Recombination and phylogeny: effects and detection

Derek Ruths* and Luay Nakhleh

Department of Computer Science,
Rice University, 6100 Main Street, 77005 TX, Houston
E-mail: druths@cs.rice.edu E-mail: nakhleh@cs.rice.edu
*Corresponding author

Abstract: The role of phylogeny in guiding comparative studies is rapidly growing in the post genomic era. Most phylogeny reconstruction methods though, assume a single tree underlying a given alignment of sequences. However, when events such as recombination occur, different regions in the alignment may have different underlying trees. In this paper, we demonstrate via simulations, the effect of recombination on the accuracy of phylogeny reconstruction methods. Our results, coupled with the significance of recombination as an evolutionary mechanism, make it imperative to devise efficient and accurate methods for detecting recombination in sequence datasets. Hence, we introduce a simple, yet effective, method for detecting recombination in a given alignment, based on incongruence among phylogenetic trees in different regions of the alignment. We have studied the performance of our method on synthetic and biological datasets, and obtained good results.

Keywords: phylogeny; recombination; maximum parsimony; phylogenetic networks.

Reference to this paper should be made as follows: Ruths, D. and Nakhleh, L. (2005) 'Recombination and phylogeny: effects and detection', *Int. J. Bioinformatics Research and Applications*, Vol. 1, No. 2, pp.202–212.

Biographical notes: Derek Ruths is a Graduate Student in the Department of Computer Science at Rice University, Houston, TX, where he also studied for his BSc Degree. His research interests include Bacterial Genomics and Computational Gene Finding.

Luay Nakhleh received his PhD Degree in Computer Science in 2004 from the University of Texas, Austin, TX. His Doctoral dissertation was about Phylogenetic Networks in Biology and Historical Linguistics. He joined the Department of Computer Science at Rice University as an Assistant Professor in July 2004. His research interests include Phylogenetics, Bacterial Genomics, and Computational Gene Finding.

1 Introduction

Phylogeny, i.e., the evolutionary history of a set of organisms, plays a major role in representing and understanding the relationship among those organisms. The rapidly growing host of applications of comparative genomics has moved phylogeny to the forefront as an indispensable tool for analysing and understanding the structure and function of genomes and various genomic regions. Further, understanding evolutionary change and its mechanisms also bears a direct impact on unravelling the genome structure and understanding phenotypic variations. One such mechanism of evolutionary change is recombination. In this paper, we use the term recombination to refer collectively to events that lead to different trees underlying different genomic regions; examples of such events are crossing over, gene conversion, horizontal gene transfer, and hybrid speciation (Linder et al., 2004).

Given the significance of recombination as an evolutionary mechanism and the central role phylogeny plays in evolutionary biology, we address two questions:

- what effects does recombination have on the accuracy of phylogeny reconstruction methods?
- how can phylogeny be used to detect recombination in a set of sequences?

Posada and Crandall addressed the first question in Posada and Crandall (2002). They studied the effect of recombination on the accuracy of phylogeny estimation through simulation studies. A set of sequences was evolved down a model tree. The effect of a recombination event was simulated by exchanging parts of the sequences between two subtrees. A phylogeny was reconstructed (using a variety of methods) and compared with the model phylogeny for its accuracy. In this paper, we extend their work by considering larger trees, to study the effect of the size of the subtree involved in recombination. Further, while Posada and Crandall studied the effect of a single recombination event, we study, as well, the effect of two (dependent) recombination events, to evaluate whether single and multiple recombination events have similar effects.

Recombination detection has been studied extensively. In Posada (2002), Posada studied the performance of 14 different recombination detection methods. Recombination detection methods fall into different categories depending on the strategies they use Posada et al. (2002). Among those categories, phylogeny-based detection methods are currently most commonly in use (Posada et al., 2002). The newest method in this category is PDM (Probabilistic Divergence Measure), of Husmeier and Wright (2004) which we briefly describe later. In this paper, we introduce a fast and simple, yet accurate, parsimony-based technique for recombination detection that is comparable to the other phylogeny-based methods.

The rest of the paper is organised as follows. In Section 2 we give a brief background on the various types of recombination events and describe in more detail, recombination involving different species (interspecific recombination), which is what we address in this paper. In Section 3 we describe the simulation study we carried out, to study the effect of recombination on phylogeny reconstruction and discuss our results. Our phylogeny-based detection method is described in Section 4. We conclude in Section 5 with final remarks and directions for future research.

2 Recombination

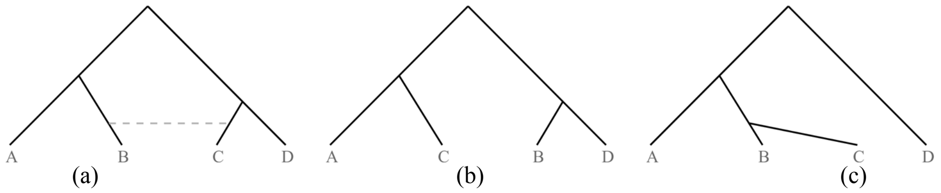
The term ‘recombination’ is used to refer to several different biological phenomena; in this paper, we use the term to refer collectively to events that cause incongruence among trees underlying the evolutionary histories of different genomics regions. Examples of such events are crossing over, horizontal gene transfer, and hybrid speciation

(Linder et al., 2004). Recombination occurs at different levels: individual, population, and species.

Recombination at the individual or chromosome level is known as meiotic recombination. During each round of sexual reproduction, the total number of chromosomes must be halved to produce the gametes. The process is called meiosis and during one phase of it the chromosome pair (sister chromatids) exchange pieces in a precise fashion known as meiotic recombination. The net result is chromatids that have two or more evolutionary histories on them. Blocks of chromosomes that share a single evolutionary history are referred to as haplotype blocks; see Wall and Pritchard (2003) for example. During sexual reproduction, the offspring inherits one chromatid from each of its two parents. Since each of these chromatids might have undergone meiotic recombination in the parent, the offspring's chromosomes may be mosaic. This event is usually referred to as sexual recombination.

Interspecific (or interspecies) recombination is a process by which genetic material is exchanged between different species lineages. In eukaryotes, meiotic and sexual recombination events are the prevalent ones. On the other hand, prokaryotes provide several possible pathways of recombination – conjugation, transformation, and transduction – that involve the non-reciprocal replacement or addition of sequences rather than their exchange. When interspecific recombination events occur, different regions in the alignment of sequences may have different underlying trees, as illustrated in Figure 1. The dashed line between the lineages of *B* and *C* in Figure 1 (a) denote a recombination event. In the case of reciprocal recombination, the tree in Figure 1(b) depicts the evolutionary history of the segment that was exchanged. The tree in Figure 1(c) depicts the evolutionary history of the segment that was transferred in a non-reciprocal recombination event from *B* to *C.* (See Posada et al., 2002; Linder et al., 2004 for a more detailed exposition of recombination.)

Figure 1 (a) A tree on four species; the dashed line corresponds to a recombination event. The solid lines represent the 'based' tree, down which the sequences without recombination evolved. In the case of reciprocal recombination, some regions in the sequences are exchanged between *B* and *C*, and their evolutionary history is depicted in (b). In the case of non-reciprocal recombination (e.g., horizontal gene transfer) in which some genetic region was transferred from *B* to C, the evolutionary history of that region is depicted in (c)

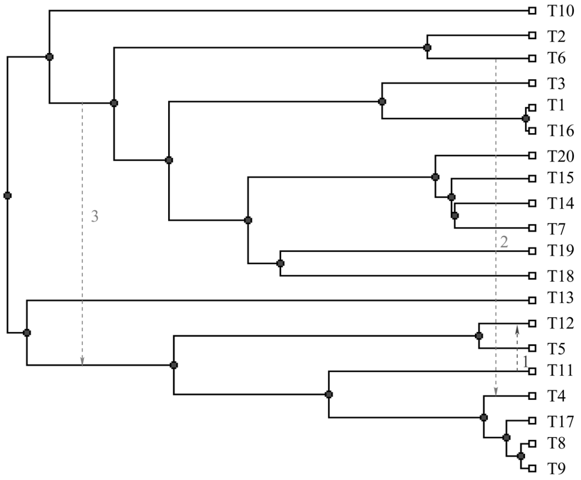


3 Effects of recombination on phylogeny reconstruction

3.1 Experimental settings

We used the r8s tool (Sanderson, 2002) to generate a random birth-death tree with 20 leaves, whose topology is shown in Figure 2. The evolutionary rate (expected number of changes in a site) along every path from the root to a leaf in the tree is 1.0. To allow for different values, we scaled the tree using three different scaling factors: 0.1, 0.3, and 0.6.[1] The r8s tool generates molecular clock[2] trees; to deviate evolution from the molecular clock, we multiplied each edge by a random number drawn with exponential distribution from the range [–1, 1].

Figure 2 The base tree which was used in the simulation study. The three dashed lines 1, 2, and 3 denote the *close, divergent,* and *ancient* recombination events, respectively. In the case of *reciprocal* recombination, genetic material is exchanged between the two endpoints of the dashed line. In the case of *non-reciprocal* recombination, the genetic material is transferred in the direction denoted by the arrow on the dashed line



In this study, we considered datasets with one or two recombination events. For 'single recombination' datasets, we considered *close* (edge #1 in Figure 2), *divergent* (edge #2 in Figure 2), or *ancient* (edge #3 in Figure 2). For 'two recombination' datasets, we considered the ancient close and ancient divergent dependent combinations. For all cases, we looked at reciprocal and non-reciprocal recombination. As illustrated in Figure 1, when a single recombination event takes place, there are two possible trees (the base and alternate) down which various sequence regions may evolve. In the same way, when two recombination events take place, there are four possible trees down which various sequence regions may evolve. Since we considered dependent combinations, there were only two trees down which sequences were evolved.

For each combination of an alternate tree, scaling factor, and sequence length (1000, 1500, 2000, 2500, and 3000), we used the Seq-Gen tool (Rambaut and Grassly, 1997) to evolve 30 sequence datasets. We used the GTR + Γ model[3] with the settings of Zwickl and Hillis (2002).

We considered five different recombination percentages: 10%, 20%, 30%, 40%, and 50%. For 'single recombination' datasets, x% recombination was simulated by evolving $(100 - x)$% sites down the base tree, and x% sites down the alternate tree, and concatenating the two datasets (as was done in Posada and Crandall, (2002). For 'two recombination' datasets, $(100 - x)$% sites were evolved down the base tree, and x% sites were evolved down the tree with the two subtrees at the end of each recombination edge swapped[4], and the two datasets concatenated.

For tree reconstruction, we used the neighbour joining (NJ) method (Saitou and Nei, 1987) and a maximum parsimony (MP) heuristic (heuristic search with branch swapping), both as implemented in the PAUP* package (Swofford, 1996). To compare the reconstructed tree T' against the model tree T, we used the Robinson-Foulds (RF) measure (Robinson and Foulds, 1981), which we now briefly review.

Let T be an unrooted tree ‘leaf labelled’ by a set S of taxa. An edge $e = (u, v)$ in T defines a bipartition of S (the set of all leaves on one side of the edge, and the set of all other leaves). Let $C(T)$ be the set of bipartitions defined by all edges in tree T. The RF measure between two trees T and T' is defined as

$$RF(T, T') = (|C(T) - C(T')| \,/\, |C(T)| + |C(T') - C(T)| \,/\, |C(T')|)/2.$$

3.2 Experimental results and analysis

We describe the results of NJ on the subset of the data shown in Figure 3 (similar results were obtained for MP). A few observations are in order. As expected, the error rate of NJ with respect to the base tree grows as the recombination percentage increases. This growth is much faster in the case of a divergent recombination event (Figure 3(a)), the reason being that a divergent recombination event spans a large part of the base tree and hence affects many edges. Notice that a recombination event affects all the edges on the path between the two endpoints of that event; thus, the longer that path (equivalently, the more divergent the recombination event), the higher the error rate. Further, the error rates of the method with respect to both trees (under which the sequences evolved) become equal at 50% (reciprocal) recombination (Figures 3(a)–(c)). At that point, the method obtains equal signals from both trees, and hence behaves similarly with respect to both trees.

Figure 3 The effect of recombination on the accuracy of the reconstructed phylogeny (using NJ) as a function of the recombination percentage in the dataset. Sequence length is 2,000, and scaling factor is 0.3. Each curve corresponds to the RF value between the constructed tree and one of the possible model trees: ‘Δ’corresponds to the base tree; ‘ ’ and ‘◊’ correspond to trees resulting from one recombination event; ‘+’ corresponds to tree resulting from both recombination events

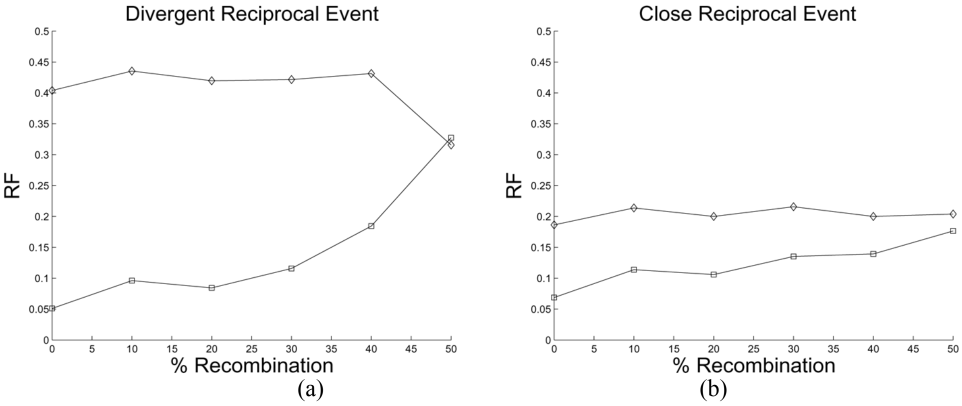
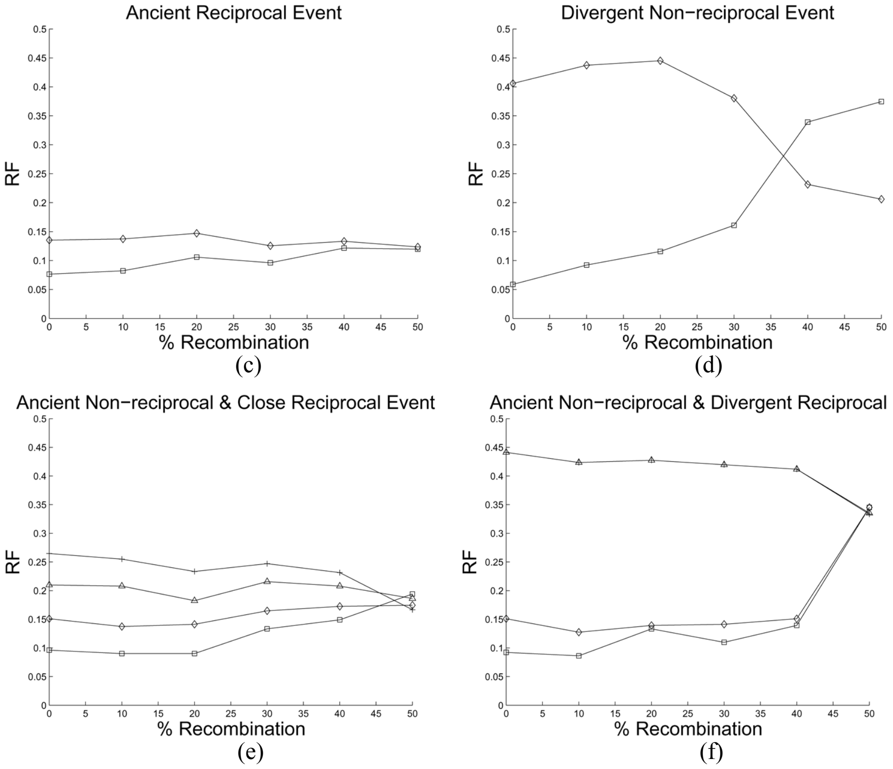


Figure 3 The effect of recombination on the accuracy of the reconstructed phylogeny (using NJ) as a function of the recombination percentage in the dataset. Sequence length is 2,000, and scaling factor is 0.3. Each curve corresponds to the RF value between the constructed tree and one of the possible model trees: ‘Δ’corresponds to the base tree; ‘ ’ and ‘◊’ correspond to trees resulting from one recombination event; ‘+’ corresponds to tree resulting from both recombination events (continued)



(c)

(d)

(e)

(f)

In the case of non-reciprocal recombination (Figure 3(d)), the growth of the error rate with respect to the base tree is steeper than that in the case of reciprocal recombination (Figure 3(a)). Further, the two curves cross around the 35% recombination rate. The reason for this difference in the behaviour of the method is that while reciprocal recombination involves the exchange of two subtrees, non-reciprocal recombination involves moving a single subtree.

In the case of two recombination events (Figures 3(e) and (f)), we observe similar behaviour. The error rate with respect to the base tree grows as the recombination rate increases. The performance of the method suffers significantly in all cases when the recombination rate is higher than 40%. The method always infers a tree that is closer to either of the two trees with the effect of a single recombination event than to the tree with the effect of two recombination events. We also observe that the error rate of the method is higher in the presence of two recombination events than in the presence of a single event.

In summary, depending on the extent of recombination in a dataset, this process may seriously confound the accuracy of phylogenetic methods. Therefore, it is imperative to design accurate methods for detecting recombination, so that it can be appropriately handled before a phylogeny reconstruction is attempted.

4 Detecting recombination

As illustrated in Section 2, recombination events result in different phylogenetic trees underlying different regions; this phenomenon is the basis for phylogeny-based recombination detection methods. We propose a phylogeny-based recombination detection method that is based on ideas from PLATO (Partial Likelihood Assessed through Tree Optimisation) (Grassly and Holmes, 1997), DSS (Difference of Sum of Squares) (McGuire et al., 1997), and PDM (Probabilistic Divergence Measure) (Husmeier and Wright, 2001, 2004). 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 neighbouring 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

- considering a likelihood approach for fitting the data to a tree
- using a distribution over trees, rather than a single tree (to capture the uncertainty of tree estimation from short sequences)
- comparing trees based on changes to their topologies.

Later, Husmeier and Wright further improved the performance of PDM by incorporating sophisticated tree clustering techniques (Husmeier and Wright, 2004). Since PDM uses a probabilistic approach, it is very slow in practice. Further, the tree space has very high dimensionality, and clustering trees may be problematic.

4.1 Our method

Our proposed method is similar to PDM in principle, yet much simpler. We slide a window of width w along the alignment, obtaining a set T_i of trees on the ith window using a maximum parsimony heuristic (heuristic search with branch swapping, as implemented in PAUP* (Swofford, 1996), and comparing the sets T and T_{i+1} of trees. The MP heuristic we use returns a set of trees, sorted by their parsimony scores. We denote by O^j the set of *all* jth best parsimony trees (with respect to their scores), and by $OPT(i)$ $(i \geq 1)$ the set $U_{i \leq j \leq i}O^j$. In the experimental study of our method, we considered $T_i = \mathrm{OPT}(j)$, and studied the performance of the method as the function of the j value (we used $j = 1, 2, 3, 4$).

Let T be a set of trees. We define the *centre* of the set, $c(T)$, to be the strict consensus[5] of all trees in the set, and the *radius*, $r(T) = \max\{RF(c(T), T)\colon T \in T\}$. Further, we define $d_{\min}(T, T) = \min\{RF(T, T')\colon T' \in T\}$. We investigated two functions for comparing the sets of trees:

- *Intersection* $(T_i, T_{i+1}) = |\{T: T \in T_{i+1} \text{ and } RF(T, c(T_i)) \leq r(T_i)\}| / |T_{i+1}|$

  $+ |\{T{:}T \in T_i \text{ and } RF(T, c(T_{i+1})) \leq r(T_{i+1})\}| / |T_i|.$

- *AvgMin* $(T_i, T_{i+1}) = \sum_{T \in Ti} d_{\min}(T, T_{i+1})/|T_i| + \sum_{T \in Ti+1} d_{\min}(T, T_i)/|T_{i+1}|.$

The rationale behind our method is as follows. Given an alignment of sequences, each of length *k*, let *i* be a site falling at a recombination breakpoint. Further, assume that the window we consider is of 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 (similar to PDM). When sets T_i and 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* 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* values.

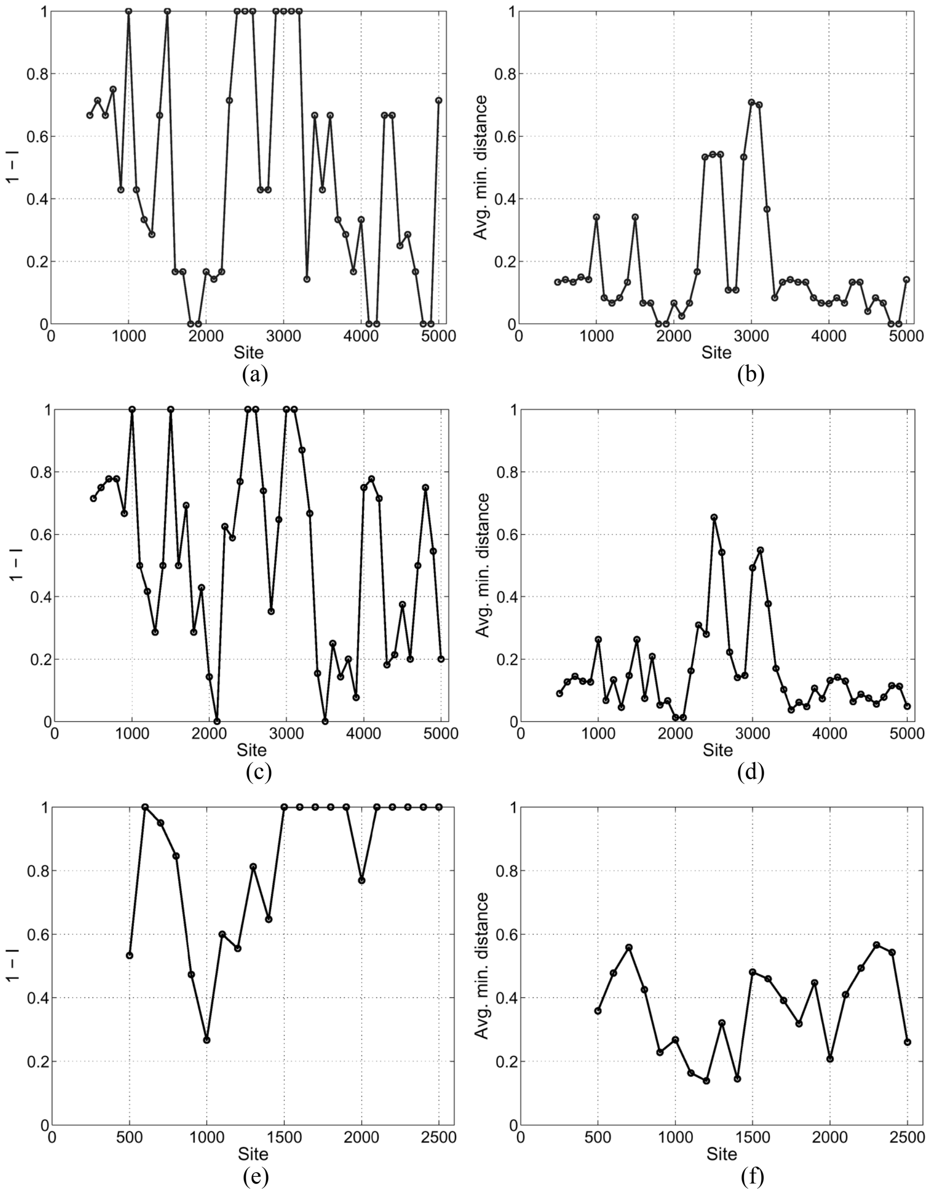
4.2 Data

To test our method, we applied it to two synthetic and one biological datasets used in Husmeier and Wright (2004). For the synthetic data, the evolution of two DNA sequence alignments, each of 5,500 nucleotides, was simulated down trees with eight leaves. The two trees differed in the average branch length. The datasets (hereafter referred to as *SD*1 and *SD*2) had two recombination events: an ancient event affecting the region between sites 1,000 and 1,500, and a recent event affecting the region between sites 2,500 and 3,000. Both datasets contained a mutational hot spot between sites 4,000 and 4,500 (sites were evolved at an increased nucleotide substitution rate) to test whether the detection method can successfully distinguish between recombination and rate variation. The biological dataset *HD* consisted of 10 Hepatitis B Virus sequences each of 3,049 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 Husmeier and Wright (2004).

4.3 Results

In the case of the *SD*1 dataset, our method detected the four recombination breakpoints (at sites 1000, 1500, 2500, and 3000) using either of the two functions (Figures 4(a) and (b)). There are clear threshold values that could be used as cutoff values between recombination/non-recombination regions: 0.8 and 0.2 for the *Intersection* and *AvgMin* functions, respectively. Similar behaviour was obtained by the two functions on the dataset *SD*2 (Figures 4(c) and (d)). However, in the case of this dataset, the *AvgMin* function could not discern the 1,000 and 1,500 site breakpoints as clearly, and the peaks were lower. The reason for this is that *SD*2 was evolved with a lower rate than that of *SD*1 and hence it was harder to analyse (which was the case for all methods described in Husmeier and Wright (2004)). On the Hepatitis B dataset, both the DSS and PDM methods detected three breakpoints around sites 600, 1,700, and 2,200. Our method had peaks at these three points based upon the two functions we used (Figures 4(e) and (f)). Nevertheless, the *Intersection* function gave the clearest signal among the two.

Figure 4 The performance of our method on the *SD*1 (top row), *SD*2 (middle row), and *HD* (bottom row) datasets. The left column corresponds to the *Intersection* function; we plot the value of one minus the *Intersection* value. The right column corresponds to the *AvgMin* function. These results we obtained using the set *OPT*(3) of MP trees (see Section 4.1). The window was slid by a step size of 100 sites



(a) (b)

(c) (d)

(e) (f)

The performance of PLATO, DSS, and PDM on the same datasets is provided in Husmeier and Wright (2004). The performance of our method is comparable to that of PDM, which performed best among those three methods. Further, since our method is parsimony-based and computes simple functions, it is much faster (orders of magnitude) than PDM, which uses intensive Bayesian analysis techniques.

5 Conclusion

Phylogeny is an indispensable tool in comparative studies, and its accuracy bears a great impact on the outcome of those studies. In this paper, we showed that, depending on its extent, recombination can be a serious confounding factor for phylogeny reconstruction. Moreover, recombination is a significant evolutionary mechanism motivating the need for accurate methods for its detection. 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 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 species tree.

References

Grassly, N.C. and Holmes, E.C. (1997) 'A likelihood method for the detection of selection and recombination using nucleotide sequences', *Molecular Biology and Evolution*, Vol. 14, pp.239–247.

Husmeier, D. and Wright, F. (2001) 'Probabilistic divergence measures for detecting inter-species recombination', *Bioinformatics*, Vol. 17, pp.S123–S131.

Husmeier, D. and Wright, F. (2004) 'Detecting interspecific recombination with a pruned probabilistic divergenec measure', *Bioinformatics*, to appear.

Linder, C.R., Moret, B.M.E., Nakhleh, L. and Warnow, T. (2004) 'Network (reticulate) evolution: biology, models, and algorithms', *The Ninth Pacific Symposium on Biocom-puting (PSB)*, A tutorial.

McGuire, G., Wright, F. and Prentice, M.J. (1997) 'A graphical method for detecting recombination in phylogenetic data sets', *Molecular Biology and Evolution*, Vol. 14, pp.1125–1131.

Posada, D. (2002) 'Evaluation of methods for detecting recombination from DNA sequences: empirical data', *Molecular biology and evolution*, Vol. 19, pp.708–717.

Posada, D. and Crandall, K.A. (2002) 'The effect of recombination on the accuracy of phylogeny estimation', *Journal of Molecular Evolution*, Vol. 54, pp.396–402.

Posada, D., Crandall, K.A. and Holmes, E.C. (2002) 'Recombination in evolutionary genomics', *Annual Review of Genetics*, Vol. 36, pp.75–97.

Rambaut, A. and Grassly, N.C. (1997) 'Seq-gen: an application for the monte carlo simulation of dna sequence evolution along phylogenetic trees', *Comp. Appl. Biosci.*, Vol. 13, pp.235–238.

Robinson, D. and Foulds, L. (1981) 'Comparison of phylogenetic trees', *Mathematical Bio-sciences*, Vol. 53, pp.131–147.

Saitou, N. and Nei, M. (1987) 'The neighbor-joining method: a new method for reconstructing phylogenetic trees', *Molecular Biology and Evolution*, Vol. 4, pp.406–425.

Sanderson, M. (2002) *r8s software package*, available from http://loco.ucdavis.edu/r8s/r8s.html.

Swofford, D.L. (1996) *PAUP: Phylogenetic Analysis Using Parsimony (and Other Methods)*, Sinauer Associates, Underland, Massachusetts, Version 4.0.

Wall, J.D. and Pritchard, J.K. (2003) 'Haplotype blocks and linkage disequilibrium in the human genome', *Nat. Rev. Genet.*, Vol. 4, No. 8, pp.587–597.

Zwickl, D. and Hillis, D. (2002) 'Increased taxon sampling greatly reduces phylogenetic error', *Systematic Biology*, Vol. 51, No. 4, pp.588–598.

Notes

[1] Scaling a tree by scaling factor x means multiplying the weight of each edge in the tree by x.

[2] The molecular clock hypothesis states that the amount of change during evolution is proportional to time. This assumption results in *ultrametric* trees, which have the property that the lengths of all paths from the root of the tree to any leaf are all equal.

[3] General time reversible model with Gamma distributed 'rates across sites'.

[4] In the case of non-reciprocal recombination, and instead of swapping, the subtree at one end of the recombination edge was made a sibling of the subtree at the other end.

[5] The strict consensus of a set of trees is a tree with maximum number of edges, whose every edge appears in every tree in the set.