Method

Open Access

A novel exploratory method for visual recombination detection

Korbinian Strimmer*, Kristoffer Forslund†, Barbara Holland‡ and Vincent Moulton†

Addresses: *Department of Statistics, University of Munich, Ludwigstrasse 33, D-80539 Munich, Germany. †The Linnaeus Centre for Bioinformatics, Uppsala University, Box 598, 7551 24 Uppsala, Sweden. ‡Allan Wilson Centre for Molecular Ecology and Evolution, Private Bag 11222, Palmerston North, New Zealand.

Correspondence: Korbinian Strimmer. E-mail: strimmer@stat.uni-muenchen.de

Published: 25 April 2003

Genome ***Biology*** 2003, **4**:R33

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/5/R33

Received: 16 December 2002
Revised: 10 March 2003
Accepted: 31 March 2003

Abstract

A versatile visual approach for detecting recombination and identifying recombination breakpoints within a sequence alignment is presented. The method is based on two novel diagrams - the highway plot and the occupancy plot - that graphically portray phylogenetic inhomogeneity along an alignment, and can be viewed as a synthesis of two widely used but unrelated methods: bootscanning and quartet-mapping. To illustrate the method, simulated data and HIV-1 and influenza A datasets are investigated.

Background

Recombination is an important evolutionary process and is one of the key factors shaping the structure of genes and genomes. Moreover, it plays a major role in contributing and maintaining genetic diversity in populations. Testing for recombination is important since sequence analysis under the assumption of a single underlying phylogeny can be severely biased by the presence of recombination [1-4]. Consequently, modeling and detection of recombination has received great attention [3,5-8].

The fact that recombination breaks down the correlation between the evolutionary history of different regions in a genome provides the rationale used by most approaches to identify recombination in molecular sequences. Generally, these tend to fall into two different classes. One class is based on the explicit reconstruction of gene trees for different parts of a sequence alignment and subsequent comparison of tree topology and branch lengths. Any differences are used as indicators for underlying recombination events. The most widely used method of this kind is bootscanning [9]. The other class only aims to determine the presence or absence of recombination, without trying to infer recombination breakpoints. This is achieved by searching for patterns in the sequence data that contradict the null-hypothesis of a single evolutionary history and does not require reconstruction of several gene trees along the sequence [10,11]. All these methods for inferring recombination involve a high degree of computation since full genealogies have to be inferred and/or distributions for test statistics have to be generated using, for example, the parametric [11] or nonparametric bootstrap [9] or Markov chain Monte Carlo simulations [12].

Here we introduce a novel and computationally inexpensive visual approach for detecting recombination and inferring breakpoints. This method emphasizes data exploration and can be regarded as a synthesis of bootscanning [9] and the quartet-mapping approach for analyzing the tree-likeness of sequence data [13,14]. In particular, it employs quartet-trees to rapidly scan for phylogenetic inhomogeneity along a sequence alignment. The information gathered during the scan is then condensed into two diagrams, which we have dubbed the highway plot and the occupancy plot.

These diagrams indicate whether and where recombination has taken place. Some examples of these plots are illustrated in Figures 1-5. In the highway plot (see for example, Figure 4, bottom left, for a highway plot constructed for a viral sequence alignment), the horizontal axis represents the sites in the alignment, whereas the vertical axis indicates changes of topology and branch length of the inferred quartet-trees along the alignment. The fact that a quartet-tree inferred for four sequences can have three possible topologies is indicated by the three 'lanes' which lie between the three horizontal lines in the plot. Each curve or 'trajectory' in the plot basically represents changes in the inferred quartet-tree along the alignment for a particular quartet of sequences. To reduce noise we employ a filtering or prescreening procedure to the trajectories, which we call 'trajectory filtering', that selects only those quartets which produce a relatively high signal.

The occupancy plot is a complementary diagram to the highway plot (see for example, Figure 4, bottom right, for the plot complementary to the highway plot in Figure 4, bottom left). This displays an 'occupancy' statistic which summarizes the positions of the trajectories within the lanes along the alignment. In particular, this indicates where substantial changes in quartet-tree topology occur along the alignment. The full mathematical details underlying construction of the highway and occupancy plots are presented in the Materials and methods section.

This approach to visual recombination detection is capable of detecting both recent and ancestral recombination events and allows the rapid exploration of an alignment for recombination. In the next section we demonstrate this by applying our new approach to a number of simulated datasets and two viral sequence alignments. Subsequently, we discuss the merits and limitations of this method and in the final section present the underlying mathematical principles and algorithms.

Results and discussion

We used several simulated and biological datasets to test the utility of visual recombination detection using highway and occupancy plots. We present a series of representative simulated datasets as well as two biological datasets as illustration of our findings. In all cases plots were similar to those we present here. Our program VisRD (see Materials and methods) was used with standard settings, that is, window size 200 base-pairs (bp), step size 10 bp, maximal change in $\theta = 0.2\ \pi$, unless stated otherwise. All data discussed here are included in the VisRD distribution (file names *.vse are given in parentheses).

Simulated data

To simulate data we follow the procedure described in [15]. Simulated alignments of nine sequences are generated by seq-gen [16] version 1.2.5, which basically evolves a sequence along two or more distinct trees with the same number of leaves and concatenates the resulting alignments. In the results presented here alignments of total length 1,000 bp were generated using the HKY model of sequence evolution and a transition/transversion ratio (κ) of 4.

Results for several representative simulated datasets are depicted in Figures 1-3. In Figure 1, data were generated along a single underlying tree to simulate absence of recombination (testno.vse). In the data for Figures 2 and 3, a recombination breakpoint was introduced by altering the underlying tree topology at position 500 (see Figure 6 of [15] for more details). In each case we employed trajectory filtering (as described in Materials and methods), using the top 5 quartets, top 20 quartets, and all possible 126 quartets (corresponding to the three rows in Figures 1-3).

As expected, the topology of the underlying tree was an important factor. In Figure 2a a nearly balanced (symmetric) underlying tree (bbc8.vse) was used, whereas for Figure 3 we used a completely unbalanced (skewed) underlying tree (bud.vse). In the case where the underlying tree is unbalanced (so that the possibility of many quartet trees with small internal branch lengths is increased) it is necessary to remove a greater number of low-ranking quartets to reduce noise.

In the absence of recombination (Figure 1) both the highway and occupancy plots clearly indicate a homogeneous phylogeny along the alignment, whereas in the other plots a breakpoint is clearly indicated. Note that the estimated positions of breakpoints shift somewhat from their actual position when the alignment scanned has been generated using an unbalanced tree (Figure 3). This shift was also observed when a bootscan analysis was made on the same alignments, and thus it is probably a general phenomenon rather than an artifact of our method.

As can be seen in all of the plots, the number of quartets used in the scan significantly affects results. Unlike occupancy plots, highway plots can display the evidence of recombination even if a large number of quartets are selected (see row 3 in Figures 2 and 3). However, they tend to be less exact than occupancy plots. If there is a recombination breakpoint, then using a small number of quartets or even a single quartet will typically result in a very clear occupancy plot from which the breakpoint position can be detected easily (for example, see the first rows in Figures 2 and 3). Larger numbers of quartets lead to a gradual weakening of this signal. However, when a small number of quartets is chosen, the plots may give rise to false positives as phylogenetically insignificant point movements may be selected by the trajectory-filtering algorithm, which can dominate the plot.

We also investigated inference of ancestral recombination and multiple recombination events (data not shown). In ancestral

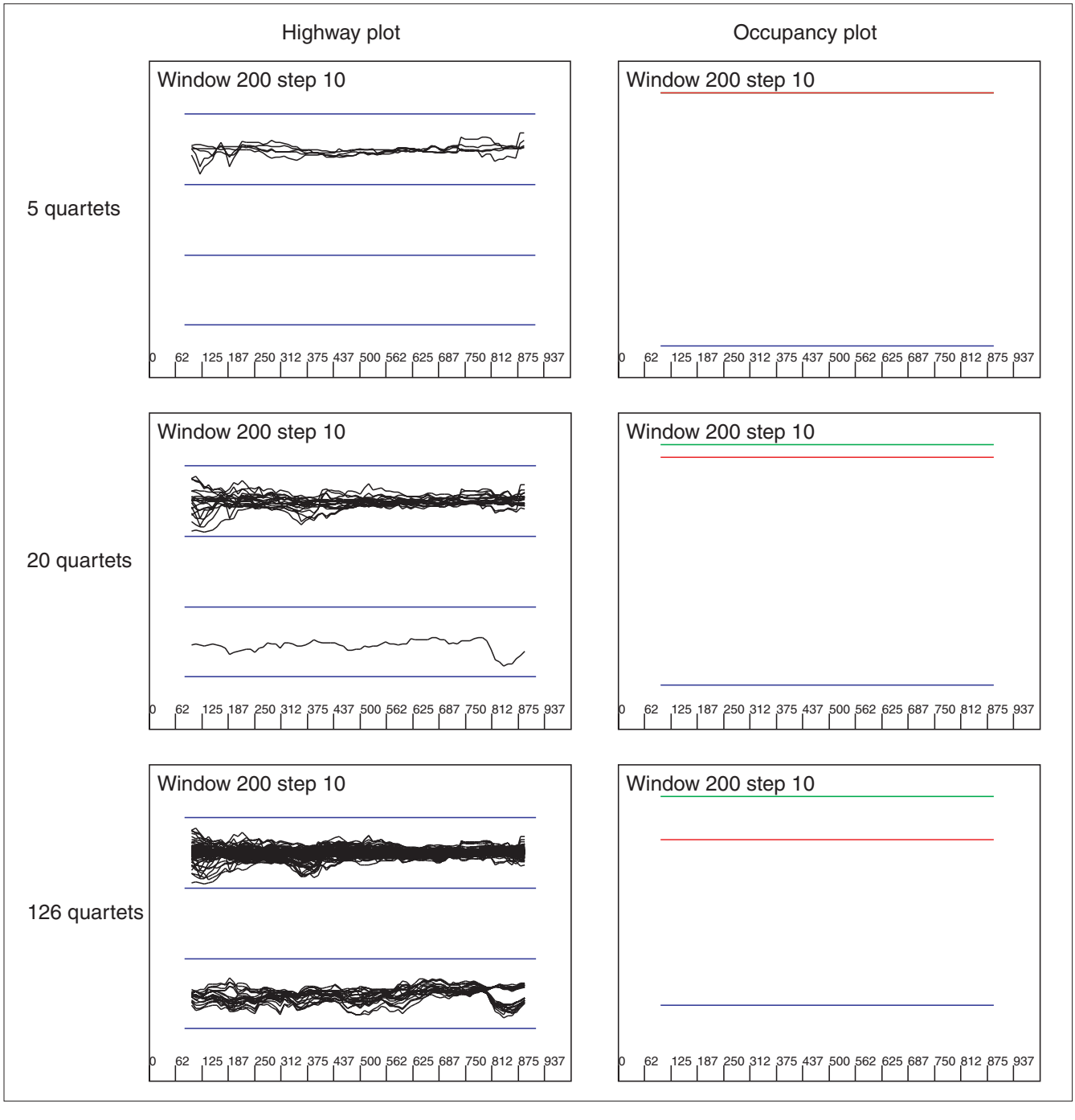

Figure 1
Highway and occupancy plots for simulated data with no recombination.

recombination only an ancestor is recombinant (that is, an internal node in the tree/network) and none of the investigated sequences is. We found that highway and occupancy plots consistently identified breakpoints for such data. The plots for multiple breakpoints were very similar to those shown in Figures 2 and 3, except that they indicated more breakpoints. Note that - as with all sliding window approaches - the window size has to be adjusted according to the number of expected breakpoints along the sequence (that is, the more recombination breakpoints the shorter the window).

Choosing suitable parameters

From the simulation studies we can also derive some guidelines for choosing the optimal setting for the number of

quartets and the windows used by the VisRD program. In our experience the default values (window size 200 bp and 20 quartets) are practical for many datasets but often also need to be adjusted.

Generally, it is desirable to use fewer rather than more quartets as this will increase the signal accuracy and remove noise connected with underlying unbalanced trees (see simulations above). However, to avoid the risk of relying on the 'wrong' quartets it is best not to use too few quartets. Similarly, in the interest of phylogenetic accuracy the window size should also be chosen to be as large as possible. To be able to locate the breakpoints more precisely and to detect multiple breakpoints the window size should not be too small.

Hence, both for the number of quartets and the window size, a trade-off between two contradictory aims must be found. In any analysis it is therefore recommended to investigate several plots with different settings and to choose those consistently giving rise to the clearest signal.

Molecular data

Recombinant HIV-1 sequence 93BR029

The human immunodeficiency virus 1 (HIV-1) 93BR029 sequence is known to be a mosaic of HIV-1 subtypes B and F [17]. We analyzed the alignment shipped with the program SimPlot v.3.2b [18], which contains nine sequences of length 1,690 bp from subtypes A-H and 93BR029 using the *gag* region of the HIV-1 genome (file gagtest.vse). Figure 4 shows several highway and occupancy plots for this alignment. We used a window size of 300 bp and took the best 5, 10 and 20 quartets to obtain the plots.

From both the highway and the occupancy plots it is evident that the sequences in this dataset underwent recombination, resulting in two observed recombination breakpoints around positions 580 and 890. Note that the trajectories of the highway plot change lanes simultaneously in all three lanes at the breakpoints.

As well as detecting breakpoint positions, the highway plots may be useful for identifying putative recombinant sequences. In particular, after trajectory filtering, a relatively small set of quartets typically remains. By looking at the frequencies with which the various taxa appear in this set of quartets - extending a similar approach for identifying recombinant sequences in a given alignment that was presented in [15] - it is possible to identify the taxa that most obfuscate a uniform phylogenetic signal across the alignment. For instance, when 20 quartets were chosen as the basis of the highway plot, the recombinant sequence 93BR029 appeared in all of the 20 quartets. The next most frequent sequence (B) was represented only in 14 of the 20 selected quartets. Similarly, 93BR029 was also the only sequence that was always present when 10 or 5 quartets were selected.

Influenza 1918 pandemic

Gene sequences from the influenza virus that caused the 1918 pandemic were analyzed in [19], where it was concluded that these sequences had been subject to recombination, and it was posited that this may have contributed to the virulence of the virus. Subsequently, these results were disputed in [20] with the argument that there is no evidence for a recombination event and that the putative recombination signal in the data was an artifact of differing evolutionary rates.

In Figure 5 we present highway and occupancy plots for the influenza alignment with 26 sequences of length 1,695 bp (influenzaA.vse). We fixed the number of quartets at 20 and explored the dataset using different window sizes between 400 bp and 800 bp. Two main things can be seen from Figure 5. First, the data contain a large amount of phylogenetic noise, and hence a large window size has to be used to increase signal quality. Second, in accordance with [20], there is no observed concerted change in trajectories in the highway plot and thus no overall evidence for a recombination breakpoint in the data.

However, there are some insular trajectories that sporadically cross lanes in the highway plot (for example, at position 850 bp) and the occupancy plot is also very noisy. Therefore we further assessed the possibility of a recombination event in the 1918 influenza sequence (South Carolina 1918) by examining the taxon frequencies in the selected 20 quartets. The South Carolina 1918 sequence was present in only 4 of the 20 quartets, whereas other sequences (for example, Kiev 1979 and Mongolia 1988) were found in 9 quartets. Therefore recombination in the South Carolina 1918 sequence can be excluded with high degree of certainty.

Comparative bootscanning, PLATO, and RecPars analysis

For validation and comparison purposes we also analyzed the molecular datasets using PLATO [21], RecPars [22] and bootscanning [9]. These three methods were chosen from the set of methods studied in [6] because all of them are - like the highway and occupancy plots introduced here - explicitly based on phylogenetic trees. Bootscanning is a visual exploratory method, and PLATO and RecPars are analytical approaches.

To analyze the data with PLATO v.2.11 we first reconstructed maximum-likelihood gene trees using TREE-PUZZLE 5.0 [23], where the best binary tree found in the set of intermediate puzzling-step trees was chosen as the candidate tree for PLATO. In the subsequent analysis, PLATO inferred six recombinant regions for the HIV-1 dataset and one recombinant region for the 1918 influenza dataset.

For the RecPars analysis we submitted the datasets to the online analysis tool at [24]. However, only the HIV-1 dataset was analyzed; the influenza dataset could not be investigated

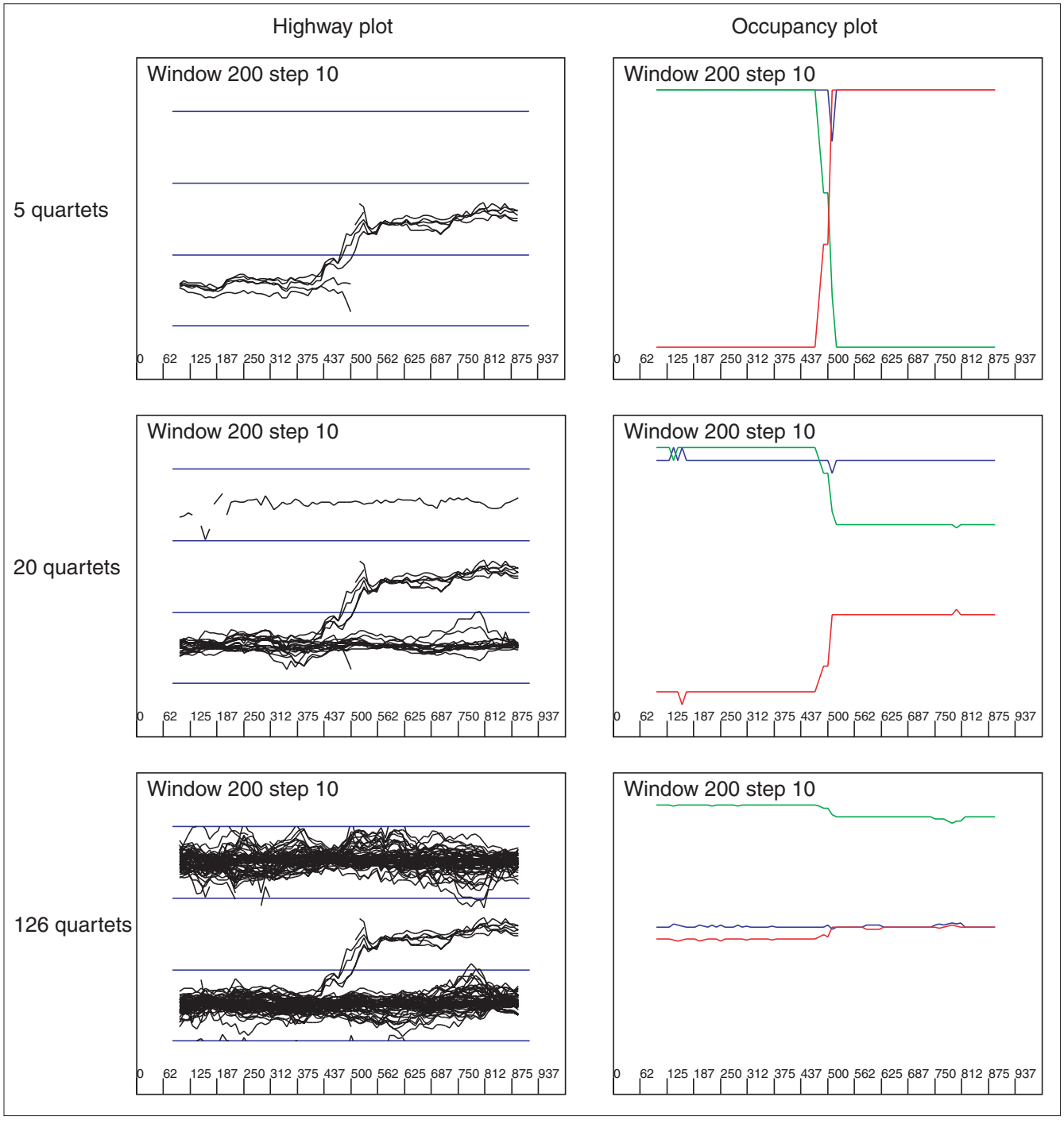

Figure 2
Highway and occupancy plots for simulated data with recombination and balanced trees.

because of its size (26 sequences). For the HIV-1 datasets, RecPars found 34 recombinant regions.

The results obtained from the PLATO and RecPars analysis seem to indicate at first sight that both datasets are highly recombinant. However, in both cases all inferred recombinant regions were fairly small, having lengths between 4 and 40 bp. Moreover, it is well-known that both methods suffer from the problem of false positives [6]. At the same time, they also give reliable results only for strong recombination signals. It therefore seems likely that the observed recombinant regions are artifacts due to the phylogenetic noise in

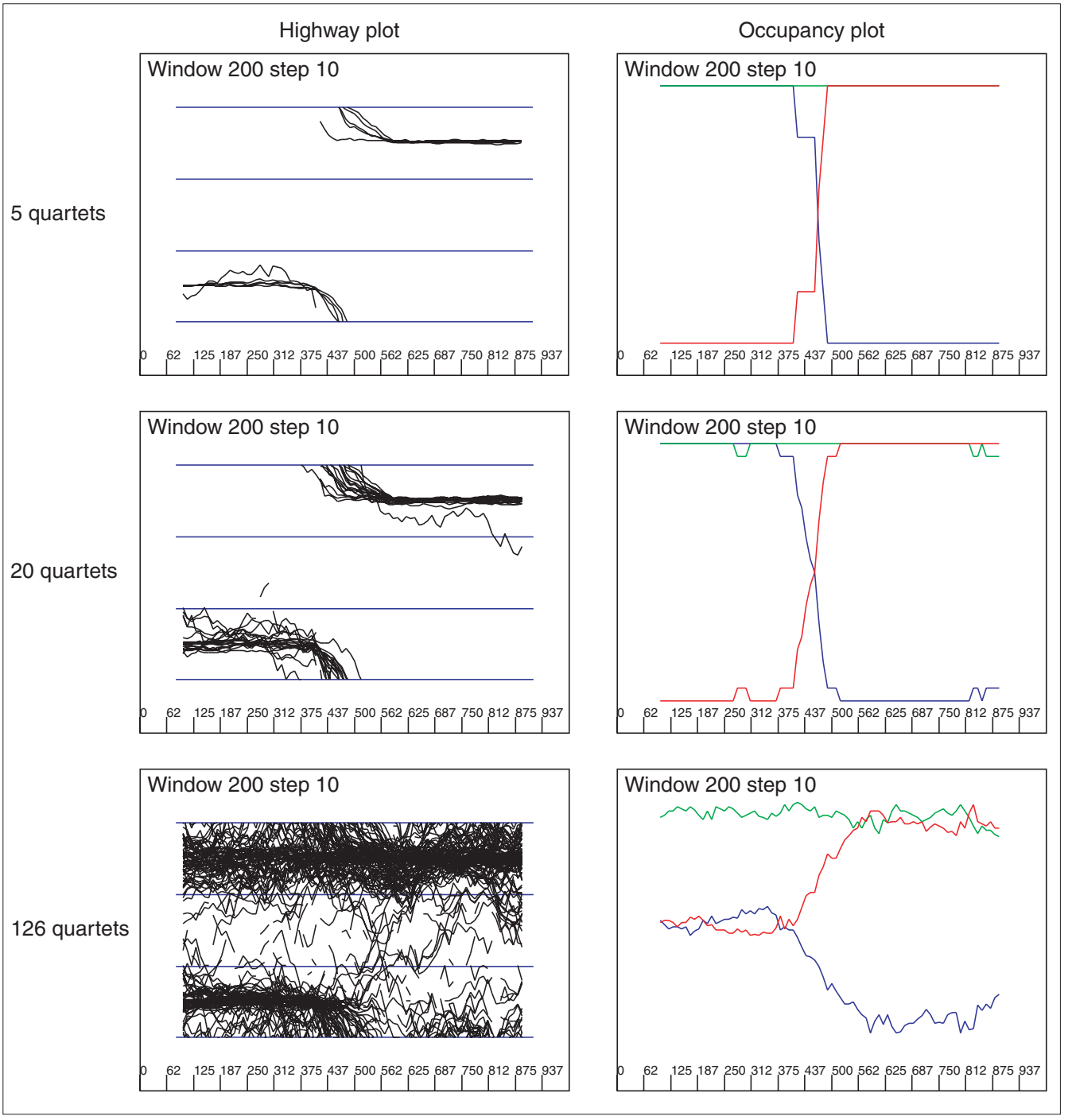

Figure 3
Highway and occupancy plots for simulated data with recombination and unbalanced trees.

our example datasets. This is further confirmed by noting that none of the recombinant regions found by PLATO and RecPars corresponds to each other, or to those detected by VisRD. However, note that the relative assessment - that the HIV-1 data are more recombinant than the influenza data - is in agreement with our analysis using VisRD.

We also analyzed the datasets using bootscanning as implemented in SimPlot v.3.2b [18]. This method detects recombination breakpoints under the explicit assumption that a certain sequence has been generated by recombination. Several values for the setting parameters were tried with very similar results. The scans presented here make use of

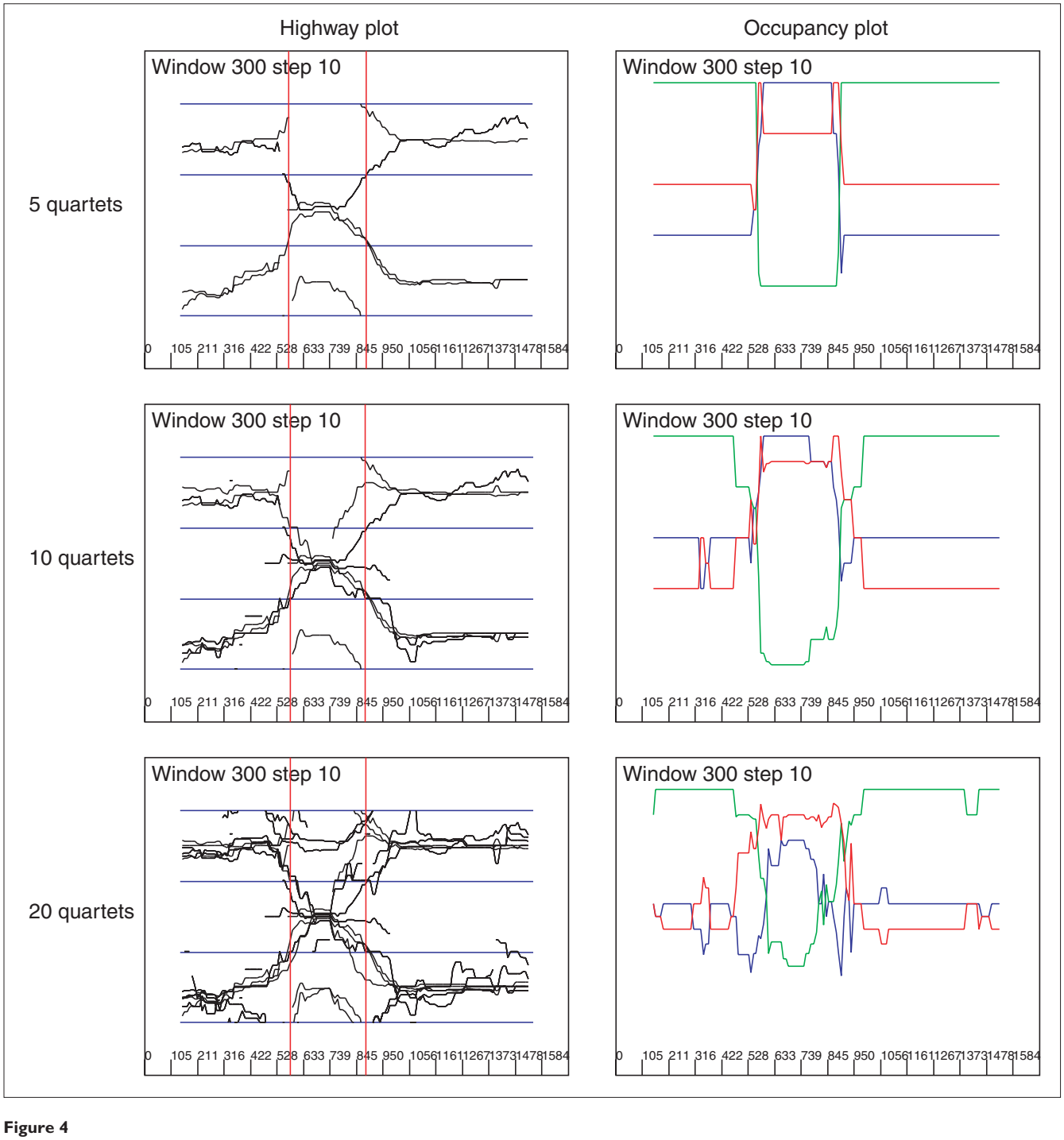

Figure 4
Highway and occupancy plots for an HIV-1 *gag* region alignment containing the recombinant sequence 93BR029. The red vertical lines indicate the two putative breakpoints.

the default settings with a window length of 200 bp and a step length of 20 bp.

The bootscan graph of the putative recombinant HIV-1 sequence 93BR029 is displayed in Figure 6a. The graph indicates one recombinant region with breakpoint positions at approximate sequence positions 580 and 890. This result is nearly identical to that obtained by quartet-scanning.

Bootscanning was also applied to the influenza A set. The resulting graphs (Figure 6b) show no clear sign of any recombination in the suggested recombinant sequence

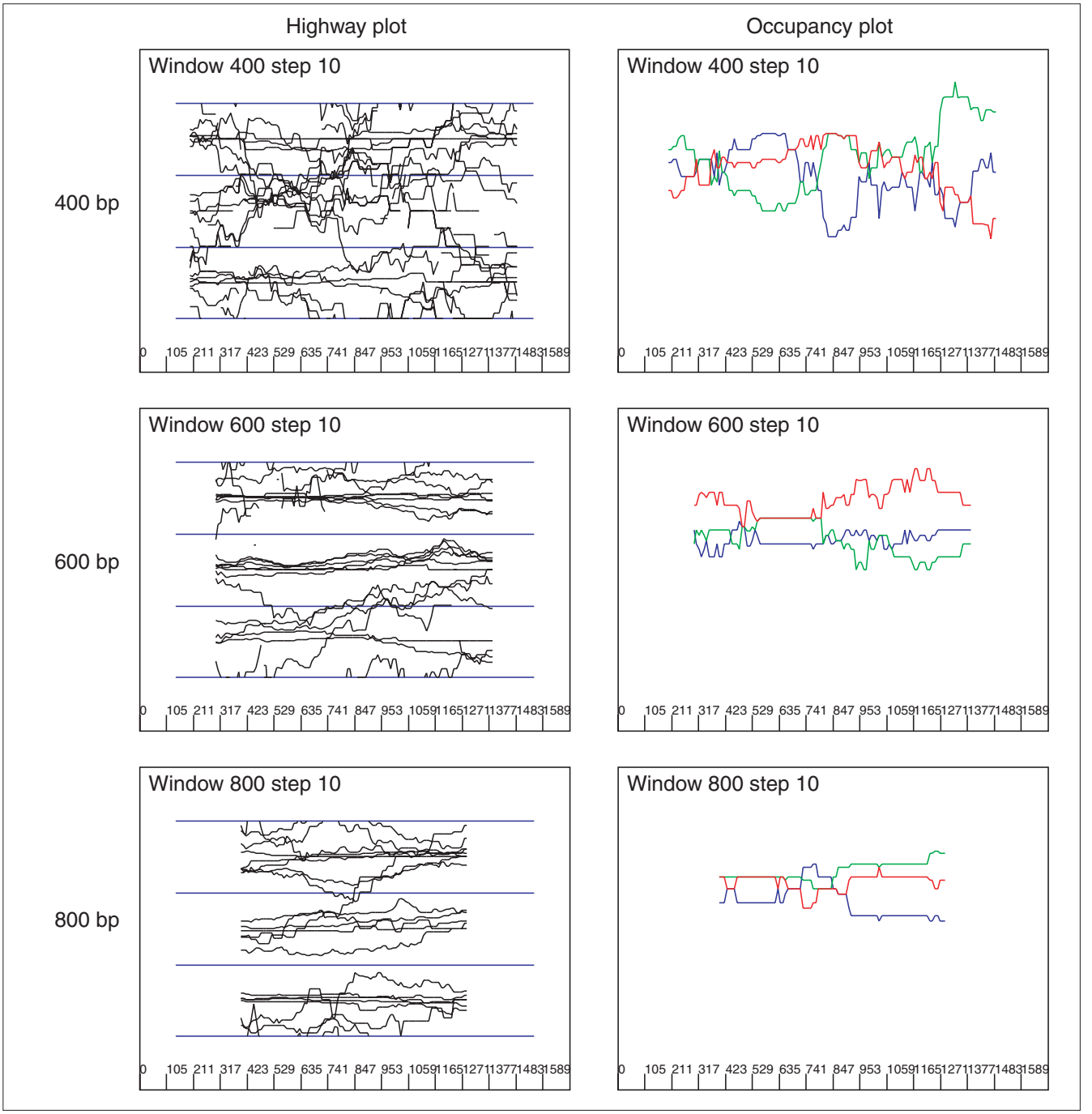

Figure 5
Highway and occupancy plots for the 1918 influenza virus dataset. The window size varies from 400 to 800 bp.

(South Carolina 1918) or in any other. Again, the results are consistent with the quartet-scanning results.

Conclusions

There are many methods available for detecting recombination [3,5-7]. Most of these approaches are based on explicit statistical models and are also computationally intensive. The highway and occupancy plots presented in this paper provide a complementary approach. In the tradition of exploratory data analysis [25], these plots identify recombination by appropriately visualizing the data. Using simulated and molecular data we have shown that this approach promises to give fast and accurate results.

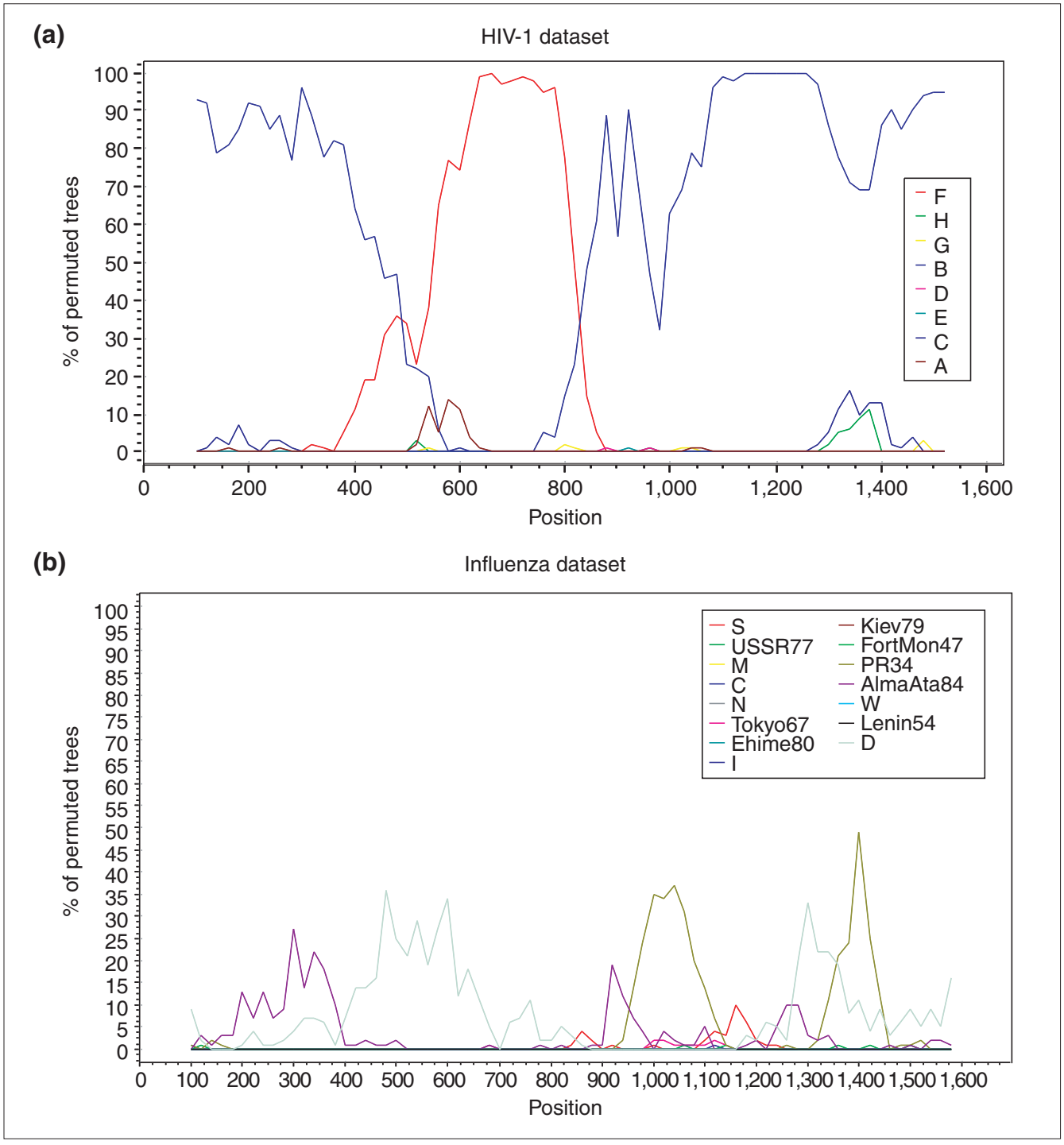

Figure 6
Bootscan plots. **(a)** HIV-1 data; **(b)** 1918 influenza virus data.

We have compared our visual method with some other widely used approaches. The most similar method is bootscanning [9], which works by detecting shifts in the bootstrap support for clades in trees computed along the sequence alignment. Generally, using our method we obtain results that are in accordance with those from bootscanning. However, our approach has some features that are complementary to bootscanning. First, our plots have the benefit that no hypothesis regarding which sequences are recombinant is required *a priori*. Second, the highway and occupancy plots

can be used for datasets with a larger number of sequences than bootscanning as we investigate only quartet trees rather than full-sized genealogies along the alignment. Third, we are able to identify not only breakpoints but, by looking at the taxon frequencies in the sampled quartets, putative recombinant sequences also.

However, one potential drawback of our method is that the visualization process can break down if the dataset contains too many recombinant sequences. In this case it is hard to separate noise from recombination signal, both for visual and statistical approaches. Furthermore, processes other than recombination, such as rapid change of evolutionary rates, could cause phylogenetic inhomogeneity, and thus lead to false positives. Finally, it may be difficult to choose suitable parameters (that is, window size and number of quartets used) when there is a lot of noise in the data. In practice it is therefore probably best to verify breakpoint estimations using several methods, combining visualization with more traditional analyses.

To summarize, the highway and occupancy plots are two versatile tools for performing a preliminary exploratory analysis of data. Both can be employed quickly with a minimum of prior assumptions. To analyze a dataset, the best approach may be to first use highway plots with different numbers of quartets to ascertain the presence of breakpoints and then estimate their positions more precisely using occupancy plots with smaller numbers of quartets. Putative recombinants may then be identified using these quartets. Not least because of the pervasiveness of recombination, we believe that visual recombination detection will have an important role in sequence analysis.

Materials and methods

Outline

Our new method for visualizing and detecting recombination consists of two main steps. First, the mode of evolution along a given sequence alignment is measured by applying quartet-mapping [13,14] to successive parts of the alignment using a sliding window approach. In this way phylogenetic (in-)homogeneity over sites is quickly assessed without requiring possibly expensive estimation of full-sized genealogies. Second, the data so generated are filtered and visualized. In particular, the highway and occupancy plots (introduced in the Background section) are constructed. These allow detection of recombination by inspection.

Quartet-mapping

Quartet-mapping is a technique for visualizing the tree-likeness of a set of aligned sequences that was introduced in [14], and is a generalization of likelihood-mapping [13]. Both quartet- and likelihood-mapping are related to the method of statistical geometry [26].

With four taxa there are exactly three possible fully resolved topologies: T_1, T_2 and T_3. For each quartet of aligned sequences a support σ_i is computed for each of the quartet trees T_i, $1 \le i \le 3$. This support can be either the likelihood of the sequences given the tree, a measure that is used in likelihood-mapping [13], or it can be computed using distance or parsimony techniques [14], as we do here. A relative support s_i is also computed for each tree T_i defined by:

$$s_i = \frac{\sigma_i}{\sigma_1 + \sigma_2 + \sigma_3},$$

so that $0 \le s_i \le 1$. The main idea behind quartet-mappings is to represent the relative support values s_1, s_2, s_3 as a vector in two-dimensional space (noting that the three components s_i are dependent on each other, as $s_1 + s_2 + s_3 = 1$). In the quartet-mapping, each vector is represented by a point in an equilateral triangle using a barycentric coordinate system (Figure 7a). For instance, the three vectors (1, 0, 0), (0, 1, 0) and (0, 0, 1), corresponding to the perfect tree topologies T_1, T_2, and T_3, respectively, are represented by the three vertices of the triangle, whereas the vector (1/3, 1/3, 1/3), assigning equal weight to all three quartet trees and hence corresponding to the star tree, is represented by the central point of the triangle. For an alignment of n sequences, there are $\binom{n}{4}$ possible quartets of sequences, so that a complete quartet-mapping triangle contains $\binom{n}{4}$ points. The distribution of these points in the triangle provides an intuitive picture of how the sequences might have evolved [13].

Visual detection of recombination

The basic strategy for visual detection of recombination is to first slide a window along the alignment in a stepwise, overlapping fashion, computing at each step a quartet tree for each quartet of sequences and its corresponding point in the quartet-mapping triangle. Subsequently, the trajectories that the resulting points along the alignment follow in the triangle are analyzed. If recombination is present, the evolutionary history of the sequences changes along the alignment and, as a consequence, the support of the topology of some of the quartet trees will change. This causes movements of the corresponding points in the quartet-mapping triangle. In general, positions in the alignment where there are large deviations will result in fluctuations that can indicate possible breakpoints.

Because there is a wealth of data in even a single quartet-mapping triangle, it is not possible to display the information gathered along the alignment without some amount of data reduction. In our approach we visualize the trajectories that the points follow in the quartet-mapping triangle in a two-dimensional graph called the highway plot, which is constructed as follows. Polar coordinates (r, θ) are chosen for the triangle, as shown in Figure 7b. A point in the triangle corresponding to a window in a sequence alignment starting

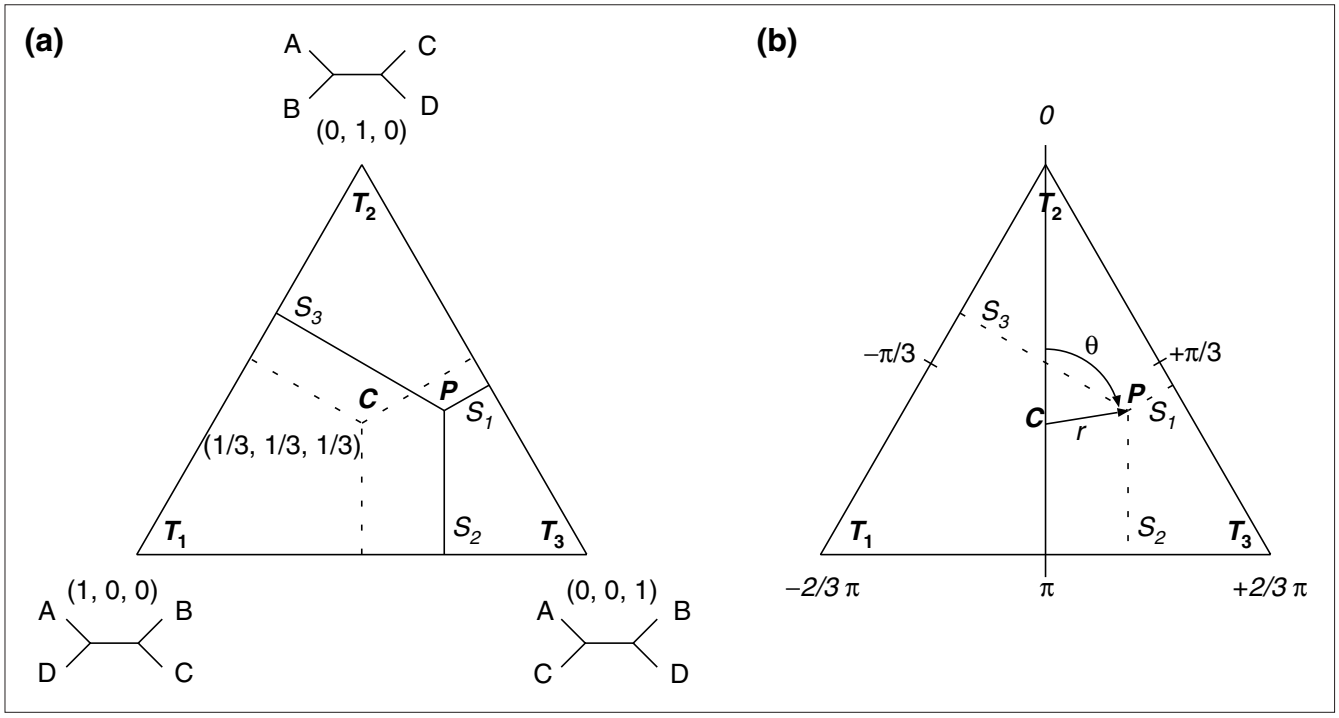

Figure 7
The quartet-mapping triangle. **(a)** Barycentric coordinate system; **(b)** polar representation. In both (a) and (b), **C** denotes the centroid (1/3, 1/3, 1/3) and **P** some general point (s_1, s_2, s_3).

at position x is then represented by a point (x, θ) in a graph which has x along the horizontal axis and θ along the vertical axis. This point is plotted on a gray scale with a darkness that increases with r. Points in the center of the triangle are transparent whereas those at its vertices are black. As many quartets are plotted simultaneously, a technique called saturation brushing is used to ensure that there is a cumulative effect on the darkness of the points of the highway plot [27,28].

In the highway plot, most trajectories will usually cluster in the middle of one of three lanes that correspond to the three corners of the triangle. In the presence of a recombination breakpoint, many trajectories within some or all of these lanes will simultaneously cross into neighboring lanes (see for example, Figure 2).

As a supplement to the highway plot, we provide an occupancy plot, which is another two-dimensional graph that is constructed as follows. For each window in a sequence alignment starting at position x the quartet-mapping triangle is divided into three equally sized regions, as indicated by the dashed lines in Figure 7a (just as in quartet- or likelihood-mapping (see Figure 3 of [13]). Subsequently, for each region the percentage of points p in this region is computed and represented by a point (x,p) on a graph that has x along the horizontal axis and p along the vertical axis. This results in a graph consisting of three curves that indicate the percentage of points occupying each region for each window along the alignment (see for example, Figures 2 and 3). Recombination events are usually indicated by rapid changes in the occupancy distribution.

Trajectory filtering

The behaviors described above for the highway and occupancy plots in the presence of recombination are rather simplistic. In reality, there are several comparatively strong sources of noise that can obscure any recombination signal. Thus one of the main obstacles in detecting recombination utilizing quartet-mappings is devising suitable filters for this intrinsic noise. This may explain why our success in detecting recombination using method proposed in [15] was extremely limited (data not shown). The only other attempt to detect recombination using a quartet approach that we know of [29] also lacked an explicit quartet filtering step.

Our approach to filtering consists of several elements. For an alignment of n sequences we either analyze all possible $\binom{n}{4}$ quartets or, if n is large ($n \geq 10$), we make a random selection of $\binom{10}{4}$ quartets and use these in all subsequent analysis.

We begin by ranking the quartets according to how far their corresponding trajectories deviate along the alignment as follows. An initial scan is made using all possible quartets in which the position of every point in the triangle is recorded

for each window. Subsequently, for each quartet the mean position of all of its corresponding points in the triangle along the alignment is computed, and, using the Euclidean metric [14], the maximum distance from each of these points to the mean position is computed. The quartets are then sorted so that those with the largest such distance come first. A recombination event will be visible mainly in the trajectories of the highest-ranking quartets, whereas the lower-ranking quartets contribute noise that can obscure the signal. Thus, when visualizing recombination signals, best results are usually achieved by selecting only a relatively small number (5-20) of top-ranking quartets (compare Figures 2 and 3, and see Results).

A second filter is implicit in the color scheme adopted for the highway plot. Points near the center of the triangle represent quartets that are nearly star-like, and thus contain little phylogenetic information. However, owing to their proximity to the center these points will have low intensity in the highway plot, so that phylogenetically weak quartets do not make a high visual contribution to the resulting plot.

Third, because of the use of polar coordinates, points near the center of the triangle can have disproportionately large effects on the highway plot, as a small change in position near the center can lead to a large variation in the θ coordinate. Further random variation can also arise as a result of sampling error when quartet-trees are reconstructed from finite-length sequences. This error depends on the window size and can also lead to large random jumps in the value of θ. To eliminate these two sources of noise, we impose a user-defined threshold that fixes the maximum allowed change in θ between the positions of the points corresponding to the quartet trees computed for two consecutive windows. The quartets removed in this step are ignored only for the current window, where they are phylogenetically uninformative, leading to little loss of relevant information.

Computer program

The highway and the occupancy plots, as well as trajectory filtering and animated quartet-mappings, are implemented in a user-friendly program called VisRD (Visual Recombination Detection or 'Wizard'). VisRD can be downloaded from [30] and is distributed under the terms of the GNU General Public License. VisRD requires Java 2, version 1.3 (or later).

Acknowledgements

We thank Andrew Rambaut for providing the 1918 flu virus alignment. K.S. was supported by an Emmy Noether research grant (STR 624/1-2) from the Deutsche Forschungsgemeinschaft (DFG). K.F. and V.M. thank The Swedish Research Council (VR). B.H. and V.M. thank The Swedish Foundation for International Cooperation in Research and Education (STINT).

References

1. Schierup MH, Hein J: **Recombination and the molecular clock.** *Mol Biol Evol* 2000, **17:**1578-1579.
2. Schierup MH, Hein J: **Consequences of recombination on traditional phylogenetic analysis.** *Genetics* 2000, **156:**879-891.
3. Posada D, Crandall KA: **Intraspecific gene genealogies: trees grafting into networks.** *Trends Ecol Evol* 2001, **16:**37-45.
4. Posada D, Crandall KA: **The effect of recombination on the accuracy of phylogeny reconstruction.** *J Mol Evol* 2002, **54:**396-402.
5. Wiuf C, Christensen T, Hein J: **A simulation study of the reliability of recombination detection methods.** *Mol Biol Evol* 2001, **18:**1929-1939.
6. Posada D, Crandall KA: **Evaluation of methods for detecting recombination from DNA sequences: computer simulations.** *Proc Natl Acad Sci USA* 2001, **98:**13757-13762.
7. Posada D: **Evaluation of methods for detecting recombination from DNA sequences: empirical data.** *Mol Biol Evol* 2002, **19:**708-717.
8. Brown CJ, Garner EC, Dunker AK, Joyce P: **The power to detect recombination using the coalescent.** *Mol Biol Evol* 2001, **18:**1421-1424.
9. Salminen M, Carr JK, Burke DS, McCutchan FE: **Identification of recombination breakpoints in HIV-1 by bootscanning.** *AIDS Res Hum Retroviruses* 1995, **11:**1423-1425.
10. Sawyer S: **Statistical tests for detecting gene conversion.** *Mol Biol Evol* 1989, **6:**526-536.
11. Worobey M: **A novel approach to detecting and measuring recombination: new insights into evolution in viruses, bacteria, and mitochondria.** *Mol Biol Evol* 2001, **18:**1425-1434.
12. McGuire G, Wright F, Prentice MJ: **A Bayesian model for detecting past recombination events in DNA multiple alignments.** *J Comput Biol* 2000, **7:**159-170.
13. Strimmer K, von Haeseler A: **Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment.** *Proc Natl Acad Sci USA* 1997, **94:**6815-6819.
14. Nieselt-Struwe K, von Haeseler A: **Quartet-mapping, a generalization of the likelihood-mapping procedure.** *Mol Biol Evol* 2001, **18:**1204-1219.
15. Holland B, Huber K, Dress A, and Moulton V: **Delta-plots: A tool for the analysis of phylogenetic distance data.** *Mol Biol Evol* 2002, **19:**2051-2059.
16. Rambaut A, Grassly NC: **Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees.** *Comput Appl Biosci* 1997,**13:**235-238.
17. Gao F, Robertson DL, Carruthers CD, Morrison SG, Jian B, Chen Y, Barre-Sinoussi F, Girard M, Srinivasan A, Abimiku AG, *et al.*: **A comprehensive panel of near-full-length clones and reference sequences for non-subtype B isolates of human immunodeficiency virus type 1.** *J Virol* 1998, **72:**5680-5698.
18. Lole KS, Bollinger RC, Paranjape RS, Gadkari D, Kulkarni SS, Novak NG, Ingersoll R, Sheppard HW, Ray SC: **Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination.** *J Virol* 1999, **73:**152-160.
19. Gibbs MJ, Armstrong JS, Gibbs AJ: **Recombination in the hemagglutinin gene of the 1918 "Spanish flu".** *Science* 2001, **293:**1842.
20. Worobey M, Rambaut A, Pybus OG, Robertson DL: **Questioning the evidence for genetic recombination in the 1918 "Spanish flu" virus.** *Science* 2002, **296:**211.
21. Grassly NC, Holmes EC: **A likelihood method for the detection of selection and recombination using nucleotide sequences.** *Mol Biol Evol* 1997, **14:**239-247.
22. Hein J: **Reconstructing evolution of sequences subject to recombination used parsimony.** *Math Biosci* 1990, **98:**185-200.
23. Schmidt HA, Strimmer K, Vingron M, von Haeseler A: **TREE-PUZZLE: maximum-likelihood phylogenetic analysis using quartets and parallel computing.** *Bioinformatics* 2002, **18:**502-504.
24. **RecPars - Parsimony analysis of DNA sequences** [http://www.daimi.au.dk/~compbio/recpars/recpars.html]
25. Tukey JW: *Exploratory Data Analysis.* Reading, MA: Addison-Wesley; 1977.
26. Eigen M, Winkler-Oswatitsch R, Dress A: **Statistical geometry in sequence space: a method of quantitative comparative sequence analysis.** *Proc Natl Acad Sci USA* 1988, **85:**5913-5917.
27. Wegman EJ: **Data mining and visualization: some strategies.** *Bull Int Statist Inst* 1999, **52:**223-226.
28. Ball P: **Picture this.** *Nature* 2002, **418:**11-13.
29. Jobb G: Eine Maximum-Likelihood Methode zum Aufspüren phylogenetischer Inhomogenitäten in molekularen Datensätzen. Master's thesis. Munich: University of Munich; 1999.
30. **VisRD: Visual Recombination Detection** [http://www.lcb.uu.se/~vmoulton/software]