VisRD 3.0 Manual: Visual Recombination Detection software package

Martin Lott

February 7, 2008

1 Introduction

Overview VisRD is a software package allowing graphical inspection of the phylogenetic content of a sequence alignment, that is primarily intended for detection of recombination and recombination breakpoints [3]. It is a quartet-based method, as such, analysis is performed on four-sequence sub-alignments.

Installation The program, VisRD is freely available for any platform from **http://www.cmp.uea.ac.uk/∼vlm/visrd/**. This page describes how to install the program and offers a number of test files to run as examples. Any queries about the installation of the program should be directed to Martin.Lott@uea.ac.uk

Running the program For real world datasets a large number of quartets (over 1000) may be required, to prevent Java running out of memory you may use the following command-line extension.

$$-XmxNNNm$$

Where NNN is the size of the memory to allocate in megabytes, 512 or 1024 is best if they are supported by your computer. Where only a taxon ranking is required you may set the number of quartets to display to 0, this will further reduce memory requirements.

Note: VisRD 3.0 is a standalone program, unlike version 2 it does not require SplitsTree.

Disclaimer This software is supplied as-is, with no warranty of any kind expressed or implied. We have made every effort to avoid errors in design and execution of this software, but we will not be liable for its use or misuse. The user is solely responsible for the validity and consequences of any results generated.

2 Getting started

Loading an alignment When the program starts, a prompt to load in a sequence alignment appears. Alignments may be either protein or nucleotide based and can be presented in either `NEXUS` or `Fasta` format. Ambiguous sites are resolved to one of their possible states based on the nature of the ambiguity. For example, a site denoted 'M' denotes either an 'A' or 'C' nucleotide; each is picked with equal probability. If a valid file is not given the program will not start.

Quartet selection On each quartet (subalignments on four sequences) there are three possible tree topologies. By default, quartets are formed from all possible four-taxa subsets, this is known as *included*. Taxa may be moved to the *included in all* and *not included* list if required by the user. To do this, select the taxa and click the appropriate button below its list.

Where pre-defined groups of taxa are known the 'use group model' option should be selected, in this case VisRD uses the extra constraint that each taxon in a given quartet must be from a different group. Initially four groups are shown but unlike previous versions of VisRD, any number of groups may be added, renamed and removed. If the start of each taxa name denotes their group *auto assign* will create those groups and automatically assign the taxa to their correct group.

Method parameters VisRD uses a sliding window approach to consider each part of the sequence alignment in isolation and thus detect recombination. Consider the mosaic sequence in figure 2, the topology of the alignment changes according to the three tree topologies shown below the mosaic sequence. A *sliding window* considers substrings of a sequence alignment of fixed size, the *window size*. We consider the first window of that size which starts at the first character in the sequence and iteratively move along the alignment using a fixed *step size*.

The "number of quartets to display" is number of quartets which will be used to produce the *highway plot*, *occupancy plot* and *quartet mapping triangle*; where only a *taxon ranking* is required this number may be set to 0 to save memory. Note that these *spinner* objects do not register that a value has changed until (1) the user clicks on another component or tab or (2) the user finishes input by pressing "Enter".

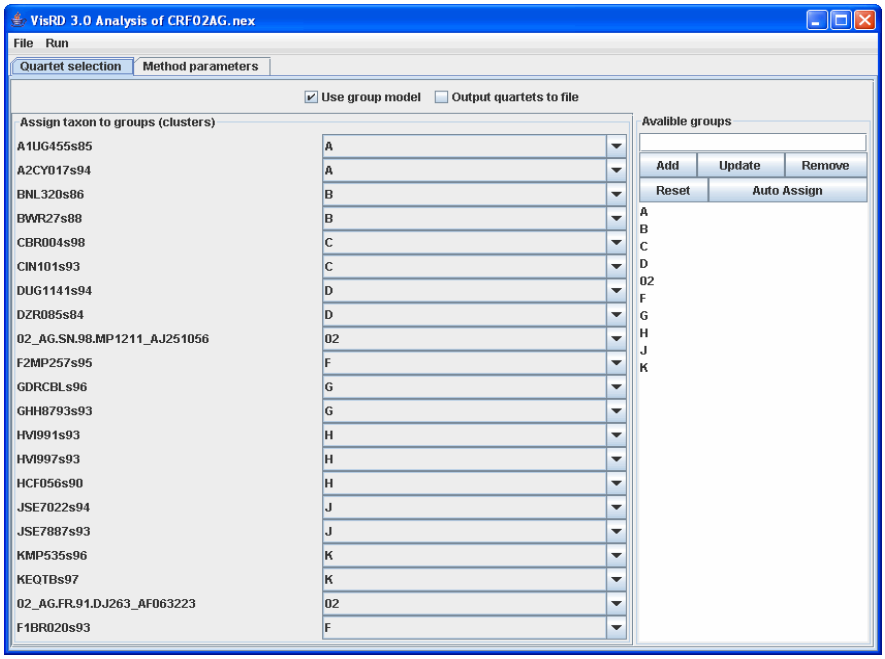

Figure 1: Use of group model to assign HIV taxa to their subtype

Scanning the alignment The *support* for each quartet, as defined by a distance metric and statistical geometry [2], is first computed by *scanning* the alignment. When a recombinant-free null distribution is required the *random* or *shannon* shuffling method may be used to shuffle sites in the alignment. The "output to file" writes those generated datasets to a file, the name of

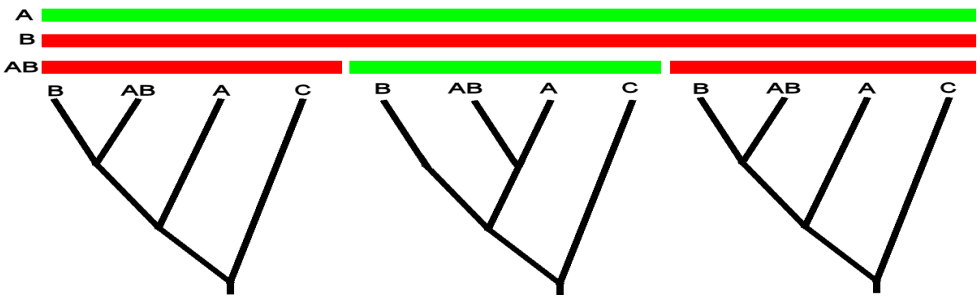

Figure 2: An example of a mosaic (recombinant) sequence AB and the tree topologies that give rise to it.

that file will be that of the input dataset with a *_shuffled* suffix.

Statistical geometry is the default weighting method, other options use the weighted statistical geometry extension introduced in [5]. That method is extended to distance-based inference, and depending on whether a nucleotide or protein sequence is loaded appropriate model options are shown. Finally, where the number of quartets is infeasible to process the user may "pick taxa at random". In this option, a subset of randomly selected quartets is used.

3 Viewing the results

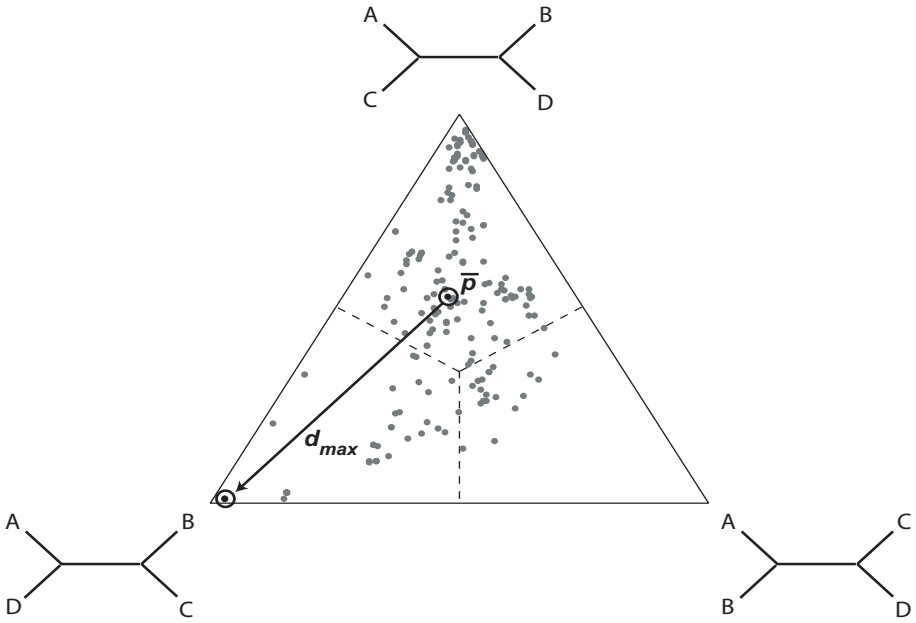

Figure 3: In the quartet-mapping triangle, each dot represents the relative support for the three unrooted topologies [5].

Quartet mapping triangle Quartets may be summarised in a quartet mapping triangle [7] (e.g. see figure 3). For each quartet, a point is used to display the support of each topology. Note that a point in the middle of an edge on this triangle corresponds to a split network [1] representing the splits of the two trees corresponding to the ends of the edge. Each frame

corresponds to the analysis of a given sequence window. A moving line close to the bottom of the graph indicates which part of the sequence is displayed.

It is possible to stop and re-start the animation at any step, and to jump to a given window as desired. By clicking the "Save EPS" button and choosing a file name, the currently shown triangle graph will be saved as a scalable encapsulated postscript file. The other controls above the graph control how many quartets are shown and which part of the sequence to use. By selecting a taxon or group under the "color quartets" down-down menu all those quartets containing that taxon/group will be colored in black for easy identification.

Highway plot The highway plot has three 'lanes', the center of each denotes maximal support for its corresponding topology. Each quartet becomes a path on the highway plot, the position of each point on the path is determined by the topology support at that window position. Where recombination occurs we expect to see a shift in topology support for all quartets which include that recombinant taxa. By clicking the graph, red vertical marker lines are added. The delta threshold is a maximum fluctuation filter which hides lines making unrealistically large jumps, setting this value to 1.0 disables the filter.

Like the quartet-mapping triangle, the highway plot also has "Color quartets" and "Save EPS" options. The line connecting the two consecutive points with largest shift in topology is colored red and the plot itself is built up gradually to stop the program crashing. Unlike previous versions, all quartets now start in the middle lane.

Occupancy plot The occupancy plot is intended to summarize the information displayed in the highway plot. Each lane in the highway plot corresponds to a line in the occupancy plot, this shows the proportion of quarters in that line at each window position. Because all quartets start in the middle lane of the highway plot the blue line will start at 1.0 whilst the red and green lines start at 0.0.

4 Additional options

Taxon ranking Depending on whether a group model is used, the rank for each taxa or each groups is presented here. Because the absolute ranks are

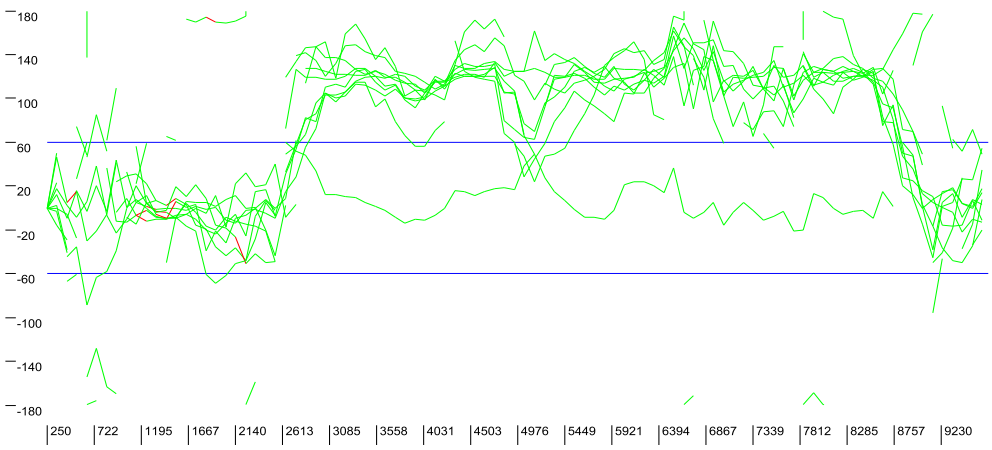

Figure 4: A Highway plot for the AB HIV recombinant 'KAL153' using standard statistical geometry. The top 50 quartets are drawn with those containing the recombinant colored green.

sensitive to the evolutionary models and topology evaluation criteria used, we also report the ranking values as percentages of highest ranking taxon or group.

Finally, where a null distribution has also been computed and the results (in CSV format) are in the same directory a P-value will be computed and displayed.

Recombination statistic A P-value may be determined to indicate whether or not the sequence alignment contains a recombinant taxa/group. This is computed by counting the number of simulated recombination-free datasets that have a higher average ranking than the real dataset. This P-value is intended to indicate the presence/absence or recombinant taxa in a dataset and has been used successfully for HIV/SIV datasets. For further details on this approach or to cite the paper please see [4].

Command-line interface Where only a taxon ranking is required the command-line interface may be used. This can be particularly useful for running large datasets remotely on a cluster. To see the full list of command line options use the `-h` switch as follows:-

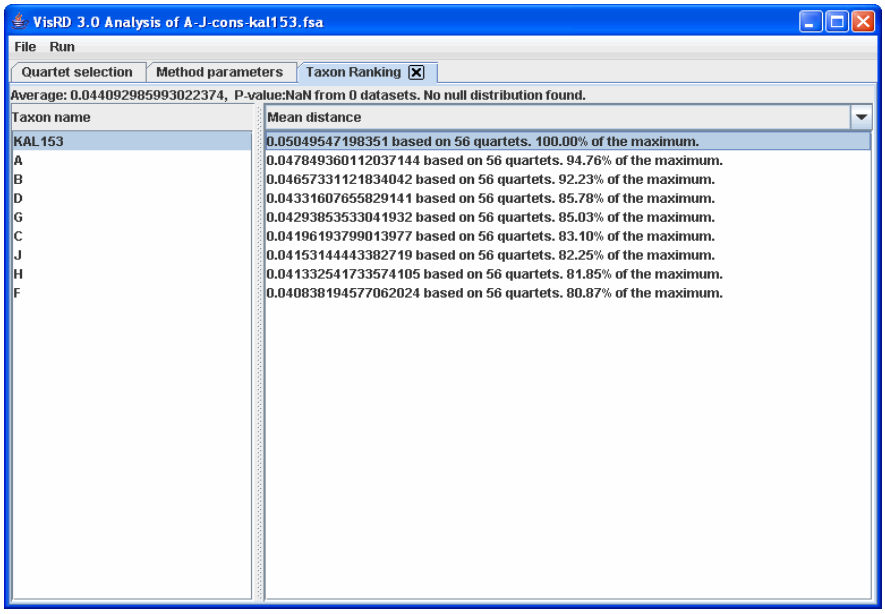

Figure 5: The main program window

```
java -Xmx1024m -jar visrd.jar -h
```

Exporting an image EPS, for use in paper Any highway plot, quartet mapping triangle or occupancy plot can be saved in EPS format for use in a paper. If such images are to be used in scientific papers we ask that you cite the original paper, [6].

5 Acknowledgements and Availability

The VisRD package was originally developed by Kristoffer Forslund and Vincent Moulton at The Linnaeus Centre for Bioinformatics with contributions from Korbinian Strimmer and Daniel Huson.

Version 3.0 was developed by Martin Lott and Vincent Moulton, whose code was adapted from Version 2.3 by Martin Lott with contributions from Philippe Lemey and Joe Parker. The current version is copyright (2002-2008) of Martin Lott, Kristoffer Forslund and Vincent Moulton. This manual is based on the manual for version 2.2 by Kristoffer Forslund.

The VisRD software is released under GNU General Public License 3.0 as set down at `http://www.gnu.org/licenses/gpl-3.0-standalone.html`

References

[1] H. J. Bandelt and A. W. Dress. Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Molecular phylogenetics and evolution*, 1(3):242–252, 1992.

[2] M. Eigen and R. Winkler-Oswatitsch. Statisitcal geometry on sequence space. *Methods in Enzymology*, 183:505–530, 1990.

[3] K. Forslund, D. H. Huson, and V. Moulton. Visrd–visual recombination detection. *Bioinformatics*, 20:3654–5, 2004.

[4] P. Lemey, M. Lott, and V. Moulton. Detecting recombination and recombinants from sequence alignments using a quartet scanning approach. *Manuscript*.

[5] K. Neiselt-Struwe and A. von Haeseler. Quartet-mapping, a generalization of the likelihood-mapping procedure. *Molecular Biology and Evolution*, 18:1204–19, 2001.

[6] K. Strimmer, K. Forslund, B. Holland, and V. Moulton. A novel exploratory method for visual recombination detection. *Genome Biology*, 4(5):R33, 2003.

[7] K. Strimmer and A. von Haeseler. Likelihood-mapping: A simple method to visualize phylogenetic content of a sequence alignment. *Proceedings of the National Academy of Sciences*, 94:6815–6819, 1997.