BIOINFORMATICS APPLICATIONS NOTE Vol. 16 no. 6 2000
Pages 562–563

RDP: detection of recombination amongst aligned sequences

Darren Martin and Ed Rybicki

Microbiology Department, University of Cape Town, Private Bag, Cape Town, South Africa

Received on September 24, 1999; revised on December 8, 1999; accepted on December 23, 1999

Abstract

Summary: *Recombination Detection Program (RDP) is a program that applies a pairwise scanning approach to the detection of recombination amongst a group of aligned DNA sequences. The software runs under Windows95 and combines highly automated screening of large numbers of sequences with a highly interactive interface for examining the results of the analyses.*

Availability: *For academic purposes RDP is available free of charge from: http://www.uct.ac.za/depts/microbiology/microdescription.htm*

Contact: *darren@molbiol.uct.ac.za*

Recombination between divergent genomes is believed to be a major mechanism by which diversity amongst viruses is generated (Robertson *et al.*, 1995). Although a number of methods have been devised for the analysis of recombination (Grassly and Holmes, 1997; Hein, 1990; Maynard Smith and Smith, 1998; McGuire *et al.*, 1997; Salminen *et al.*, 1995; Sawyer, 1989; Siepel *et al.*, 1995; Weiller, 1998), the vast majority of computer programs that have been devised to automate these methods lack an interactive user interface, are incompatible with the most common personal computer operating systems, and are relatively inaccessible to casual users. We have written Recombination Detection Program (RDP) as a means of addressing these problems. It runs under Windows95/98/NT and couples a high degree of analysis automation with an interactive and detailed graphical user interface.

For the detection of recombination, RDP utilises a pairwise scanning approach. Beginning with a multiple sequence alignment in Phylip, DNAMAN, FASTA, GCG or CLUSTAL formats, the software examines every possible combination of three sequences for evidence of recombination in a three-step procedure. In the first step all phylogenetically non-informative sites are discarded from the group of three sequences to obtain three information-rich sub-sequences. In every group of three sequences there are two sequences, A and B, that are more closely related to one another than to a third sequence, C. Non-informative sites are:

1. identical in all three sequences
2. different in all three sequences
3. unique to A or B and are not present in any member of a group of reference sequences.

The method of reference sequence selection is user-defined and is based on the relative positions of the three selected sequences in an UPGMA dendrogram. In the second analysis step a window of user-defined width is moved along the aligned sub-sequences one nucleotide at a time and an average percentage identity for each of the three possible sequence pairs is calculated at each position. Sequences of possibly recombinant origin are identified as regions where the percentage identities of sequences A and C or B and C are higher than for sequences A and B. In the third analysis step the probability that the nucleotide arrangement in the identified region that results in sequences A or B appearing more closely related to sequence C may have occurred by chance, is approximated using the binomial distribution adapted from Rice (1995):

$$P = G \times \frac{L}{N} \times \sum_{m=M}^{N} \left(\frac{N!}{m!(N-m)!} \right) p^m \times (1-p)^{N-m}$$

where G is the number of possible combinations of three sequences, L is the length of the information rich subsequences, N is the length of the putatively recombinant region, M is the number of nucleotides in common between either A or B and C in the putatively recombinant region, and p is the proportion of nucleotides in common between either A or B and C in the entire subsequence. If the value of P is lower than a user-definable cut-off figure, information on the potential recombination is stored for later access before the next combination of three sequences is selected and analysed. Once every combination of three sequences has been analysed, an interactive graphical interface for examination of the analysis results enables the user to access the stored information.

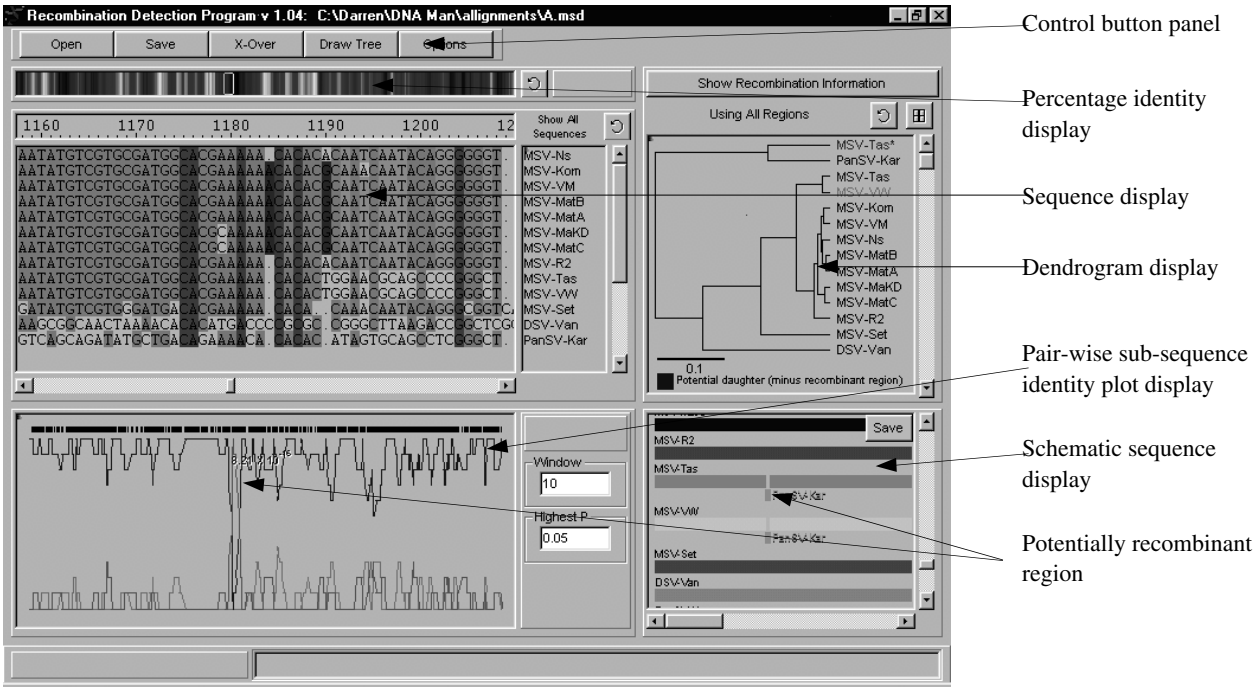

Fig. 1. The RDP user interface. An example is displayed of the output obtained following analysis of the sequences indicated in the sequence display and selection of a potentially recombinant region in the schematic sequence display for further analysis.

Because reference sequences are selected based on their positions relative to the selected triplets within an UPGMA dendrogram there are situations when RDP is unable to correctly discriminate between daughter and parental sequences. These situations may occur if the parental and daughter sequences are all nearest neighbors in the dendrogram, if only one parental sequence is present in the alignment or where daughter sequences have obtained too few of their phylogenetically informative nucleotides from a single parent for them to be situated in the dendrogram at a position that properly reflects their evolutionary history. In all these cases, however, the program will still approximate the correct recombination breakpoints.

The program's interface is divided into a number of sections that display both general information relating all of the aligned sequences to one another, and specific information on the relationships of user-specified putatively recombinant sequences to sequences closely related to their potential parents (Figure 1).

We have used RDP to simultaneously analyse 86 full length HIV and SIV genomes, and have been able to determine the composition of all previously identified inter-subtype HIV-1 recombinants (Robertson *et al.*, 1995; Salminen *et al.*, 1995; Siepel *et al.*, 1995).

References

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.

McGuire,G., Wright,F. and Prentice,M.J. (1997) A graphical method for detecting recombination in phylogenetic data sets. *Mol. Biol. Evol.*, **14**, 1125–1131.

Maynard Smith,J. and Smith,N.H. (1998) Detecting recombination from gene trees. *Mol. Biol. Evol.*, **15**, 590–599.

Rice,J.A. (1995) *Mathematical statistics and data analysis*. Duxbury Press, Belmont, pp. 36–38.

Robertson,D.L., Hahn,B.H. and Sharp,P.H. (1995) Recombination in AIDS viruses. *J. Mol. Evol.*, **40**, 249–259.

Salminen,M., 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. Hum. Retrovirus.*, **11**, 1423–1425.

Sawyer,S. (1989) Statistical tests for detecting gene conversion. *Mol. Biol. Evol.*, **6**, 526–538.

Siepel,A.C., Halpern,A.L., Macken,C. and Korber,B.T.M. (1995) A computer program designed to screen rapidly for HIV type 1 intersubtype recombinant sequences. *AIDS Res. Hum. Retrovirus.*, **11**, 1413–1416.

Weiller,G.F. (1998) Phylogenetic profiles: a graphical method for detecting recombinations in homologous sequences. *Mol. Evol. System.*, **15**, 326–335.