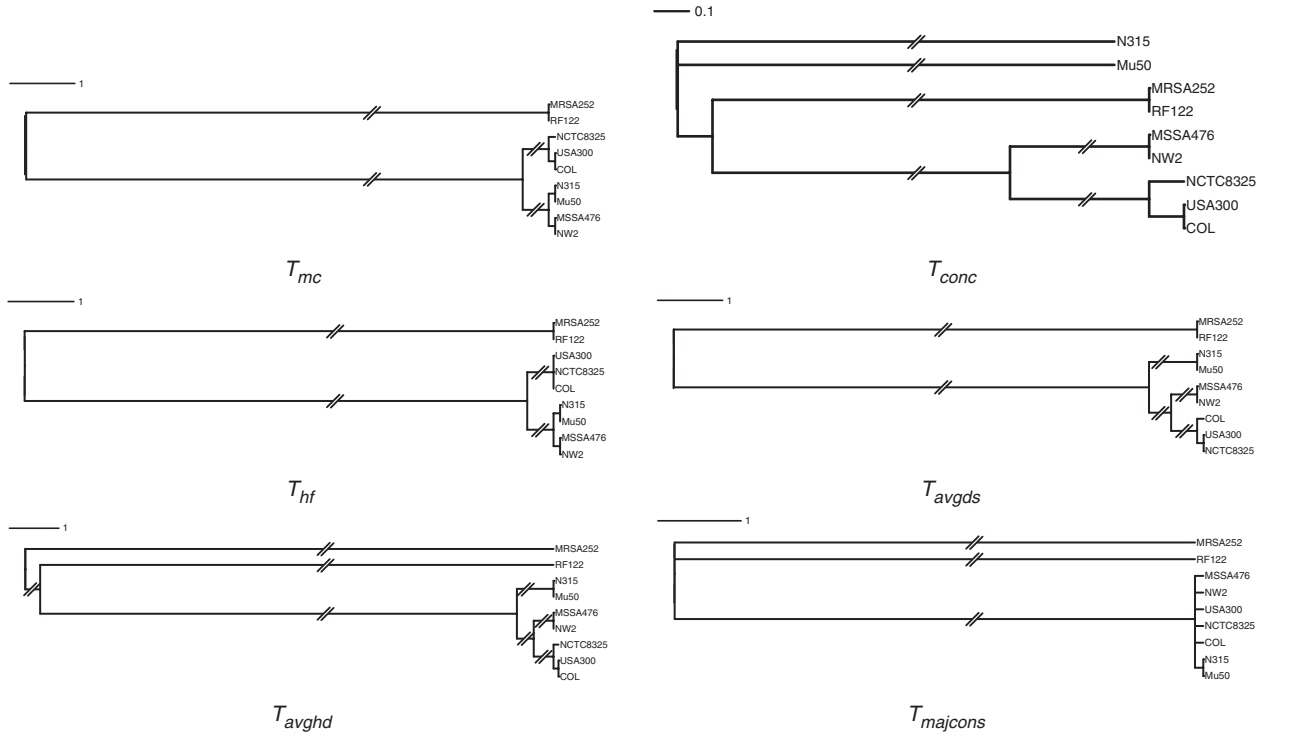

Fig. 11. Strain trees with times assigned by MILP, where T_{mc} is the best maximal clique tree and the rest are as defined in Section 3.3. The lengths of the 'shortened' branches were divided by 10^5, so that the resolution of the trees can be shown clearly.

*USA*300), *COL*) and (3) ((*NCTC*8325, *COL*) *USA*300), the MILP formulation led to a fully binary strain tree that has exactly the same coalescence scenarios among all gene trees. Notice that the majority consensus tree $T_{majcons}$ is the optimal among all 29 trees in terms of the coalescence scenarios. However, this tree has two problems. First, in terms of missing clades, it is one of the least optimal, as shown in Figure 9. Further, it is highly unresolved, containing only two internal branches, as shown in Figure 11.

The concatenation tree, T_{conc} is the best of all trees in terms of minimizing the number of shallow coalescence events, yet is the worst in terms of the sum of the depth of all deep coalescence events. Further, it is the only tree that had the wrong outgroup. This indicates that concatenation of gene sequences and reconstructing a strain tree from the resulting 'supergene' may result in very inaccurate trees, particularly when there is a massive extent of discordance among gene trees, a fact that has already been established through extensive experimental studies (Kubatko and Degnan, 2007). While it seems from Figure 11 that T_{conc} indicates very large divergence time between N315 and Mu50, this is but a reflection of time estimation given that these two strains did not form a single clade in the concatenation tree. To solve this problem, we will consider in future development of our tool all possible refinements of any non-binary strain-tree topology candidate.

The other two trees, T_{avgds} and T_{avghd} are very similar in terms of topology, as shown in Figure 11, and both fall 'in the middle' in terms of optimality, as shown in Figures 9 and 10. Therefore, our proposed evolutionary history of all nine strains of *S.aureus* is the tree T_{mc}, shown in Figure 11.

5 CONCLUSIONS AND FUTURE WORK

In this article, we introduced a three-phase method for efficient inference of an optimal strain tree from genome-scale multilocus data. We have implemented all phases of our method and analyzed nine strains of *S.aureus*. Our hypothesis for the 'vertical' evolutionary history of these nine strains is the tree T_{mc}, shown in Figure 11. It is very important to note that even though the closely related set of strains has a very high degree of sequence identity at the nucleotide level, our method was able to infer a fully resolved evolutionary tree for them. Further, the method computed and evaluated each of 24 possible strain trees within an hour, which is efficient, considering that we used about 1900 loci from nine strains in this analysis.

Two immediate future directions that we will pursue are (1) studying the performance of our method in extensive simulations and (2) investigating the evolutionary diameter of a dataset within which the method reliably returns good strain, or even species, trees. It is worth mentioning that our method can be adapted in a straightforward manner to handle multiallelic loci in the data.

ACKNOWLEDGEMENTS

The authors wish to thank the three anonymous reviewers for very helpful comments on the manuscript.

Funding: L.N. and C.T. were supported in part by DOE grant DE-FG02-06ER25734 and NSF grant CCF-0622037. H.I. and R.S. were supported in part by NSF grant CCF-0622037.

Conflict of Interest: none declared.

REFERENCES

Altschul,S. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Baba,T. *et al.* (2002) Genome and virulence determinants of high virulence community-acquired MRSA. *Lancet*, **359**, 1819–1827.

Degnan,J. and Rosenberg,N. (2006) Discordance of species trees with their most likely gene trees. *PLoS Genetics*, **2**, 762–768.

Diep,B. *et al.* (2006) Complete genome sequence of USA300, an epidemic clone of community-acquired meticillin-resistant Staphylococcus aureus. *Lancet*, **367**, 731–739.

Edwards,S. *et al.* (2007) High-resolution species trees without concatenation. *Proc. Natl Acad. Sci. USA*, **104**, 5936–5941.

Errington,J. *et al.* (2001) DNA transport in bacteria. *Nat. Rev. Mol. Cell Biol.*, **2**, 538–544.

Gill,S. *et al.* (2005) Insights on evolution of virulence and resistance from the complete genome analysis of an early methicillin-resistant Staphylococcus aureus strain and a biofilm-producing methicillin-resistant Staphylococcus epidermidis strain. *J. Bacteriol.*, **187**, 2426–2438.

Herron-Olson,L. *et al.* (2007) Molecular correlates of host specialization in staphylococcus aureus. *PLoS ONE*, **2**, e1120.

Holden,M. *et al.* (2004) Complete genomes of two clinical Staphylococcus aureus strains: evidence for the rapid evolution of virulence and drug resistance. *Proc. Natl Acad. Sci. USA*, **101**, 9786–9791.

Kubatko,L. and Degnan,J. (2007) Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst. Biol.*, **56**, 17–24.

Kuroda,M. *et al.* (2001) Whole genome sequencing of meticillin-resistant Staphylococcus aureus. *Lancet*, **357**, 1225–1240.

Maddison,W. and Knowles,L. (2006) Inferring phylogeny despite incomplete lineage sorting. *Syst. Biol.*, **55**, 21–30.

Milkman,R. and Stoltzfus,A. (1988) Molecular evolution of the Eschericia coli chromosome. II. Clonal segments. *Genetics*, **120**, 359–266.

Nei,M. and Kumar,S. (2000) *Molecular Evolution and Phylogenetics*. Oxford University Press, Oxford.

Ochman,H. and Wilson,A. (1987) Evolution in bacteria: evidence for a universal substitution rate in cellular genomes. *J. Mol. Evol.*, **26**, 74–86.

Retchless,A. and Lawrence,J. (2007) Temporal fragmentation of speciation in bacteria. *Science*, **317**, 1093–1096.

Robinson,D. and Foulds,L. (1981) Comparison of phylogenetic trees. *Math. Biosciences*, **53**, 131–147.

Rokas,A. *et al.* (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, **425**, 798–804.

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

Semple,C. and Steel,M.A. (2003) *Phylogenetics.* Vol. 24. Oxford University Press, Oxford.

Stoltzfus,A. *et al*. (1988) Molecular evolution of the Escherichia coli chromosome. I. Analysis of structure and natural variation in a previously uncharacterized region between trp and tonB. *Genetics*, **120**, 345–358.

Swofford,D.L. (2003) *PAUP*: Phylogenetic Analysis using Parsimony* (* and others methods), Version 4. Sinauer Associates, Sunderland, MA.