Detecting past recombination events in *Potato virus Y* genomic sequences using statistical methods

D. Husmeier & F. Wright

Molecular evolution of RNA viruses: Plant RNA viruses evolve more rapidly than the DNA of their plant hosts due to a higher mutation rate. Mutational processes include nucleotide substitution that replaces one nucleotide by another, and insertions and/or deletions that result in sequence length changes. In addition to mutation, inter-strain recombination can produce new mosaic strains[9]. It is important to detect recombination events because, firstly, phylogenetic studies of inter-strain relationships that assume no recombination are very likely to be incorrect, and secondly, because recombination may have important consequences for virus control strategies. We will illustrate methods to detect evidence of past recombination events in four strains of *Potato virus Y* (PVY). As we shall see, detecting recombination in this data set is not simple.

PVY is a member of the Potyvirus genus and, like other potyviruses, has a single stranded positive sense RNA genome. Four complete length strains were available from the EMBL/GenBank sequence database (accession number/ lengths: U09509/ 9698bp, D00441/ 9704bp, M95491/ 9703bp, and X97895/ 9701bp). These were labelled Singh (abbreviated to Si), Robaglia (Ro), Hungarian (Hu) and Baulcombe (Ba), respectively. The analysis was carried out on a 9692bp alignment after removing all positions with gaps (totalling 22bp). Part of the alignment is shown in Figure 1.

Figure 1 Part of the PVY alignment with recombination breakpoint RB1 predicted at approximately position 2422. Manual inspection supports this: Hungarian and Baulcombe are very similar up to 2407, whereas Hungarian is clearly more similar to Singh and Robaglia (rather than Baulcombe) from 2467 onwards.

Phylogenetic trees and recombination: In the absence of recombination, the relationships among the four PVY strains would be best represented by a single phylogenetic tree, consisting of a branching order (*topology)* plus branch lengths, based on the entire alignment (see Fig. 2). However, inspection of the tree reveals that the branch length leading to Hu is very short (0.0004 substitutions per position), compared to the the branch leading to Ba (0.0713), suggesting a relative rate of nucleotide substitution of 173.8. This is a very high ratio even assuming the action of natural selection, and suggests that the phylogenetic tree model, assuming no recombination, is not appropriate.

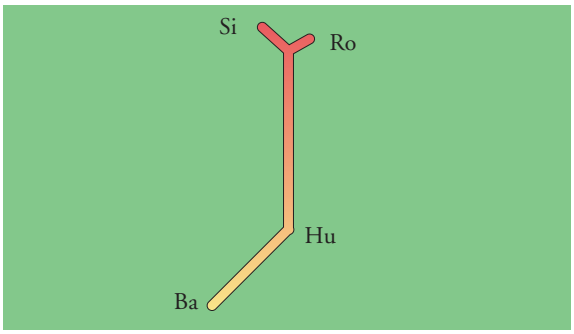

Figure 2 Tree for alignment, assuming no recombination.

If there is recombination, phylogenetic trees can be estimated for each region of the alignment after these have been located by recombination breakpoints. This article will focus on statistical methods to detect these breakpoints.

Evidence of among-site rate heterogeneity (as seen by the presence of conserved and variable regions) is almost always found in alignments, and complicates

the detection of recombination. Methods vary in how much they deal with rate heterogeneity: some do not distinguish rate heterogeneity from topology heterogeneity, some remove the effect of differences in *mean* rate among positions of the alignment, and some try to exclude the effect of rate heterogeneity completely and look only for changes in topology.

Detecting recombination breakpoints is difficult if (1) the recombination event occurred long ago, (because nucleotide substitution accumulation will make 'ancient' recombination events difficult or impossible to detect), if (2) recombination has occurred between similar strains, and if (3) recombination breakpoints lie close to each other.

Statistical methods for detecting recombination breakpoints: Different methods are appropriate for small (e.g. 4 sequences), medium (e.g. 10 sequences), or large alignments (e.g. 50 sequences). A method for each of these categories has been developed by BioSS. Once recombination breakpoints have been detected, reconstructing the history of recombination events among sequences is then done by interpreting the output of phylogenetic analyses of each recombinant region.

Hidden Markov model method: A hidden Markov model (HMM) approach can be applied to the problem of detecting recombination in small alignments[1,2,5]. The mean distance between recombination breakpoints is modelled by the probability of a recombination event as we move along the sequence alignment. For the four sequences, there are three possible tree (unrooted) topologies (Fig. 3):

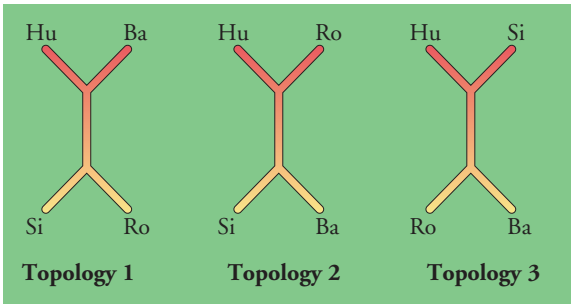

Figure 3 The three possible topologies for the four PVY strains.

Note that topologies depict branching order only, and do not show branch lengths to scale. The transitions between the three topologies are assigned probabilities. The hidden states of the HMM represent the different phylogenetic tree topologies: we observe the sequences but cannot directly see the "hidden" tree topologies. The parameters of the model, namely the branch lengths associated with each topology and the recombination probability, are optimised in a maximum likelihood sense by applying the expectation maximization (EM) algorithm. The HMM method focuses only on topology changes in the alignment, and attempts to reduce rate heterogeneity effects. Statistical significance is assessed by (posterior) probabilities assigned to each topology for each position in the alignment.

DSS sliding window method: The TOPAL[6,7] program utilises fast approximate distance-based phylogenetic methods and can be used with large alignments. The method slides a fixed-size window (e.g. 500bp wide) along the alignment, comparing the left-hand window (W_L) with the right-hand window (W_R). In W_L, the matrix of pairwise genetic distances among the sequences is calculated, and a phylogenetic tree is then estimated by minimising the sum-of-squares (SS_L) between the observed distances and the distances based on the tree. A distance matrix is then calculated for W_R, and the W_L topology is fitted to it, yielding a second sum-of-squares value (SS_R). When the W_R topology has changed due to recombination, the W_L topology will be a poor fit to the W_R distance matrix. Putative recombination breakpoints can be observed by plotting the difference between SS_L and SS_R (DSS statistic) against the window centre. The influence of *mean* rate heterogeneity is removed from the analysis, but the DSS statistic will still be inflated when branch lengths change non-uniformly among branches as we move along the alignment. Recent improvements allow the statistical significance of DSS peaks to be estimated with parametric bootstrapping[8].

MCMC sliding window method: Markov chain Monte Carlo (MCMC) approaches have revolutionised Bayesian modelling and analysis, and recently MCMC techniques have been applied to phylogenetic analysis[4]. We have developed a method for detecting recombinants based on a marginal posterior distribution analysis[3]. A fixed-size window (e.g. 500 bp) is moved along a sequence alignment. For every position, the posterior probability of tree topologies conditional on the subsequence alignment selected by the moving window is determined by a MCMC simulation. On moving into a recombinant region, this marginal posterior distribution of topologies can be expected to change. This can be quantified by probabilistic divergence measures, for example, a local measure (AS) comparing the distributions on two adjacent

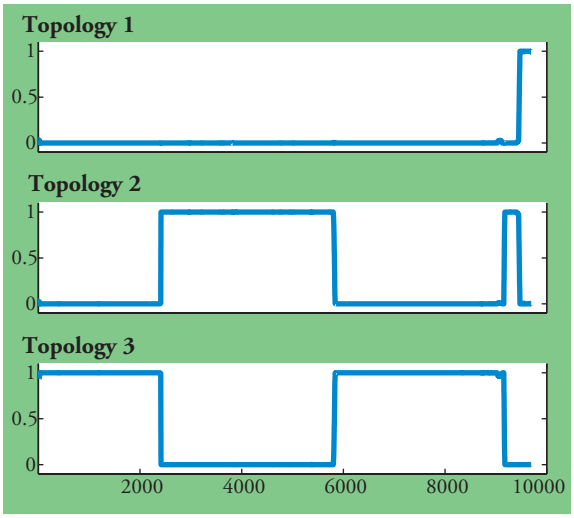

Figure 4 Output from HMM analysis, showing the probability of each topology (as described in Figure 3) at each position in the alignment.

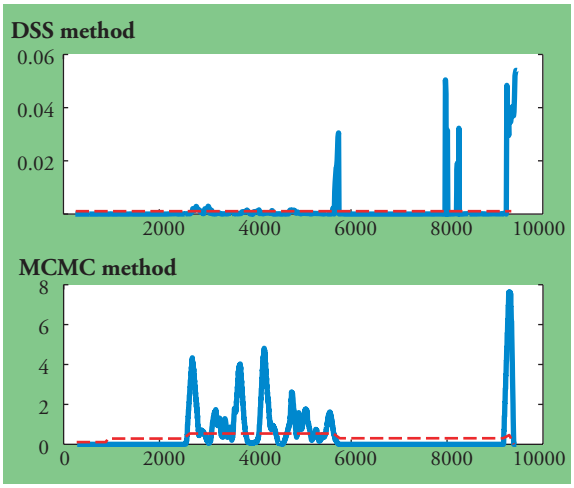

Figure 5 Output from DSS and MCMC analyses, showing the DSS and AS statistics respectively, plotted along the alignment. The dashed line is the statistical significance threshold for peaks.

windows. This divergence measure is then plotted along the alignment. The MCMC approach is currently limited to analyses of about 10 sequences. The MCMC method attempts to focus only on topology changes and to exclude all rate heterogeneity effects.

A limited simulation study[3] has shown that the MCMC method outperforms the DSS method. We expect, in general, that the HMM method should be best at predicting the position of recombination breakpoints, although the DSS method is perhaps better than both HMM and MCMC at detecting recombination events that do not lead to topology changes.

Results of recombination breakpoint analyses: The HMM graphical output is shown in Figure 4, giving the probability of each of the three topologies for each position in the alignment. The HMM method detects four putative recombination breakpoints at approximately 2422 bp (denoted RB1), 5837 bp (RB2), 9178 bp (RB3) and 9511 bp (RB4). Note the methods do not place confidence intervals around the predicted breakpoints. In comparing positions, we have chosen 200 bp as a significant difference for this data set.

Figure 5 shows the DSS and MCMC output. The MCMC method detects RB1 (but not RB2) plus more than four breakpoints between RB1 and RB2. In addition, it detects RB3 but not RB4. The DSS method detects RB2 (but not RB1), RB3, detects some evidence of RB4, and detects six weakly significant breakpoints between RB1 and RB2. In addition, it predicts two new breakpoints at approximately 7970 bp (RBx) and 8250 bp (RBy).

Looking at phylogenetic trees for the identified regions helps interpretation. For example, Figure 6 shows the trees (with branch lengths approximately to scale) for the nonrecombinant regions before and after RB1. We can see that Hu has changed position relative to the other three strains, resulting in a change in *topology*. The topology in the 'Start-RB1' region is significantly better than the other two possible topologies. In contrast, there is no significant best topology for the 'RB1-RB2' region: the three strains Si, Ro and Hu are similar and the main feature of the tree is the long branch connecting to Ba. Within the RB1-RB2 region, the DSS and MCMC methods detect fluctuations between the three topologies.

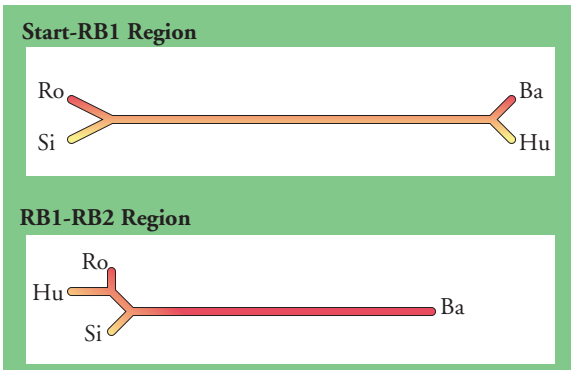

Figure 6 Phylogenetic trees for regions before and after recombination breakpoint RB1.

Inspection of the phylogenetic tree for the recombinant RBx-RBy region (predicted by DSS) revealed a long branch connecting Si to the Ro strain suggesting a recombination event that did not change the tree topology.

Interpreting mosaic structure: We will restrict the interpretation of mosaic structure to the six recombination breakpoints (RB1 to RB4, plus RBx and RBy), plus two breakpoints found within the RB1-RB2 region by DSS and MCMC. These partition the alignment into nine nonrecombinant regions. Looking at the distribution (not shown) of pairwise genetic distances among strains within each region, we appear to have two types of strains. Between members of each type, the pairwise genetic distances are small (0.01-0.08 substitutions per position). Between pairs not belonging to the same type, the genetic distances are large (0.17 to 0.22).

Si and Ba strains can be chosen as typical members of each of the two types because they are separated by a large distance in phylogenetic trees calculated for each region in the alignment. One interpretation of the mosaic structure is that Si and Ba are similar or identical to the 'parental' types and that Ro and Hu have been produced by homologous recombination. Si-type strains are coloured yellow, and Ba-type strains green, in Figure 7.

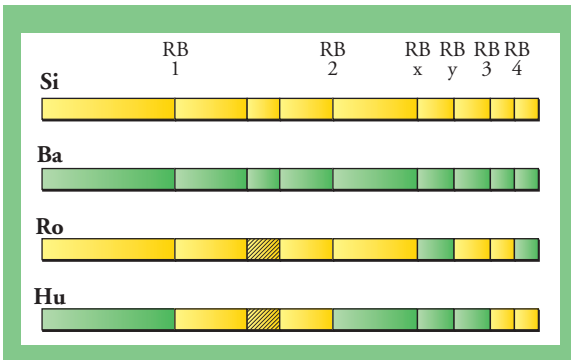

Figure 7 Predicted mosaic structure of PVY strains.

The shaded subregion within the RB1-RB2 region was detected by both DSS and MCMC methods and appears to be due to a recombination event among the Si-type strains (involving an exchange between Ro and Hu).

Conclusions:
The analysis of the four PVY strains exposes some of the weaknesses of current methods. The PVY data showed evidence of recombination events that did not cause topology changes, and contained a region (RB1-RB2) where the dominant topology was not significantly better than the two alternatives. In addition, some recombination events appeared to have occurred between similar strains, and some recombination breakpoints are positioned close together.

The results of the three methods, while in some agreement, do differ. The RBx-RBy region appears to be a genuine recombinant region that was missed by HMM and MCMC methods due to a lack of a change in topology. On the other hand, the subregion within RB1-RB2 does involve a topology change but is possibly a false positive.

The window-based methods (DSS and MCMC) had problems detecting recombination breakpoints when these were close enough for two to fit within the window, or when they were positioned close to the end of the alignment. Although not investigated here, reducing the window size below 500 bp may help. However, the choice of window size length is not trivial: too small and a number of non-significant peaks will be obtained; too large and recombination breakpoints may be missed. We intend to reanalyse a larger PVY dataset with other window sizes.

We are currently working to improve methodology, in particular to reduce the influence of rate heterogeneity on the detection of past recombination events. An automatic method of reconstructing the evolutionary

history of a sample of sequences based on the recombination breakpoint prediction would also be useful.

Acknowledgements:
Dirk Husmeier is funded by the BBSRC/EPSRC Bioinformatics initiative (grant BIO10494).

References:

[1] Husmeier, D. & Wright, F. (2001). In Bornberg-Bauer, E., Rost, U., Stoye, J. and Vingron, M. (eds.). *Proceedings of the 15th German Conference on Bioinformatics*, Berlin. Logos Verlag,19-26. ISBN 3-89722-498-4.

[2] Husmeier, D. & Wright, F. (2001). *Journal of Computational Biology* 8, 401-427.

[3] Husmeier, D. & Wright, F. (2001). *Bioinformatics* 17, S123-S131.

[4] Larget, B. & Simon, D.L. (1999). *Molecular Biology and Evolution* 16, 750-759.

[5] McGuire, G., Wright, F. & Prentice, M.J. (2000). *Journal of Computational Biology* 7, 159-170.

[6] McGuire, G., Wright, F. & Prentice M.J. (1997). *Molecular Biology and Evolution* 14, 1125-1131.

[7] McGuire, G. & Wright, F. (1998). *Bioinformatics* 14, 219-220.

[8] McGuire, G. & Wright, F. (2000). *Bioinformatics* 16, 130-134.

[9] Worobey, M. & Holmes, E.C. (1999). *Journal of General Virology* 80, 2535-2543.