

## Sister-Scanning: a Monte Carlo procedure for assessing signals in recombinant sequences

Mark J. Gibbs\*, John S. Armstrong and Adrian J. Gibbs

Research School of Biological Sciences, The Australian National University,  
GPO Box 475, Canberra ACT 2601, Australia

Received on January 13, 2000; revised on March 9, 2000; accepted on March 14, 2000

### Abstract

**Motivation:** To devise a method that, unlike available methods, directly measures variations in phylogenetic signals in gene sequences that result from recombination, tests the significance of the signal variations and distinguishes misleading signals.

**Results:** We have developed a method, that we call 'sister-scanning', for assessing phylogenetic and compositional signals in the various patterns of identity that occur between four nucleotide sequences. A Monte Carlo randomization is done for all columns (positions) within a window and Z-scores are obtained for four real sequences or three real sequences with an outlier that is also randomized. The usefulness of the approach is demonstrated using tobamovirus and luteovirus sequences. Contradictory phylogenetic signals were distinguished in both datasets, as were regions of sequence that contained no clear signal or potentially misleading signals related to compositional similarities. In the tobamovirus dataset, contradictory phylogenetic signals were separated by coding sequences up to a kilobase long that contained no clear signal. Our re-analysis of this dataset using sister-scanning also yielded the first evidence known to us of an interspecies recombination site within a viral RNA-dependent RNA polymerase gene together with evidence of an unusual pattern of conservation in the three codon positions.

**Availability:** A program package, SiScan, for use under MS-DOS can be downloaded from <http://life.anu.edu.au> with test data and instructions.

**Contact:** [mgibbs@rsbs.anu.edu.au](mailto:mgibbs@rsbs.anu.edu.au);  
[johna@rsbs.anu.edu.au](mailto:johna@rsbs.anu.edu.au); [gibbs@rsbs.anu.edu.au](mailto:gibbs@rsbs.anu.edu.au)

### Introduction

We face three main questions when considering evidence of recombination in a set of aligned nucleotide or amino-acid sequences: (Q1) Is there clear evidence of recombination? (Q2) Where were the recombination sites? (Q3) Which sequences evolved through recombination? A

series of papers describing methods, attest to the difficulty of answering these questions (Stephens, 1985; Sawyer, 1989; Hein, 1990; Fitch and Goodman, 1991; Maynard Smith, 1992; Robertson *et al.*, 1995; Salminen *et al.*, 1995; Grassly and Holmes, 1997; Weiller, 1998; Gao *et al.*, 1999). Most of these methods attempt to answer the first and second questions together and the last question is dealt with subsequently by comparing phylogenetic trees, but it is clear that all three questions are linked. Phylogenetic signals and trees must be assessed to test the evidence of recombination and to identify recombinant sequences, but regions in the alignment with coherent signals need to be identified first and to do this it is necessary to test regions for possible signals. This preliminary analysis is necessary because if a tree or signal is found using a sequence that contains regions that have different phylogenetic histories, then the tree will probably be wrong or the signal will be contaminated (e.g. Gibbs and Weiller, 1999)

When attempting to detect recombination, the problem is usually simplified by examining small sub-sets from the alignment of the sequences of interest. Evidence of recombination may be detected by comparing as few as two or three sequences (Miller *et al.*, 1988; Maynard Smith, 1992), but most methods compare sets of four sequences. Alignments of four sequences can contain informative sites (where two taxa have the same nucleotide and the other two have a different nucleotide), and, by using a known outlier as the fourth sequence, the root of a cluster of three sequences may be located. Hence, comparisons with a fourth sequence may help identify some misleading signals that arise when sequences are evolving at different rates (Felsenstein, 1978; Siddall, 1998). Robertson *et al.* (1995) used a method in which an alignment of four sequences was split between two windows and for each position of the boundary of the windows, the significance of the distribution of informative sites between the windows was tested. The method was largely successful, but, like several other early methods, it oversimplified the problem by assuming that one phylogenetic signal would be abruptly replaced by another at a point in the alignment

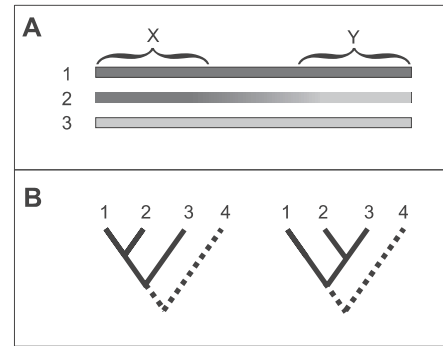
\*To whom correspondence should be addressed.

(i.e. at the recombination site). This is equivalent to assuming that the sequences consist of discrete, adjacent regions each containing a clear phylogenetic signal. In the case of the method used by Robertson *et al.* (1995), the adjacent windows represented these discrete regions. This assumption is probably invalid for many sequences as they contain regions where there is no clear phylogenetic signal or where there are misleading signals (see below). Furthermore, recombination sites may be overprinted by other mutations, so that the signals change gradually and it may not be possible to pinpoint recombination sites (e.g. Gibbs and Weiller, 1999). The methods of Gao *et al.* (1999) and Salminen *et al.* (1995) partly addressed some of these problems. In the first, evolutionary distances were calculated for pairs of sequences within the window and by sliding the window across the alignment and producing distance values at each position, localized variations in this measure were detected. Regions with no clear signal could be detected, but Gao *et al.* (1999) did not directly test the significance of the signal variations. In the method of Salminen *et al.* (1995), bootstrap samples drawn from the sequences in the window were used to infer the support for the three possible resolved trees and these bootstrap values were plotted. A weakness of this method is that bootstrap estimates of support are strongly influenced by the tree-building method and the substitution model used, and even when these parameters are optimized the level of support may be over- or underestimated (Hillis and Bull, 1993).

Here we present a method for directly measuring several kinds of signals within a window, together with a test of the signals that may be used to assess the evidence for recombination and to detect misleading signals arising from localized compositional similarities. To our knowledge, no other methods directly address this last problem. We have analysed simulated sequences, as well as two sets of viral sequences, to demonstrate the power of the method and investigate its properties.

## Rationale

Figure 1 depicts evidence of recombination based on three sequences. On one side of a recombination site, across region X, sequences 1 and 2 are sister taxa, but on the other side of the site, across region Y, sequences 2 and 3 are sister taxa. With reference to Figure 1, we rephrase Q1 (above) as follows: (Q4) Do two or more regions in an alignment contain opposing phylogenetic signals and are the signals significant in those regions regardless of any compositional similarities? We define the opposing signals as exclusive sister pairings of taxa (sequences), as shown in Figure 1, and answer the question by testing the significance of the signals found in different regions of the alignment. The tested regions do not have to be



**Fig. 1.** A diagram illustrating an example of recombination. a. The bars represent three hypothetical sequences including a recombinant, and the similarly shaded regions of the bars represent the most closely related regions in the sequences. b. Rooted trees describing the relationships of the hypothetical sequences on either side of a recombination site. Sequence 4 is an outlier used to root the trees.

adjacent. Signals are defined as patterns of nucleotide identity in the alignment that support the possible sister pairings. The strength of the evidence of recombination is indicated by the significance of the opposing signals, and, where possible, recombination sites are located by delineating the regions with opposing signals. The effects of compositional similarities are tested by replacing one of the real sequences in the alignment with a randomized sequence with the same composition as one or more of the real sequences. Including such a randomized outlier changes the frequency of some of the patterns of identity, and if a signal is diminished when this is done, so that it no longer appears to be significant, then it is likely that the signal is due to a compositional similarity.

## Algorithm

As with other methods, we define a region using a window that slides over an alignment of four sequences. An outlier sequence should be included to identify the sister pairings, and in our method this may be a real sequence or a locally randomized sequence (see below).

1. Count the number of positions within a window that conform to each of the patterns shown in Table 1. Pass the window over the alignment with a constant step length and make the same calculation at each window position.
2. For each window, sum the counts of positions with patterns where two sequences are identical, e.g. sequences 1 and 2 are identical in patterns 2, 8, 11 and 12 so the counts of positions with these patterns are summed (see Table 2, S 4–9);

**Table 1.** Definitions of the nucleotide identity patterns used to classify the positions in an alignment. If two or more taxa have the same nucleotide at a position they are linked by equals signs. If the nucleotide belonging to one taxon differs from those of the other taxa, it is preceded by a tilde sign. For informative sites (see text), the taxa with the same nucleotide are grouped and the two pairs are separated by a tilde sign

Pattern	Nucleotide identity between sequences (taxa) 1, 2, 3 and 4
P1	1 ~ 2 ~ 3 ~ 4
P2	1 = 2 ~ 3 ~ 4
P3	1 = 3 ~ 2 ~ 4
P4	1 = 4 ~ 2 ~ 3
P5	2 = 3 ~ 1 ~ 4
P6	2 = 4 ~ 1 ~ 3
P7	3 = 4 ~ 1 ~ 2
P8	1 = 2 ~ 3 = 4
P9	1 = 3 ~ 2 = 4
P10	1 = 4 ~ 2 = 3
P11	1 = 2 = 3 ~ 4
P12	1 = 2 = 4 ~ 3
P13	1 = 3 = 4 ~ 2
P14	2 = 3 = 4 ~ 1
P15	1 = 2 = 3 = 4

- For each window, sum the counts of each kind of informative site (where there are two pairs of identical nucleotides) and the patterns where two sequences are identical and which differ from an informative site by only one nucleotide substitution (quasi-informative sites), e.g. pattern 8 represents an informative site and patterns 2 and 7 could be obtained from pattern 8 by one substitution, so the counts for patterns 2, 7 and 8 are summed (see Table 2, S 1–3).
- For each window, generate four randomized sequences by assigning a new nucleotide at each position chosen at random, without replacement, from the nucleotides that occur at the position in the real sequences (Monte Carlo sampling to produce vertical randomization). Count the various patterns for the randomized sequences as described in steps 1 and 2, and calculate the sums of patterns. Repeat this step 100 times to create a population of scores from randomized sequences.
- For each window, calculate *Z*-scores for each of the patterns and sums of patterns.
- Plot the *Z*-scores for each position of the window.

We developed two variants of the method. In the first, four real sequences are analysed. In the second, three real sequences are analysed with a fourth sequence generated by randomization. The randomized sequence is generated

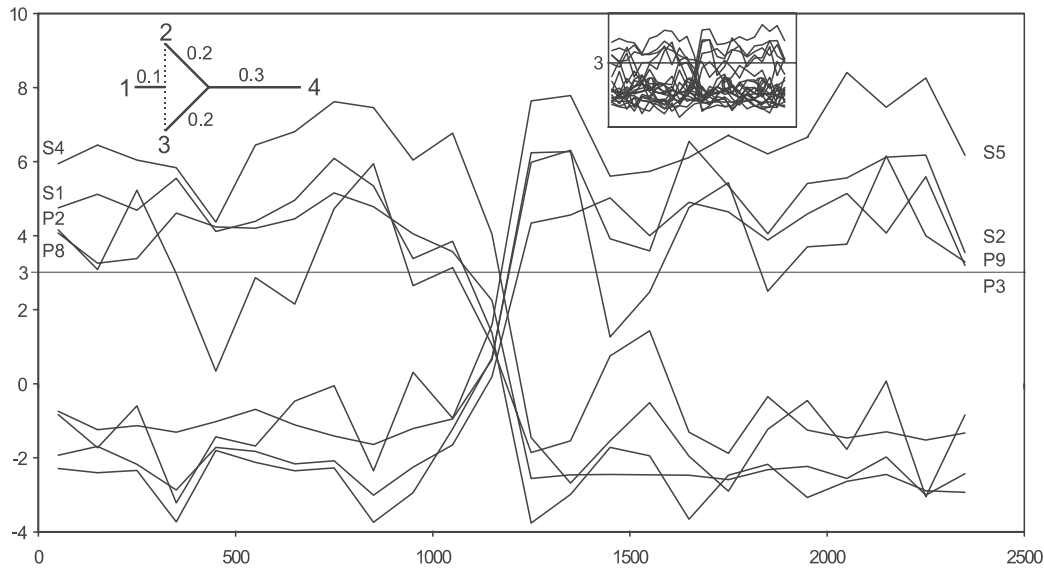
**Table 2.** Definitions of sums of patterns. Note, positions with pattern 11 are often excluded from sums 4, 5 and 7 (see text). Total pair-wise identity scores equate to sums S4 to S9

Sum of patterns	Patterns from Table 1 added to the sum
S1	$\sum$ 2, 7, 8
S2	$\sum$ 3, 6, 9
S3	$\sum$ 4, 5, 10
S4	$\sum$ 2, 8, 11, 12 (1 = 2)
S5	$\sum$ 3, 9, 11, 13 (1 = 3)
S6	$\sum$ 4, 10, 12, 13 (1 = 4)
S7	$\sum$ 5, 10, 11, 14 (2 = 3)
S8	$\sum$ 6, 9, 12, 14 (2 = 4)
S9	$\sum$ 7, 8, 13, 14 (3 = 4)

for each position of the window by randomizing the positions of the nucleotides within the window in just one of the real sequences (horizontal randomization). Thus, the composition of the randomized sequence depends on the composition of a segment of real sequence defined by the window (local composition). We have also included an option in SiScan version 1.01 that allows the fourth sequence to be generated by horizontally randomizing the positions of nucleotides in two of the real sequences within the window and then selecting a nucleotide for each position at random from the two randomized sequences.

### Test datasets

- Sets of simulated sequences were generated using Seq-Gen 1.1 (Rambaut and Grassly, 1997). Each set consisted of four sequences 1000–5000 nucleotides long that had ‘evolved’ using the Jukes and Cantor (1969) model of substitution to match a tree with branch lengths which were specified as the probability of a substitution per site. Recombinant sequences were made by splicing simulated sequences in a word processing program.
- An alignment of nucleotide sequences was assembled to match the amino-acid alignment [used by Gibbs and Cooper (1995)] of the virion protein and read-through protein sequences of cucurbit aphid-borne yellows polerovirus (CABYV), beet western yellows polerovirus (BWYV) and pea enation mosaic enamovirus (PEMV1).
- The complete genomic sequences of 20 tobamoviruses were aligned. The sequences were separated into five distinct regions: (i) 5′ terminal untranslated regions (UTRs); (ii) 3′ UTRs; (iii) replicase genes, that include the methyltransferase, helicase and polymerase domain-encoding



**Fig. 2.** Changes in Z-score (y-axis) for selected patterns and sums of patterns for four simulated sequences plotted against position in the alignment (x-axis). The four sequences included a recombinant (taxon 1) and the two parental sequences from which it was derived (taxa 2 and 3). Z-scores for the complete set of patterns and sums of patterns are shown in the inset, as is the tree used when generating the sequences. Branch lengths equate to the probability of a substitution per site. A window 100 nucleotides long was used with a step length of 100 positions. Z-scores are plotted at the mid point of each window.

sequences; (iv) movement protein genes and (v) virion protein genes. Untranslated regions were directly aligned using CLUSTALV. Amino-acid sequences translated from the gene sequences were aligned also using CLUSTALV (Higgins *et al.*, 1992) and gaps were added to the nucleotide sequences so that they matched the amino-acid alignments using the program ADDGAPS (kindly supplied by Dr Georg Weiller). Finally, the gapped nucleotide sequences of the UTRs and genes were rejoined to produce the fully aligned genomic sequences. The tobamovirus and luteovirus sequence datasets described above are available with the SiScan 1.01 package.

### Implementation

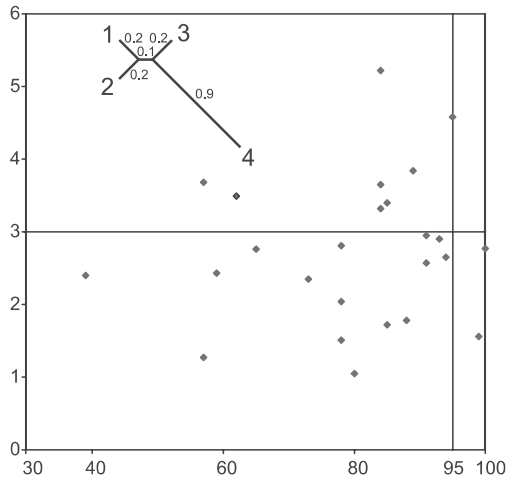
SiScan 1.01 operates under MS-DOS. It requires data-files containing aligned nucleotide sequences in NBRF (NBRF/PIR) format and produces tables of raw counts of patterns, summed counts of patterns, Z-scores for these counts and sums, compositional data for the sequences and counts of excluded positions. These tables are also produced as comma-separated variables in Microsoft Excel format.

### Results and discussion

#### *Simulated sequences*

Z-scores were calculated using SiScan for a set of simulated sequences that consisted of a recombinant, the two parental sequences from which it was generated and an outlier (Figure 2). Z-scores  $< -3$  were obtained for some patterns because they occurred less frequently than would be expected given the composition of the sequences. Z-scores  $> 3$  were obtained for patterns that supported either one of the two opposing phylogenies. These significant scores were obtained from more than 90% of the windows when they were 50 nucleotides long, from more than 50% of the windows when they were 20 nucleotides long and from 2.6% of the windows when they were 5 nucleotides long. No Z-scores  $> 3$  were obtained for patterns that supported relationships other than those expected on either side of the recombination point, except when windows 5–10 nucleotides long were used. When such short windows were used, scores supporting the true phylogenies were found in almost twice as many windows as scores supporting alternative relationships. The artificial recombination point was located to within 10 nucleotides, using windows 10 nucleotides long.

SiScan results were compared with those obtained by bootstrapping using non-recombinant simulated sequences. Bootstrap values were calculated from maximum likelihood trees from 100 bootstrap samples using



**Fig. 3.** Z-scores (y-axis) for pattern 2 plotted against bootstrap values (x-axis) for a set of four simulated sequences. The tree used when the sequences were generated is shown (inset). A window 200 nucleotides long was used with a step length of 200 positions.

PAUP version 4.0b2 (written by David Swofford) and the Jukes and Cantor model of substitution. A dataset generated using a tree with terminal branches of 0.5 and an internal branch of 0.4 was analysed using a window 200 nucleotides long. Significant Z-scores from patterns supporting the true tree were obtained for 15 out of 25 windows, whereas bootstrap values of  $>95$  supporting the true tree were found for 17 out of 25 windows. This result was not unexpected, as the bootstrap values were obtained from trees inferred using the substitution model that was used to generate the sequences (Hillis *et al.*, 1994; Nei *et al.*, 1995). However, Z-scores  $>3$  were obtained for some windows that yielded bootstrap values of less than 95, suggesting that the randomization test used in SiScan is sometimes more sensitive than the optimal bootstrapping procedure. We confirmed this characteristic using a second simulated dataset (Figure 3). Z-scores  $>3$  were obtained from patterns that supported the true phylogeny from 8 out of 25 windows, but only 3 out of 25 windows yielded bootstrap values  $>95$ .

As expected, pattern 11 (Table 1) predominates in datasets that include three relatively closely related sequences and a distant or randomized outlier. Pattern 11 is included in three of the sums of counts (Table 2: S4, S5 and S7) and so when a distant or randomized outlier is used, these sums of counts may be dominated by pattern 11 counts. For this reason, we included an option in the program for excluding sites with pattern 11 from an analysis. To test for other possible biases of the method, we analysed sets of simulated sequences that contained no phylogenetic signal. These were generated using an unresolved tree with four branches each one unit long.

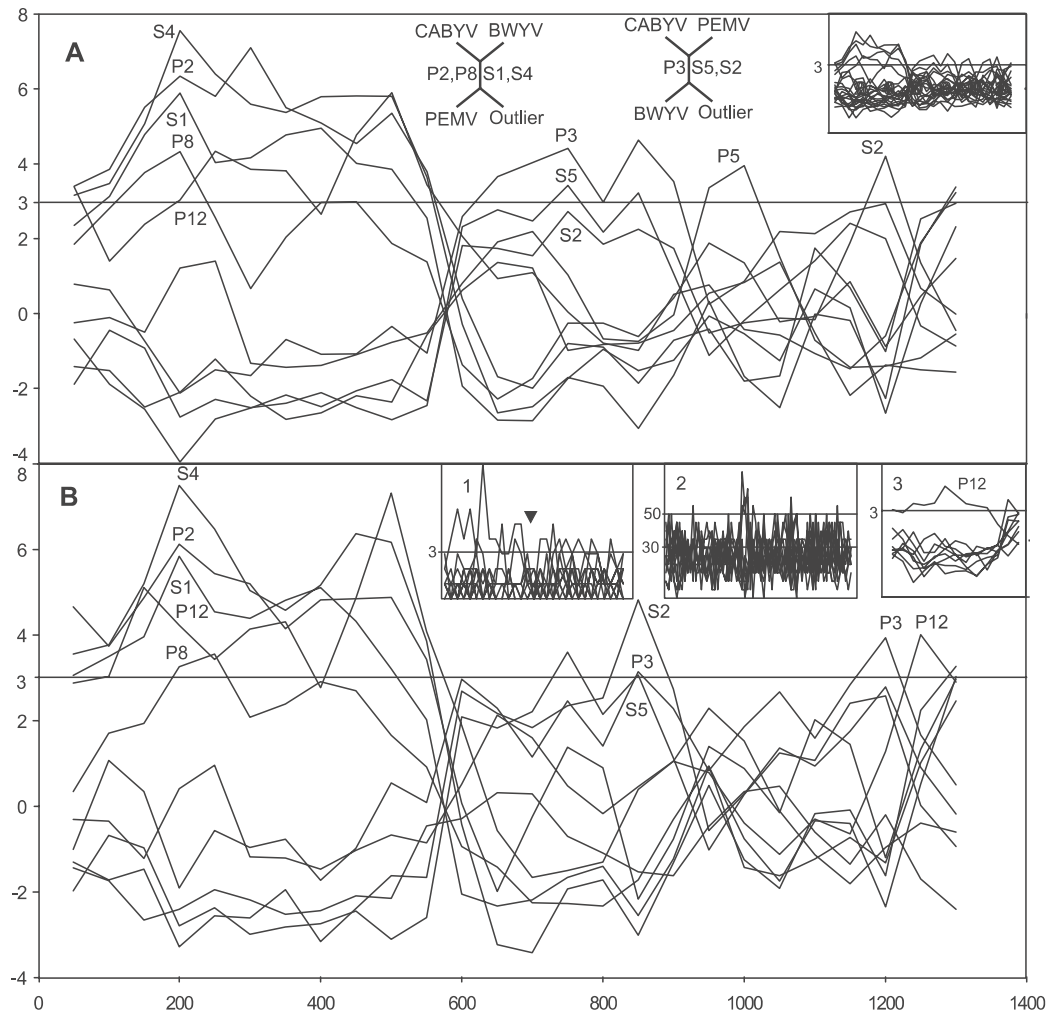
Z-scores were obtained for each pattern and sum of patterns from these sequences from 100 non-overlapping windows 100 nucleotides long. The average scores for these patterns and sums ranged from 0.13 to  $-0.18$  and their standard deviations ranged from 1.14 to 0.92. Only seven Z-scores  $> 3$  and two  $< -3$  were obtained.

Using a Pentium 2, 333 MHz processor, SiScan 1.01 took approximately 6 s to analyse 50 windows when the window length was set at 100 nucleotide positions and a randomized outlier was generated from one of the sequences. Under the same conditions, it took approximately 8 s to analyse 500 windows when the window length was set at 10 nucleotide positions.

#### *Luteovirus sequences*

Gibbs and Cooper (1995) found evidence of recombination in the history of three luteoviruses by analysing the phylogenies of their virion and read-through protein amino-acid sequences. They suggested that CABYV evolved from an ancestral recombinant virus that was produced through two recombinational events and that BWYV and PEMV probably belonged to the parental virus lineages.

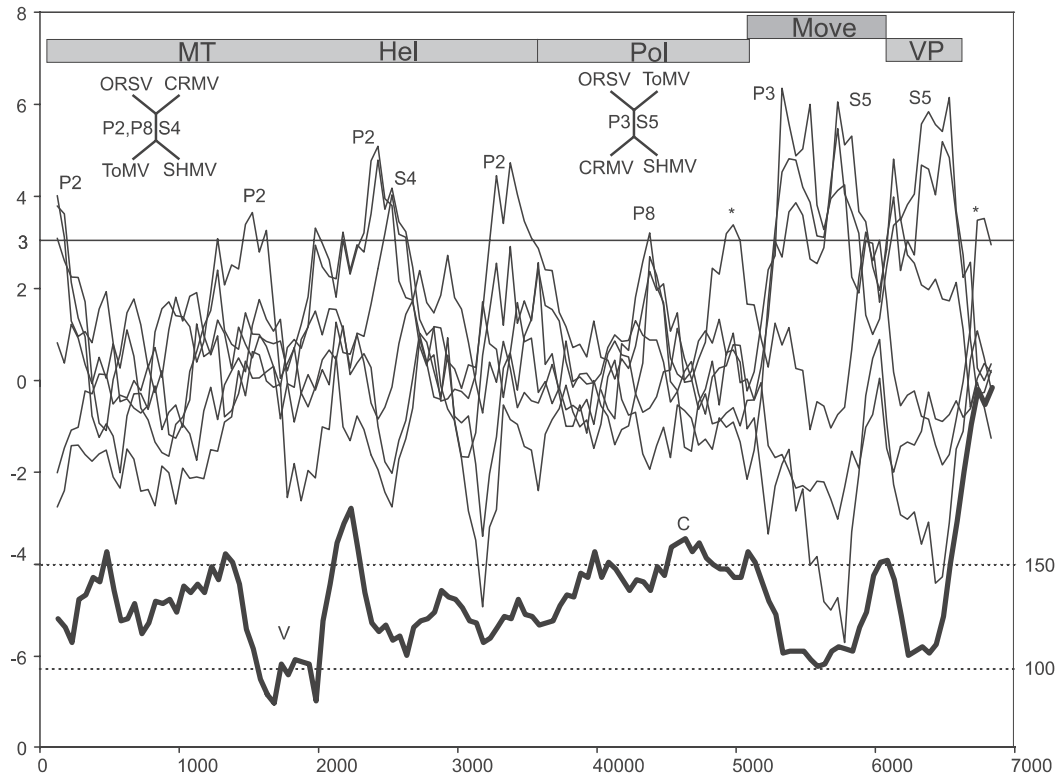
A plot of Z-scores calculated with SiScan (Figure 4A) confirmed some of the evidence of recombination. The first recombination site found using SiScan, between positions 550 and 600, was also found by Gibbs and Cooper (1995). SiScan plots showed clear support for grouping CABYV with BWYV on one side of the site but for grouping CABYV with PEMV1 on the other side of this site. The plot shows that the affinities of the sequences changed again at about position 925 (Figure 4A) with support for grouping PEMV1 with BWYV on the 3' side of the site. However, the signal that supports this grouping (at position 1000 in Figure 4A) was found to be sequence independent. A randomized outlier sequence was generated from the CABYV sequence for the comparisons plotted in Figure 4A. When the randomized outlier was instead generated from the BWYV sequence, no Z-scores  $>3$  were found for patterns supporting the grouping of BWYV with PEMV1 (Figure 4B). Thus, we concluded that the signal shown in Figure 4A was due to a local compositional similarity. Signals at the 3' end of the alignment represented by patterns 3 and 12, and sum 2, also appeared to be due to local compositional similarities. Gibbs and Cooper (1995) found that CABYV and BWYV sequences were more closely related on the 3' side of the second possible recombination site. The difference between their results and ours can be explained because we removed sites including gaps from the alignment and in doing this, we deleted the positions across this region that carried the phylogenetic signal. The location of the first recombination point in the luteovirus sequences, between positions 550 and 600 (Figures 4A and B) was not defined more



**Fig. 4.** Plots of Z-scores (y-axis) for patterns and sums of patterns for a set of luteovirus sequences and randomized outlier sequences. Luteovirus sequences were used in the order: 1, CABYV; 2, BWYV; 3, PEMV. A window 100 nucleotides long was used with a step length of 50 positions for both plots. Positions that included a gap or that conformed to pattern 1 or 15 were excluded from the analysis. The randomized outlier was derived from the CABYV sequence for the calculations plotted in panel A and was derived from the BWYV sequence for the calculations plotted in panel B. Plots for each pattern or sum of patterns is labeled according to Tables 1 and 2. Z-scores for the complete set of patterns and sums, except patterns 1 and 15, are shown in the inset in A. The two alternative trees and the patterns and sums of patterns that support them are also shown. B inset 1: Z scores for all patterns between positions 300 and 900 in the alignment found using a window of 20 was used with a step length of 20. A region between positions 550 and 610 (see text) is marked with a black triangle. B inset 2: The percentage nucleotide composition of the three luteovirus sequences also measured using SiScan. The strongest central peaks occur at about position 600 and represent the percentage cytosine in each sequence (see text). B inset 3: Z-scores for pattern 12 and the patterns and scores that measure the relative similarity between the PEMV sequence and the other real sequences (lower curves) for positions 1–700 of the alignment, i.e. patterns 3, 5, 9, 10 and 11 and sums 2, 3, 5 and 7. The randomized outlier was derived from the PEMV sequence for the calculations plotted in this inset.

accurately using shorter window lengths, but instead this region was found to contain no phylogenetic signal (Figure 4B inset 1). The region is cytosine rich in the three sequences (Figure 4B inset 2), and this strongly conserved cytosine stretch appears to obscure the exact point of recombination.

Pattern 12 occurred more frequently than expected between positions 200 and 600 in the plots (Figure 4A and B). This could be interpreted as evidence of a compositional similarity between the CABYV and BWYV sequences. However, that was not the case, as we discovered when the randomized outlier was generated from the



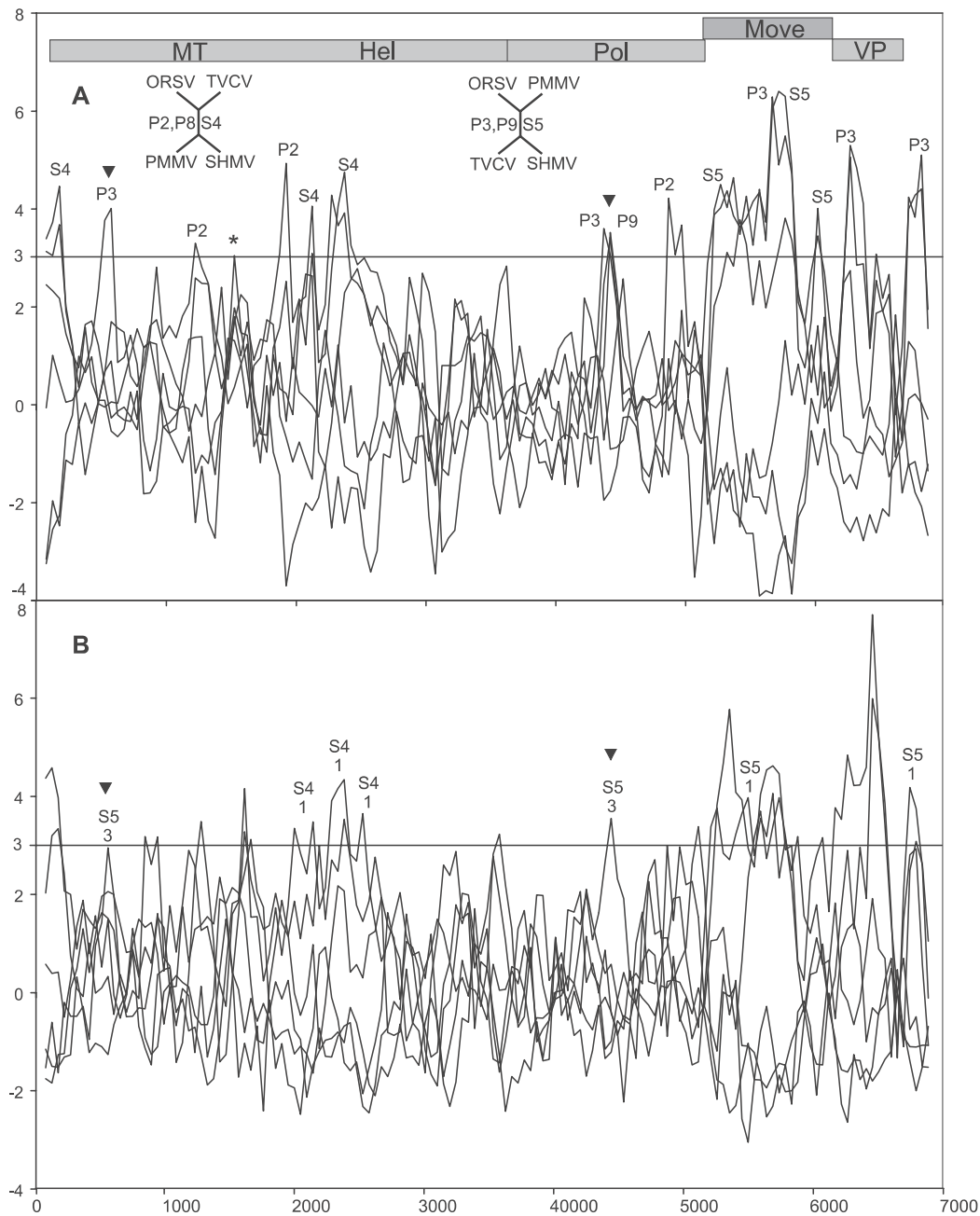
**Fig. 5.** Plots of selected Z-scores (y-axis) for an alignment of the complete genomic sequences of a set of four tobamovirus sequences. The tobamovirus sequences used were those of: 1, ORSV; 2, Chinese rape mosaic tobamovirus (CRMV; subgroup 3); 3, tomato mosaic tobamovirus (ToMV; subgroup 1); 4, sunn-hemp mosaic tobamovirus (an outlier). A window 250 nucleotides long was used with a step length of 50 and positions including gaps were excluded from the analysis. A map of the ORSV genome is shown in which blocks represent genes. The methyltransferase encoding, helicase encoding and polymerase encoding domains of the replicase gene are labeled MT, Hel and Pol. The movement protein and virion protein genes are labeled Move and VP. The two alternative trees and the patterns and sums of patterns that support them are also shown. The lower curve shown in bold is the counts of positions per window in which the ORSV, CRMV and ToMV sequences match, i.e. patterns 11 and 15. The scale for these counts is shown on the right-hand side of the figure. Peaks in the various plots are labeled according to the list of patterns shown in Tables 1 and 2. The two peaks marked with asterisks were not found in plots made using a randomized outlier and hence, were assumed to be due to compositional similarities.

PEMV1 sequence. Pattern 12 was also significantly supported when the PEMV1-derived outlier was used, even though there was no support for grouping the real PEMV1 sequence with the CABYV and BWYV sequences (Figure 4B inset 3). Thus, the effect was independent of sequence and composition. We then examined the actual counts and found that, given the frequency of identities between the real sequences, we would expect on average 6.7 positions in 100 to have pattern 12 over the region after alignment with a completely random sequence. If all four sequences were random we would expect less than 1 position in 100 to have the pattern. In the real alignment, we found that on average 9.0 positions in 100 had pattern 12 across the region. We concluded that when two of the three real sequences are very similar and these sequences are aligned with a randomized outlier, the randomized

outlier is likely to match the closely similar sequences at a significant number of sites.

#### *Tobamovirus sequences*

Lartey *et al.* (1996) found evidence of interspecies recombination in the history of tobamoviruses by inferring phylogenies from five sets of aligned amino-acid sequences from the viruses, i.e. those of the methyltransferase, helicase and polymerase domains within the replicase protein and those of the movement and coat proteins. The replicase sequences of odontoglossum ringspot tobamovirus (ORSV) were grouped with those of the tobamoviruses that infect crucifers (subgroup 3), whereas the movement and coat protein sequences of ORSV were grouped with those of the tobamoviruses that infect *Solanaceae* (subgroup 1). Bootstrap values for some of the trees supported



**Fig. 6.** Plots of selected Z-scores (y-axis) for an alignment of the complete genomic sequences of a set of four tobamovirus sequences. The tobamovirus sequences used were those of: 1, ORSV; 2, pepper mild mottle tobamovirus (PMMV; subgroup 1); 3, turnip vein-clearing tobamovirus (TVCV; a subgroup 3); 4, sunn-hemp mosaic tobamovirus (an outlier). A window 150 nucleotides long was used with a step length of 50 and positions including gaps were excluded from the analysis. (A) The ORSV genome map is shown, as are the two alternative trees and the patterns and sums of patterns that support them. Peaks in the various plots are labeled according to the list of patterns shown in Tables 1 and 2. The two peaks marked with black triangles are from patterns that support the right-hand tree. The peak marked with an asterisk was not found in plots made using a randomized outlier and hence, it was assumed to be due to some compositional similarity. (B) Plots of Z-scores for sums 4 and 5 for the same alignment but using only the first, second or third codon positions. Peaks that represent signals from codon positions 1 and 3 are labeled. The two peaks where the signal is predominantly in codon position 3 nucleotides are marked with black triangles.

the opposing phylogenies and Lartey *et al.* (1996) concluded that ORSV probably had a recombinant ancestor with a single recombination site close to the 3' end of the replicase gene.

Plots of Z-scores calculated with SiScan concurred with the phylogenetic analysis (Figure 5), with a recombination site located between positions 4375 and 5275, which is close to the terminus of the replicase gene. SiScan plots showed, however, that the tobamovirus sequences contained a complex pattern of signals. Regions that contained significant signals were interspersed with ones that contained no clear signal, and some coding stretches with no clear signal were several hundred nucleotides long. A plot of the number of positions where the subgroups 1 and 3 and ORSV sequences were identical (Figure 5 lower curve in bold) showed that some of the low signal regions were relatively strongly conserved (e.g. the region marked 'c') whereas others were not (e.g. the region marked 'v'). Several regions with sequence-independent signals (compositional similarities) were also found. The main opposing signals were found regardless of which combination of sequences from subgroups 1 and 3 was used, but other significant signals were also found when some combinations were tested. Perhaps the most interesting of these signals was found using a subset that included the sequences of pepper mild mottle tobamovirus (PMMV; subgroup 1) and turnip vein-clearing tobamovirus (TVCV; subgroup 3). A Z score plot (Figure 6a) showed that the ORSV sequence grouped with the subgroup 1 virus (PMMV) sequence rather than the subgroup 3 virus (TVCV) sequence at two regions in the replicase gene (the peaks at positions 550 and 4400). Matching significant signals were found in plots made using a randomized outlier and a database search made using the program BLASTN (Altschul *et al.*, 1997) confirmed the affinities detected around position 4400.

The signals around positions 550 and 4400 are distinct from the main opposing signals suggesting that they were generated by distinct recombinational events and, on either side of these regions, the ORSV sequence is grouped with the TVCV sequence (Figure 6a) suggesting that both aberrant regions were generated by double cross-over recombinational events. Thus, it appears that our analysis with SiScan provides evidence for as many as five recombinational events. The fact that the additional evidence of recombination was not found with all combinations of subgroups 1 and 3 sequences suggests that some of the events occurred after the subgroups diversified.

Further analysis with SiScan suggested, however, that the signals grouping the ORSV and PMMV sequences in the replicase gene might be not be due to a common ancestry. Our analysis involved the use of an option in SiScan 1.01 that permits the patterns from all first, second

or third positions in the sequence to be counted independently so that the signals present in the nucleotides at different codon positions may be assessed. Figure 6b shows that the signal around position 4400 is largely carried by nucleotides in the third codon position, and the same tendency occurs around position 550. We confirmed this result by examining the alignment and by bootstrapping. One possibility is that the nucleotide sequences at this position have been selected for some property in addition to that of coding for amino acids. Nucleotide sequences that fold into particular structures or that successfully interact with various proteins may have been favoured, and the signal for this second function may be carried in the third codon position because of its amino-acid coding redundancy. Thus, these signals may be related to function rather than phylogeny and they may result from convergence. An analysis of RNA secondary structures may help determine which of the two possible processes, recombination or convergence, is responsible for the aberrant signals, or whether wet bench experiments may be needed. In any case, it appears that SiScan may be used to detect a second class of potentially misleading signals.

#### Broader applications

Figure 6b shows that the main signal grouping, the ORSV and TVCV sequences, is largely carried by nucleotides in the first codon position. This may be significant because a tree built using the second codon position, which is usually the most conserved, rather than the first codon position, might contain more errors or weak branches. For similar reasons, it may be important to recognize those regions in the tobamovirus sequences that contain no clear signal or that contain misleading signals related to compositional similarities. Other sequences, including those from cellular organisms, must also contain regions that carry such biases or that lack signal. However, the kind of analysis we have partly automated using SiScan, where the signals in different regions of an alignment are assessed and tested independently, is rarely done when sequences are used to infer phylogenies. Thus, we propose that SiScan may have broader applications.

#### References

- Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
- Fitch,D.H. A. and Goodman,M. (1991) Phylogenetic scanning: a computer-assisted algorithm for mapping gene conversions and other recombinational events. *Bioinformatics*, **7**, 207–215.
- Felsenstein,J. (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.*, **27**, 401–410.
- Gao,F., Bailes,E., Robertson,D.L., Chen,Y., Rodenburg,C.M., Michael,S.F., Cummins,L.B., Arthur,L.O., Peeters,M., Shaw,G.M., Sharp,P.M. and Hahn,B.H. (1999) Origin of

- HIV-1 in the chimpanzee *Pan troglodytes troglodytes*. *Nature*, **397**, 436–441.
- Gibbs, M.J. and Cooper, J.I. (1995) A recombinational event in the history of luteoviruses probably induced by base-pairing between the genomes of two distinct viruses. *Virology*, **206**, 1129–1132.
- Gibbs, M.J. and Weiller, G.F. (1999) Evidence that a plant virus switched hosts to infect a vertebrate and then recombined with a vertebrate-infecting virus. *Proc. Natl. Acad. Sci. USA*, **96**, 8022–8027.
- Grassly, N.C. and Holmes, E.C. (1997) A likelihood method for the detection of selection and recombination using nucleotide sequences. *Mol. Biol. Evol.*, **14**, 239–247.
- Hein, J. (1990) Reconstructing evolution of sequences subject to recombination using parsimony. *Math. Biosci.*, **98**, 185–200.
- Higgins, D.G., Bleasby, A.J. and Fuchs, R. (1992) CLUSTAL V: improved software for multiple sequence alignment. *CABIOS*, **8**, 189–191.
- Hillis, D.M. and Bull, J.J. (1993) An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst. Biol.*, **42**, 182–192.
- Hillis, D.M., Huelsenbeck, J.P. and Swofford, D.L. (1994) Hobgoblin of phylogenetics? *Nature*, **369**, 363–364.
- Jukes, T.H. and Cantor, C.R. (1969) Evolution of protein molecules. In Munro, H. (ed.), *Mammalian Protein Metabolism* Academic Press, pp. 21–132.
- Lartey, R.T., Voss, T.C. and Melcher, U. (1996) Tobamovirus evolution: gene overlaps, recombination and taxonomic implications. *Mol. Biol. Evol.*, **13**, 1327–1338.
- Maynard Smith, J.M. (1992) Analysing the mosaic structure of genes. *J. Mol. Evol.*, **34**, 126–129.
- Miller, W.A., Waterhouse, P.M. and Gerlach, W.L. (1988) Sequence and organisation of barley yellow dwarf virus genomic RNA. *Nucleic Acids Res.*, **16**, 6097–6111.
- Nei, M., Takezaki, N. and Sitnikova, T. (1995) Assessing molecular phylogenies. *Science*, **267**, 253–255.
- Rambaut, A. and Grassly, N.C. (1997) Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *CABIOS*, **13**, 235–238.
- Robertson, D.L., Hahn, B.H. and Sharp, P.M. (1995) Recombination in AIDS viruses. *J. Mol. Evol.*, **40**, 249–259.
- Salminen, M.O., Carr, J.K., Burke, D.S. and McCutchan, F.E. (1995) Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. *AIDS Res. and Human Retroviruses*, **11**, 1423–1425.
- Sawyer, S.A. (1989) Statistical tests for detecting gene conversion. *Mol. Biol. Evol.*, **6**, 526–538.
- Siddall, M.E. (1998) Success of parsimony in the four taxon case: long-branch repulsion by likelihood in the Farris Zone. *Cladistics*, **14**, 209–220.
- Stephens, J.C. (1985) Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. *Mol. Biol. Evol.*, **2**, 539–556.
- Weiller, G.F. (1998) Phylogenetic profiles: a graphical method for detecting recombinations in homologous sequences. *Mol. Biol. Evol.*, **15**, 326–335.