The Kernel of Maximum Agreement Subtrees

Krister M. Swenson, Eric Chen, Nicholas D. Pattengale, and David Sankoff

Abstract—A Maximum Agreement SubTree (MAST) is a largest subtree common to a set of trees and serves as a summary of common substructure in the trees. A single MAST can be misleading, however, since there can be an exponential number of MASTs, and two MASTs for the same tree set do not even necessarily share any leaves. In this paper, we introduce the notion of the Kernel Agreement SubTree (KAST), which is the summary of the common substructure in all MASTs, and show that it can be calculated in polynomial time (for trees with bounded degree). Suppose the input trees represent competing hypotheses for a particular phylogeny. We explore the utility of the KAST as a method to discern the common structure of confidence, and as a measure of how confident we are in a given tree set. We also show the trend of the KAST, as compared to other consensus methods, on the set of all trees visited during a Bayesian analysis of flatworm genomes.

Index Terms—Phylogenetics, consensus tree, agreement subtree, MAST.

1 Introduction

Phylogeny inference done on genetic data using maximum parsimony, maximum likelihood, and Bayesian analysis usually yields a set of most likely trees (phylogenies). A typical approach used by biologists to discern the commonality of the trees is to apply a consensus method which yields a single summary tree containing edges that are well represented in the set. For example, the majority rules consensus tree contains only the edges (bipartitions of the leaf set) that exist in a majority of input trees, and is in some sense the optimal balance between including edges that are false, and including edges that are true [2]. Consensus methods are also commonly used for their original purpose [3]; they summarize the information provided from *different* data sets, as in the case when gene trees from different genes provide conflicting phylogenies. There are other uses [4] but these are the two that we consider in this paper.

If one desires a more conservative summary, they may use the strict consensus tree, which has an edge if and only if the edge exists in all of the input trees. Yet even for this extremely conservative consensus method, there has been debate as to its validity and the conditions under which it should be used [5], [6], [7]. In particular, Barrett et al. [5] showed an example where a parsimony analysis of two data sets yields a consensus tree that is at odds with the tree obtained by combining the data. Nelson [6] replied with an argument that the error was not the act of taking the consensus, but the act of pooling the data.

The issue at the heart of this debate is, essentially, that of *wandering* or *rogue* leaves (taxa). Indeed, one or many leaves appearing in different locations of otherwise identical trees have created the problems noticed by Barrett et al., and can also reduce the consensus tree to very few, if any, internal edges. On the other hand, at the time of this debate, Finden and Gordon [8] had already characterized Maximum Agreement SubTrees (MASTs): maximum cardinality subsets of the leaves for which all input trees agree. By calculating a MAST, one avoids Barrett's issue because all MASTs agree with the parsimonious tree they computed on the combined data. As we will see, a single MAST can be misleading however, as there can exist two MASTs (on a single set of trees) which share almost no leaves. Further, there are potentially an exponential (in the number of leaves) number of MASTs for a single set of trees [9]. For Barrett's example we will see that our new method appropriately excludes the contentious part of the tree, and so may be more fit than traditional consensus methods for comparing trees obtained from different analyses.

Wilkinson was the first to directly describe the issues surrounding rogue leaves and develop an approach to try to combat them [4]. Since then, a large body of work by Wilkinson and others has grown on the subjects of finding a single representative tree [4], [10], [11], [12], [13], [14] or something other than a tree (forest, network, etc.) [15], [16], [17], [18], [19]. A full review of this work is out of the scope of this paper so we refer the reader to the chapters of Bryant [20] or Bonizzoni et al. [21], the earlier work of Wilkinson [4], [10], and Pattengale et al. [14]. Despite the myriad of options we notice a distinct lack of an efficiently computable base-line method for reporting subtrees of high confidence; a method analogous to the strict consensus, but less susceptible to rogue leaves.

Thus, we introduce the Kernel Agreement SubTree (KAST) to summarize the information shared by all (potentially exponential) MASTs: the KAST is the intersection of all

- *K.M. Swenson is with the Department of Mathematics and Statistics, University of Ottawa and the Laboratoire de Combinatoire et d'informatique Mathématique (LaCIM) at the Université du Québec à Montréal (UQAM), 8337 Ave. Casgrain, Montreal, QC H2P2K7, Canada. E-mail: akswenson@uottawa.ca.*
- *E. Chen is with the Department of Biology, University of Ottawa, Pavillon Gendron, Pièce 160, 30 Marie Curie, Ottawa, ON K1N 6N5, Canada. E-mail: lupi123@gmail.com.*
- *N.D. Pattengale is with the Sandia National Laboratories, PO Box 5800, Albuquerque, NM 87185. E-mail: npcomplete@gmail.com.*
- *D. Sankoff is with the Department of Mathematics and Statistics, University of Ottawa, 585 King Edward Ave., Room KED 303B, Ottawa, Ontario K1N 6N5, Canada. E-mail: sankoff@uottawa.ca.*

Manuscript received 11 July 2011; revised 6 Nov. 2011; accepted 13 Nov. 2011; published online 4 Jan. 2012.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBBSI-2011-07-0187.
Digital Object Identifier no. 10.1109/TCBB.2012.9.

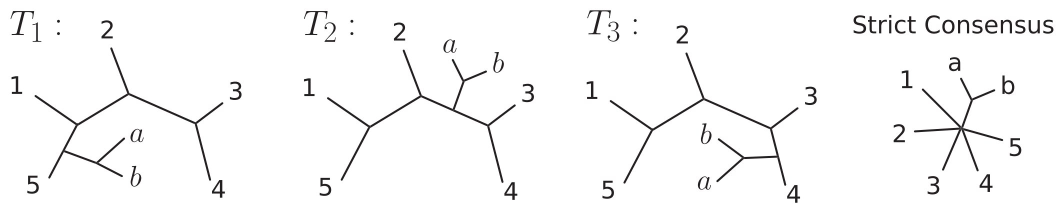


Fig. 1. The effect of adding a tree to the input set. The MASTs for $\{T_1, T_2\}$ are {1,2,3,4,5}, $\{a,b,1,3,4\}$, $\{a,b,2,3,4\}$, and $\{a,b,3,4,5\}$, yielding the KAST $\{3,4\}$. The MAST for $\{T_1, T_2, T_3\}$ is $\{1,2,3,4,5\}$, yielding the KAST $\{1,2,3,4,5\}$. The strict consensus of $\{T_1, T_2, T_3\}$ has only one internal edge.

MASTs. Like the strict consensus, the KAST gives a summary of the common structure of highest confidence, except that it excludes the rogue leaves that confound traditional consensus methods. In other words, the strict consensus gives only the edges with 100 percent support of the input trees, while a MAST criterion gives a single subset of the leaves M, where all edges have 100 percent support, considering only those leaves. Where there may be many MASTs, the KAST is the unique subtree that has 100 percent support irrespective of the subset of leaves we consider, and may thus be considered a subtree of confidence.

The KAST has the benefits of having a simple definition, of summarizing the subtree of confidence by reporting a single tree, and unlike the other known subtree methods can be computed in polynomial time (when at least one input tree has bounded degree). Note that we do not use the term *kernel* in the machine learning sense (as in [22]).

When speaking of a reconstruction method that produces many most probable trees, Barrett et al. [5] called for "conservatism" and suggests the use of the strict consensus. In Section 5, we show the utility of the KAST as a means to get a conservative summary of many most probable trees. We then show the utility of the KAST in the original setting of consensus methods: on trees obtained through different analyses. In each setting, we use the KAST not only to find subtrees of confidence, but as an indicator of randomness in the input trees. We then explore the trend of the KAST, as well as other commonly used consensus methods, on the set of phylogenies visited during a Bayesian analysis of flatworm phylogenies.

The paper is organized as follows: we continue by formally defining the problem in Section 1.1 and showing properties of the MAST and KAST in Section 1.2. We then present Bryant's algorithm for computing the MAST in Section 2, on which our algorithm to compute the KAST (Section 3) is based. Section 4 reports experimental values for the expected size of the KAST on various sets of trees generated at random while Section 5 shows how the KAST can be used to find subtrees of confidence, and report subsets of trees for which we are confident.

1.1 Definitions

Consider a set of trees $\mathcal{T} = \{T_1, T_2, \ldots, T_k\}$ and a set of labels L such that each $x \in L$ labels exactly one leaf of each T_i. We will restrict a tree to a subset L' of its leaf set L; $T_i|_{L'}$ is the minimum homeomorphic subtree of T_i which has leaves L' (i.e., remove all leaves in $L \setminus L'$ and contract all degree 2 nodes). An *agreement subtree* for $\mathcal{T}$ is a subset $L' \subseteq L$ such that $T_1|_{L'} = T_2|_{L'} = \cdots = T_k|_{L'}$. A *maximum agreement subtree* (MAST) is an agreement subtree of maximum size. The set of all maximum agreement subtrees is denoted $\mathcal{M}$.

Definition 1.1. *The* Kernel Agreement SubTree *is the intersection of all MASTs (i.e.,* $\cap_{T\in\mathcal{M}}T$*).*

See Fig. 1 for an example.

As usual, node a is an *ancestor* of b if the path from b to the root passes through a. b is a *descendant* of a. For nodes a and b, the least common ancestor $lca(a,b)$ is the ancestor of a and b that is a descendant of all ancestors of a and b.

1.2 Properties of a MAST and the KAST

In Section 3, we show that the KAST can be computed in the same time as the fastest known algorithm to compute the MAST, by a convenient use of dynamic programming. We devote this section to showing desirable properties of the KAST by contrasting it with the MAST. First, we look at the role the KAST can play in Barrett's example [5]. The rooted trees obtained by his parsimony analyses are $T_1 = (A,(B,(C,D)))$ and $T_2 = (A,((B,C),D))$ (written in Newick format). The set of maximum agreement subtrees for T_1 and T_2 is

$$\{(A,(B,D)),(A,(D,C)),(A,(B,C))\}.$$

While the consensus methods are forced to give trees on the full leaf set, the KAST has only a single leaf A, which indicates that there is not enough information to imply a subtree of confidence. This is the result we would prefer to see, given the circumstances. We see more examples in Section 5 that show a KAST, which finds substantial common substructure, yet does not falter by including subtrees that are at odds with biological observation.

Take a tree set $\mathcal{T}$ with a MAST of size m. Adding a tree T to $\mathcal{T}$ cannot result in a MAST larger than m. This is due to the fact that an agreement subtree of $\mathcal{T} \cup \{T\}$ must also be an agreement subtree of $\mathcal{T}$. On the other hand, the signal for a particular kernel can become apparent when more trees that agree are added to the set.

Property 1.2. *The KAST on tree set* $\mathcal{T}$ *can be smaller than that of* $\mathcal{T} \cup T$*, for some tree* T.

Fig. 1 shows an example exhibiting this property. The KAST on input tree set $\{T_1, T_2\}$ has two leaves (is essentially empty) whereas the subtree on leaves $\{1,2,3,4,5\}$ is amplified by the addition of the tree T_3 to the set. We also see in Section 4 that the KAST size can often increase when adding somewhat similar trees to a set.

We finish with a few negative results about the MAST. The first shows that for some sets of trees there may exist two MASTs that have nearly nothing in common. In other words, the MAST is not necessarily a good indicator of the common subtrees of confidence between two trees.

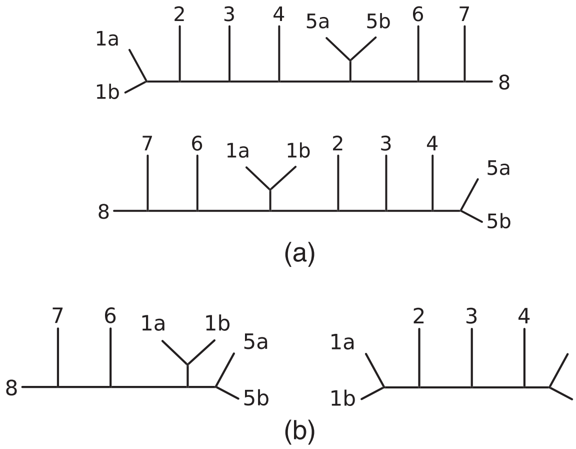


Fig. 2. (a) A tree set displaying Property 1.4. (b) The two MASTs for the tree set (a).

Property 1.3. *There exists a family of tree sets that yields at least two MASTs, the intersection of which is size 2.*

Take the caterpillar trees

$$(1,(2,(3,\ldots,(n-1,n)\ldots))),$$

and

$$(n/2,(n/2+1,\ldots,(n,(n/2-1,\ldots,(2,1)\ldots))\ldots)),$$

for even n. Two of the MASTs for these trees are $\{1,2,\ldots,n/2,n/2+1\}$ and $\{n/2,n/2+1,\ldots,n-1,n\}$.

The next property shows that the number of MASTs and the size of them are not correlated with how similar they are. In other words, those values are not good indicators of their quality. We will see experimental evidence corroborating this fact in Section 4.

Property 1.4. *There exists a family of tree sets that yields exactly two MASTs of size $\Omega(n)$, but the KAST is of size 4.*

For this example, we use trees that are nearly caterpillars. We write them as caterpillars, except S_1 denotes a subtree $(1a,1b)$ while $S_{\frac{m}{2}+1}$ denotes a subtree $((m/2+1)a,(m/2+1)b)$. As depicted in Fig. 2, the first tree is

$$(S_1,(2,(3,\ldots,(S_{\frac{m}{2}+1},\ldots,(m-1,m)\ldots)\ldots))),$$

and the second is

$$\left(m,\left(m-1,\ldots,\left(\frac{m}{2}+2,\left(S_1,\left(2,\ldots,\left(\frac{m}{2},S_{\frac{m}{2}+1}\right)\ldots\right)\right)\right)\ldots\right)\right),$$

where $m=n-2$. The only two MASTs are now

$$\{1a,1b,2,\ldots,m/2,(m/2+1)a,(m/2+1)b\},$$

and

$$\{1a,1b,(m/2+1)a,(m/2+1)b,m/2+2,\ldots,m-1,m\}.$$

2 Finding the MAST

The current fastest known algorithms for the MAST problem are due to Farach et al. [23] and Bryant [24]. Let d_i be the maximum degree (number of children) of tree $T_i\in\mathcal{T}$. These algorithms run in $O(kn^3+n^d)$ time where $n=|L|$, k is the number of trees in the input, and d is the minimum over all d_i, $1\leq i\leq k$. While either of these algorithms can be adapted to compute the KAST, we find it instructive to describe the algorithm of Bryant. We are comprehensive in our description. However, we refer the reader to Bryant's dissertation [24] for a more precise description of the algorithm.

Take $a,b\in L$ and call $\mathcal{T}(a,b)$ the set of all agreement subtrees where the $lca(a,b)$ is the root of the tree. Let $\mathcal{M}(a,b)\subseteq\mathcal{T}(a,b)$ be the set of maximum agreement subtrees where $lca(a,b)$ is the root, and $MAST(a,b)$ be the number of leaves in any member of $\mathcal{M}(a,b)$. We devote the rest of this section to computing $MAST(a,b)$ since the size of the MAST is simply the maximum $MAST(a,b)$ over all possible a and b.

Take three leaves $a,b,c\in L$. $ac|b$ denotes a *rooted triple* where $lca(a,c)$ is a descendant of $lca(a,b)$. When $lca(a,b)$ is the root we say that c is *on a's side of the root* with respect to b. Leaves a, b, and c form a *fan triple*, written (abc), if $lca(a,b)=lca(a,c)=lca(b,c)$. Define R to be the set of rooted triples common to all trees in $\mathcal{T}$ and F to be the set of fan triples common to all trees in $\mathcal{T}$. Bryant showed that an agreement subtree in $\mathcal{T}$ is equivalent to a subset of the set of rooted and fan triples R and F.

The algorithm to compute $MAST(a,b)$ hinges upon the fact that the triples on a's side of the root, and the triples on b's side of the root can be addressed independently. Consider the set $X=\{x:xa|b\in R\}\cup\{a\}$ such that $lca(a,b)$ is the root. In this case, X corresponds to the leaves in a subtree on a's side of the root. Define $MAST_a=max\{MAST(a,x):x\in X\}$ to be the MAST of the leaves in a subtree on a's side of the root. $MAST_b$ is defined similarly, where $X=\{x:a|bx\in R\}\cup\{b\}$.

If F is empty (i.e., the root of every tree in $\mathcal{T}$ is binary), then we have simply,

$$MAST(a,b)=MAST_a+MAST_b.$$

Otherwise, consider the maximum size subset $C\subseteq L$ such that for all $c\in C$ we have $(abc)\in F$. Again, $MAST_c$ is the MAST that considers only the vertices x such that $xc|b$ (or equivalently, $xc|a$). The triples corresponding to a $MAST_c$ are not the same as those for $MAST_a$ and $MAST_b$. However, $MAST_c$ and $MAST_{c'}$ for $c,c'\in C$ could correspond to the same triples. To avoid conflict, construct a graph $G(C)$ as follows: for each $c\in C$ create a vertex with weight $MAST_c$. Make an edge between v and w if and only if $(bvw)\in F$ (equivalently $(avw)\in F$). In other words, v and w have the potential to appear in a subtree of the root that does not include a or b. A maximum weight clique S in this graph is the MAST of all potential subtrees that do not include a or b. So $MAST(a,b)$ can be written

$$MAST(a,b)=MAST_a+MAST_b+\sum_{s\in S}MAST_s,$$

where S is a maximum weight clique in $G(C)$ and $MAST_s$ is defined similarly to $MAST_a$ but with $X=\{x:a|sx\in R\}\cup\{s\}$.

3 Finding the KAST

$KAST(a,b)$ is the intersection of all MASTs in $\mathcal{M}(a,b)$ (the MASTs where $lca(a,b)$ is the root). In this section, we show how to compute $KAST(a,b)$ through a modification of the algorithm of Section 2, without changing the asymptotic running time.

The following theorem hints that the independence of subsolutions that gives rise to the MAST dynamic programming algorithms, will similarly give rise to a KAST algorithm.

Theorem 3.1. *Take leaves x and y such that $lca(x,y)$ is not the root for some MAST. Then $lca(x,y)$ is not the root for every MAST containing both x and y.*

Proof. Assume that the theorem does not hold, namely that there is another MAST containing both x and y where x occurs on the other side of the root from y. Since the second MAST has root $lca(x,y)$ (because x and y are on either side of the root), this implies that the second MAST could be a valid subtree in the first MAST, a contradiction. □

Let $\mathcal{M}_a$ be the set of all MASTs on the leaf set $\{x : xa|b \in R\}$. In other words, $\mathcal{M}_a$ is the collection of sets of leaves that correspond to all $MAST_a$. Call $L(\mathcal{M}_a)$ the set of leaves in any MAST in $\mathcal{M}_a$ (i.e., $L(\mathcal{M}_a) = \{z \in M : M \in \mathcal{M}_a\}$). Symmetrically, $\mathcal{M}_b$ corresponds to the leaf set on the leaves $\{x : a|bx \in R\}$ and $L(\mathcal{M}_b) = \{z \in M: M \in \mathcal{M}_b\}$. We begin by showing how to find $KAST(a,b)$ for binary trees.

Theorem 3.2. *If the trees $T_1, T_2, \ldots, T_k$ are binary, then*

$$KAST(a,b) = (\cap_{T \in \mathcal{M}_a} T) \cup (\cap_{T \in \mathcal{M}_b} T).$$

Proof. If $a = b$ then this is trivially true. Assume by induction that $KAST(c,d)$ can be calculated where $lca(c,d)$ is a descendant of $lca(a,b)$.

Recall that $MAST(a,b) = MAST_a + MAST_b$ when the trees in $\mathcal{T}$ are binary and that $\mathcal{M}_a$ is the set of MASTs that include only the leaves a and x such that $lca(a,b)$ is an ancestor of $lca(a,x)$. It follows that $\mathcal{M}_a$ and $\mathcal{M}_b$ have the following property:

$$L(\mathcal{M}_a) \cap L(\mathcal{M}_b) = \emptyset.$$

So $KAST(a,b)$ depends on $\mathcal{M}_a$ and $\mathcal{M}_b$ independently.

Bryant showed that any leaf included in $\mathcal{M}_a$ or $\mathcal{M}_b$ will necessarily exist in some MAST for $\mathcal{T}$ (a corollary of Theorem 6.8 in [24]). Since the KAST contains only the leaves that exist in every MAST, then $KAST(a,b)$ must be equal to the intersection of all MASTs in $\mathcal{M}_a$, along with the intersection of all MASTs in $\mathcal{M}_b$. □

So the algorithm to compute $KAST(a,b)$ takes the intersection over all sets $KAST(a,x)$ such that $ax|b \in R$ and $MAST(a,x)$ is maximum. It does the same for b's side of the root, and then takes the union of the result.

We now present the main result of this section. Recall from Section 2 the graph $G(C)$ where C is the set of triples satisfying $(abc) \in F$, and that $MAST(a,b) = MAST_a + MAST_b + \sum_{s \in S} MAST_s$.

Theorem 3.3. $KAST(a,b) = (\cap_{T \in \mathcal{M}_a} T) \cup (\cap_{T \in \mathcal{M}_b} T) \cup (\cap_{S \in \mathcal{K}} (\cup_{s \in S} (\cap_{T \in \mathcal{M}_s} T)))$ *where $\mathcal{K}$ is the set of all maximum weight cliques on graph $G(C)$.*

Proof. If $a = b$ then this is trivially true. Assume by induction that $KAST(c,d)$ can be calculated where $lca(c,d)$ is a descendant of $lca(a,b)$.

Take any maximum weight clique $S \in \mathcal{K}$. Bryant showed that for $S = \{s_1, \ldots, s_m\}$, $\cup_{i=1}^{m} T_i$ where $T_i \in \mathcal{M}_{s_i}$, is a MAST on the set of leaves $\{c : (abc) \in C\}$. By the definition of $G(C)$ we know that $L(\mathcal{M}_{s_1})$, $L(\mathcal{M}_{s_2}), \ldots, L(\mathcal{M}_{s_m}), L(\mathcal{M}_a)$, and $L(\mathcal{M}_b)$ are pairwise disjoint. Further, any leaf in the sets $L(\mathcal{M}_{s_i})$, $L(\mathcal{M}_a)$, or $L(\mathcal{M}_b)$ are necessarily included in some MAST for $\mathcal{T}$ (a corollary of Theorem 6.8 in [24]). So the leaves in a $KAST(a,b)$ could have only the leaves that are in every MAST in $\mathcal{M}_{s_i}$ (i.e., $(\cup_{s \in S} (\cap_{T \in \mathcal{M}_s} T))$), for all $1 \leq i \leq m$. But each clique in $\mathcal{K}$ represents a different MAST, so only the leaves that are in every clique will be in the KAST. Finally, this set is disjoint from $L(\mathcal{M}_a)$ and $L(\mathcal{M}_b)$ for the same reason that $L(\mathcal{M}_a)$ and $L(\mathcal{M}_b)$ are disjoint from each other. □

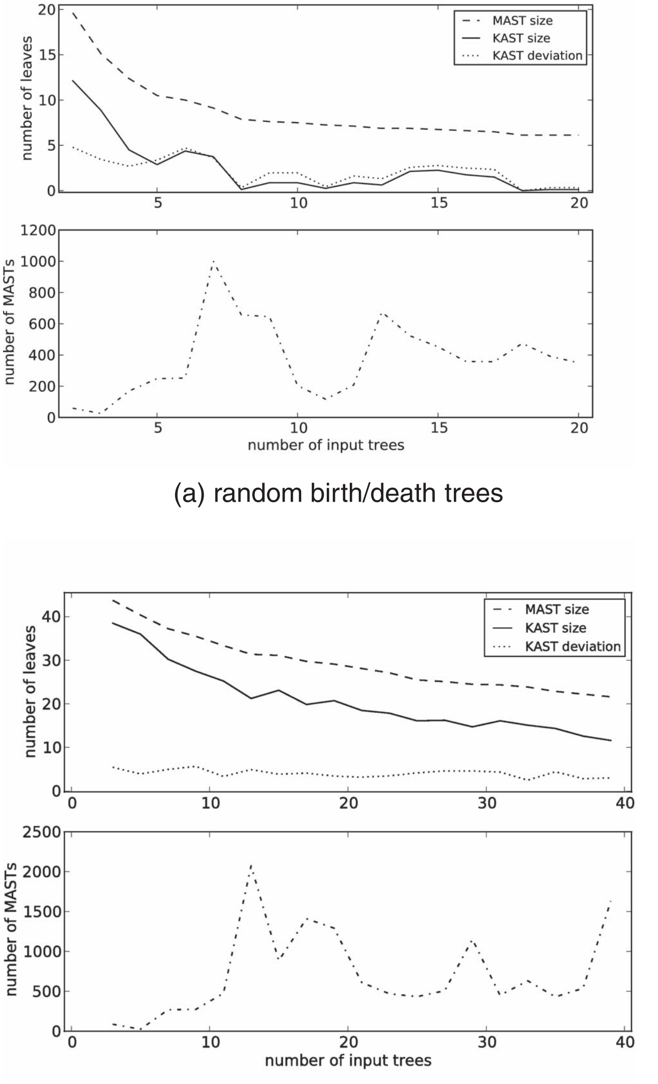
4 Experiments

We implemented the KAST from code that computes the MAST in the phylogenetic package RAxML [25]. In this section, we report empirical evidence about the expected size of the KAST and MAST under three different models. The first model builds a tree set $\mathcal{T}$ of random trees constructed through a birth/death process, while the second starts with a random birth/death tree and then incrementally adds new trees to the set, each a single Nearest Neighbor Interchange (NNI) move ([26], [27]) from the last. This way we see how the expected sizes react to adding drastically dissimilar, or fairly similar trees to the set $\mathcal{T}$. The birth/death process has parameters 1 for birth and 1/2 for death but is terminated when the desired number of leaves is reached (if it terminates before this, then the process is repeated).

Fig. 3 shows that as the size of $\mathcal{T}$ (the tree set) increases, the size of the KAST decreases precipitously in the case of the birth/death model, whereas it decreases more gracefully in the case of the NNI model. Each plot is generated from an initial birth/death tree on 50 leaves, where new trees are added to the tree set according to the prescribed model. This process was repeated 10 times and the average is reported. Plots with various numbers of leaves are similar except that the curve is scaled on the "number of leaves" axis proportionately.

The third model starts with a tree set comprised of three trees, each with the same topology on 40 leaves. Rogue leaves are then introduced by placing a new leaf in a random location, independently in each of the three trees. Fig. 4 shows the average of 10 runs. The figure clearly shows that the size of the majority and strict consensus trees are very small in the presence of only a few rogue leaves. When only six rogue leaves are present, nearly a quarter of the edges of the original tree are lost on the majority consensus, whereas almost all are present in the KAST. On the other hand, the KAST maintains three quarters of all possible edges even when the number rogue leaves is equal to the number of leaves in the original topology.

With regards to the number of MASTs, the plots in Figs. 3 and 4 show an erratic curve while the curve for the



(a) random birth/death trees

(b) each tree 1 random NNI move from the last

Fig. 3. The expected sizes and standard deviations of the MAST and KAST for sets of dissimilar (a) and similar (b) trees.

KAST is stable, confirming that the phenomenon described in Property 1.4 is not a rarity.

5 Applications

We now demonstrate the application of the KAST to biological data, for finding subtrees of confidence, as well as finding subsets of the input tree set of confidence. To do this, we gleaned phylogenies from the literature that are known to have an agreed upon structure, except for a few contentious leaves. Our intention is not to provide biological insight, but to confirm the utility of the KAST by comparing our results to familiar phylogenies. The real utility of the KAST will be on phylogenies that are much larger, so large as to make it difficult for a humans to process.

5.1 Analyses on Flatworm Phylogenies

In a recent publication by Philippe et al. [28], the proposed phylogeny describes the Acoel and the Nemertodermatids and Xenoturbellid as a sister clade to Ambulacraria, which is vastly different from the previous publications. The competing hypotheses are depicted in Fig. 5. In earlier publications both Nemertodermatids and Acoels are the outgroups with Xenoturbellid leaf grouping either with the Ambulacraria or with the Nemertodermatids and Acoels. Setting aside the interpretation and biological ramifications of the new proposed tree topology, it is a good real-world example for observing the effects of KAST on contentious trees.

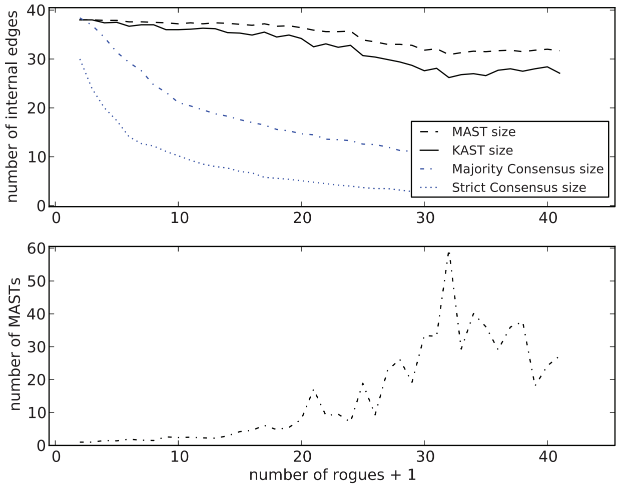


Fig. 4. The number of internal edges of the MAST, KAST, strict consensus, and majority consensus trees as a function of the number of “rogue” leaves in the tree set.

There are two main objectives that we wish to explore through the use of this example. The first objective is to determine if the KAST of a set of phylogenetic trees can identify a subset that we are confident in. The second objective is to show the KAST as a measurement of how confident we are in the hypothesis of these trees.

5.1.1 Confidence in the Phylogenetic Tree

From their publication, Philippe et al. [28] presents three trees, two from prior publications and one from their own experimental result. With only 10 common leaves among the three trees it is very easy to identify the similarity between them by eye (Fig. 5).

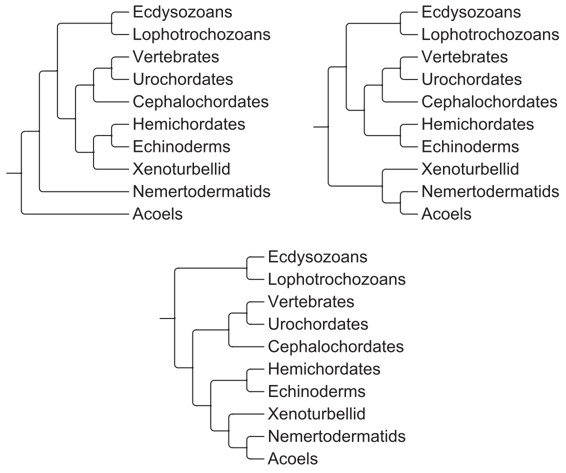


Fig. 5. Phylogenies from Fig. 1 of Philippe et al. [28] Xenoturbellid, Nemertodermatids, and Acoels wander.

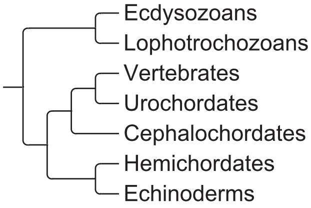


Fig. 6. The KAST (and MAST) of the phylogenies of Fig. 5.

A quick observation can find that Ecdysozoans and Lophotrochozoans of Protostomia forms a clade, Vertebrates and Urochordates and Cephalochordates of Chordata forms a clade, and Hemichordates and Echinoderms of Ambulacraia forms a sister clade to Chordata; and that Xenoturbellid, Nemertodermatids, and Acoels are the rogue leaves. The KAST of these three trees agrees with this observation (Fig. 6). This suggests that the KAST is able to find a subtree that is biologically obvious.

With bigger trees it will be harder to identify the similarity by eye. We would argue that the KAST can be an important tool in identifying or verifying these similarities.

5.1.2 Phylogeny Reconstruction

To see if the KAST could identify a subset of trees that we are confident of, in the context of phylogeny reconstruction, we tried to replicate the analysis of Philippe et al. [28]. The aligned mitochondrial gene set was taken from the supplementary material section and used as input for the Bayesian analysis tool that they used: PhyloBayes 3.2 [29], [30], with the CAT model [28], [29] as the amino acid replacement model and default settings for everything else. We ran 10,000 cycles and discarded the first 1,000 as burnins, as they did. The consensus tree by majority rule was then obtained by CONSENSE [31] using all remaining 9,000 trees. The majority consensus tree (Fig. 7) we found is in agreement with the $\mathrm{CAT} + \Gamma$ model tree from their supplementary Fig. 1, which is available in the online supplemental material, we obtain a node with degree three that they do not.

To test the validity of the conservative tree produced by the KAST, the kernel of the 9,000 tree set is calculated. Of the 10 species in the KAST, three sponge species and two jellyfish species group together as predicted, the two annelida species group together as predicted, the three echinoderms also group together as predicted, and the topology of these phyla are also organized in the biologically obvious fashion (Fig. 8). This corroborates the notion that the topology of KAST is the base-line topology.

Next, we test the variability of the KAST within the tree set. Philippe et al. [28] sampled once every 10 cycles, to simulate this, we sample 900 random trees in the 9,000 tree set and calculate the KAST. We replicate this 1,000 times and calculate the symmetric Robinson-Foulds distances (because the KAST is binary, we divide it by two) between every pair of KASTs generated. The average distance between these KASTs is 0.73 with an average KAST size of 10.73.

We also calculate the KAST on varying sizes of tree sets to test how sample size effects the KAST size. Samples of 5,000, 2,500, 1,200, and 500 trees all have KAST size of 11. Starting with the samples of 250 trees the size of KAST start to increase, with samples of 30 trees having size 16. While the KAST from the whole 9,000 tree set is obviously more conservative, the KAST from the smaller samples agree with all known competing hypotheses while including up to half the leaves.

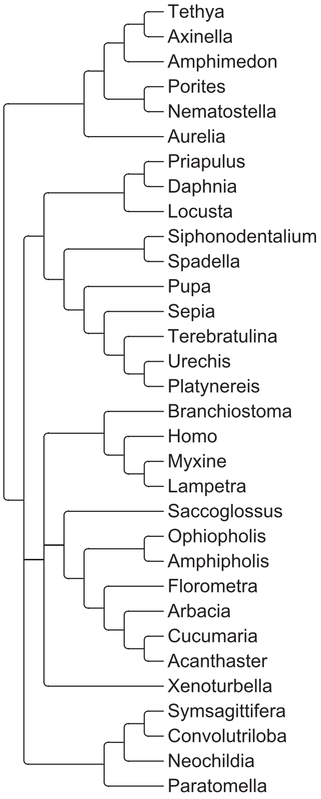


Fig. 7. The majority rule consensus tree using the last 9,000 trees visited by PhyloBayes. The topology is essentially the same as Supplementary Fig. 1, which is available in the online supplemental material, from Philippe et al. [28].

The situation improves if only the k most frequently visited trees are considered [13]. Fig. 9 shows the number of edges obtained by the various methods for values of k from 1 to 20. When the six most frequent topologies are considered, the KAST has the same number of edges as the majority consensus; they each have 28 of the possible 30 edges. The majority consensus tree does not change after the consideration of the nine most frequent topologies, indicating that the addition of other topologies does not strengthen or weaken edges' relative presence in the tree set. The frequency for $k = 1, 2, \ldots, 15$ is as follows: 182, 180, 145, 127, 112, 109, 105, 104, 88, 76, 71, 69, 63, 63, 62. The significant drop in the

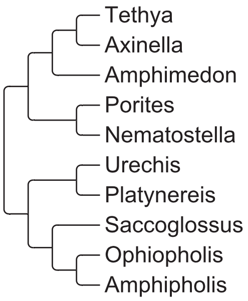


Fig. 8. The KAST of the 9,000 trees from the Bayesian analysis program PhyloBayes.

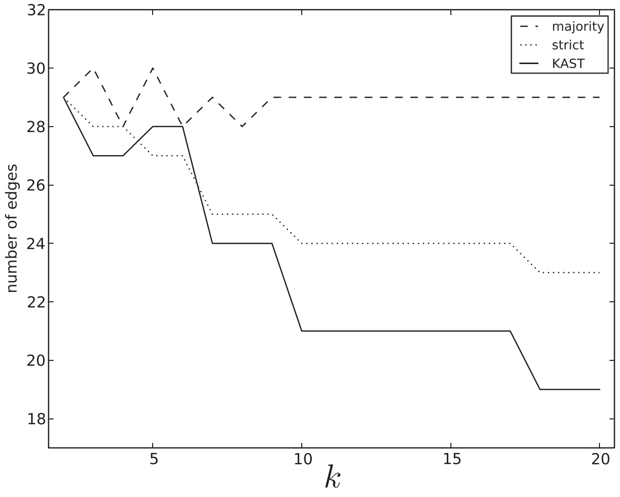


Fig. 9. The number of edges in the KAST, majority, and strict consensus trees on the k most frequent topologies visited in the last 9,000 iterations by PhyloBayes.

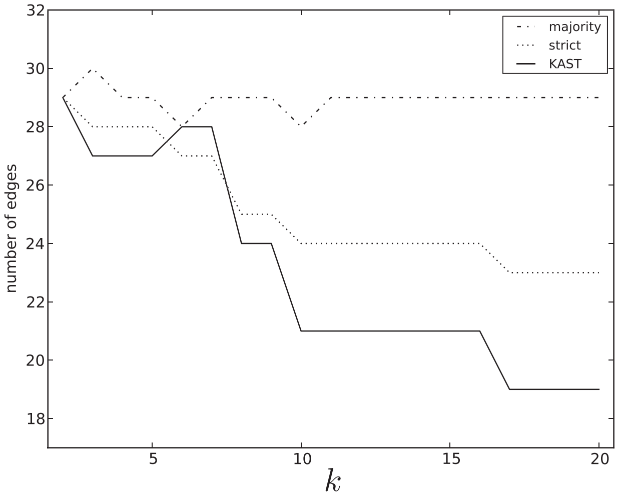


Fig. 10. The number of edges in the KAST, majority, and strict consensus trees on the k most frequent topologies visited in all 10,000 iterations by PhyloBayes.

number of occurrences of the ninth most frequent topology corresponds to the stability of the majority consensus. The KAST of the nine most frequent topologies has 24 edges. The drastic fall in the size of the KAST after $k = 6$ indicates that the seventh most frequent topology is significantly different than the first six. The size of the KAST continues to drop, predictably (see Fig. 3), as more of the less frequently visited (so less significant) topologies are considered.

A comparison of Figs. 9 and 10 shows that the removal of 1,000 burn-in cycles changes little in terms of the characteristics of the KAST, strict, and majority consensus trees. In this case, the majority consensus tree of the whole set is the majority consensus tree of the set without the first 1,000. The proportion of the 10 most frequent topologies visited to all topologies visited is roughly the same in this case; it is roughly 15 percent for both. However, if the Bayesian analysis is recomputed on all taxa except for the two excluded from the KAST of the six most frequent topologies (Xenoturbella and Sepia Esculenta), we find that the 10 most frequent topologies account for about half of all those visited. Further, the majority consensus tree for this analysis on 30 taxa has the same number of internal edges as the analysis on the full set of 32. This is evidence that the KAST has detected the genomic data with confounding information, in this case the Xenoturbella and Sepia Esculenta contain the phylogenetically confounding information.

5.2 Analyses on γ-Proteobacteria Phylogenies

Finally, we test the KAST on the phylogeny of γ-proteobacteria that has been the subject of pains-taking study. We refer the reader to Herbeck et al. [32] for a discussion of previous work. For our purposes, we concentrate on the studies related to 12 particular species used in Lerat et al. [33], who reconstructed a phylogeny based on hundreds of genes. Since then there have been other attempts to reconstruct the phylogeny based on the syntenic data of the whole genome [34], [35], [36], [37].

We turn our attention to two studies that produced trees in discordance with that of Lerat et al. Belda et al. [36] produced two trees, one using Maximum Likelihood on amino acid sequences and the other using reversal distance on the syntenic information (they used the breakpoint distance as well, which produced the same tree as the inversion distance). The likelihood analysis gave a tree that agreed with Lerat's. The inversion distance gave a tree that has significant differences to that of Lerat; the KAST between the two has 9 of 12 leaves. However, we will see that when we add certain trees from the study of Blin et al., the KAST size is 10. Further, the leaves excluded are Wigglesworthia brevipalpis and Pseudomonas aeruginosa; the former identified by Herbeck et al. [32] as troublesome to place, and the latter being the outgroup that they used to root their trees.

Blin et al. [35] used model free distances (breakpoints, conserved intervals, and common intervals) on the syntenic data to reconstruct their phylogenies. They produced many trees with the various methods on two different data sets. The syntenic data that yielded the interesting phylogeny for our purposes was produced from coding genes along with ribosomal and transfer RNAs. Blin et al. noticed that their trees computed on this data, using conserved and common intervals, were less similar to the Lerat tree than the others. The KAST confirms this: the KAST on the set of all published trees other than these two is 10 while the inclusion of either one (they are the same) yields a KAST of size 5. Our experimental data tell us that a sequence of six trees, each produced by a random NNI operation from the last, will yield a KAST of size 5 while six unrelated trees would produce a KAST of size 2. This provides a hint that we could have higher confidence in the set of trees that don't include those two trees.

6 Conclusion

We claim that the utility of the KAST is two-fold. The first is that the KAST is a safe summary of the subtree of confidence for a set of trees. The second is that the size of the KAST is correlated with how related the set of trees is. The KAST is not as susceptible to rogue leaves as the very conservative strict consensus, and is not as misleading as the MAST can be. Furthermore, unlike the other methods that attempt to characterize tree structure in the presence of rogue leaves, our measure is computable in polynomial time.

Acknowledgments

The first author would like to thank Andre Aberer for use and discussions about his code to compute the MAST, and Bernard Moret for discussions about the utility of the KAST.

References

[1] K.M. Swenson, E. Chen, N.D. Pattengale, and D. Sankoff, "The Kernel of Maximum Agreement Subtrees," *Proc. Seventh Int'l Conf. Bioinformatics Research and Applications (ISBRA '11),* pp. 123-135, 2011.

[2] M.T. Holder, J. Sukumaran, and P.O. Lewis, "A Justification for Reporting the Majority-Rule Consensus Tree in Bayesian Phylogenetics," *Systematic Biology,* vol. 57, no. 5, pp. 814-821, 2008.

[3] E.N. Adams, "Consensus Techniques and the Comparison of Taxonomic Trees," *Systematic Zoology,* vol. 21, pp. 390-397, 1972.

[4] M. Wilkinson, "Common Cladistic Information and its Consensus Representation: Reduced Adams and Reduced Cladistic Consensus Trees and Profiles," *Systematic Biology,* vol. 43, no. 3, pp. 343-368, 1994.

[5] M. Barrett, M.J. Donoghue, and E. Sober, "Against Consensus," *Systematic Zoology,* vol. 40, no. 4, pp. 486-493, 1991.

[6] G. Nelson, "Why Crusade Against Consensus? A Reply to Barret, Donoghue, and Sober," *Systematic Biology,* vol. 42, no. 2, pp. 215-216, 1993.

[7] M. Barrett, M.J. Donoghue, and E. Sober, "Crusade? A Reply to Nelson," *Systematic Biology,* vol. 42, no. 2, pp. 216-217, 1993.

[8] C.R. Finden and A.D. Gordon, "Obtaining Common Pruned Trees," *J. Classification,* vol. 2, no. 1, pp. 255-267, 1985.

[9] E. Kubicka, G. Kubicki, and F. McMorris, "On Agreement Subtrees of Two Binary Trees," *Congressus Numeratium,* vol. 88, pp. 217-224, 1992.

[10] M. Wilkinson, "More on Reduced Consensus Methods," *Systematic Biology,* vol. 44, pp. 435-439, 1995.

[11] M. Wilkinson, "Majority-Rule Reduced Consensus Trees and Their Use in Bootstrapping," *Moleculer Biology and Evolution,* vol. 13, no. 3, pp. 437-444, 1996.

[12] J.L. Thorley, M. Wilkinson, and M. Charleston, "The Information Content of Consensus Trees," *Studies in Classification, Data Analysis, and Knowledge Organization,* ser. Advances in Data Science and Classification, A. Rizzi, M. Vichi, and H. Bock, eds., pp. 91-98, Springer, 1998.

[13] K.A. Cranston and B. Rannala, "Summarizing a Posterior Distribution of Trees Using Agreement Subtrees," *Systematic Biology,* vol. 56, no. 4, pp. 578-590, 2007.

[14] N.D. Pattengale, A.J. Aberer, K.M. Swenson, A. Stamatakis, and B.M.E. Moret, "Uncovering Hidden Phylogenetic Consensus in Large Datasets," *IEEE/ACM Trans. Computational Biology and Bioinformatics,* vol. 8, no. 4, pp. 902-911, July/Aug. 2011.

[15] H. Bandelt and A. Dress, "Split Decomposition: A New and Useful Approach to Phylogenetic Analysis of Distance Data," *Moleculer Phylogenetics and Evolution,* vol. 1, no. 3, pp. 242-252, 1992.

[16] D.H. Huson, "Splitstree: Analyzing and Visualizing Evolutionary Data," *Bioinformatics,* vol. 14, no. 1, pp. 68-73, 1998.

[17] D. Bryant and V. Moulton, "Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks," *Moleculer Biology and Evolution,* vol. 21, no. 2, pp. 255-265, 2004.

[18] O. Gauthier and F.-J. Lapointe, "Seeing the Trees for the Network: Consensus, Information Content, and Superphylogenies," *Systematic Biology,* vol. 56, no. 2, pp. 345-355, 2007.

[19] B. Redelings, "Bayesian Phylogenies Unplugged: Majority Consensus Trees with Wandering Taxa," http://www.duke.edu/br51/wandering.pdf, 2012.

[20] D. Bryant, "A Classification of Consensus Methods for Phylogenetics," *Bioconsensus,* ser. DIMACS Series in Discrete Math. and Theoretical Computer Science, vol. 61, pp. 163-184, AMS Press, 2002.

[21] P. Bonizzoni, G.D. Vedova, R. Dondi, and G. Mauri, "The Comparison of Phylogenetic Networks: Algorithms and Complexity," *Bioinformatics Algorithms: Techniques and Applications,* Wiley Interscience, pp. 143-173, 2008.

[22] K. Shin and T. Kuboyama, "Kernels Based on Distributions of Agreement Subtrees," *AI 2008: Advances in Artificial Intelligence,* W. Wobcke and M. Zhang, eds., vol. 5360, pp. 236-246, Springer, 2008.

[23] M. Farach, T.M. Przytycka, and M. Thorup, "On the Agreement of Many Trees," *Information Processing Letters,* vol. 55, no. 6, pp. 297-301, 1995.

[24] D. Bryant, "Building Trees, Hunting for Trees, and Comparing Trees," PhD dissertation, Dept. of Math., Univ. of Canterbury, 1997.

[25] A. Stamatakis, "RAxML-VI-HPC: Maximum Likelihood-Based Phylogenetic Analyses with Thousands of Taxa and Mixed Models," *Bioinformatics,* vol. 22, no. 21, pp. 2688-2690, 2006.

[26] D.F. Robinson, "Comparison of Labeled Trees with Valency Three," *J. Combinatorial Theory,* vol. 11, no. 2, pp. 105-119, 1971.

[27] G.W. Moore, M. Goodman, and J. Barnabas, "An Iterative Approach from the Standpoint of the Additive Hypothesis to the Dendrogram Problem Posed by Molecular Data Sets," *J. Theoretical Biology,* vol. 38, no. 3, pp. 423-457, 1973.

[28] H. Philippe, H. Brinkmann, R.R. Copley, L.L. Moroz, H. Nakano, A.J. Poustka, A. Wallberg, K.J. Peterson, and M.J. Telford, "Acoelomorph Flatworms are Deuterostomes Related to Xenoturbella," *Nature,* vol. 470, no. 7333, pp. 255-258, Feb. 2011.

[29] N. Lartillot and H. Philippe, "A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process," *Moleculer and Biology Evolution,* vol. 21, no. 6, pp. 1095-1109, June 2004.

[30] N. Lartillot, H. Brinkmann, and H. Philippe, "Suppression of Long-Branch Attraction Artefacts in the Animal Phylogeny Using a Site-Heterogeneous Model," *BMC Evolution Biology,* vol. 7, Suppl. 1, Mar. 2006.

[31] J. Felsenstein, *Phylogenetic Inference Package (PHYLIP), Version 3.5,* Univ. of Washington, 1993.

[32] J.T. Herbeck, P.H. Degnan, and J.J. Wernegreen, "Nonhomogeneous Model of Sequence Evolution Indicates Independent Origins of Primary Endosymbionts Within the Enterobacteriales (*gamma*-Proteobacteria)," *Moleculer and Biology Evolution,* vol. 22, no. 3, pp. 520-532, 2005.

[33] E. Lerat, V. Daubin, and N.A. Moran, "From Gene Trees to Organismal Phylogeny in Prokaryotes: The Case of the γ-Proteobacteria," *PLoS Biology,* vol. 1, no. 1, p. e19, 2003.

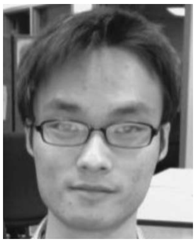
[34] J. Earnest-DeYoung, E. Lerat, and B. Moret, "Reversing Gene Erosion: Reconstructing Ancestral Bacterial Genomes from Gene-Content and Gene-Order Data," *Proc. Fourth Int'l Workshop Algorithms in Bioinformatics (WABI '04),* pp. 1-13, 2004.

[35] G. Blin, C. Chauve, and G. Fertin, "Genes Order and Phylogenetic Reconstruction: Application to γ-Proteobacteria," *Proc. Int'l Conf. Comparative Genomics (RCG '05),* pp. 11-20, 2004.

[36] E. Belda, A. Moya, and F. Silva, "Genome Rearrangement Distances and Gene Order Phylogeny in γ-Proteobacteria," *Moleculer Biology and Evolution,* vol. 22, no. 6, pp. 1456-1467, 2005.

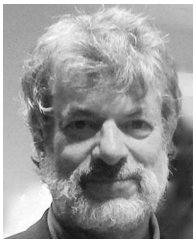
[37] K. Swenson, W. Arndt, J. Tang, and B. Moret, "Phylogenetic Reconstruction from Complete Gene Orders of Whole Genomes," *Proc. Sixth Asia Pacific Bioinformatics Conf. (APBC '08),* pp. 241-250, 2008.

Krister M. Swenson received the PhD degree in computer science from the Ecole Polytechnique Fédérale de Lausanne, Switzerland, in 2009. He is currently a postdoctoral fellow at the Université de Montréal and McGill University in Montréal, Quebec.



Eric Chen received the graduate degree in molecular biology and biochemistry from Simon Fraser University and worked on a bioinformatic analysis of the malaria parasites for his undergrad research project. He is currently working toward the master's degree in Dr. Sankoff's lab at the University of Ottawa, working on genome rearrangement studies.

Nicholas D. Pattengale received the PhD degree in computer science from the University of New Mexico in 2010. He is an algorithms researcher and software engineer at Sandia National Laboratories in Albuquerque, New Mexico. His research interests include computational molecular biology (primarily phylogenetics) has been in deriving efficient algorithms for phylogenetic postanalysis.



David Sankoff received the PhD degree in mathematics from McGill University, and has been a member of the Centre de recherches mathematiques in Montreal for many years. He currently holds the Canada Research Chair in mathematical genomics in the Mathematics and Statistics Department at the University of Ottawa, and is cross appointed to the Biology and the Computer Science Departments. His research interests include comparative genomics, particularly probability models, statistics, and algorithms for genome rearrangements.