Bioinformatics, computer lab VII
University of Potsdam WS 2019/2020, S. Hartmann

Phylogenetic reconstruction

The objective of today's computer exercise is to be able to compute a phylogenetic tree using the maximum parsimony method, and to evaluate and interpret the result.

The data

Today you will use a parsimony approach to compute a phylogeny from a multiple sequence alignment. Because parsimony may result in incorrect phylogenies when highly divergent sequences are used, only the five most conserved sequences of last week's alignment will be used for today's exercise. The data has been converted into phylip format, which is required for the software you'll use today.

☞ Why may parsimony result in incorrect phylogenies when highly divergent sequences are used?

☞ Download the file `globins.phy` from Moodle. This file contains five globin sequences. Look at it with `less` in a Linux terminal or with a text-only editor. Describe the phylip format.

1 Computing the phylogeny

You will use the program `protpars` from the Phylip package to

- compute the most parsimonious phylogeny for the input alignment
- print the mutational steps required for each alignment column (according to the most parsimonious tree)
- print the inferred sequence for each internal node (according to the most parsimonious tree for each alignment column)

☞ Using `cd` in the terminal, move to the directory in which the file "globins.phy" is. Then start the protpars program by typing `phylip protpars`. Next, you will be asked for the name of the input file; enter "globins.phy".

☞ You will then see a text-based menu showing the different parameters for the parsimony analysis. Make the following changes to the default settings:

- type the letter "O" (followed by a return) to indicate that the phylogeny should be outgroup-rooted. In the alignment, the fifth sequence is the one from lamprey (a fish) that will serve as an outgroup: type "5".
- type the number "1" to see the alignment at the top of the output file
- type the number "4" to see the mutational steps for the alignment columns in the output file
- type the number "5" to see the inferred ancestral sequence at internal nodes in the output file
- type a period "." to turn off the "dot-display" mode for the sequences (which makes the results much more readable)
- type the number "6" so that no additional file with best tree will be generated.

When you have made all these changes to the default settings, the menu on your screen should look like this:

```
Setting for this run:
 U                 Search for best tree?  Yes
 J   Randomize input order of sequences?  No. Use input order
 O                        Outgroup root?  Yes, at sequence number 5
 T              Use Threshold parsimony?  No, use ordinary parsimony
 C               Use which genetic code?  Universal
 W                       Sites weighted?  No
 M           Analyze multiple data sets?  No
 I          Input sequences interleaved?  Yes
 0   Terminal type (IBM PC, ANSI, none)?  ANSI
 1    Print out the data at start of run  Yes
 2  Print indications of progress of run  Yes
 3                        Print out tree  Yes
 4         Print out steps in each site  Yes
 5  Print sequences at all nodes of tree  Yes
 .  Use dot-differencing to display them  No
 6       Write out trees onto tree file?  No

Are these settings correct? (type Y or the letter for one to change)
```

☞ Now type "Y" to start the computation. When protpars is done with the analysis, it tells you that the file 'outfile' was generated, and then it exits.

☞ The name 'outfile' is a little too generic, so please rename it to something more meaningful.

2 Analyzing and interpreting the output

The alignment

☞ The first section of the output file contains the alignment. For better readability, alignment columns are shown in blocks of ten, separated by spaces. As shown here, the dots represent conservation with the sequence in the first row: in the first position, the sequence 'human_b' has a gap, and the dots in the other human and horse sequences indicate that these also contain a gap there. The lamprey has an "M" in the first alignment position.

The inferred tree

The next section shows the most parsimonious phylogeny that was inferred. The sequence names are shown at the tips of the tree, and the internal nodes are given IDs by which we can identify them.

NOTE: underdeath the tree it says "remember: (although rooted by outgroup) this is an unrooted tree!". Phylip has many quirks, and this is one of them: a rooted tree was computed, but an unrooted tree is displayed. Just imagine some dashes in front of the 4, and all is well.

☞ Summarize the difference between a *species tree* and a *gene tree*. Which of these is shown here?

☞ Describe the evolution of the five globin gene sequences based on the phylogeny you computed.

Mutational steps

The next section shows the number of mutational steps that are required for each alignment column. For example, the most parsimonious tree shown above would require 2 mutational steps for the 10^{th} alignment column, 4 for the 11^{th}, and 1 for the 20^{th} column. Make sure you understand how to read this table.

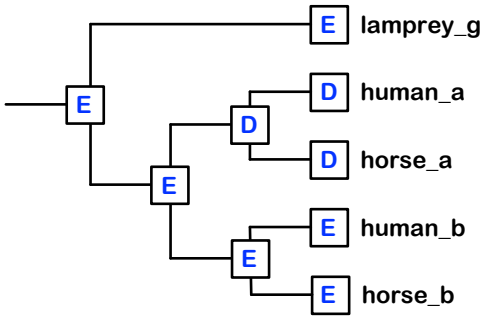
☞ How many mutational steps does the most parsimonious tree require for columns 16, 23, and 43 in the alignment?

Ancestral states

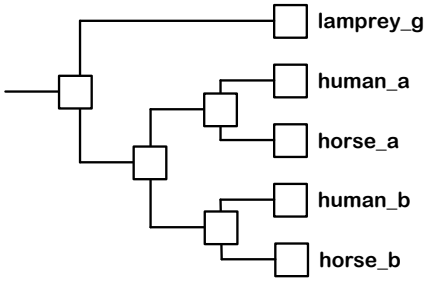
The last section of the output file lists the alignment of the input sequences, plus the inferred ancestral sequences for each of the internal nodes in the tree. The node IDs are the same as in the most parsimonious tree shown above. The character-state at each node (as listed in the second column, the one containing internal node IDs and sequence names) is shown for each of the alignment columns, making it easy to see changes from one node to the next.

The diagram shown below illustrates this for the 16^{th} column. Make sure you understand how to read and interpret this diagram, and how the information in the output was used to generate it.

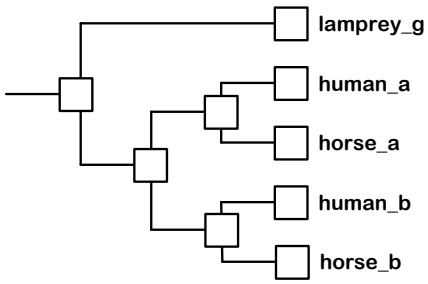
☞ Look up how many mutational steps are required for this column. Given the diagram – does that make sense, and where on the tree did the mutation probably occur?



☞ Complete the following diagram for columns 23 and 43: which amino acids are found in the globin sequences, and which are inferred to be the ancestral nucleotides at the internal nodes?



sequenced and ancestral states
of alignment column 23



sequenced and ancestral states
of alignment column 43

☞ Compare this with the number of mutational steps that are required for these two alignment columns (section 'Mutational steps'). Can you explain the result? (Hint: do mutations occur at the level of amino acids or at the level of nucleotides?)