Recombination Analysis Tool (RAT): a program for the high-throughput detection of recombination

Graham J. Etherington1,∗, Jo Dicks2 and Ian N. Roberts1

1Department of Food Safety Science, Institute of Food Research and 2Computational Biology Group, John Innes Centre, Norwich Research Park, Colney Lane, Norwich NR4 7HA, UK

Received on May 28, 2004; revised on August 3, 2004; accepted on August 18, 2004

Advance Access publication August 27, 2004

ABSTRACT

Motivation: Recombination can be a prevailing drive in shaping genome evolution. RAT (Recombination Analysis Tool) is a Java-based tool for investigating recombination events in any number of aligned sequences (protein or DNA) of any length (short viral sequences to full genomes). It is an uncomplicated and intuitive application and allows the user to view only the regions of sequence alignments they are interested in.

Results: RAT was applied to viral sequences. Its utility was demonstrated through the detection of a known recombinant of HIV and a detailed analysis of Noroviruses, the most common cause of viral gastroenteritis in humans.

Availability: RAT, along with a user’s guide, is freely available from http://jic-bioinfo.bbsrc.ac.uk/bioinformatics-research/staff/graham_etherington/RAT.htm

Contact: Graham.Etherington@bbsrc.ac.uk

INTRODUCTION

Recombination is the process whereby two separate molecules of DNA or RNA exchange regions of their genome. The exchange is often in homologous regions of the genome and occurs in both single-stranded and double-stranded DNA and RNA molecules. Thus, recombination is a major mechanism driving evolutionary change. There is a continuing need to define, as accurately as possible, sequences where recombination events may have occurred.

There are five generally recognized methods of detecting recombination in DNA sequences: similarity methods (using the tendency of neighbouring nucleotides to be more compatible than sites farther apart), distance methods (using the estimation of genetic distances between sequences), phylogenetic methods (identifying incongruous tree topologies from different parts of a sequence or genome), compatibility methods (testing for partition phylogenetic incongruence and do not require the phylogeny of the sequences to be known) and substitution distribution (examines the sequences for a significant grouping of substitutions) (Posada et al., 2002). There are also various applications available for examining recombination. SIMPLOT (Lole et al., 1999), PHYPRO (Weiller, 1998), RDP (Martin and Rybicki, 2000) and TOPALi (Milne et al., 2004) look for changes in patterns of genetic diversity. LARD (Holmes et al., 1999), PLATO (Grassly and Holmes, 1997) and BOOTSCAN (Salminen et al., 1995) look for incongruent phylogenetic trees. RETICULATE (Jakobsen and Easteal, 1996) calculates compatibility matrices. PIST (Worobey, 2001) looks for excessive convergent evolution.

However, many current applications have important limitations, e.g. they are not cross-platform tools, are not intuitive to use, do not accept multiple sequence alignments, do not search automatically, use complex, time-consuming algorithms, do not have a user interface, require the sequences to be in a particular format, cannot identify small areas of recombination or do not accept protein sequences. In this paper, we introduce a new program and demonstrate its utility by reference to recombination in viral sequences.

SYSTEMS AND METHODS

A full and detailed explanation of how to use RAT and also the algorithms used may be found on the download page, given in the abstract above.

The Recombination Analysis Tool (RAT) is a cross-platform, Java-based application intended for high-throughput, recombination analysis of both DNA and protein multiple sequence alignments, in any one of seven different file formats. It uses the distance-based method of recombination detection. All of RAT’s operations are carried out through the main RAT GUI, and all output can be saved as data files (.txt, .xls, .csv), or as .jpg files in the case of graphical outputs. RAT is intuitive and easy to use, and only requires a
minimum of input from the user. All the parameters have
default values, all of which may be changed by the user. The
user may use RAT to examine sequences individually using the
Single-sequence viewer, or may use the Auto Search option
that searches for recombination given a user-defined search
criterion.

Single-sequence viewer
If a recombinant sequence (the ‘Test Sequence’) is already
known or suspected, it may be chosen from a drop-down list
of all the sequences in the alignment and then the Sequence
Viewer launched to view the similarity of all other sequences
to the Test Sequence.

Auto Search
The Auto Search option can be used to search through every
sequence for possible recombination events. There are three
parameters involved; lower threshold (the genetic distance
that a sequence must be under to qualify as a suspect recom-
binant), upper threshold (the genetic distance that a sequence
must jump to/from when compared to the lower threshold) and
the number of contributing sequences (the maximum number
of sequences allowable to contribute to a recombinant).

The resulting output from RAT Auto Search is a JTextArea
report. If RAT finds an area of sequence that matches the
input parameters, it prints out a report on the sequence, allow-
ing the user to view the sequence involved in the Sequence
Viewer (but with only the contributing sequences checked and
displayed).

Testing and verification
Using KAL 153 (accession number AF193276), a known
recombinant HIV-1 strain from the region of the former Soviet
Union (Liitsola et al., 2000), a multiple sequence alignment of
27 similar HIV-1 strains was obtained by means of a BLAST
search (Altschul et al., 1990). First, the recombinant sequence,
KAL 153, was examined in the Single-sequence viewer. It was
evident that KAL 153 was a recombinant sequence between
strains 97BL006 (AF193275) and UKR1216 (AF193278).
97BL006 started off with <80% similarity to KAL153 and
then rose sharply to almost 95% similarity. Conversely,
UKR1216 started off more than 99% similar to KAL153, but
then dropped to 80% similarity at the same point.

In order to test the recombination search feature of RAT, the
alignment was searched for sequences that started at <82%
similarity and then jumped to >92% similarity. The resulting
Auto Search report for the 27 HIV sequences showed six ‘hits’
that involved two or more contributing sequences. Upon visual
inspection, two of the six hits showed clear signs of recom-
bination (i.e. two or more sequences showing recombination
crossover points). Of the two, one was the known recombin-
ant KAL 153 (Fig. 1), with the known contributing sequences
to the recombination event also being successfully identified
(97BL006 and UKR1216). The second sequence showing
clear evidence of recombination was isolated as 98BY10443
(AF414006), also from the former Soviet Union. The con-
tributing sequences for this strain were the same as for KAL
153, so 98BY10443 was presumed to be a close relative of
KAL 153. The remaining four ‘hits’ did not show a clear
crossover point and so were presumed to be due to random
sequence heterogeneity although further statistical analysis
may be worthwhile.

IMPLEMENTATION
All full-length Norovirus genomes, along with other long
stretches of Norovirus genomic DNA, were obtained from
GenBank, aligned using ClustalW (Thompson et al., 1994)
and edited to remove gaps. The Auto Search option on RAT
was used to examine recombination.

RESULTS
Recombinant Noroviruses and their contributing sequences
were identified as follows (work that found similar res-
ults is referenced): Arg320 (AF190817) between Mexico
virus (U22498) and Lordsdale virus (X86557) (Jiang et al.,
1999), Norovirus Mc37 (AY237415) between Saitama U1
(AB039775) or Gifu’96 (AB045603) in the ORF1 region
and an unknown virus in the ORF2 region (Hansman et al.,
2004), WUG1 (AY081723) between Southampton virus
(107418) and Norovirus BS5 (AF093797) (Katayama et al.,
2002), Norovirus MD 145-12 (AY032605) between Saita-
amu U1 (AB039775) or Gifu’96 (AB045603) in the ORF1
region and an unknown virus in the ORF2 region, Gifu’96
between Lordsdale virus and Hawaii virus (U07611) and
Snow Mountain virus (AY134748) between Lordsdale virus
and Melksham virus (X81879). The latter three recombinants
have not been reported before.

DISCUSSION
Our results demonstrate the utility of RAT in finding viral
recombinants and indicate that large-scale recombination
occurs in the Noroviruses. We also show a ‘hot spot’ for
recombination between the end of ORF 1 (nonstructural
polyprotein) and the start of ORF 2 (capsid protein).

When compared to other programs available, RAT is a much
easier and intuitive tool to use and allows the examination
of large amounts of data with the minimum amount of user
intervention. The lack of complex algorithms allows users to
speedily obtain a clear insight into the history of recombi-
nation by automatically searching through sequence data. RAT
is especially useful as an initial, fast, exploratory tool to form
the basis for a more detailed analysis. We anticipate that RAT
will be equally useful in revealing recombination events in
higher organisms as well as in viruses. When analyzing data,
it is possible that RAT will throw up false-positives, but these
can be quickly dispelled by visual inspection. When positive
The result of using the Auto Search option in RAT. The output for sequence KAL 153 is shown, along with its contributing sequences 97BL006 and UKR1216. The lines on the graph represent the genetic distance (y-axis) of each sequence checked in the sequence list pane (left). The x-axis represents the location on the sequence. Sequences in the sequence graph pane (right) that are not of interest can be unchecked in the sequence list pane and removed by clicking on the ‘Redraw’ button. Sequences can be recovered simply by rechecking them and clicking ‘Redraw’ again. The colour of the line on the graph is the same as the colour of its sequence name on the left.

Results are found, further work may be needed to find statistical support for recombination, and to this end, future versions of RAT including statistical tools are planned.

ACKNOWLEDGEMENTS

This work was supported by a grant from the Biotechnology and Biological Sciences Research Council.

REFERENCES


