



RDP2: recombination detection and analysis from sequence alignments

D. P. Martin^{1,*}, C. Williamson¹ and D. Posada²

¹Institute of Infectious Diseases and Molecular Medicine, University of Cape Town, Cape Town 7000, South Africa and ²Department of Biochemistry, Genetics and Immunology, University of Vigo, 36200 Vigo, Spain

Received on April 20, 2004; revised on June 28, 2004; accepted on August 13, 2004

Advance Access publication September 17, 2004

ABSTRACT

Summary: RDP2 is a Windows 95/XP program that examines nucleotide sequence alignments and attempts to identify recombinant sequences and recombination breakpoints using 10 published recombination detection methods, including GENECONV, BOOTSCAN, MAXIMUM χ^2 , CHIMAERA and SISTER SCANNING. The program enables fast automated analysis of large alignments (up to 300 sequences containing 13 000 sites), and interactive exploration, management and verification of results with different recombination detection and tree drawing methods.

Availability: RDP2 is available free from the RDP2 website (<http://darwin.uvigo.es/rdp/rdp.html>)

Contact: darren@science.uct.ac.za

Supplementary information: Detailed descriptions of RDP2 and the methods it implements are included in the program manual, which can be downloaded from the RDP2 website.

A major problem encountered while using standard phylogenetic methods in studies involving recombining organisms is that the evolutionary history of a recombinant sequence cannot be described with a single phylogenetic tree. A single recombinant sequence in an alignment can seriously influence the branching order and branch lengths of the phylogenetic trees constructed using the alignment (Posada and Crandall, 2002). In addition, recombination compromises the validity of several phylogenetic inferences one can make by examining trees (Schierup and Hein, 2000a,b). A number of computational tools for detecting and quantifying various aspects of recombination have therefore been developed (for a list of available recombination detection programs see <http://www.umber.embnnet.org/~robertson/recombination/index.shtml>). A comparison of the recombination detection power of 14 of these methods using simulated and real datasets indicated that while some always performed better than others, no single method can be adjudged to be best in detecting recombination

under all conditions (Posada and Crandall, 2001; Posada, 2000).

Sharing major components of its user interface and the RDP recombination detection method with its predecessor, RDP, RDP2 implements a variety of additional non-parametric recombination detection methods (i.e. methods that do not make use of population genetic models and make no attempt to estimate the population recombination rate; Table 1). Among the new inclusions are many methods that have performed well in comparative tests (Drouin *et al.*, 1999; Posada and Crandall, 2001; Posada, 2000). We have focused on published methods that can be used to (1) identify recombinant sequences, (2) identify recombination breakpoints and (3) identify parental sequences. The program can use any combination of six methods to automatically (RDP, GENECONV, MAXIMUM χ^2 , BOOTSCAN, CHIMAERA and SISTER SCANNING) identify recombinant and parental sequences, estimate breakpoint positions and calculate probability scores for potential recombination events. Once all potential recombination events are identified, RDP2 sorts analysis results and attempts to determine the number of unique recombination events identifiable in an alignment. RDP2 can be set to automatically (1) filter out unique events detected by fewer than a specified number of methods, (2) identify consensus daughter and parental sequences using all evidence for a single actual recombination event (often involving many potential parental and daughter sequence combinations detected using multiple methods) and (3) use all evidence for a single actual event to determine most probable breakpoint positions using a modified maximum χ^2 approach (Maynard-Smith, 1992).

RDP2 permits exploration and checking of analysis results in a highly interactive and user-friendly way. For any detected recombinations event, informations such as the method used to detect the event, breakpoint positions, parental sequences, probability values, degrees of agreement with results obtained using other detection methods, raw plot data, informative sites in the alignment and phylogenetic trees, can be displayed by simply clicking on a graphical representation of the event. Once an event is selected for more detailed study, checking

*To whom correspondence should be addressed.

Table 1. A brief description of recombination detection methods implemented in RDP2

Method (a.k.a.)	Sequence comparisons ^a	Variable (V)/ All (A) sites scanned ^b	Sliding window	Automated scans ^c	References
RDP (RDP method)	T	V	+	+	Martin and Rybicki (2000)
GENECONV (Sawyer's runs test)	T/D	V	–	+	Padidam <i>et al.</i> (1999)
BOOTSCAN	T	A	+	+	Salminen <i>et al.</i> (1995)
MAXIMUM χ^2 (MaxChi)	T/D	V	+	+	Maynard-Smith (1992)
CHIMAERA	T	V	+	+	Posada and Crandall (2001)
SISTER SCANNING (SiScan)	T/F	A	+	+	Gibbs <i>et al.</i> (2000)
LARD	T	A	–	–	Holmes <i>et al.</i> (1999)
DISTANCE PLOT (SimPlot)	T/D	A	+	–	Lole <i>et al.</i> (1999)
TOPAL	T	A	+	–	McGuire and Wright (2000)
RETICULATE (compatibility matrix)	F	V	–	–	Jakobsen and Easteal (1996)

^aT, every possible combination of three sequences in an alignment scanned; D, every possible combination of two sequences in an alignment scanned with variable sites inferred from full alignment; and F, full alignment or substantial part thereof (4+ sequences) scanned with variable sites inferred only from the sequences being scanned.

^bThe exact subset of sites scanned will differ between methods and can also differ for the same method with different program settings.

^cOnly six methods can be used to automatically identify recombinant sequences and breakpoints from an alignment. Methods can also be run in either a manual or a checking mode allowing users to test specific recombination hypotheses.

the evidence for recombination using 10 different recombination detection methods (besides the six automated methods these also include LARD, TOPAL, RETICULATE and DISTANCE PLOTS) is achieved by simply selecting the methods from a menu. To further aid in evaluating evidence for recombination, RDP2 can also use PHYLIP components simultaneously (Felsenstein, 1989; Olsen *et al.*, 1994) to display phylogenetic trees (UPGMA, bootstrapped neighbor-joining, least squares or maximum-likelihood) constructed from different portions of an alignment.

As the amount of detectable recombination in an alignment increases, the complexity of correctly inferring which sequences are parental and which are recombinant increases as well. RDP2 encourages user verification of its analysis results and permits user acceptance and rejection of potential recombination events (useful for tracking the progress of an analysis), and interactive 'correction' of apparent parental and daughter sequence misidentification.

We have not placed any restrictions on the size of alignments that can be examined using RDP2. For example, automated analyses using all the detection methods together on a PC with 256 MB RAM and a 1 GHz Celeron Processor can take 5 min for a 50 sequence alignment of 3 kb long sequences and less than 48 h for a 316 sequence alignment of 13 kb long sequences.

ACKNOWLEDGEMENTS

We would like to thank Stanley Sawyer, Andrew Rambaut, Ingrid Jakobsen, Joseph Felsenstein, Gary Olsen, Adrian Gibbs and John Armstrong for either agreeing to have their programs distributed using RDP2 or providing pieces of code in RDP2. We also thank The National Research Foundation

of South Africa (D.P.M.), US National Institutes of Health (D.P.) and the 'Ramón y Cajal' programme of the Spanish government (D.P.) for partially funding the development and distribution of RDP2.

REFERENCES

- Drouin, G., Prat, F., Ell, M. and Clarke, G.D.P. (1999) Detecting and characterizing gene conversions between multigene family members. *Mol. Biol. Evol.*, **16**, 1369–1390.
- Felsenstein, J. (1989) PHYLIP—Phylogeny Inference Package (Version 3.2). *Cladistics*, **5**, 164–166.
- Gibbs, M.J., Armstrong, J.S. and Gibbs, A.J. (2000) Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics*, **16**, 573–582.
- Holmes, E.C., Worobey, M. and Rambaut, A. (1999) Phylogenetic evidence for recombination in Dengue virus. *Mol. Biol. Evol.*, **16**, 405–409.
- Jakobsen, I.B. and Easteal, S. (1996) A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. *Comput. Appl. Biosci.*, **12**, 291–295.
- Lole, K.S., Bollinger, R.C., Paranjape, R.S., Gadarki, D., Kulkarni, S.S., Novak, N.G., Ingersoll, R., Sheppard, H.W. and Ray, S.C. (1999) Full-length human immunodeficiency type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *J. Virol.*, **73**, 152–160.
- Martin, D. and Rybicki, E. (2000) RDP: detection of recombination amongst aligned sequences. *Bioinformatics*, **16**, 562–563.
- Smith, J.M. (1992) Analyzing the mosaic structure of genes. *J. Mol. Evol.*, **34**, 126–129.
- McGuire, G. and Wright, F. (2000) TOPAL 2.0: improved detection of mosaic sequences within multiple alignments. *Bioinformatics*, **16**, 130–134.
- Olsen, G.J., Matsuda, H., Hagstrom, R. and Overbeek, R. (1994) fastDNAML: a tool for construction of phylogenetic trees of

- DNA sequences using maximum likelihood. *Comput. Appl. Biosci.*, **10**, 41–48.
- Padidam,M., Sawyer,S. and Fauquet,C.M. (1999) Possible emergence of new geminiviruses by frequent recombination. *Virology*, **265**, 218–225.
- Posada,D. (2002) Evaluation of methods for detecting recombination from DNA sequences: empirical data. *Mol. Biol. Evol.*, **19**, 708–717.
- Posada,D. and Crandall,K.A. (2001) Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. *Proc. Natl Acad. Sci. USA*, **98**, 13757–13762.
- Posada,D. and Crandall,K.A. (2002) The effect of recombination on the accuracy of phylogeny estimation. *J. Mol. Evol.*, **54**, 396–402.
- 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. Hum. Retroviruses.*, **11**, 1423–1425.
- Schierup,M.H. and Hein,J. (2000a) Consequences of recombination on traditional phylogenetic analysis. *Genetics*, **156**, 879–891.
- Schierup,M.H. and Hein,J. (2000b) Recombination and the molecular clock. *Mol. Biol. Evol.*, **17**, 1578–1579.