Introductory Workshop
to ML Tree Reconstruction
Exercises

April 4, 2008

1 Introduction

In this section we will use a dataset of 14 HIV sequences to reconstruct phylogenetic trees of the HIV subtypes A (aU09127), B (aL02317, aAF025763, aU08443, aAF042106) C (cAF067158, cU09126), D (dU27399, dU43386), G (gU27426, gU27445), and some outgroup sequences (HIV1groupO, SIVcpzUS, SIVcpzGAB).

The alignment (`hivALN.phy`) has a length of 2352 bp.

2 Prerequisites

We will use an adopted LiveDVD based on the KNOPPIX Linux DVD version 5.3. First, switch on the PC and immediately put the DVD into the DVD drive o the the PC can boot Linux.

When the boot process is finished you are sitting in front of a typical X-Windows based user interface often used on Unix and Linux machines. It does in principal react like an MS Windows system, but programs are typically called differently.

Furthermore, the Unix/Linux does offer a powerful environment to write so-called shell scripts to automate workflows where for instance many files have to be processed.

2.1 The programs

We will used the following programs for the analysis:

- IQPNNI, a maximum likelihood tree reconstruction program (`http://www.cibiv.at/software/iqpnni/`)
- TREE-PUZZLE, a maximum likelihood tree reconstruction program (`http://www.tree-puzzle.de`)
- FigTree, a tree viewer (`http://tree.bio.ed.ac.uk/software/figtree`)
- the PHYLIP Package, a PHYLogenetic Inference Package (`http://evolution.genetics.washington.edu/phylip.html`)

From the PHYLIP package we will use the programs `seqboot` and `consense`. Furthermore, we want to use the following UNIX/BASH commands:

- `for-do-done` loops (to process a list of files),
- `ls` (to get the contents of the current directory),
- `less` (to view the contents of a text file page-wise),
- `cd` (to change the directory),
- `cp` (to copy files),
- `mkdir` (create a new directory),
- `cat` (to write the content of a file),
- `grep` (to find words in files),
- `man` (show a command's manual if available),
- `split` (to split file with the bootstrap samples into separate files),
- piping `>` (redirecting the output of a program into a file),
- `|` (redirecting the output of a program as input of another one),
- etc.

Please open a Terminal by clicking on the 'monitor' on the task bar of your Desktop. Try a few of the commands above.

If one wants to know something about Unix commands one can often use the command `man comandname`, where `comandname` is the name of the command you need information for. `man` will print a (short) manual if available.

2.2 Your desktop

On your desktop you will find a number of programs, some which are mentioned above (like FigTree) but also two directories `workshop-data` and `workshop-documents` where the former contains the necessary alignments and the latter some interesting documents and papers.

Furthermore, there are two web-links to CIBIV and to a folder where you can find these exercises as well as handouts. I might put other information after the workshop if requested.

2.3 Your desktop

Before you can start open a terminal (if you haven't done so above) it will start in your home directory (`/home/knoppix`). Change into the data directory with '`cd Desktop/workshop-data`'.

Us `ls` to find out what file are there, a look into the `hivALN.phy` with '`less hivALN.phy`' (you can quit with 'q').

3 Phylogenetic signal

Although this might be an important test for your dataset with should be done prior to the actual analysis, you might postpone it in this workshop.

3.1 Likelihood mapping

Perform a likelihood mapping plot with the TN93 model using the `puzzle` program. Start the Program with `puzzle -wtstv hivALN.phy`. You will get a menu. Change the type of analysis to `likelihood mapping` and adjust the model of evolution.

How does the result differ when you also test `banana.phy`?

How does the amount of unresolved quartets change, if one changes the the complexity of the evolutionary model by using uniform, Γ-distributed rates, or mixed rates?

3.2 Transition:transversion saturation plot

Obviously, transition:transversion plots can only applied to DNA data. Use the HIV example file (`hivALN.phy`).

To produce a file with the necessary values start `puzzle` with the `-wtstv` commandline option, which causes `puzzle` to write a `hivALN.phy.tstv` file.

Its contents can be plotted with `R` using

```
tstvtab = read.table("ali.phy.tstv", header=T)  # read data
attach(tstvtab)                                 # use headers as names
pdf(file="tstv.pdf")                            # open PDF file
maxsubst=max(ts,tv)                             # find maximum
plot(distance,ts,col=2,ylab="observed substitutions",ylim=c(0,maxsubst))
points(distance,tv,col=3)                       # plot
dev.off()                                       # close PDF file
detach(tstvtab)                                 # release names
q()                                             # quit R program
```

Just start `R`, the you type in the above commands, but use `hivALN.phy.tstv` as data file.

What does the saturation plot show for this dataset?

4 Phylogenetic ML tree with IQPNNI

Reconstruct a maximum likelihood tree for the `hivALN.phy` dataset using the `iqpnni` program ('`iqpnni hivALN.phy`').

Set the model of evolution to TN93 and start the analysis with `y`. (Open a new file called `params` with a graphical editor (e.g., KWrite) list all parameters you typed to run `iqpnni` into the file named, including all `enter`-strokes and the '`y`' and 'enter' at the end. Save the file.

With such a parameter file, you can easily re-run an analysis with:

`iqpnni` *`hivALN.phy`* `< params`

which we will do later.

Visualize the tree, which can be found in a `*.treefile` file, with the program `FigTree`.

5 Bootstrap

Just reconstructing one ML tree, usually doesn't tell us anything about, whether and which branches might be reliable. A common tool to obtain support values is bootstrapping (and also Bayesian analysis with MCMC or sometimes quartet puzzling with TREE-PUZZLE). Here we will perform a bootstrap:

Hint: due to lack of time,do 10-20 bootstraps only!!!

- Generate the (at least) 100 bootstrap samples with the program `seqboot`. Just start `seqboot` and the program will ask you for the alignment input file. Enter `hivALN.phy`, start with `y`, and enter an odd random number

seed (just a random number you think of). All 100 bootstrap samples are written to a single file `outfile`.

- Use a graphical editor to find out how many lines one a single bootstrap sample has. (You might also use the command '`grep -n 2352 hivALN.phy`' - what does it do?)

  Split the `outfile` into file containing only one sample alignment, each (e.g. with '`split --suffix-length=3 --numeric-suffixes --lines=NUMBER outfile hivALNboot.`' - NUMBER has to be substituted with the number lines determined above. try to find out with '`man grep`' what the command does.)

- Use a little shell programming (loops) to run `iqpnni` on each of the 100 bootstrap alignment using the `params` file from above:

  - `for i in hivALNboot.0??; do`
  - `echo $i`
  - `cat param | iqpnni $i`
  - `done`

  What does this little program do?

- Collect the 100 trees from the tree files in one single trees-file (e.g. with '`cat hivALNboot.0??.iqpnni.treefile > hivALNboot.trees`'). Use `puzzle` with the trees-file and the original alignment ('`puzzle hivALN.phy hivALNboot.trees`').

- Visualize the bootstrap tree in `hivALNboot.trees.tree` with the program `FigTree`. Is the tree fully resolved?

6 TREE-PUZZLE analysis

A faster method to construct support values is to use TREE-PUZZLE. The support values behave similar but they are not the same and should not be mistaken as bootstrap values. Note, that bootstrap values are much better understood.

Run the command '`puzzle hivALN.phy`' to start. Change the model of evolution to TN93. And start the analysis with '`y`'.

After program has finished, view the tree in `hivALN.phy.tree` with FigTree. How does the tree compare with the bootstrap tree?

Scan also through the puzzle report file `hivALN.phy.puzzle`, e.g. with `less` or KWrite. The file will tell you a lot of information about your dataset.

Are there identical sequences in the dataset? Are there sequences that have a high amount of unresolved quartets (might indicate that something is wrong with the alignment or the sequence ist too similar or too divergent)? How many constant and/or variable sites are in the alignment?

7 Further comments and information

If you haven't finished, you just might try this exercise at home. In the `workshop-documents` folder there are also 2 documents including each theoretical information and a hands-on session, one solely on TREE-PUZZLE, the other on ML analyses using IQPNNI and TREE-PUZZLE.