



Probabilistic Divergence Measures for Detecting Interspecies Recombination

Dirk Husmeier and Frank Wright

Biomathematics & Statistics Scotland, SCRI, Invergowrie, Dundee, DD2 5DA, UK

ABSTRACT

This paper proposes a graphical method for detecting interspecies recombination in multiple alignments of DNA sequences. A fixed-size window is moved along a given DNA sequence alignment. For every position, the marginal posterior probability over tree topologies is determined by means of a Markov chain Monte Carlo simulation. Two probabilistic divergence measures are plotted along the alignment, and are used to identify recombinant regions. The method is compared with established detection methods on a set of synthetic benchmark sequences and two real-world DNA sequence alignments.

Contact: dirk@bioss.ac.uk

INTRODUCTION

The recent advent of multiple-resistant pathogens has led to an increased interest in interspecies recombination as an important, and previously underestimated, source of genetic diversification in bacteria and viruses. The discovery of a surprisingly high frequency of mosaic RNA sequences in HIV-1 suggests that a substantial proportion of AIDS patients have been coinfecting with HIV-1 strains belonging to different subtypes, and that recombination between these genomes can occur *in vivo* to generate new biologically active viruses (Robertson et al., 1995). A phylogenetic analysis of the bacterial genera *Neisseria* and *Streptococcus* has revealed that the introduction of blocks of DNA from penicillin-resistant non-pathogenic strains into sensitive pathogenic strains has led to new strains that are both pathogenic and resistant (Smith, 1992). Thus interspecies recombination, illustrated in Figure 1, raises the possibility that bacteria and viruses can acquire biologically important traits through the exchange and transfer of genetic material.

In the last few years, a plethora of methods for detecting interspecies recombination have been developed – following up on the seminal paper by Smith (1992) – and it is beyond the scope of this article to mention them all. Instead, we will focus on statistical phylogenetic procedures. Here, the idea is to compute a phylogeny-based score function for varying subsets (‘windows’) of the alignment, and to record deviations between these subsets or between a subset and the whole alignment as possible indications for pu-

tative recombination events.

PLATO (Grassly and Holmes, 1997) first finds the phylogenetic tree that maximises the likelihood of the whole DNA sequence alignment, and then systematically looks for subsets with a low likelihood under this model by computing the statistic

$$Q = \frac{\sum_{t=bW}^{(b+1)W-1} L_t}{W} / \frac{\sum_{t=1}^{bW-1} L_t + \sum_{t=(b+1)W}^N L_t}{N - W} \quad (1)$$

where L_t denotes the log likelihood of the t th column vector of the alignment, W is the size of the window, and N is the length of the alignment. This measure is calculated for all possible positions b along the sequence alignment and for varying window sizes, typically $5 \leq W \leq N/2$. The maximum values of Q are associated with regions showing low likelihoods under the global maximum likelihood model, which are candidates for putative recombination events. Parametric bootstrapping is applied to generate the null distribution of the maximised Q value and thus to test whether Q is significantly larger than one, that is, whether Q is greater than what one would expect by chance.

TOPAL (McGuire et al., 1997) is a graphical method to detect sporadic recombination events. The idea is to slide a fixed-size moving window along a DNA sequence alignment. On the first half of the window, a distance matrix is calculated according to some Markov model of nucleotide substitution, and a phylogenetic tree is estimated using the least squares method. A distance matrix is then calculated for the second half of the window, and the topology estimated from the first half is fitted to it, again using least squares (see Figure 2, top). Obviously, when the topology in the right window has changed as a result of recombination, the topology in the left window will be a poor fit to the distance matrix from the right. Consequently, by plotting the difference of the two sum-of-squares statistics – termed the ‘DSS’ statistic – against the centre of each corresponding window, the alignment can be scanned for recombination. The significance of DSS peaks can be estimated with parametric bootstrapping, as reported in (McGuire and Wright, 2000).

LARD (Holmes et al., 1999) compares the null hypothesis H_0 that no recombination event has occurred with the

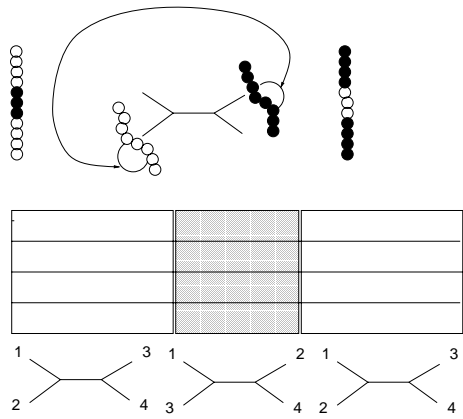


Fig. 1. Illustration of interspecies recombination. The transfer or exchange of DNA subsequences between different strains results in a change of the tree topology in the affected region of the DNA sequence alignment.

alternative hypothesis H_1 that there are two different trees at either side of some breakpoint b . For H_0 , a single phylogenetic tree is optimised with maximum likelihood on the whole sequence alignment, whereas for H_1 two different trees are optimised. For each value of b the difference between the log likelihood scores, $\Delta L = L(H_1) - L(H_0)$, is computed. The value of b that maximises ΔL is the candidate breakpoint for a putative recombination event. To test the significance of this event, a Monte Carlo simulation is applied to assess whether ΔL is significantly larger than zero.

While for all of these methods positive results have been reported, they are not without problems. PLATO uses a single reference tree to calculate the site likelihoods. If the recombinant regions are large relative to the whole data set, then the reference tree – determined with maximum likelihood from the entire sequence alignment – is some type of average of the dominant tree and the recombinant trees, and the site likelihood values in the different regions might not be differentiated properly. LARD assumes a block structure of the sequence alignment, with the two regions to the left and the right of a tentative breakpoint stemming from different phylogenies. This requires some prior knowledge about the location of putative recombination events, since the method is likely to fail if a recombinant region lies in the middle of the sequences. Finally TOPAL optimises the tree in each window with a distance method. By mapping the high-dimensional space of nucleotide sequences onto the low-dimensional space of pairwise genetic distances between strains, a certain amount of information is discarded. Also, the original algorithm failed to distinguish between recombination and rate variation, although this has partly

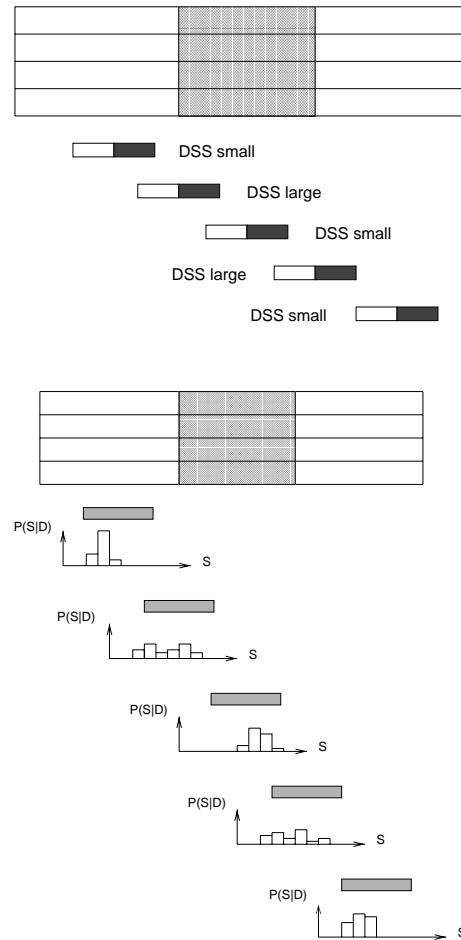


Fig. 2. Top: TOPAL. A sliding window is moved along the alignment. A tree is estimated from the data in the left part of the window, and a score function for assessing the quality of the fit is compared between the two parts of the window. Bottom: The method proposed in this article slides a window along the alignment and computes, for each position, the marginal posterior distribution over tree topologies (using MCMC).

been redeemed by a normalisation of the DSS values (McGuire and Wright, 2000).

In this paper, we will present a new method which focuses on the very entity that gets affected by the recombination event: the topology of the phylogenetic tree (see Figure 1). As with TOPAL, a window is moved along the sequence alignment. However, rather than optimising the tree and computing differences in the branch lengths, we sample trees from the posterior distribution, conditional on the given window, by means of a Markov chain Monte Carlo (MCMC) simulation. Marginalising over the branch lengths, we obtain the posterior distribution of tree topologies for each window position, from which we can compute divergence measures to scan for recombination.

METHOD

Consider a given alignment of DNA sequences, \mathbf{D} , from which we select a consecutive subset \mathbf{D}_t of predefined width W , centred on the t th site of the alignment. Let k be an integer label for tree topologies, and define

$$P_k(t) := P(k|\mathbf{D}_t) = \int \int P(k, \mathbf{w}, \boldsymbol{\theta}|\mathbf{D}_t) d\mathbf{w} d\boldsymbol{\theta} \quad (2)$$

This is the marginal posterior distribution of tree topologies k , conditional on the ‘window’ \mathbf{D}_t , which includes a marginalisation over the branch lengths \mathbf{w} and the parameters of the nucleotide substitution model, $\boldsymbol{\theta}$ (see, e.g., (Durbin et al., 1998), chapter 8). In practice the integral in (2) will be solved numerically by means of a Markov chain Monte Carlo simulation (Larget and Simon, 1999), which yields a sample of triples $\{k_{ti}, \mathbf{w}_{ti}, \boldsymbol{\theta}_{ti}\}_{i=1}^N$ simulated from the joint posterior distribution, $P(k, \mathbf{w}, \boldsymbol{\theta}|\mathbf{D}_t)$. We then replace the true posterior distribution by the empirical distribution

$$P(k, \mathbf{w}, \boldsymbol{\theta}|\mathbf{D}_t) \approx \frac{1}{N} \sum_{i=1}^N \delta_{k, k_{ti}} \delta(\mathbf{w} - \mathbf{w}_{ti}) \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_{ti}) \quad (3)$$

Inserting (3) into (2) gives:

$$P_k(t) = \frac{1}{N} \sum_{i=1}^N \delta_{k, k_{ti}} = \frac{N_k(t)}{N} \quad (4)$$

where $N_k(t)$ denotes the number of times a tree has been found to have topology k .

The basic idea of a graphical method for detecting recombinant regions is to move the window \mathbf{D}_t along the alignment and to monitor the distribution $P_k(t)$. We would then, obviously, expect a shift in the distribution as we move into a recombinant zone (see Figure 2, bottom). The question, then, is how to easily monitor such a shift and how to estimate its significance. To this end we consider the Kullback-Leibler distance as the natural divergence measure in probability space:

$$KL(P, Q) = \sum_k P_k \ln \left(\frac{P_k}{Q_k} \right) \quad (5)$$

in which P and Q denote probability distributions. To estimate the divergence between the local and the global distributions, define P as in (4) and Q as the average distribution (averaged over all T window positions):

$$d[P(t), \bar{P}] = KL[P(t), \bar{P}]; \quad \bar{P}_k = \frac{1}{T} \sum_{t=1}^T P_k(t) \quad (6)$$

We will refer to this as the KL measure and note that, since $\text{Support}(P_k(t)) \subseteq \text{Support}(\bar{P})$, its non-singularity is guaranteed. To determine the divergence between two local distributions $P(t)$ and $P(t + \Delta t)$ – locally defined over two adjacent windows with centre positions t and $t + \Delta t$ – we choose the following modified divergence measure (first suggested by Sibson; see, e.g., Krzanowski and Marriott (1995), chapter 14):

$$d[P(t), P(t + \Delta t)] = \frac{1}{2} \left[KL \left(P(t), \frac{P(t) + P(t + \Delta t)}{2} \right) + KL \left(P(t + \Delta t), \frac{P(t) + P(t + \Delta t)}{2} \right) \right] \quad (7)$$

Note again that $\text{Support}[P(t), \text{Support}[P(t + \Delta t)]] \subseteq \text{Support} \left[\frac{P(t) + P(t + \Delta t)}{2} \right]$ guarantees the non-singularity of $d[P(t), P(t + \Delta t)]$.

To estimate whether the observed divergence measures are significantly different from zero, note that under the null hypothesis, $P = Q$, the Kullback-Leibler divergence is asymptotically χ^2 distributed (Hoel, 1984):

$$P = Q, \quad M \gg \nu := |\text{Support}(P)| \\ \implies 2MKL(P, Q) \sim \chi^2(\nu - 1) \quad (8)$$

where M is the number of independent samples from which P and Q are determined, and $|\text{Support}(P)|$ denotes the cardinality of the support of P . Note that consecutive samples of an MCMC simulation are usually not independent, so the total sample size N has to be replaced by the equivalent *independent* sample size M . For example, if the autocorrelation function is exponential with an autocorrelation time $\tau \gg 1$, then $M = \frac{N}{2\tau}$.

The local divergence measure (7) depends on the distance between two windows, Δt . It seems natural to choose two consecutive windows, as in TOPAL. However, by allowing a certain overlap between the windows the spatial resolution of our detection method can be improved. We found that the best results can be obtained by averaging over different degrees of window overlap:

$$\bar{d} = \frac{1}{A} \sum_{a=1}^A d[P(t), P(t + a\Delta t)] \quad (9)$$

where $d(\cdot)$ was defined in (7), and we average over all window overlaps between 50% and 90%. A demonstration is given in Figure 3, which was obtained from a synthetic DNA sequence alignment, as described in the caption of Figure 4. We will refer to the divergence measure of (9) as ASD (average Sibson divergence). We thus have two probabilistic divergence measures: KL (6) as a global measure akin to PLATO, and ASD (9) as a local measure in the vein of TOPAL.

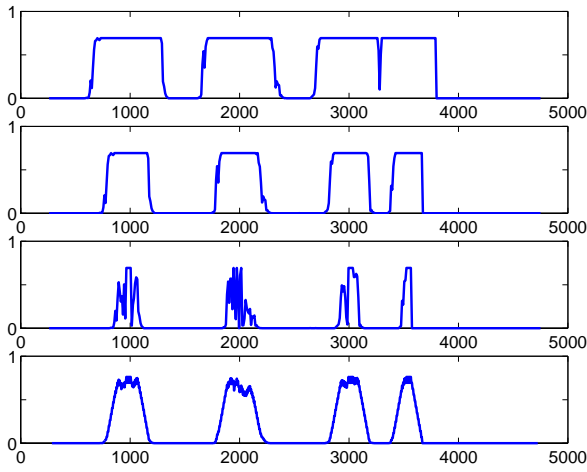


Fig. 3. Effect of averaging the local divergence measure. From top to bottom: 1) No overlap between adjacent windows; 2) 50% overlap between adjacent windows; 3) 90% overlap between adjacent windows; 4) averaged divergence measure, using (9). The true recombinant regions are between positions 1000 bp and 2000 bp, and between positions 3000 bp and 3500 bp.

MATERIAL

Synthetic Data To assess the performance of the proposed method, DNA sequences subject to recombination were obtained by simulating their evolution down known phylogenies, shown in Figure 4, using the Kimura 2-parameter model (transition-transversion ratio = 2) of nucleotide substitution. (See (Durbin et al., 1998) for an introduction to mathematical models of evolution.) A variety of different recombination scenarios was simulated. In experiment series A, partial sequences were evolved down different topologies, as indicated in the top of Figure 4, and then spliced together. This reflects the swapping of branches, that is, the *exchange* of DNA subsequences. In experiment series B, the sequences were simulated along the phylogeny as far as a particular depth (half the length of the indicated branch). At this point, a region from a sequence replaced the corresponding region in another sequence, as indicated in the bottom of Figure 4. The sequences were then evolved along the remaining part of the phylogeny. This simulates the *transfer* of genetic material between different strains. In both experiment series, two different recombination events of different detection difficulty were simulated, and the process was repeated for a variety of evolution rates (that is, unit branch lengths). In experiment series B, we also included a differently diverged region, where the branch lengths had been increased by a factor 3. This tests whether the proposed method can distinguish between rate variation and recombination. Details of the data are

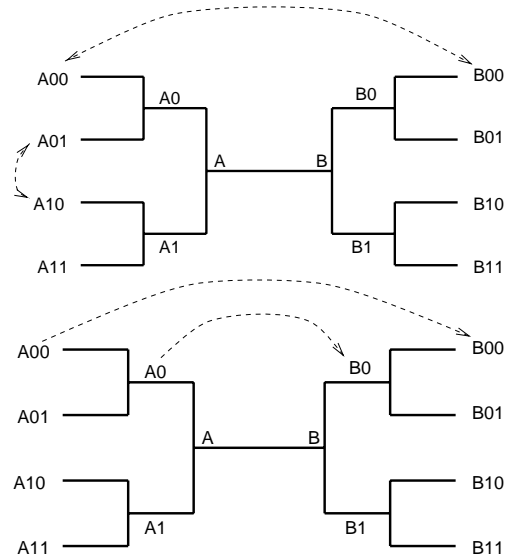


Fig. 4. Phylogenetic trees for the synthetic data. The top tree (simulation A) shows recent recombinations between closely related taxa ($A01 \leftrightarrow A10$) and distantly related taxa ($A00 \leftrightarrow B00$), where the indicated lineages are swapped. The bottom tree (simulation B) shows a recent ($A00 \leftrightarrow B00$) and an ancient recombination event between distantly related taxa, where DNA subsequences are transferred at a single point in time (half-way along the branch).

Exp	Recombination			Rate Variation	Basic Rate
	Recent, close	Recent, distant	Ancient		
A.1	1000-2000	3000-4000	—	—	0.1
A.2	1000-2000	3000-4000	—	—	0.025
A.3	1000-2000	3000-4000	—	—	0.01
B.1	—	2500-3000	1000-1500	4000-4500	0.1
B.2	—	2500-3000	1000-1500	4000-4500	0.05
B.3	—	2500-3000	1000-1500	4000-4500	0.01

Table 1. Details of the synthetic data sets, where w denotes the unit branch lengths of the trees in Figure 4.

shown in Table 1.

Hepatitis B Virus Hepatitis B is caused by a DNA virus with a short genome of only 3200 bp. Evidence for recombination was first found by Bollyky et al. (1996), and in this paper we investigate a subset of five strains with the following Genbank identifiers (accession numbers in brackets): HPBADW1 (D00329), HPBADW2 (D00330), HPBADWZCG (M57663), HBVDNA (X68292), HPBADRC (D00630). The sequences were aligned with ClustalW, using the default parameters. Columns with

gaps were discarded, giving a total alignment length of 3049 bp.

Dengue Virus Dengue fever is a common vector-borne disease of humans with more than 100 million cases recorded each year in Africa, South America, and Southeast Asia. While it was originally believed that the causative RNA virus accumulated genetic variation only through nucleotide substitution, recent studies suggest that genetic exchange between strains is an increasing possibility (Holmes et al., 1999). In our study we investigated the mosaic structure of seven strains, for which contiguous C, prM/M, and E gene sequences were available. These seven strains were the same as those studied by (Holmes et al., 1999), with simplified names (GenBank accession numbers in brackets) Brazil (S64849), Singapore (M87512), Jamaica (D00501), Philippines (D00503), Thailand (D00502), Nauru (M23027), and French Guiana (not available on GenBank, taken from (Holmes et al., 1999)). We used the multiple sequence alignment from (Holmes et al., 1999), which has a length of 2295 bp[†].

RESULTS

In the applications described below, the proposed new recombination detection scheme was applied as follows. A window of width 500 bp was moved along the alignment with a fixed step size of $\Delta t = 10$ bp. The MCMC sampling was done with BAMBE[‡], described by Larget and Simon (1999). For each new window position, the system was equilibrated over T_{eq} Metropolis-Hastings steps, starting from a random initialisation and using global[§] tree manipulations for the proposal moves. This was followed by a sampling period over T Metropolis-Hastings steps, using local tree manipulations and sampling tree configurations in intervals of ΔT Metropolis-Hastings steps. The MCMC simulations were tested for convergence by inspecting the evolution of the total log likelihood and its autocorrelation function as well as by testing the results for consistency, that is, by ensuring that the results were invariant with respect to a further increase of the equilibration and sampling periods. We found that, for the synthetic problem, a choice of $T_{eq} = 10,000$, $T = 20,000$, and $\Delta T = 40$ was sufficient. On the real-world problems, we multiplied these values by a factor of 10. These, however, are conservative estimates that may be considerably reduced.

For a comparison, we applied TOPAL[¶], version 2, to the data sets, sliding a window of the same size, 500 bp,

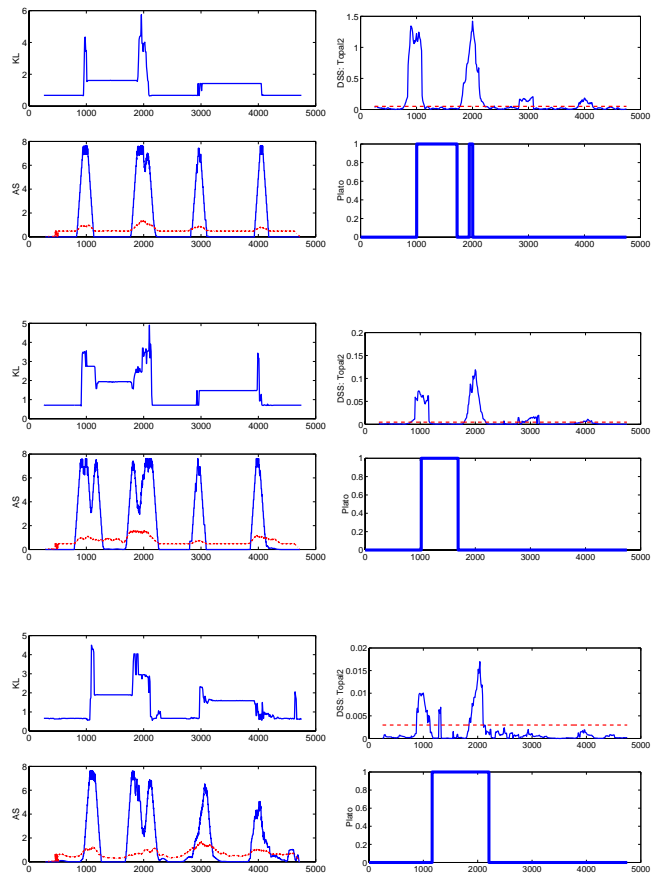


Fig. 5. Detection of recombination for the first synthetic problem, experiment series A. *Top four subfigures:* Large evolution rate, $w = 0.1$. *Middle four subfigures:* Medium evolution rate, $w = 0.025$. *Bottom four subfigures:* Small evolution rate, $w = 0.01$. The left subfigures in each group show the results obtained with the new method. *Top left:* KL measure (peaks) with the 95% critical region for the null hypothesis (dotted horizontal curve). For a comparison, the *top-right* subfigure shows the DSS statistic of TOPAL (solid curve) with the 95% critical region for the null hypothesis (dashed horizontal line), and the *bottom-right* subfigure shows the critical regions predicted by PLATO. The recombinant regions are situated between 1000 and 2000 bp (involving distantly related strains) and between 3000 and 4000 bp (involving closely related strains).

with the same step size, 10 bp, along the alignment. The default option^{||} was chosen, and we tested the significance of the DSS peaks with parametric bootstrapping, as described by McGuire and Wright (2000). We also applied PLATO^{**}, version 2.11, to the data sets, using the default

[†]The total length of the alignment in (Holmes et al., 1999) is 2325 bp, of which we discarded the columns with gaps.

[‡]<http://www.mathcs.duq.edu/larget/bambe.html>.

[§]In BAMBE, two different kinds of proposal moves – local and global – are used, which are described by Larget and Simon (1999).

[¶]www.bioss.sari.ac.uk/~frank/Genetics/topal.html.

^{||}We chose the more accurate *least-square* tree optimisation rather than *neighbour joining* and avoided the unstable power option (that is, power=0). However, we replaced the (default) Jukes-Cantor evolution model with the Kimura 2-parameter model (transition-transversion ratio = 2).

^{**}evolve.zoo.ox.ac.uk/software/Plato/Plato2.html.

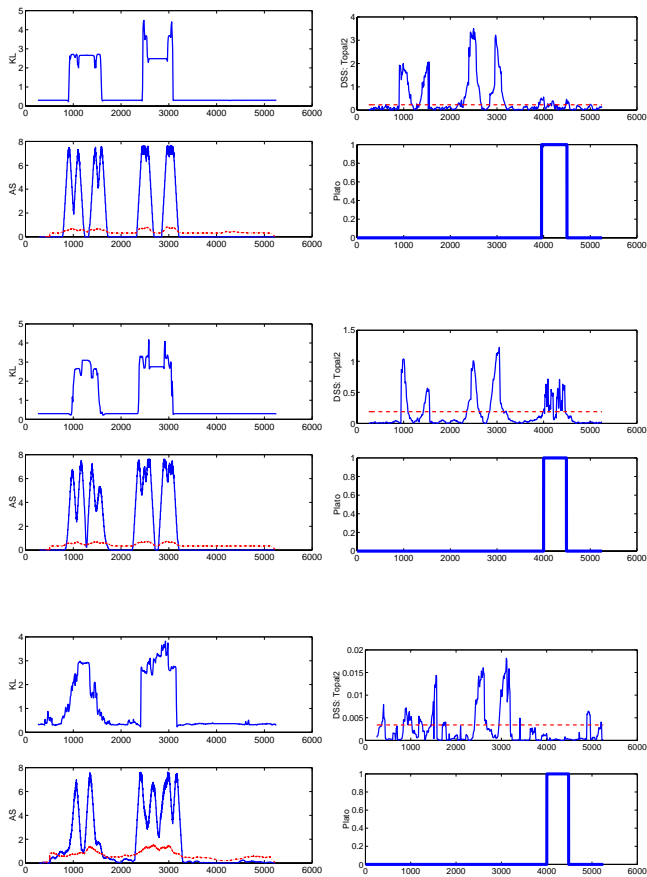


Fig. 6. Detection of recombination for the second synthetic problem, experiment series B. *Top four subfigures:* Large evolution rate, $w = 0.1$. *Middle four subfigures:* Medium evolution rate, $w = 0.05$. *Bottom four subfigures:* Small evolution rate, $w = 0.01$. The left subfigures in each group show the results obtained with the new method. *Top left:* KL measure. *Bottom left:* ASD measure (peaks) with the 95% critical region for the null hypothesis (dotted horizontal curve). For a comparison, the *top-right* subfigure shows the DSS statistic of TOPAL (solid curve) with the 95% critical region for the null hypothesis (dashed horizontal line), and the *bottom-right* subfigure shows the critical regions predicted by PLATO. The recombinant regions are located between 1000 and 1500 bp (ancient recombination event) and between 2500 and 3000 bp (recent recombination event). The region between 4000 and 4500 bp has diverged at an increased evolution rate (factor 3) without being subject to recombination.

options. This tests all windows from five base pairs up to half the sequence length for significantly low values of the likelihood, where the reference tree was obtained with maximum likelihood on the whole data set (using DNAML of the PHYLIP^{††} package with the F84 model of nucleotide substitution). We did not test LARD on

^{††}evolution.genetics.washington.edu/phylip.html.

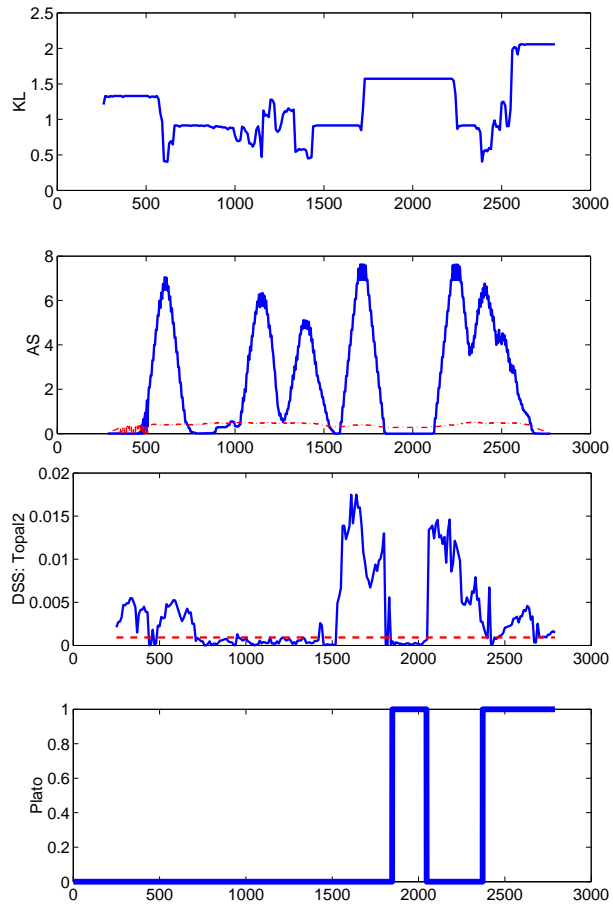


Fig. 7. Recombination in Hepatitis B virus. From top to bottom: 1) Global (KL) divergence measure; 2) Local (ASD) divergence measure; 3) DSS statistic of TOPAL; 4) critical regions predicted by PLATO.

the synthetic benchmark sequences because it was not devised to detect recombinant regions in the middle of an alignment (and therefore would presumably perform poorly), but rather to precisely locate breakpoints between *two* regions of different phylogenetic history.

Synthetic Data Figure 5 shows the results obtained with the various methods on the first synthetic benchmark problem. The different subgroups of the figure were obtained from phylogenetic trees with different unit branch lengths w . As we move from the top to the bottom, w decreases, which increases the difficulty of the detection. Also, note that the second recombination event (between closely related strains) is more difficult to detect than the first (between distantly related strains). PLATO detects the first recombinant region but fails to detect the second. TOPAL detects both recombinant regions when the unit branch length w is large, but fails to detect the second region as

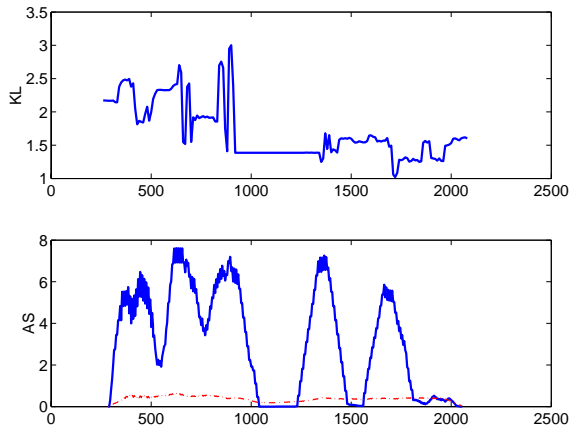


Fig. 8. Recombination in Dengue virus. *Top*: Global (KL) divergence measure. *Bottom*: Local (ASD) divergence measure.

the evolutionary distances decrease. Our new methods detects both regions for all branch lengths, although the accuracy slightly degrades as the branch lengths decrease. Note that the spatial resolution of the detection by TOPAL and our new method are comparable, and typical of the same size as the discrepancy between the prediction by PLATO and the true location of the recombinant zones.

Figure 6 demonstrates the performance of the various methods on the second synthetic problem. The level of difficulty is increased by a shortening of the recombinant zones and the existence of a confounding region, which has evolved at a higher rate (‘mutation hotzone’) *without* being subject to recombination. PLATO, in fact, fails to detect the recombinant zones and gets confounded by this differently diverged section. TOPAL fails to detect the first (ancient, and therefore more difficult) recombination event when the evolutionary distances are small (top-right subfigure in the bottom group) and fails, in one case (top-right subfigure in the middle group), to discriminate between recombination and rate variation. The new scheme, on the contrary, detects both recombination events irrespective of the branch lengths and succeeds in distinguishing between recombination and among-site rate variation.

Recombination in Hepatitis B Virus A plot of the KL measure (Figure 7, top) indicates, via step transitions, four breakpoints in the sequence alignment (600 bp, 1720 bp, 2250 bp, 2550 bp). These breakpoints coincide with peaks in the ASD measure (Figure 7, second from the top), and the widths of these peaks are indicative of the uncertainty in the location of the breakpoints. Within this uncertainty, the prediction resulting from our new method accords with the prediction by TOPAL (Figure 7, third from the top), while PLATO (Figure 7, bottom) misses out the

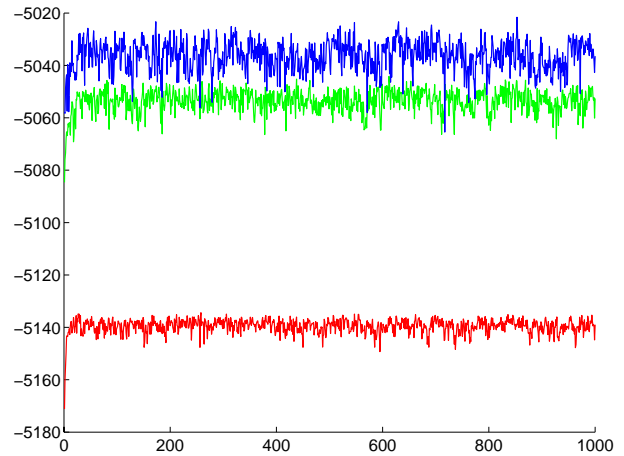


Fig. 9. Recombination in Dengue virus. Log likelihood scores sampled from the MCMC trajectories for the null-hypothesis (*bottom*), the segmentation suggested by LARD (*middle*), and the segmentation suggested by the KL divergence measure of Figure 8 (*top*).

first breakpoint. However, the ASD measure shows two further peaks at about 1175 bp and 1415 bp. It still has to be investigated whether these peaks indicate breakpoints of another true recombination event, or whether they are artifacts resulting from a noise amplification of the KL signal.

Recombination in Dengue Virus Holmes et al. (1999) applied LARD^{††} to detect a recombination breakpoint in the DNA sequence alignment of the seven Dengue virus strains described above. They identified one significant recombination breakpoint at nucleotide 1146, which we also found with TOPAL. This breakpoint, however, is not confirmed with our new detection method. Although the KL measure, plotted in Figure 8, top, is rather noisy, and the resolution of the ASD measure, plotted in Figure 8, bottom, is poor, these measures suggest that there is *no* breakpoint at nucleotide 1146, but rather two breakpoints at around nucleotides 920 and 1370^{§§}.

To assess this prediction, we first obtained the log likelihood scores for a single tree (H_0), the segmentation predicted with LARD (H_1), and the segmentation suggested by inspecting the KL measure in Figure 8 (H_2). This was done with DNAML of the PHYLIP package, using the F84 model of nucleotide substitution. The scores thus obtained were: H_0 : -5140, H_1 : -5051, and H_2 : -5035. Since the improvement of H_2 over H_1 can be due to over-fitting (H_2 is the more flexible model), we followed a Bayesian-

^{††}<http://evolve.zoo.ox.ac.uk/software/Lard/Lard.html>.

^{§§}The ASD measure indicates two further breakpoints at about 400 bp and 1750 bp, the nature of which still has to be investigated.

like approach and sampled the log likelihood values from the MCMC trajectories, using BAMBE with the HKY evolution model. The result is shown in Figure 9 and suggests that H_2 gives significantly higher log likelihood scores than H_1 . This is formally confirmed with a ranksum test, which rejects the hypothesis of equal distributions at a 99.9% significance level. Although this is not a proper Bayesian hypothesis test[¶], it gives a certain indication that the prediction made with our new approach is to be preferred over that in (Holmes et al., 1999). Note that this implies that there are two recombination events rather than one.

DISCUSSION

We have proposed a new phylogenetic method for the detection of interspecies recombination in DNA sequence alignments, and we have compared our new approach with two established methods: TOPAL and PLATO. On a benchmark sequence alignment generated from various synthetic evolution and recombination scenarios, the new approach outperformed PLATO and TOPAL in two respects. First, it could clearly distinguish between recombination and rate variation. Second, it detected *all* recombination events, whereas TOPAL and PLATO failed as the detection difficulty increased (short branch lengths and recombination between closely related strains). On the Hepatitis B virus sequences, the new approach reproduced (within the indicated uncertainty) the breakpoints detected with TOPAL, while it also predicted two additional breakpoints (the nature of which still has to be investigated). A substantial disagreement between the prediction of our new scheme with another established method, LARD, was found on the Dengue virus sequences. Although the KL measure was rather noisy and the spatial resolution of the ASD measure poor, they indicated a different segmentation of the alignment, which was found to have a higher likelihood than the segmentation predicted with LARD.

In spite of these fairly positive results, we need to point out certain restrictions of our scheme. Figures 5 and 6 demonstrate that as the evolutionary distances decrease (from top to bottom in both figures), the spatial resolution of the ASD signal deteriorates. Also, spurious peaks tend to occur, since ASD is a differential measure and therefore susceptible to amplification of noise in the more robust KL signal. Consequently, it still has to be investigated whether the additional peaks found in the ASD signal in Figures 7 and 8 indicate true recombination breakpoints or just artifacts caused by noise amplification. The sampling of tree topologies with MCMC increases the computational

costs, and the typical CPU time required on a SUN Ultra-10 was 5 hours (to be compared with 1.5 hour for PLATO, 10 minutes for TOPAL without bootstrapping, and 16 hours for TOPAL with bootstrapping). The spatial resolution for identifying the breakpoints of recombinant regions is limited by the size of the sliding window. A decrease of the window size increases the vagueness of the posterior distributions and thus leads to degraded results. As the number of strains becomes larger, the window size has to be increased to ensure that the posterior distributions are reasonably informative. This reduces the spatial resolution of the detection scheme and thus soon reaches the limit of practical viability as the number of strains exceeds a value of about ten.

Nevertheless, in spite of these limitations, the positive results on the synthetic benchmark data make us confident that our new scheme can make significant contributions in identifying the mosaic structure of DNA sequence alignments and may identify recombination events hitherto undetected with the established methods.

REFERENCES

- Bollyky, P. L., A. Rambaut, P. H. Harvey, and E. C. Holmes (1996). Recombination between Sequences of Hepatitis B Virus from Different Genotypes. *Journal of Molecular Evolution* 42, 97–102.
- Durbin, R., S. R. Eddy, A. Krogh, and M. G. (1998). *Biological sequence analysis. Probabilistic models of proteins and nucleic acids*. Cambridge, UK: Cambridge University Press.
- Grassly, N. C. and E. C. Holmes (1997). A Likelihood Method for the Detection of Selection and Recombination Using Nucleotide Sequences. *Molecular Biology and Evolution* 14(3), 239–247.
- Hoel, P. G. (1984). *Introduction to Mathematical Statistics*. Singapore: John Wiley and Sons.
- Holmes, E. C., M. Worobey, and A. Rambaut (1999). Phylogenetic Evidence for Recombination in Dengue Virus. *Molecular Biology and Evolution* 16(3), 405–409.
- Krzanowski, W. J. and F. H. C. Marriott (1995). *Multivariate Analysis*, Volume 2. Arnold. ISBN 0-340-59325-3.
- Larget, B. and D. L. Simon (1999). Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution* 16(6), 750–759.
- McGuire, G. and F. Wright (2000). TOPAL 2.0: improved detection of mosaic sequences within multiple alignments. *Bioinformatics* 16(2), 130–134.
- McGuire, G., F. Wright, and M. Prentice (1997). A Graphical Method for Detecting Recombination in Phylogenetic Data Sets. *Molecular Biology and Evolution* 14(11), 1125–1131.
- Robertson, D. L., P. M. Sharp, F. E. McCutchan, and B. H. Hahn (1995). Recombination in HIV-1. *Nature* 374, 124–126.
- Smith, J. M. (1992). Analyzing the Mosaic Structure of Genes. *Journal of Molecular Evolution* 34, 126–129.

[¶]A proper Bayesian hypothesis test is based on the marginal likelihood $P(\mathbf{D}|H_2)$. This requires an integration over the whole parameter space, which is intractable.