Bioinformatics Advance Access published August 7, 2006

BIOINFORMATICS

Robust inference of positive selection from recombining coding sequences

Konrad Scheffler *, Darren P. Martin , Cathal Seoighe *

Computational Biology Group, Institute of Infectious Disease and Molecular Medicine, University of Cape Town, Private Bag, Rondebosch 7701, South Africa
Tel: +27-21-406 6176; Fax: +27-21-650 5192

Associate Editor: Keith A Crandall

ABSTRACT

Motivation: Accurate detection of positive Darwinian selection can provide important insights to researchers investigating the evolution of pathogens. However, many pathogens (particularly viruses) undergo frequent recombination and the phylogenetic methods commonly applied to detect positive selection have been shown to give misleading results when applied to recombining sequences. We propose a method that makes maximum likelihood inference of positive selection robust to the presence of recombination. This is achieved by allowing tree topologies and branch lengths to change across detected recombination breakpoints. Further improvements are obtained by allowing synonymous substitution rates to vary across sites.

Results: Using simulation we show that, even for extreme cases where recombination causes standard methods to reach false positive rates above 90%, the proposed method decreases the false positive rate to acceptable levels while retaining high power. We applied the method to two HIV-1 data sets for which we have previously found that inference of positive selection is invalid due to high rates of recombination. In one of these (*env* gene) we still detected positive selection using the proposed method, while in the other (*gag* gene) we found no significant evidence of positive selection.

Availability: A *HyPhy* batch language implementation of the proposed methods and the HIV-1 data sets analysed are available at http://www.cbio.uct.ac.za/pub_support/bioinf06. The *HyPhy* package is available at http://www.hyphy.org, and it is planned that the proposed methods will be included in the next distribution. *RDP2* is available at http://darwin.uvigo.es/rdp/rdp.html.

Contact: konrad@cbio.uct.ac.za, cathal@science.uct.ac.za

1 INTRODUCTION

The standard phylogenetic approach to inferring positive Darwinian selection in protein-coding sequences is based on the codon models first proposed by Muse and Gaut (1994) and Goldman and Yang (1994), which have since been developed into a set of robust methods that detect positive selection while allowing for selective pressure to vary across sites (Nielsen and Yang, 1998; Yang *et al.*, 2000; Wong *et al.*, 2004). These methods, however, assume that the phylogenetic tree topology and branch lengths are constant across all sites in the sequence – an assumption which is invalid when the sequences have been affected by recombination. Indeed, it has been shown (Anisimova *et al.*, 2003; Shriner *et al.*, 2003) that the presence of recombination can cause these methods to fail with type I (false positive) error rates as high as 90%. In a recent study (Scheffler and Seoighe, submitted), we quantified the percentage of false positive inferences as a function of recombination rate and demonstrated that inferred positive selection on two example HIV data sets is invalidated by the presence of recombination.

Recombination can contribute to false inference of positive selection by causing the branch lengths (Figure 1(a)) and tree topologies (Figure 1(b)) to differ between sites. In order to devise a robust method of inferring positive selection we investigated the impact of allowing tree topology and branch length parameters to change across recombination breakpoints. In a real analysis we anticipate that a subset of recombination breakpoints might be undetected. In order to improve the performance of our method in the presence of a subset of undetected recombination breakpoints we included a variable synonymous substitution rate in our models, which allows the total tree length to vary from site to site. Sequences can evolve under a variable synonymous substitution rate due to mutation rate variation or due to site-specific selection acting on synonymous changes, but synonymous rate variation could also be detected as a result of recombination events that alter branch lengths. Incorporating synonymous rate variation in the model can therefore account for some of the misestimated branch lengths that result from recombination events that alter branch lengths but not tree topology. In general, we expect these recombination events to be more difficult to detect than those that cause a substantial change in tree topology. We evaluated the performance of the method by simulation and applied it to investigate whether the HIV data sets mentioned above can be inferred to be evolving under positive selection when recombination is taken into account.

2 MATERIALS AND METHODS

We generated a number of data sets using the Codonrecsim program written by Rasmus Nielsen (Anisimova *et al.*, 2003) that simulates recombined coding sequence alignments. It does this by simulating under a phylogenetic model of evolution using the discrete model (M3) of site-to-site rate variation proposed by Yang *et al.* (2000), but with the evolution taking place along genealogies simulated under the coalescent model with recombination (Hudson, 1983). This means that sites that have a recombination breakpoint between them do not evolve along the same phylogenetic tree. Barton and Etheridge (2004) have shown that selection has little effect on genealogies, which justifies neglecting selection when simulating genealogies under the coalescent model with recombination.

We performed two suites of simulation experiments, one using 10-taxon and one using 30-taxon data sets (Table 1). In each suite we simulated neutrally evolving data sets (i.e. $\omega = 1$, mimicking pseudogene evolution) to estimate false positive rates and data sets evolving with site-to-site rate

*to whom correspondence should be addressed

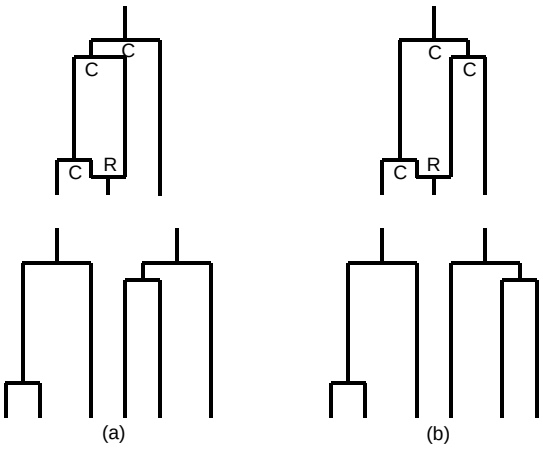

Fig. 1. Recombination graphs (Hudson, 1983) (above) and corresponding trees (below) illustrating (a) a recombination event that changes the tree length but not the topology and (b) a recombination event that changes both the tree length and the topology. In the recombination graphs, the letter C indicates coalescent events while the letter R indicates recombination events.

Table 1. Simulation parameters used to create data sets.

Data set	Nr of taxa	ρ^a	θ^b	Selection^c
Small, neutral	10	0.05	3.6	no
Small, positive selection	10	0.05	3.6	yes
Small, neutral (no recomb.)	10	0	3.6	no
Small, pos. sel. (no recomb.)	10	0	3.6	yes
Large, neutral	30	0.01	0.36	no
Large, positive selection	30	0.01	0.36	yes
Large, neutral (no recomb.)	30	0	0.36	no
Large, pos. sel. (no recomb.)	30	0	0.36	yes

[a] ρ: population-scaled recombination rate, $\rho = 4N_e r$. [b] θ: population-scaled mutation rate, $\theta = 4N_e \mu$. [c] Selection: The discrete model of Yang *et al.* (2000) was used; "no" indicates that $\omega = 1$ at all sites, while "yes" indicates $\omega_1 = 0.08$, $p_1 = 0.659$, $\omega_2 = 0.61$, $p_2 = 0.206$, $\omega_3 = 2.55$, $p_3 = 0.135$, where ω values are the non-synonymous/synonymous rate ratios and p values are the proportion of sites for which the corresponding ω values apply.

variation and positive selection (using the parameters inferred by Anisimova *et al.* (2003) on their hepatitis D antigen data set under the 3-component discrete model (Yang *et al.*, 2000)) to estimate power. Each simulated alignment was 500 codons long, and each data set consisted of 100 replicates. The transition/transversion rate ratio (κ) was set to 2 and the codon equilibrium frequencies to those empirically estimated for the Hepatitis D antigen data set. We chose mutation and recombination rate parameters that produced high false positive rates when using the standard method (see below) to infer positive selection on the neutral data sets. For the 30-taxon data sets the population-scaled recombination rate (ρ) was 0.01 and the population-scaled mutation rate (θ) was 0.36, resulting in an average of 43.2 recombination events in the entire genealogy and an expected number of 1.43 mutations per codon. For the 10-taxon data sets ρ was 0.05 and θ was 3.6, resulting in an average of 247.11 recombination events in the entire genealogy and an expected number of 10.18 mutations per codon (the very high values for the 10-taxon data sets serve to illustrate that the method works well even in extreme cases). To verify that the proposed method does not have an adverse effect when used on unrecombined data, we also simulated data sets with exactly the same parameters but with zero recombination rate.

Finally, we analysed the HIV-1 subtype C *env* and *gag* data of our recent study (Scheffler and Seoighe, submitted). These data sets contain 10 taxa each, with accession numbers *AY118165-AY118166, AF286227, AY158533-AY158535, AF411967, AF391234-AF391235* and *AF391238* for the *env* sequences (1053 codons in length) and *AY118165-AY118166, AF286227, AY158533-AY158535, AF411967, AY162223-AY162224* and *AF391254* for the *gag* sequences (590 codons in length).

3 ALGORITHM

In this study we report results for four methods of detecting positive selection, using different combinations of the two strategies investigated:

Standard: This is the baseline method, which assumes that topology, relative branch lengths and total tree length are constant over all sites.

Synonymous rate variation: This method assumes that topology and relative branch length are constant over all sites, but allows total tree length to vary from site to site.

Partitioning: This method uses recombination breakpoints (either detected or the actual simulated breakpoints) to divide the alignment into partitions, each of which is assumed to include no further recombination breakpoints. Topology, relative branch lengths and total tree length are forced to be constant over all sites within a partition, but allowed to vary between partitions.

Synonymous rate variation with partitioning: This method combines the previous two methods: topology and relative branch lengths are assumed constant over all sites within a partition, but allowed to vary between partitions. Total tree length is allowed to vary from site to site irrespective of partitioning.

We implemented the above methods using the batch language of the HyPhy package (Kosakovsky Pond *et al.*, 2005).

3.1 Baseline (standard) method

We detected positive selection by comparing the discrete "nearly neutral" and "selection" models M1a and M2a of Wong *et al.* (2004). We used the PAUP* program (Swofford, 2002) to estimate the maximum likelihood topologies under the HKY85 model (Hasegawa *et al.*, 1985). To save computation time, we did not estimate the branch lengths separately for each model, but instead used the branch lengths estimated under the M0 (single rate) model (Yang *et al.*, 2000). We report a sequence as being under positive selection at the 5% or 1% level if model M2a provides a significantly better fit than model M1a as measured by a likelihood ratio test with the appropriate significance level.

3.2 Allowing synonymous rate variation

In the methods that model synonymous rate variation we added a synonymous substitution rate parameter to the baseline method. We treat the synonymous rate s as belonging to one of a number of discrete rate classes, similar to the treatment of the non-synonymous/synonymous rate ratio ω, so that the expression for the instantaneous substitution rate from codon i to codon j at site

h becomes:

$$q_{ij}^{(h)} = \begin{cases} 0, & \text{for difference at more than one position,} \\ \pi_j s^{(h)}, & \text{for synonymous transversion,} \\ \kappa\pi_j s^{(h)}, & \text{for synonymous transition,} \\ \omega^{(h)}\pi_j s^{(h)}, & \text{for non-synonymous transversion,} \\ \omega^{(h)}\kappa\pi_j s^{(h)}, & \text{for non-synonymous transition.} \end{cases} \quad (1)$$

Here, κ is the transition/transversion rate ratio and π_j is the codon equilibrium frequency of codon j. $\omega^{(h)}$ and $s^{(h)}$ denote, respectively, the non-synonymous/synonymous rate ratio and synonymous rate at site h.

The synonymous rate is drawn from a discrete distribution with three rate categories (we obtained no noticeable difference in results when using four categories, data not shown), with rates scaled such that the average synonymous rate over all sites is 1. This distribution is identical to that used for the ω parameter in the discrete model M3 of Yang *et al*. (2000), except that the latter is unscaled. Thus each site, in addition to belonging to one of the ω categories, also belongs to a synonymous rate category. This can also be viewed as providing three different tree scales: the evolution at each site is modelled as following the same tree topology and relative branch lengths, but the tree may be scaled differently for different sites.

Note that our parameterisation of site-to-site rate variation is different from that used by Kosakovsky Pond and Muse (2005), which uses the synonymous rate only for synonymous changes and hence is not a direct measure of total tree length ($s^{(h)}$ is absent from the expression for the instantaneous rate of non-synonymous transitions and transversions). Whereas Kosakovsky Pond and Muse (2005) apply parametric models to the distribution of synonymous and of nonsynonymous rates, our parameterisation applies the same parametric models to the distribution of synonymous rates and of selective strengths.

3.3 Detecting recombination breakpoints

For the methods using partitioning by detected recombination we estimated the positions of recombination breakpoints using the non-parametric RDP (Martin and Rybicki, 2000), GENECONV (Padidam *et al.*, 1999), and MAXIMUM CHI SQUARED (Maynard Smith, 1992) methods as implemented in RDP2 (Martin *et al.*, 2005). See Poke *et al*. (2006) for a description of how these methods work. Default program settings were used throughout except that a Bonferroni corrected P-value cutoff of 0.01 was used to minimise the probability of falsely inferring recombination. All breakpoints detected by any of the three methods were taken into consideration.

3.4 Allowing different tree topologies for different sequence fragments

Once the recombination breakpoints have been detected, we use them to partition the alignment into separate segments (Figure 2). When the number of segments exceeds a preset maximum N (20 in this study), we use only the N longest unbroken segments and discard the remaining data. The rationale behind this is that when the segments between recombination breakpoints are very short, they contain very little phylogenetic information and therefore the tree topology and branch length parameters cannot be estimated accurately for the partitions. Moreover, such small partitions contribute very little information so that discarding them should be less costly than introducing additional uncertainty resulting from estimating additional branch length and topology parameters for the partition. In the present study, data was discarded only for the simulated data, which had very high rates of recombination. The number of breakpoints detected in the real data sets we examined was lower than the maximum in both cases.

Next, topologies and branch lengths are estimated as in the baseline method, except that a separate topology and set of branch lengths is used for each segment. The remaining model parameters, however, are shared across all segments. In particular, the parameters of models M1a and M2a describing the rate categories are estimated only once for all segments.

To allow fairer comparison with the unpartitioned methods, we present the results for the simulation experiments not only for the full unpartitioned sequence (Figure 2, top), but also for an unpartitioned analysis of the sites in the largest unrecombined segments only (Figure 2, middle). This latter result provides a more direct comparison with the partitioned analysis (Figure 2, bottom) which uses the same subset of the codons.

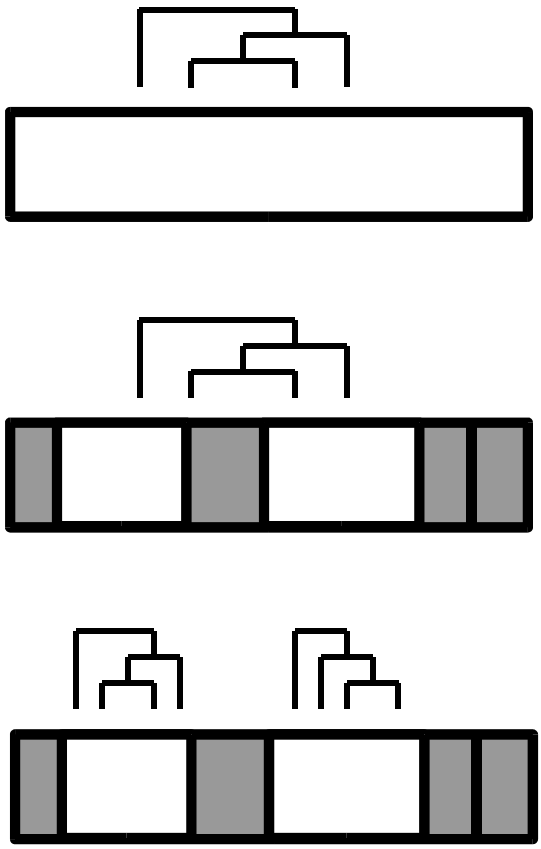

Fig. 2. Strategy for partitioning sequences according to recombination breakpoints. Full sequence (top): the entire sequence is used and described by a single tree topology and set of branch lengths, ignoring recombination breakpoints. Largest unrecombined segments, unpartitioned (middle): only codons in the largest N segments that contain no recombination breakpoints (vertical lines) are used (illustrated here by the white regions, with $N = 2$). As for the full sequence analysis, these codons are described by a single tree topology and set of branch lengths. Partitioning using largest unrecombined segments (bottom): each of the largest N segments that contain no recombination breakpoints is modelled using a separate topology and set of branch lengths.

4 RESULTS AND DISCUSSION

4.1 Simulation experiments

We investigated power and false positive rates using the simulated data sets (summarised in Table 1). For each data set we first considered the effect of allowing the synonymous rate to vary across sites and of separating the tree topology and branch length parameters between the segments defined by recombination breakpoints, given that the locations of the recombination breakpoints are known. This was done by retrieving the recombination breakpoints used in the simulations. We then present the power and false positive rates for the more realistic case in which the breakpoint locations are not known, but are instead inferred using a set of breakpoint detection algorithms (Martin *et al.*, 2005).

4.1.1 True breakpoints The neutral simulations provide a worst case (but nevertheless realistic) scenario with which to investigate false positive rates. We found (Table 2) that allowing the synonymous substitution rate to vary across sites brought about a large decrease in false positives relative to the standard method, but still left the false positive rate unacceptably high. Partitioning according to the true breakpoints (Table 2), on the other hand, brought false positive levels down to close to the desired rate. In this case, synonymous rate variation with partitioning did not give further improvement over partitioning alone. The decrease in false positives when partitioning has two causes. First, the fact that some data is discarded inevitably leads to a reduction in power: this can be seen by comparing the full sequence results with the largest unrecombined segments (LUS) results for the unpartitioned methods. Second, the partitioning itself causes a further reduction, which is the desired effect: the magnitude of this effect can be seen by comparing the results for the partitioning methods with the LUS results of the corresponding unpartitioned methods. Therefore, in order to see the effect of partitioning the phylogeny parameters between unrecombined segments or allowing the synonymous rate to vary on the false positive rates, the results obtained using these methods should be compared to those obtained by applying the standard method to the LUS.

The positive selection simulations provide a means to investigate power (Table 3). Again, some caution is required here because positive results could be artefacts of recombination rather than instances where the method detected the signal of positive selection. Nevertheless, when the false positive rate obtained on the corresponding set of neutral simulations is low, we can conclude that the result obtained on the positive selection simulations is a good indication of power.

For the case in which we assume that the true recombination breakpoints are known, the power was higher on the large data set than on the small data set. This was partly because the recombination levels were so high in the small data set that the average segment length (for the 20 largest unrecombined segments) was below 8 codons. In fact, given that tree topologies and branch lengths were inferred on such short segments, it is surprising that the method retains any power to discriminate between data sets with and without positive selection (as demonstrated by the higher rate of positives in the positive selection data sets than in the neutral data sets). This shows that, even when recombination creates what might appear to be a hopelessly fragmented evolutionary history, it can still be possible to perform reasonable inferences provided recombination is taken into account.

Inferring trees and branch lengths on very short segments for the partitioning method caused a large decrease in power for the small data sets, and possibly also a small increase in false positives. This is particularly noticeable for the partitioning method (without synonymous rate variation) applied to the small positive selection data set, on which we obtained only 6% power at the 5% significance level. To confirm that this severe drop in power was caused by misestimation of tree topologies and branch lengths on the short segments we repeated the analysis, but with the topology and branch lengths for each segment fixed to the true (simulated) values. This resulted in 99% power (at both 5% and 1% significance levels), which is, as expected, identical to the result obtained for the corresponding unrecombined simulations. When the true topology was fixed but the branch lengths estimated as usual, the power was $14\%(10\%)$ at the $5\%(1\%)$ significance level. Thus the decrease in power can be attributed to inaccurate estimation of the branch lengths, which appears to become particularly acute when the segment lengths are this short. We caution that extremely short segment lengths (e.g. resulting from extremely high recombination rates such as in this simulation) may cause the proposed method to lack power.

4.1.2 Detected breakpoints In real data, the true breakpoints are unknown and have to be detected by a recombination detection method. This has the disadvantage that there may be inaccuracy in the breakpoints detected, but may also have advantages in that recombination events that have little or no effect (for instance because they occur between closely related taxa and do not change the tree topology, as in Figure 1(a)) will remain undetected, and thus will not have any negative effect on the power of the method. This could explain the results in Tables 4 and 5 where we found that using the detected breakpoints resulted in better performance (both a lower rate of false positives and higher power) than using the true breakpoints. In particular, the average segment lengths for the small data sets were longer, due to the suppression of many presumably unimportant (and difficult to detect) recombination breakpoints. The longer segment lengths yielded improvements of the results obtained by methods using partitioning on these data sets.

Using the detected breakpoints, the power obtained using partitioning with synonymous rate variation on the small data set was even higher than that obtained on the large data set. This can be explained by the fact that the diversity in this data set was much higher so that, once the false signal caused by recombination has been compensated for, the data set contains more information that can be used to obtain inferences about selective pressure.

It is reassuring that modelling synonymous rate variation had very little effect on the recombination-free sequences: false positives were essentially unchanged while power decreased slightly. Partitioning had no effect: trivially, when the true breakpoints were used, there were no breakpoints to take into account so that the partitioning methods were identical to the corresponding unpartitioned methods. Recombination detection resulted in only a few falsely detected breakpoints (in three and eight of the 100 replicates for the small neutral and small positive selection data sets respectively, and in none of the large data sets), but the inference of positive selection after partitioning gave a different result from that obtained without partitioning only for one replicate in the small positive selection data

Table 2. Number of false positive inferences out of 100 replicates obtained at the 5% (1%) significance level by different methods on the simulated neutral data sets when using the true recombination breakpoints.

Data set	Standard method		Synonymous rate variation		Partitioning	Synonymous rate variation with partitioning	Avg # LUS codons[b]
	Full sequence	LUS[a]	Full sequence	LUS[a]			
Small, neutral	94 (93)	69 (57)	27 (11)	17 (9)	11 (4)	13 (6)	157.79
Small, neutral (no recombination)	12 (8)	12 (8)	11 (7)	11 (7)	12 (8)	11 (7)	500
Large, neutral	90 (81)	85 (72)	37 (28)	34 (26)	5 (2)	5 (2)	410.53
Large, neutral (no recombination)	9 (2)	9 (2)	8 (1)	8 (1)	9 (2)	8 (1)	500

[a] LUS: using sites from largest unrecombined segments only. [b] Average number of codons contained in the largest unrecombined segments.

Table 3. Power (number of true positive inferences out of 100 replicates) obtained at the 5% (1%) significance level by different methods on the simulated positive selection data sets when using the true recombination breakpoints.

Data set	Standard method		Synonymous rate variation		Partitioning	Synonymous rate variation with partitioning	Avg # LUS codons[b]
	Full sequence	LUS[a]	Full sequence	LUS[a]			
Small, positive selection	73 (68)	52 (41)	18 (9)	10 (4)	6 (5)	36 (26)	157.79
Small, positive selection (no recombination)	99 (99)	99 (99)	91 (89)	91 (89)	99 (99)	91 (89)	500
Large, positive selection	100 (100)	100 (100)	70 (46)	54 (31)	80 (68)	48 (32)	410.53
Large, positive selection (no recombination)	100 (100)	100 (100)	90 (74)	90 (74)	100 (100)	90 (74)	500

[a] LUS: using sites from largest unrecombined segments only. [b] Average number of codons contained in the largest unrecombined segments.

set, and only at the higher of the two significance levels listed. Hence the proposed methods do not have negative effects when applied to unrecombined data.

4.2 Analysis of viral data sets

Next, we used the four methods to analyse the HIV-1 subtype C data sets for which we have previously shown (Scheffler and Seoighe, submitted) that the recombination levels are high enough to cause false inference of positive selection. Indeed, the standard method inferred positive selection on both data sets at very high levels of significance.

For the *env* data (Table 6) we detected twelve recombination breakpoints. We found that both modelling synonymous rate variation and partitioning (using thirteen segments and discarding no data) caused reductions both in the significance level of the result and in the magnitude of positive selection inferred under the M2a model (as seen from the value of the ω_2 parameter), but that even when using both synonymous rate variation and partitioning we still detected positive selection at a highly significant level. We conclude that these sequences are likely to have evolved under both recombination and positive selection.

For the *gag* data (Table 7) we detected only four recombination breakpoints. This time, although partitioning (using five segments and discarding no data) without modelling synonymous rate variation did not remove the evidence of positive selection, the result was no longer significant when the synonymous rate was allowed to vary and even less so when synonymous rate variation and partitioning were combined. We conclude that, when recombination is taken into account, there is no convincing evidence that these sequences have evolved under positive selection.

5 CONCLUSIONS

Our simulation results reveal that modelling synonymous rate variation tends to make inference of positive selection more conservative:

Table 4. Number of false positive inferences out of 100 replicates obtained at the 5% (1%) significance level by different methods on the simulated neutral data sets when using the detected recombination breakpoints.

Data set	Standard method		Synonymous rate variation		Partitioning	Synonymous rate variation with partitioning	Avg # LUS codons[b]
	Full sequence	LUS[a]	Full sequence	LUS[a]			
Small, neutral	94 (93)	93 (92)	27 (11)	25 (17)	11 (8)	2 (2)	479.35
Small, neutral (no recombination)	12 (8)	12 (8)	11 (7)	11 (7)	12 (8)	11 (7)	500
Large, neutral	90 (81)	90 (81)	37 (28)	36 (28)	10 (5)	6 (3)	498.26
Large, neutral (no recombination)	9 (2)	9 (2)	8 (1)	8 (1)	9 (2)	8 (1)	500

[a] LUS: using sites from largest unrecombined segments only. [b] Average number of codons contained in the largest unrecombined segments.

Table 5. Power (number of true positive inferences out of 100 replicates) obtained at the 5% (1%) significance level by different methods on the simulated positive selection data sets when using the detected recombination breakpoints.

Data set	Standard method		Synonymous rate variation		Partitioning	Synonymous rate variation with partitioning	Avg # LUS codons[b]
	Full sequence	LUS[a]	Full sequence	LUS[a]			
Small, positive selection	73 (68)	75 (69)	18 (9)	17 (12)	91 (80)	83 (75)	463.36
Small, positive selection (no recombination)	99 (99)	99 (99)	91 (89)	91 (89)	99 (99)	91 (90)	500
Large, positive selection	100 (100)	100 (100)	70 (46)	70 (47)	97 (97)	67 (49)	498.90
Large, positive selection (no recombination)	100 (100)	100 (100)	90 (74)	90 (74)	100 (100)	90 (74)	500

[a] LUS: using sites from largest unrecombined segments only. [b] Average number of codons contained in the largest unrecombined segments.

Table 6. Results for *env* data

Method	p-value	$2\Delta \ln L$	Parameter values[a]	
Standard	0	360.8	$\omega_0 = 0.045$,	$p_0 = 0.60$,
			$\omega_1 = 1$,	$p_1 = 0.30$,
			$\omega_2 = 5.37$,	$p_2 = 0.10$.
Synonymous rate variation	1.24e-9	41.0	$\omega_0 = 0.064$,	$p_0 = 0.59$,
			$\omega_1 = 1$,	$p_1 = 0.27$,
			$\omega_2 = 4.15$,	$p_2 = 0.14$.
Partitioning	0	206.2	$\omega_0 = 0.061$,	$p_0 = 0.60$,
			$\omega_1 = 1$,	$p_1 = 0.29$,
			$\omega_2 = 3.95$,	$p_2 = 0.11$.
Synonymous rate variation with partitioning	1.66e-4	17.4	$\omega_0 = 0.071$,	$p_0 = 0.54$,
			$\omega_1 = 1$,	$p_1 = 0.34$,
			$\omega_2 = 2.81$,	$p_2 = 0.11$.

[a] Parameter values for ω distribution under M2a model

Table 7. Results for *gag* data

Method	p-value	$2\Delta \ln L$	Parameter values[a]	
Standard	1.3e-9	40.91	$\omega_0 = 0.047$,	$p_0 = 0.77$,
			$\omega_1 = 1$,	$p_1 = 0.19$,
			$\omega_2 = 4.03$,	$p_2 = 0.05$.
Synonymous rate variation	0.063	5.52	$\omega_0 = 0.055$,	$p_0 = 0.74$,
			$\omega_1 = 1$,	$p_1 = 0.21$,
			$\omega_2 = 2.85$,	$p_2 = 0.05$.
Partitioning	1.0e-7	32.21	$\omega_0 = 0.056$,	$p_0 = 0.76$,
			$\omega_1 = 1$,	$p_1 = 0.20$,
			$\omega_2 = 4.42$,	$p_2 = 0.04$.
Synonymous rate variation with partitioning	0.16	3.61	$\omega_0 = 0.072$,	$p_0 = 0.78$,
			$\omega_1 = 1$,	$p_1 = 0.00$,
			$\omega_2 = 1.44$,	$p_2 = 0.22$.

[a] Parameter values for ω distribution under M2a model

both false positives and power go down. However, the levels of false positives observed in these simulations were still unacceptably high despite being much lower than when constant synonymous rates were assumed.

Using tree topology and branch lengths inferred separately for segments defined by detected recombination breakpoints caused a dramatic reduction in the false positive rate. For example, in the 10-taxon data set we obtained an improvement from 94% false positives on the neutral simulations and 73% power on the positive selection simulations to 11% false positives on the neutral simulations and 91% power on the positive selection simulations. By combining partitioning with synonymous rate variation the false positive rate dropped further to an acceptable 2%, albeit at the cost of some reduction in power. The final power of 83% was nevertheless higher than the original power of 73%.

One of the most encouraging aspects of the simulation results was the performance of the partitioning methods using the detected recombination breakpoints. In the current set of simulations these methods performed better than the methods that used the simulated breakpoints, most likely because of the small segment lengths obtained when all of the recombination breakpoints were used. These results imply that the method we propose is not highly susceptible to inaccuracy in the detected breakpoints and that the majority of the benefit derived from partitioning appears to be obtained from the subset of most easily detectable recombination breakpoints.

We have not investigated the accuracy of site-specific selection detection using the proposed methods. In their simulation studies, Anisimova *et al.* (2003) and Shriner *et al.* (2003) found that site-specific analyses using standard phylogenetic methods are much more robust to recombination than whole-sequence analyses. This is consistent with our preliminary investigations (data not shown), in which we failed to find high levels of site-specific false positive inference using standard methods. More recently, Kosakovsky Pond *et al.* (2006) have found that under some conditions site-specific inference using a fixed effects likelihood method can also give highly misleading results in the presence of recombination. These authors found that the effects of recombination on site specific inference can be alleviated by analysing unrecombined segments separately and we therefore recommend that the method presented here should also be used for site-specific inference of positive selection when recombination is suspected.

Our results indicate that the proposed methods are able to filter out false inferences of positive selection on recombined sequences, but also have the power required to infer positive selection in such

sequences when the signal of positive selection does exist. Furthermore we show that there is no evidence of a disadvantage of applying partitioning to sequences when the sequences have not in fact undergone recombination. In such cases few, if any, recombination breakpoints were detected and inferring the tree topology and branch length parameters separately for the resulting large unrecombined segments appeared to have no effect on the power or false positive rates. We therefore recommend that a method such as the one we describe, that includes a screen for recombination and separation of phylogeny parameters between recombination breakpoints be applied routinely when phylogenetic methods are used to infer positive selection in sequences for which recombination is possible.

6 ACKNOWLEDGEMENTS

This study was supported by the South African National Bioinformatics Network and by the National Institute of Allergy and Infectious Disease and the National Institutes of Health through the Centre for the AIDS Programme of Research in South Africa (grant no. 1U19AI51794). We also thank Rasmus Nielsen for making his Codonrecsim program available to us, Fourie Joubert and David Posada for use of the Linux clusters at the University of Pretoria, South Africa and the University of Vigo, Spain, and Sergei Kosakovsky Pond for help with the HyPhy package and offering to incorporate the proposed methods into future distributions of HyPhy.

REFERENCES

Anisimova,M., Nielsen,R. and Yang,Z. (2003) Effect of Recombination on the Accuracy of the Likelihood Method for Detecting Positive Selection at Amino Acid Sites. *Genetics,* **164** (3), 1229–1236.

Barton,N.H. and Etheridge,A.M. (2004) The Effect of Selection on Genealogies. *Genetics,* **166** (2), 1115–1131.

Goldman,N. and Yang,Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol,* **11** (5), 725–736.

Hasegawa,M., Kishino,H. and Yano,T. (1985) Dating the human-ape split by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution,* **22**, 160–174.

Hudson,R. (1983) Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology,* **23**, 183–201.

Kosakovsky Pond,S.L., Frost,S.D.W. and Muse,S.V. (2005) HyPhy: hypothesis testing using phylogenies. *Bioinformatics,* **21** (5), 676–679.

Kosakovsky Pond,S.L. and Muse,S.V. (2005) Site-to-Site Variation of Synonymous Substitution Rates. *Mol Biol Evol,* **22** (12), 2375–2385.

Kosakovsky Pond,S.L., Posada,D., Gravenor,M.B., Woelk,C.H. and Frost,S.D. (2006) Automated Phylogenetic Detection of Recombination Using a Genetic Algorithm. *Mol Biol Evol,* , msl051.

Martin,D.P. and Rybicki,E. (2000) RDP: detection of recombination amongst aligned sequences. *Bioinformatics,* **16**, 562–563.

Martin,D.P., Williamson,C. and Posada,D. (2005) RDP2: recombination detection and analysis from sequence alignments. *Bioinformatics,* **21**, 260–262.

Maynard Smith,J. (1992) Analysing the mosaic structure of genes. *Journal of molecular evolution,* **34**, 126–129.

Muse,S. and Gaut,B. (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol,* **11** (5), 715–724.

Nielsen,R. and Yang,Z. (1998) Likelihood Models for Detecting Positively Selected Amino Acid Sites and Applications to the HIV-1 Envelope Gene. *Genetics,* **148** (3), 929–936.

Padidam,M., Sawyer,S. and Fauquet,C.M. (1999) Possible emergence of new geminiviruses by frequent recombination. *Virology,* **265**, 218–225.

Poke,F., Martin,D., Steane,D., Vaillancourt,R. and Reid,J. (2006) The impact of intragenic recombination on phylogenetic reconstruction at the sectional level in Eucalyptus when using a single copy nuclear gene (cinnamoyl CoA reductase). *Molecular Phylogenetics and Evolution,* **39**, 160–170.

Shriner,D., Nickle,D.C., Jensen,M.A. and Mullins,J.I. (2003) Potential impact of recombination on sitewise approaches for detecting positive natural selection. *Genet Res.,* **81** (2), 115–21.

Swofford,D. (2002) *PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods). Version 4.* Sinauer Associates, Sunderland, Massachusetts.

Wong,W.S.W., Yang,Z., Goldman,N. and Nielsen,R. (2004) Accuracy and Power of Statistical Methods for Detecting Adaptive Evolution in Protein Coding Sequences and for Identifying Positively Selected Sites. *Genetics,* **168** (2), 1041–1051.

Yang,Z., Nielsen,R., Goldman,N. and Pedersen,A.M.K. (2000) Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites. *Genetics,* **155** (1), 431–449.