RECOMP: A PARSIMONY-BASED METHOD FOR DETECTING RECOMBINATION

DEREK RUTHS LUAY NAKHLEH

Department of Computer Science, Rice University, Houston, Texas 77005, USA.
{druths,nakhleh}@cs.rice.edu

The central role phylogeny plays in biology and its pervasiveness in comparative genomics studies have led researchers to develop a plethora of methods for its accurate reconstruction. Most phylogeny reconstruction methods, though, assume a single tree underlying a given sequence alignment. While a good first approximation in many cases, a tree may not always model the evolutionary history of a set of organisms. When events such as interspecific recombination occur, different regions in the alignment may have different underlying trees. Accurate reconstruction of the evolutionary history of a set of sequences requires recombination detection, followed by separate analyses of the non-recombining regions. Besides aiding accurate phylogenetic analyses, detecting recombination helps in understanding one of the main mechanisms of bacterial genome diversification. In this paper, we introduce RECOMP, an accurate and fast method for detecting recombination events in a sequence alignment. The method slides a fixed-width window across the alignment and determines the presence of recombination events based on a combination of topology and parsimony score differences in neighboring windows. On several synthetic and biological datasets, our method performs much faster than existing tools with accuracy comparable to the best available method.

1. Introduction

Phylogeny, i.e., the evolutionary history of a set of organisms, plays a major role in representing and understanding relationships among the organisms. The rapidly-growing host of applications of comparative genomics has moved phylogeny to the forefront, rendering it an indispensable tool for analyzing and understanding the structure and function of genomes and genomic regions. Further, understanding evolutionary change and its mechanisms also bears direct impact on unraveling the genome structure and understanding phenotypic variations. One such mechanism of evolutionary change is *interspecific recombination*—the exchange of genetic material among different organisms across species boundaries.

Accurate detection of recombination is important for at least two major reasons. Studies have shown that the presence of recombination events has negative effects on the quality of the reconstructed phylogenetic tree.[9,12] Therefore, accurate reconstruction of the evolutionary history of a set of sequences that contains recombination events necessitates first detection of recombination events and then individual analyses of the non-recombined regions. Further, recombination plays a significant role in bacterial genome diversification. Whereas eukaryotes evolve mainly though lineal descent and mutations, bacteria obtain a large proportion of their genetic diversity through the acquisition of sequences from distantly related organisms, via horizontal gene transfer (HGT) or recombination.[6] Further,

recombination is one of the processes by which bacteria develop resistance to antibiotics.[1,7]

In light of their effects on the accuracy of phylogenetic methods and their significance as a central evolutionary mechanism, developing accurate methods for detecting recombination is imperative. Many methods have been proposed for this problem (for example, Posada studied the performance of 14 different recombination detection methods[8]). Recombination detection methods fall into various categories, depending on the strategies they employ.[10] Among those categories, phylogeny-based detection methods are currently the most commonly used.[10] Recombination events result in different phylogenetic trees underlying different regions of the sequence alignment, and it is this observation that forms the basis for phylogeny-based recombination detection methods. The most recent methods include PLATO (Partial Likelihood Assessed through Tree Optimization),[2] DSS (Difference of Sum of Squares),[5] and PDM (Probabilistic Divergence Measure).[3,4] Central to all these methods is the idea of sliding a window along the alignment of sequences, fitting data in each window to a phylogeny, and comparing phylogenies in neighboring windows.

Ruths and Nakhleh addressed the limitations of these methods, and introduced preliminary measures for recombination detection.[12] In this paper, we extend our previous work by considering both the topologies of trees and their parsimony scores across adjacent windows of the alignment. We introduce a new phylogeny-based framework, RECOMP (RECOMbination detection using Parsimony), that uses parsimony-based tree reconstruction and evaluation, coupled with measurement of topological differences. We have implemented and studied the performance of four different measures (within the RECOMP framework) on synthetic as well as biological datasets. Our results show that RECOMP's accuracy is comparable to the most accurate existing methods, and is much faster.

The rest of the paper is organized as follows. In Section 2 we briefly describe interspecific recombination and review the most recent phylogeny-based methods for its detection. In Section 3, we describe our new method, RECOMP. We describe our experimental settings and results in Section 4, and conclude in Section 5 with final remarks and directions for future research.

2. Phylogeny-based Recombination Detection

Interspecific (or inter-species) recombination is a process by which genetic material is exchanged between different species lineages. When interspecific recombination events occur, different regions in the sequence alignment may have different underlying trees, as illustrated in Figures 1 and 2. The sequence alignment depicted in Figure 1 has three non-recombining regions I, II, and III, defined by a recombination event that involves the exchange of region II sequences between organisms B and D. The phylogenetic tree shown in Figure 2(a) models the evolutionary history of regions I and III of the alignment, whereas the phylogenetic tree in Figure 2(b) models the evolutionary history of region II of the alignment.

The scenario depicted in these two figures illustrates that recombination events may result in different phylogenetic trees underlying different regions; this phenomenon is the basis for phylogeny-based recombination detection methods. Three of the most recent and

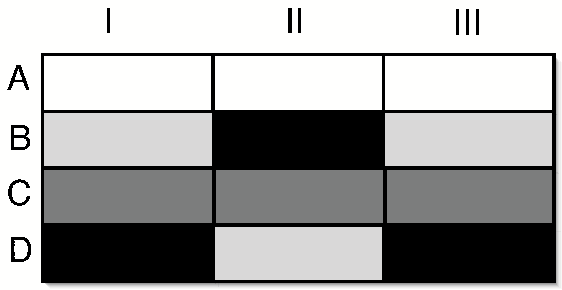


Figure 1. An alignment of four sequences whose evolutionary history contains a recombination event that involves the exchange of sequences in region II between organisms B and D.

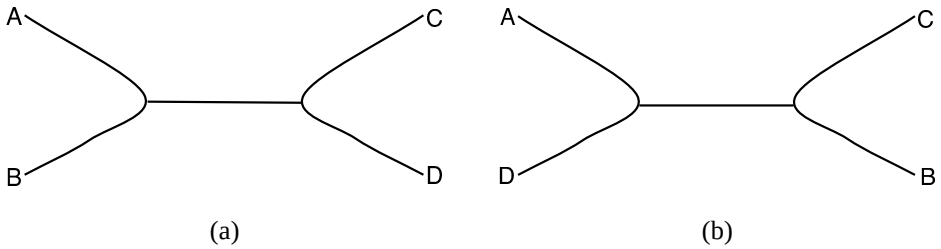


Figure 2. (a) The phylogenetic tree underlying regions I and III of the alignment in Figure 1. (b) The phylogenetic tree underlying region II of the alignment in Figure 1.

accurate phylogeny-based recombination detection methods are PLATO (Partial Likelihood Assessed through Tree Optimization),[2] DSS (Difference of Sum of Squares),[5] and PDM (Probabilistic Divergence Measure).[3,4] Central to all these methods is the idea of sliding a window along the alignment of sequences, fitting data in each window to a phylogeny, and comparing phylogenies in neighboring windows.

PLATO computes the likelihood of various regions of the sequence alignment from a single reference tree. The idea is that recombination regions will have a low likelihood score. The main problem with this approach is that the reference tree may be inaccurate since it is estimated from the whole sequence alignment.

DSS improves upon PLATO by sliding a window along the alignment, computing a tree on the first half of the window, and estimating the fit of the second half of the window to that tree (using a distance-based measure). The main problem with this approach is that it uses distance-based methods; such methods are inaccurate, especially given short sequences (which is the case when using DSS).

PDM addresses the shortcomings of DSS by (1) considering a likelihood approach for fitting the data to a tree, (2) using a distribution over trees, rather than a single tree (to capture the uncertainty of tree estimation from short sequences), and (3) comparing trees based on changes to their topologies. Later, Husmeier and Wright further improved the performance of PDM by incorporating sophisticated tree clustering techniques.[4] Since

PDM uses a probabilistic approach, it is very slow in practice. Further, since the tree space has very high dimensionality, clustering trees may be problematic.

3. RECOMP

Our proposed method is similar to PDM in principle, yet much simpler and faster, and comparable in accuracy. We slide a window of width w along the alignment, obtaining a set $\mathcal{T}_i$ of trees on $\mathcal{S}_i$, the set of sequences in the i^{th} window, using a maximum parsimony heuristic (heuristic search with branch swapping, as implemented in PAUP*[13]), and comparing the sets $\mathcal{T}_i$ and $\mathcal{T}_{i+1}$ of trees. The MP heuristic we use returns a set of trees, sorted by their parsimony scores: some trees may have an identical parsimony score. We denote the set of *all* j^{th} ($j = 1, 2, \ldots$) best parsimony trees (with respect to their scores, sometimes called the j^{th} *level*) by LVL^j, and the set of trees in the top k levels by $OPT(k)$ ($k \geq 1$), formally the set $\cup_{1 \leq \ell \leq k} LVL^\ell$. In the experimental study of our method, we considered $\mathcal{T}_i = OPT(k)$, and studied the performance of the method as a function of the k value (we used $k = 1, 2, 3, 4$).

Let $\mathcal{T}$ be a set of trees. We define the *center* of the set, $c(\mathcal{T})$, to be the strict consensus[a] of all trees in the set, and the *radius*, $r(\mathcal{T}) = \max\{RF(c(\mathcal{T}), T) : T \in \mathcal{T}\}$, where RF denotes the Robinson-Foulds distance between a pair of trees.[11] Further, we define $d_{min}(T, \mathcal{T}) = \min\{RF(T, T') : T' \in \mathcal{T}\}$ and $d_{max}(T, \mathcal{T}) = \max\{RF(T, T') : T' \in \mathcal{T}\}$. We write $P(S, T)$ to denote the parsimony score of tree T leaf-labeled by set S of sequences, and $P(S, \mathcal{T})$ to denote $\min_{T \in \mathcal{T}} P(S, T)$. We investigated four functions for comparing the sequences in two adjacent windows W_i and W_{i+1}:

- **Intersection(W_i,W_{i+1})** = $\frac{|\{T : T \in \mathcal{T}_{i+1} \text{ and } RF(T, c(\mathcal{T}_i)) \leq r(\mathcal{T}_i)\}|}{|\mathcal{T}_{i+1}|} + \frac{|\{T : T \in \mathcal{T}_i \text{ and } RF(T, c(\mathcal{T}_{i+1})) \leq r(\mathcal{T}_{i+1})\}|}{|\mathcal{T}_i|}$.
- **AvgMin(W_i,W_{i+1})** = $\frac{\sum_{T \in \mathcal{T}_i} d_{min}(T, \mathcal{T}_{i+1})}{|\mathcal{T}_i|} + \frac{\sum_{T \in \mathcal{T}_{i+1}} d_{min}(T, \mathcal{T}_i)}{|\mathcal{T}_{i+1}|}$.
- **AvgMax(W_i,W_{i+1})** = $\frac{\sum_{T \in \mathcal{T}_i} d_{max}(T, \mathcal{T}_{i+1})}{|\mathcal{T}_i|} + \frac{\sum_{T \in \mathcal{T}_{i+1}} d_{max}(T, \mathcal{T}_i)}{|\mathcal{T}_{i+1}|}$.
- **ParsDiff(W_i,W_{i+1})** = $|P(\mathcal{S}_{i+1}, \mathcal{T}_{i+1}) - P(\mathcal{S}_i, \mathcal{T}_{i+1})|$.

Further, we normalized the values computed by each of the four functions as follows. Let m and n be the minimum and maximum values, respectively, obtained by a function across all windows for a given sequence alignment. We normalize each value x computed by the function on the alignment by

$$\frac{x - m}{n - m}.$$

Therefore, the four functions return values in the range $[0, 1]$. The rationale behind the functions is as follows. Given an alignment of sequences, each of length L, let i be a site falling at a recombination breakpoint. Further, assume that the window we consider is of

[a] The strict consensus of a set of trees is the maximally resolved tree (i.e., the tree that has a maximum number of edges) in which every edge is also an edge of every tree in the set.

width w. Then, the tree T on which sites $(i-w)\ldots(i-1)$ is different from tree T' on which sites $i\ldots i+(w-1)$ evolved. Due to the inaccuracy of phylogeny reconstruction methods, and the potential errors in evolutionary assumptions made, T and T' may be unattainable; hence the need for considering sets of trees, rather than a single tree. When sets $\mathcal{T}_i$ and $\mathcal{T}_{i+1}$ correspond to sequence regions that fall on different sides of a recombination breakpoint, we expect the trees to differ between the two sets, which implies a lower **Intersection** value, and higher **AvgMin**, **AvgMax**, and **ParDiff** values. When the two sets of tree correspond to sequence regions that fall on the same side of any recombination event, we expect a higher **Intersection** value, and lower **AvgMin**, **AvgMax**, and **ParDiff** values. For consistency purposes, we always report $1-I$, where I is the value computed by the comparison function.

The outline of the RECOMP method is as follows:

RECOMP(S**,**w**,**t**)**

for $i=0$ **to** $L-w$

$X_i = f(W_i, W_{i+1})$;

$i = i+t$;

Plot X.

The sequence alignment is denoted by S, the window size by w, and the step size by t. The parameter L denotes the length of the sequences in S, f can be any of the aforementioned four functions, and W_i denotes the sequence alignment in window i. The output of RECOMP is a graphical representation of the output of the functions. Choosing a threshold that distinguishes the recombination sites can be determined by inspecting the graphical output of RECOMP (as is the case with all phylogeny-based methods that have graphical output). Further, such a threshold can be automatically computed by a careful training of the method on datasets with characteristics similar to those of the dataset under investigation.

4. Empirical Performance

4.1. *Data*

To test our method, we applied it to the three synthetic and one biological datasets used in another paper.[4] For the three synthetic datasets $SD1$, $SD2$, and $SD3$, the evolution of three DNA sequence alignments, each of 5500 nucleotides, was simulated down trees with 8 leaves. Each of the two datasets $SD1$ and $SD2$ contains two recombination events: an ancient event affecting the region between sites 1000 and 1500, and a recent event affecting the region between sites 2500 and 3000. Further, they both contain a mutational hot spot between sites 4000 and 4500 (sites were evolved at an increased nucleotide substitution rate) to test whether the detection method can successfully distinguish between recombination and rate variation. The average branch lengths of the phylogenetic trees underlying datasets $SD1$ and $SD2$ are 0.1 and 0.01, respectively. The third synthetic dataset, $SD3$, contains two recombination events: an ancient event affecting the region between sites 1000 and

2000, and a recent recombination event between sites 3000 and 4000. The branch lengths of the phylogenetic tree underlying dataset $SD3$ were drawn from a uniform distribution on the interval $[0.003, 0.005]$.

The biological dataset, HD, consists of 10 Hepatitis B Virus sequences each of 3049 nucleotides, with evidence for recombination events (the dataset contained two recombinant strains and eight nonrecombinant strains). For more details on the datasets, the reader is referred to the original paper.[4]

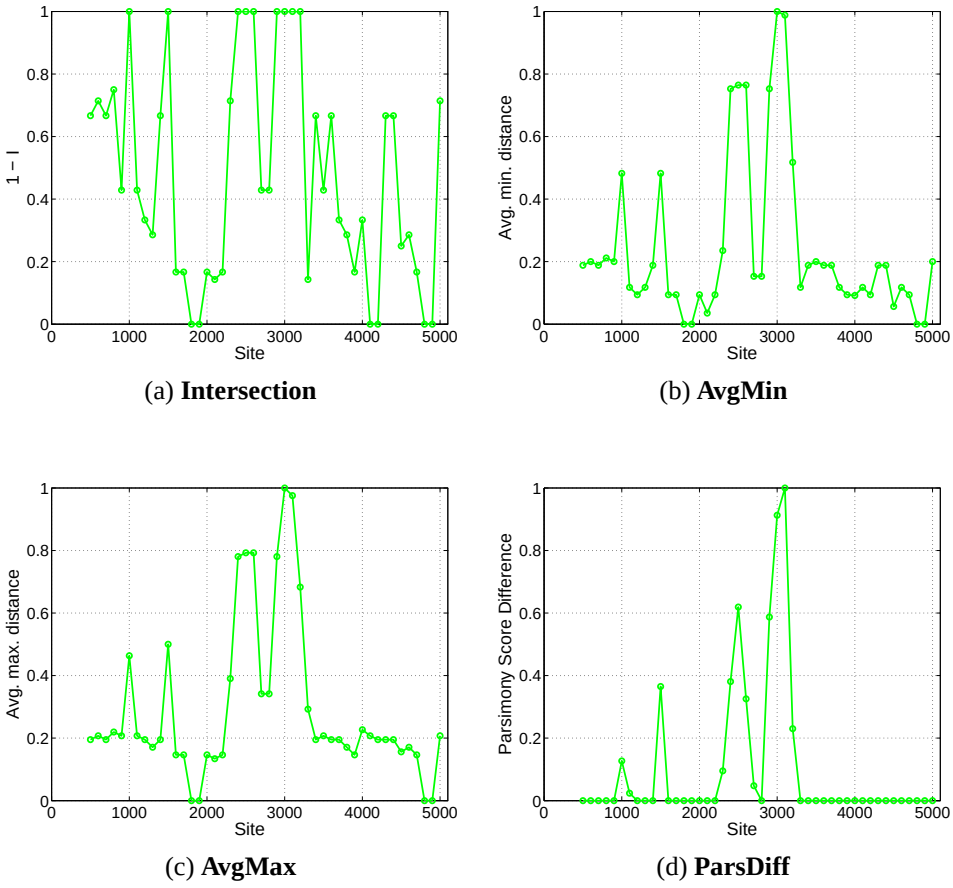
4.2. *Results*

We ran RECOMP with all four functions on the four datasets. We considered four different values of k (1,2,3, and 4) for sets $OPT(k)$ of trees, two window sizes 300 and 500, and step size 100. We describe our results of the four functions on all datasets when using $OPT(3)$ for window size 500, which produced the best results among all parameter settings. These results are shown in Figures 3—5 for the three synthetic datasets, and in Figure 6 for the biological dataset.

In the case of the $SD1$ dataset, our method detected the four recombination breakpoints (at sites 1000, 1500, 2500, and 3000) based on all four functions (Figure 3). There are clear threshold values that could be used as cutoff values between recombination/non-recombination regions: 0.8 for the **Intersection** function, 0.4 for the **AvgMin** and **AvgMax** functions, and 0.1 for the **ParsDiff** function. Clearly, the signal for a recombination breakpoint at sites 2500 and 3000 is stronger than that at sites 1000 and 1500. The reason for this is that the recombination event involving the region between sites 2500 and 3000 occurred between more distantly related taxa, which results in larger topological differences and parsimony score differences among trees across recombination breakpoints. Observe that the **ParsDiff** function is very robust, in this case, to the mutational hotspots: it correctly predicts no recombination in the mutational hotspot region between sites 4000 and 4500. The **Intersection** function has the strongest signal of recombination at all four recombination breakpoints (sites 1000, 1500, 2500, and 3000); however, the function is sensitive to mutational hotspots, and exhibits large fluctuations.

Similar behavior was obtained by the four functions on the dataset $SD2$ (Figure 4). However, in the case of this dataset, the **AvgMin** and **AvgMax** functions showed a weak signal for the ancient recombination event between sites 1000 and 1500. The **Intersection** and **ParsDiff** function still showed clear signal for recombination at all four breakpoints. Once again, the **ParsDiff** outperformed all other three functions in robustness with respect to mutational hotspots. The $SD2$ dataset was evolved with a lower rate of evolution than that of $SD1$ and hence was harder to analyze (which is the case for the other existing methods[4]).

The $SD3$ dataset was evolved down the tree with the lowest rate of evolution among all three synthetic datasets, and hence was the hardest for the methods to analyze (which is the case for the other existing methods[4]). As with the other two datasets, detecting the recent recombination event is easier, which is shown in the performance of all four functions in Figure 5. In particular, all four functions had a weak signal of recombination at site 2000.



(a) **Intersection** (b) **AvgMin**

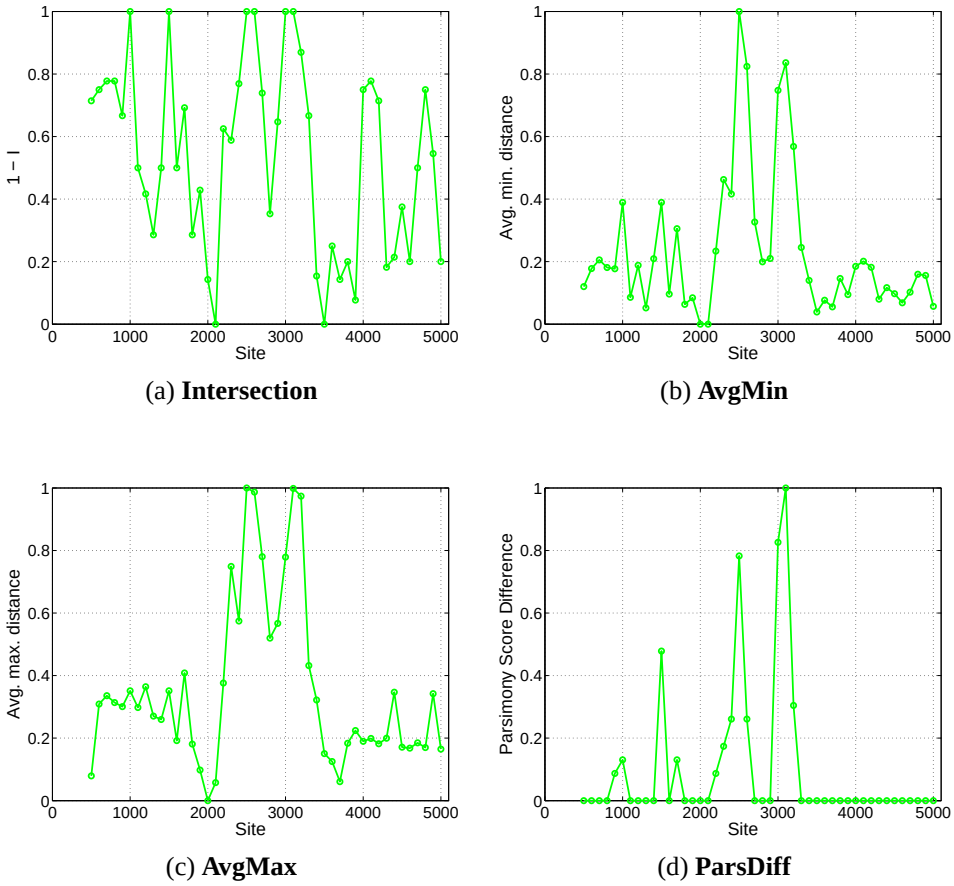
(c) **AvgMax** (d) **ParsDiff**

Figure 3. Results of the four functions on $SD1$.

Yet again, most of the sites in this alignment were synonymous, which made it hard for all methods to detect recombination.

On the Hepatitis B dataset, both the DSS and PDM methods detected three breakpoints around sites 600, 1700, and 2200. Our method shows peaks at these three points, based upon the four functions we used (Figure 6). Nevertheless, the **Intersection** and **ParsDiff** functions gave the clearest signal among the two.

The performance of PLATO, DSS, and PDM on the same datasets is provided by Husmeier and Wright.[4] The performance of our method is comparable to that of PDM, which performed best among those three methods. Further, since our method uses a fast MP heuristic, calculates parsimony scores (which is computable in polynomial time), and computes simple functions, it is much faster (orders of magnitude) than PDM, which uses compute-intensive Bayesian analysis techniques.



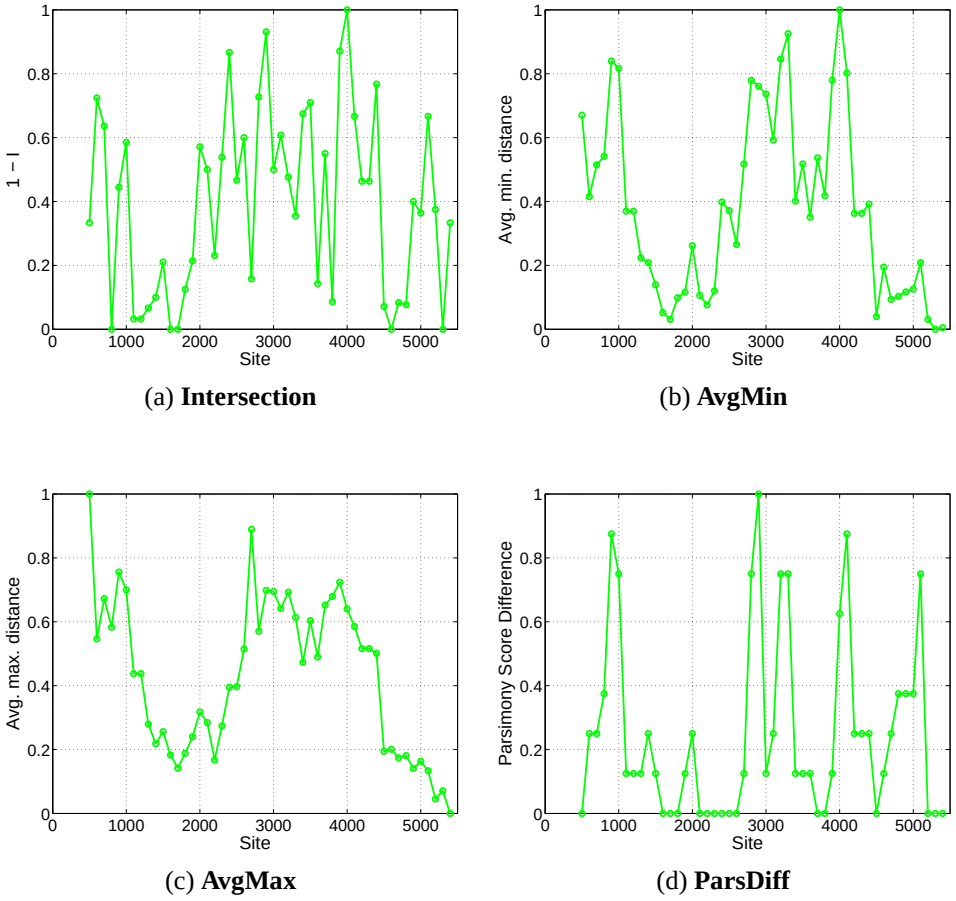
(a) **Intersection** (b) **AvgMin**

(c) **AvgMax** (d) **ParsDiff**

Figure 4. Results of the four functions on $SD2$.

5. Conclusions and Future Work

In this paper, we introduced a simple, effective and fast parsimony-based method for detecting recombination. In experimental studies involving both synthetic and biological datasets, our method produced very good results—comparable to those of the best known methods (and ran orders of magnitude much faster). Our future work includes exploring ways to improve the performance of our method in the presence of mutational hot spots. Further, we are interested in devising methods for detecting the locations of the recombination events on the organismal tree. An open-source, stand-alone implementation of RECOMP is currently available for download and use. It is implemented in the Sequoia software suite as both a command-line tool as well as a Java library which allows its incorporation into larger programs.



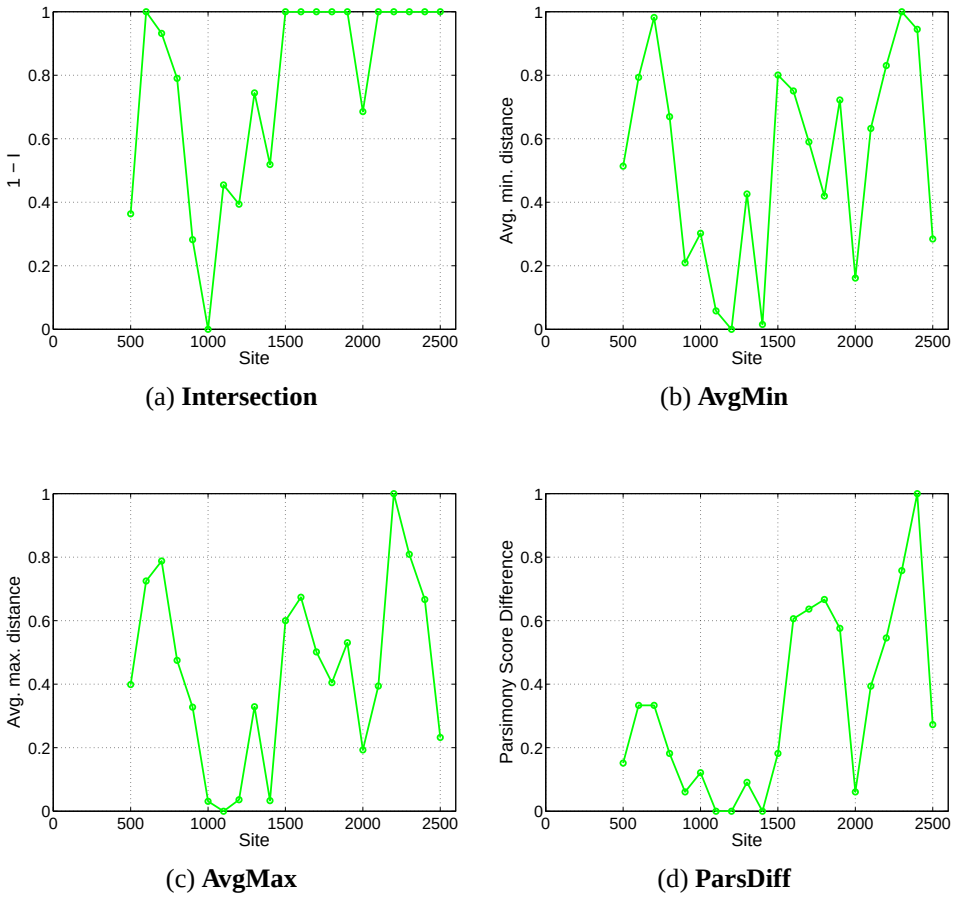
(a) **Intersection** (b) **AvgMin**

(c) **AvgMax** (d) **ParsDiff**

Figure 5. Results of the four functions on $SD3$.

References

1. M.C. Enright, D.A. Robinson, G. Randle, E.J. Feil, H. Grundmann, and B.G. Spratt. The evolutionary history of methicillin-resistant Staphylococcus aureus (MRSA). *Proc. Natl. Acad. Sci. USA*, 99(11):7687–7692, 2002.
2. N.C. Grassly and E.C. Holmes. A likelihood method for the detection of selection and recombination using nucleotide sequences. *Molecular Biology and Evolution*, 14:239–247, 1997.
3. D. Husmeier and F. Wright. Probabilistic divergence measures for detecting interspecies recombination. *Bioinformatics*, 17:S123–S131, 2001.
4. D. Husmeier and F. Wright. Detecting interspecific recombination with a pruned probabilistic divergenec measure. *Unpublished manuscript*, 2004.
5. G. McGuire, F. Wright, and M.J. Prentice. A graphical method for detecting recombination in phylogenetic data sets. *Molecular Biology and Evolution*, 14:1125–1131, 1997.
6. H. Ochman, J.G. Lawrence, and E.A. Groisman. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784):299–304, 2000.
7. I.T. Paulsen *et al*. Role of mobile DNA in the evolution of Vacomycin-resistant Enterococcus



(a) **Intersection** (b) **AvgMin**

(c) **AvgMax** (d) **ParsDiff**

Figure 6. Results of the four functions on HD.

faecalis. *Science*, 299(5615):2071–2074, 2003.

8. D. Posada. Evaluation of methods for detecting recombination from DNA sequences: empirical data. *Molecular biology and evolution*, 19:708–717, 2002.
9. D. Posada and K.A. Crandall. The effect of recombination on the accuracy of phylogeny estimation. *Journal of Molecular Evolution*, 54:396–402, 2002.
10. D. Posada, K.A. Crandall, and E.C. Holmes. Recombination in evolutionary genomics. *Annual Review of Genetics*, 36:75–97, 2002.
11. D. Robinson and L. Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53:131–147, 1981.
12. D. Ruths and L. Nakhleh. Recombination and Phylogeny: Effects and Detection. *The International Journal of Bioinformatics Research and Applications*, In press, 2005.
13. D. L. Swofford. PAUP*: Phylogenetic analysis using parsimony (and other methods), 1996. Sinauer Associates, Underland, Massachusetts, Version 4.0.