RecDetec

Detecting Recombination and
Phylogenetic Information Along Alignments

RecDetec Manual
preliminary Version (April 2013)

Heiko A. Schmidt
Center for Integrative Bioinformatics Vienna (CIBIV),
Max F. Perutz Laboratories (MFPL)
A-1030 Vienna, Austria.
email: `heiko.schmidt @ univie.ac.at`

General Information

RecDetec is a computer program to analyze alignments of, typically viral, genomes to analyze the phylogenetic information content and to visualize traces of recombination applying maximum likelihood phylogenetics and bootscanning.

RecDetec is available free of charge from

- `http://www.cibiv.at/software`

RecDetec is written in Java. Thus, it will run on most personal computers and workstations if Java is installed. The package contains precompiled executables for tree reconstruction for the most prominent operating system, namely Linux, MacOSX and Windows.

The software tools included are the ML programs TREE-PUZZLE (Strimmer and von Haeseler, 1996; Schmidt *et al.*, 2002) and IQPNNI (Vinh and von Haeseler, 2004; Minh *et al.*, 2005a) for tree building and split operations. Furthermore, RecDetec uses Figtree for visualizing trees (`http://tree.bio.ed.ac.uk/software/figtree/`).

Contents

Chapter 1

Legal Stuff

Chapter 2

Installation

The content of the distributions do not differ between the operating systems. However, installation procedures differ slightly between UNIX-like systems (MacOSX and Linux) and Windows. These are outlined below.

There are two distributions which only differ by their packaging software, `recdetec-VERSION.tar.gz` (for Linux or MacOSX) and `recdetec-VERSION.zip` (for Windows or MacOSX). If you know how to handle ZIP or TAR files with your operating system, you can download the one you prefer.

2.1 Java

To run RecDetec you need Java installed. In most cases it already comes with your operating system. In all other cases you need to install it.

There are more than one Java distribution available, but the presumably most common one is available from `http://www.java.com`. Follow the installation instructions which come with the Java package or ask your local IT expert for help.

2.2 Linux/MacOSX Distribution

Get the file `recdetec-VERSION.tar.gz` from `http://www.cibiv.at/software/recdetec`

Extract it (using the `tar xvzf recdetec-VERSION.tar.gz` command or another program capable of handling `tar.gz` format.

After extracting this file (if you have problems, please ask your local IT expert), you will find a folder called `recdetec-VERSION` on your hard disk. Move the folder to the location where you want to have RedDetec installed, e.g. in your home directory.

The folder `recdetec-VERSION` contains a run-script called `recdetec.sh`. Edit `recdetec.sh` to change the content of the variable `RECDETEC_EXE_DIR` to contain the correct location of your `recdetec-VERSION` folder, e.g.

SET RECDETEC_EXE_DIR= /recdetec-VERSION

To run RecDetec you just double-click `recdetec.sh` or run it from the command line in a terminal.

To make `recdetec.sh` more accessible you may want to create a link on the Desktop or copy it into your `~/bin` or `/usr/local/bin`. (If you have problems, please ask your local IT expert.)

2.2.1 Windows Distribution

Get the file `recdetec-VERSION.zip` from `http://www.cibiv.at/software/recdetec` (where VERSION is the current version number). After un-zipping this file (if you have problems, please ask your local Windows expert), you will find a folder called `recdetec-VERSION` on your hard disk. Move the folder to the location where you want to have RedDetec installed, e.g. on your Desktop.

The folder `recdetec-VERSION` contains a run-script called `recdetec.bat`. Edit `recdetec.bat`, e.g., with the Windows program `textedit` or `notepad` and change the content of the variable `RECDETEC_EXE_DIR` to contain the correct location of your `recdetec-VERSION` folder, e.g.

SET RECDETEC_EXE_DIR=C:/Documents and Settings/USERNAME/Desktop/recdetec-VERSION

To run RecDetec you just double-click `recdetec.bat`.

To make `recdetec.bat` more accessible you may want to create a shortcut on the Desktop or copy it there (especially if you did not install the folder on the Desktop).

Chapter 3

Introduction

RecDetec is a software to perform bootscan analyses (Salminen *et al.*, 1995) using ML phylogeny reconstruction (Vinh and von Haeseler, 2004; Minh *et al.*, 2005b) to obtain the bootstrap (Felsenstein, 1985) support values along an alignment of aligned genomes. RecDetec can also use TREE-PUZZLE (Schmidt *et al.*, 2002) to obtain support values (QP-scan) which is typically faster than a full ML bootstrap analysis.

Bootscanning computes support values in windows along a multiple sequence alignment whaich are then used to visualize possible recombinations by changing clustering patterns of reference and query sequences.

RecDetec can assess groups of sequences whether they are well supported throughout the alignment or whether there are regions where the support is lost or if a group is not supported at all.

Another important feature of RecDetec is that it allows for assessing the phylogenetic information along an alignment using the fractions of informative, partly and non-informative sites (e.g. Swofford *et al.*, 1996) as well as the phylogenetic fractions of resolved, partly and unresolved quartets Strimmer and von Haeseler (1997). This way one can easily detect genomic regions which are too variable or to conserved to be suitable for pylogenetic analysis.

Chapter 4

Analysis

4.0.2 Start RecDetec

To run RecDetec, start the run-script either from the command-line (Linux, MacOSX) or by double-clicking (Windows, but also possible on Linux or MacOSX). The RecDetec graphical user interface (GUI) should appear on you screen.

4.0.3 Load Alignment

Load an alignment by clicking the button in the `Load Project/Alignment File` field or use the `File > Load Alignment File` menu. Several alignment formats are possible like Phylip or FASTA format.

The names of the loaded sequnces are now shown in the Groups field of the GUI.

We will use the alignment `example/example.phy` from the recdetec-VERSIOn folder as an example below.

4.0.4 Define Sequence Groups

To get meaningful results one typically has to group the sequences. Such groups should have the same evolutionary background, that means, they should go back to a common ancestor which could be a recombinant or a pure (i.e. unrecombined) form.

To group sequences click on the name and a `Edit Group` window will open. In that window you can change the name of the group, add sequences (`>` button) or remove them from the group (`<` button). To finish a group press OK. Groups with all sequences removed will disappear. The `Redistribute colors` button will assign each group a unique color. With the button in front af the name one can assign a color manually.

By un-ticking a sequence in the Groups field it will be discarded completely from the analysis.

For the example, group sequnces A1/A1 as group A, B1/B2 as B, C1/C2 as C and D1/D2 as D. R remains separate.

4.0.5 Perform ML Bootscan

To draw ML bootscan plots, first a ML bootstrap reconstruction has to be run. This is indicated by a grayed-out `ML bootscan` button. Choose ML bootscan as the method and press `START`. I progress bar will indicate the progress of the reconstruction.

If reconstruction results have already obtained, the according visualization buttons will be shown in green.

Press the `ML bootscan` button. Next you will be asked which which of the groups should be your query group. For the example choose R and press OK. Next you can choose sequences to be ignored in the bootscan plot. This way interfering recombinants (and their signal) can be ignored when analysing another recombinant.

The resulting plot will show a clear signal that the beginning and the end of sequence R is more related to group A and the middle is more related to group B. The curves of different color do represent support for relationship of the query group to the according reference groups (see legend next to the plot). If the curves change for different regions there might have been a recombination. However, not every change needs to be caused by recombination. The following points should be checked. The support should be high and at least shoud increase above 65%. A recombinant region should be long enough – e.g. one or two single windows showing a change are usually caused by noise (see below for more information). The recombinant regions should exhibit phylogenetic information, if region are too conserved or to diverged the signal might be caused by chance. The query and reference groups should be well supported on their own – if a group is not well supported, the signal might be caused by chance due to lack of signal (see below for more information).

The data of this plot can be saved a CSV file which can be used in Excel or R to refine the diagram or to combine several plots. Furthermore, the diagrams can be saved as encapsulated Postscript (EPS).

4.0.6 Perform QP-scan

To draw QP-scan plots, first a phylogenetic reconstruction with TREE-PUZZLE has to be run. This is indicated by a grayed-out `QP-scan` button. Choose QP scan as the method and press `START`. I progress bar will indicate the progress of the reconstruction.

If reconstruction results have already obtained, the according visualization buttons will be shown in green.

Press the `QP scan` button. Next you will be asked which which of the groups should be your query group. For the example choose R and press OK. Next you can choose sequences to be ignored in the bootscan plot. This way interfering

recombinants (and their signal) can be ignored when analysing another recombinant.

The resulting plot will show a clear signal that the beginning and the end of sequence R is more related to group A and the middle is more related to group B. The curves of different color do represent support for relationship of the query group to the according reference groups (see legend next to the plot). If the curves change for different regions there might have been a recombination. However, not every change needs to be caused by recombination. The following points should be checked. The support should be high and at least shoud increase above 65%. A recombinant region should be long enough – e.g. one or two single windows showing a change are usually caused by noise (see below for more information). The recombinant regions should exhibit phylogenetic information, if region are too conserved or to diverged the signal might be caused by chance. The query and reference groups should be well supported on their own – if a group is not well supported, the signal might be caused by chance due to lack of signal (see below for more information).

The data of this plot can be saved a CSV file which can be used in Excel or R to refine the diagram or to combine several plots. Furthermore, the diagrams can be saved as encapsulated Postscript (EPS).

4.0.7 Plot group support

For bootscan analysis to work it is crucial that the groups reflect sequences of the same evolutionary context. Thus, such groups should be supported by the underlying data.

The support values obtained for the clustering of the members of a group can be used to assess how stable a group is along the alignment. To visualize the phylogenetic stability of a group press one of the `Group Support` buttons. They will visualize the ML bootstrap support (`ML`) or Quartet Puzzling support (`QP`) values, respectively. If the button is grayed, then the respective phylogenetic reconstriction has not been run yet.

As an illustrative example, please group sequences A1/A2/R in one group. (If you have run the reconstructions already, you do not have to re-do them. Otherwise run them now by choosing the Method and click `START`.)

After choosing a group support button, one has to choose the group one wants to assess. Afterwards, one has the choice to exclude certain sequences, which is importand to exclude interfering signal by recombinant sequences or groups. If the support curve does show a drop like the one analyzing the A1/A2/R group, highlights that there is something wrong with this user-defined group which has to be examined further: There might be recombinant sequence for which a member of this group has acted as parent, the group could contain sequences of a different evolutionary background (e.g., like here, a recombinant R and two pure strains A1/A2), but also a lack of phylogenetic information could be possible. The scenarios can be elucidated using bootscan or QP-scan plots, viewing the phylogenetic trees reconstructed along a alignment (press the right

mouse button in the region of interest in the plot), or using the `Informative Sites/Likelihood mapping` plots.

One can also plot the supports of all groups at once. However, one cannot exclude putative or known recombinants this way.

4.0.8 Analyze Phylogenetic Signal

Using the `Informative Sites` button one can visualize the fractions of (parsimony) informative sites along the alignment. A parsimony informative site is an alignment column that contains at least two different nucleotides and at least two of the nucleotides occur at least twice (Swofford *et al.*, 1996). We call a parsimony informative site partly informative, if some nucleotides nucleotides occur only once, or if gaps or other ambiguous characters (like N) occur at that site. All other alignment columns are (parsimony) uninformative. Among these, two types of constant sites are defined. Completely constants sites contain only one nucleotide in all sequences, while constant sites can contain gaps or ambiguous characters (like N) besides one single nucleotide.

Using the `Likelihood mapping` button visualizes the amounts of resolved, partly, and unresolved quartet trees as resulting from a likelihood mapping analysis (Strimmer *et al.*, 1997).

High amounts of informative sites and resolved quartets point to genomic regions well suited for phylogenetic analysis.

If the fraction of partly and unresolved quartets are high, the phylogenetic information is low. This is typically caused by too similar sequences (indicated by low amounts of informative sites) or too divergent sequences suffering from saturation effects (high amounts of informative sites).

4.0.9 Export diagram data

The data of this plot can be saved a CSV file which can be used in Excel or R to refine the diagram or to combine several plots. Furthermore, the diagrams can be saved as encapsulated Postscript (EPS).

Chapter 5

Other Features

5.1 Prepare to run phylogenetics reconstructions on a cluster

RecDetec can also be run from the commandline, e.g. to produce all necessary files to distribute the phylogenetic reconstruction on a compute cluster or on multi-core PCs. The commandline flag `-cluster` causes RecDetec to write the files to a folder `recdetec_cluster_DATE_TIME` (where `DATE` and `TIME` are the current date and time.

The necessary scripts to run the analysis are found in the subdirectory `scripts` of the `recdetec_cluster_DATE_TIME` folder.

To re-import the reconstruction into RecDetec one has to type `recdetec.sh -cluster_info recdetec.info` where `recdetec.info` is the file containing all necessary information, which is produced with the other files for a cluster run.

5.2 Other Software needed

To run RecDetec a current Java distribution has to be installed. Java is available from `http://www.java.com`.

Chapter 6

RecDetec References and Further Reading

A manuscript about the software is currently in prepararion.

Chapter 7

Acknowledgments and Credits

RecDetec includes the following software: TREE-PUZZLE 5.3.rc, IQPNNI 3, and FigTree.

Finally we thank the Austrian Science Fund (FWF) for financial support.

Bibliography

Felsenstein, J. (1985) Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, **39**, 783–791.

Minh, B. Q., Vinh, L. S., von Haeseler, A. and Schmidt, H. A. (2005a) pIQPNNI – parallel reconstruction of large maximum likelihood phylogenies. *Bioinformatics*, **21**, 3794–3796.

Minh, B. Q., Vinh, L. S., Schmidt, H. A. and von Haeseler, A. (2005b) IQPNNI: Parallelization and improvements. In *Proceedings of the 1st Young Vietnamese Scientists Meeting (YVSM 2005)*, Nha Trang, Vietnam.

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.

Schmidt, H. A., Strimmer, K., Vingron, M. and von Haeseler, A. (2002) TREE-PUZZLE: Maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, **18**, 502–504.

Strimmer, K., Goldman, N. and von Haeseler, A. (1997) Bayesian probabilities and quartet puzzling. *Mol. Biol. Evol.*, **14**, 210–213.

Strimmer, K. and von Haeseler, A. (1996) Quartet puzzling: A quartet maximum–likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.*, **13**, 964–969.

Strimmer, K. and von Haeseler, A. (1997) Likelihood–mapping: A simple method to visualize phylogenetic content of a sequence alignment. *Proc. Natl. Acad. Sci. USA*, **94**, 6815–6819.

Swofford, D. L., Olsen, G. J., Waddell, P. J. and Hillis, D. M. (1996) Phylogeny reconstruction. In Hillis, D. M., Moritz, C. and Mable, B. K. (eds.), *Molecular Systematics*, 2nd edition, pp. 407–514, Sinauer Associates, Sunderland, Massachusetts.

Vinh, L. S. and von Haeseler, A. (2004) IQPNNI: Moving fast through tree space and stopping in time. *Mol. Biol. Evol.*, **21**, 1565–1571.