DistR Documentation

By: Rachel Bevan
e-mail: rachel@mcb.mcgill.ca

November 10, 2005

Introduction

DistR uses the DistR estimation method to compute the evolutionary rates of different loci based upon two sets of input: trees/distances between species for the loci of interest; multiple sequence alignments of the loci of interest. Rates can be computed for both nucleotide and amino acid data. For a detailed explanation of the algorithm and applications please see Bevan, R., Lang, B.F., and Bryant D. (2005) Calculating the Evolutionary Rates of Different Genes: A Fast, Accurate Estimator with Applications to Maximum Likelihood Phylogenetic Analysis. Systematic Biology. 54(6):900-915.

Implementation

DistR is a command-line controlled program written in C. It should compile easily on any UNIX/Linux workstation or Mac machine. The source files in the directory 'src' are: main.cpp, distR.cpp, matrixOperations.cpp, fileio.cpp, distanceList.cpp, wrapper.cpp, phylogeny.cpp, global.cpp, bit_set.cpp, simple_nexus.cpp. The include files in the directory 'include' are: bit_set.h,treerates.h, matrixOperations.h, distanceList.h, fileio.h, global.h, phylogeny.h,simple_nexus.h, tnt_array1d.h, tnt_array1d_utils.h, tnt_array2d.h, tnt_array2d_utils.h, tnt.h, tnt_i_refvec.h, tnt_math_utils.h, jama_qr.h, tnt_version.h. All of the .h files beginning with tnt are part of the Template Numerical Toolkit, which can be downloaded fully from http://math.nist.gov/tnt/download.html. The jama_qr.h file (which depends upon the tnt files) is part of the JAMA/C++ Linear Algebra Package, which can be downloaded from the same website.

Compiling on UNIX/Linux/Mac

To compile DistR, start in the main directory and type make. This will make the program, and the executable will be placed in the directory 'bin'. The default compiler is c++, with flags for debugging (-g -Wall) and deprecation of code checks.

Running the program

Typing "distR -h" on the command line will give the following help menu for the commands.

```
Options:
     -h        This help screen
     -t File   Specify file which contains a list of tree files in NEWICK format
               Must use either this option, or the -d, or -b options to specify
               trees

     -d File   Specify file which contains a list of distance matrix files in
               NEXUS format
               Must use either this option, or the -t, or -b options to specify
               trees.

     -p File   Specify file which contains a list of alignment files in PHYLIP
               format - both interleaved and sequential are acceptable
               Must use either this option, or the -n, or -b options to specify
               alignments.

     -n File   Specify file which contains a list of alignment files in NEXUS format
               Must use either this option, or the -p, or -b options to specify
               alignments.

               Each line in the list of alignments file, contains the file name
               of the alignment corresponding to the tree/distance matrix in the
               appropriate file
     -b File   Specify file which contains a list of NEXUS files that have both
               a distance matrix and an alignment
               This option allows for specification of both tree distances and
               alignment in same file.  Note:  I have removed this option due to
               bugs in the code with more complex nexus files.  If you wish to
               use this option simply uncomment the code in the switch statement
               under option b, and uncomment the second myGetOpt command.
```

```
Please note that if no alignment files are provided the program defaults to
PHYLIP formatted alignment files which the file extension (i.e. the last '.*')
of the tree/distance files specified by -t/-d option are changed to .phy

Also note that the user MUST specify either a list of trees in newick format
with the -t option OR a list of distance matrices in NEXUS format with the -d
option
```

Input and Output

Sequence files can be in phylip, fasta or simple nexus format. Distance files can be in either newick or nexus format. Please note that only simply nexus format will work properly, such as the examples given in the bin/ directory.

For example, suppose there are three protein multiple sequence alignments in three separate files called protein1.phy, protein2.phy and protein3.phy. The format of the alignment is phylip. Suppose that three trees have been estimated based upon these sequences, in files tree1.newick, tree2.newick and tree3.newick. To run DistR it is necessary to create a file listing the names of the protein files, and a file listing the names of the tree files. It is important that corresponding tree/alignment files be listed on the same line for the program to work.

i.e. Mytree.infile:
tree1.newick
tree2.newick
tree3.newick

Myalignment.infile:
protein1.phy
protein2.phy
protein3.phy

Given the above two input files, the program will run as follows:
distR -t Mytree.infile -p Myalignment.infile

Another possibility is to specify just the list of trees. If you have the corresponding proteins in phylip format, specified by a '.phy' extension, then the program can determine the names of the protein files based on the names of the tree files. This will only work if both the tree files and the protein files have the same start (before the first '.').

i.e.
Mytree.infile2
atp6.phy-gb_phyml_tree
atp8.phy-gb_phyml_tree
atp9.phy-gb_phyml_tree

Given the protein files atp6.phy, atp8.phy and atp9.phy the command:

```
distR -t Mytree.infile2
```

will give the protein rates.

If the distance information is in nexus format, a file of nexus files can be specified:

i.e.
Mydistancemat.infile
atp6.phy-gb_distmat
atp8.phy-gb_distmat
atp9.phy-gb_distmat

and the following command used: distR -d Mydistancemat.infile

The program automatically distR an output file called ‘tree.rates’ which lists the name of the protein followed by it’s evolutionary rate. The distances are automatically output in nexus format to a file called ‘distances.nexus’. A file called ‘distance.counts’ is also output, giving summary statistics on the number of missing distances and the number of estimated distances.