Integrative Biology 200A *University of California, Berkeley*
PRINCIPLES OF PHYLOGENETICS *Spring 2010*
Revised by Nick Matzke

Lab 4
Parsimony tree estimation with TNT

TNT *(Tree analysis using New Technology; Goloboff, Farris and Nixon 2000) [http://www.cladistics.com] is a program available for Windows, MacOS or Linux. It has very efficient tree-searching algorithms for large data sets of 300 to 500 tax a. Parsimony is the only available optimality criterion. It implements many new heuristic search methods, such as the ratchet and sectorial searches. It can also be used for tree manipulation and diagnosis. As it is optimized for large matrices it is probably not the best program to use for data sets with fewer taxa.*

Setup: Download and install TNT (google e.g. "TNT cladistics" to find it on the web)

The Parsimony Ratchet

Most real data matrices have too many taxa (i.e. more than about 25 taxa) to be analyzed by exact methods therefore a search for the most parsimonious trees must be conducted. In many cases the shortest trees—or more precisely the trees that we think are the shortest—are easily located. In other instances the shortest trees are difficult to locate. It is not possible to predict, from the matrix, the ease in which the shortest trees will be found (if they ever will), or ascertain that one has found the shortest trees. The only criterion that can be used is reproducibility: if numerous searches, with different search parameters, of the matrix produce the same result(s) then one must assume that the shortest trees—or at least the shortest trees that will ever be found—have been located. Effective search parameter must be determined empirically.

In a "conventional" search a Wagner tree (or some other starting tree) is calculated and then a branch swapping algorithm (of some kind) is applied to the tree. Usually multiple starting points are utilized to minimize the possibility of becoming stuck in local optimal (or sub–optimal) portions of "tree space". The search stops when all the trees retained in memory have been swapped and no shorter trees have been found. Given a finite amount of time, the best way to maximize the exploration of tree space is to limit the number of trees retained during branch swapping. In most cases only the shortest trees found during the first phase are swapped, but in some cases some percentage of the shortest length trees are swapped.

Nixon (1999) proposed a new tree search method called the parsimony Ratchet (Nixon 1999). The ratchet can be viewed as the application of a Markov Chain to tree search. The ratchet procedure starts by searching for the best tree. Then it resamples the data with replacement or jacknifes it and randomly constrains some nodes. It searches tree space again with the newly defined parameters. It returns to the original settings and repeats the whole process multiple times. By reweighting the characters the ratchet produces a more radical search

of tree space, which is still constrained by the data.

TNT

TNT can do a number of different heuristic searches in addition to the standard ones included in most phylogeny packages. The more advanced searches are included under new technology searches, and can be used alone or in combination.

Sectorial- Explores rearrangements of local clades while leaving the rest of the tree unmodified. It does this successively for different clades chosen at random.

Ratchet- The same as described above

Drifting- Like the ratchet it alternates between normal searches and more liberal ones. Instead of reweighting the characters during the liberal searches it accepts new trees based on the fit between the new tree and those already in memory.

Tree Fusion- Mixes trees that are already in memory making new synergistic trees. If the trees come from different searches then the scores can be improved quickly and drastically.

Setup:

1. Using your skills at google and innate intelligence, find the TNT website and download and unzip TNT.
    a. Macs: download either "Mac32" or "Mac64" (no limit to number of taxa) – either should probably work.
    b. Windows: download "Win (no taxon limit, bin only)" – I think. There is also a menu-based Windows version which you can play with if you like.
2. Put the unzipped TNT folder in a place you will find it.
    a. TNT is one of those programs where it easiest if you keep the program and the data file inputs and outputs in the same directory.
3. Google to find the TNT Wiki, and keep that page open for future reference

Some wisdom on using specialist scientific applications:

Often, when you are trying to use some new random program some scientists have written to do some analysis, complications arise. It is good to have in mind what the typical challenges are, and in what order they need to be dealt with:

1. First you have to find/download/install the program. Hopefully the program has an executable or zipfile appropriate for your system, which you can just download and install with minimal trouble. Good programs will have excecutables available for Windows, Mac 10.4, 10.5, etc.

- But sometimes, all you will have is raw code which you have to compile yourself. Sometimes, compiling is easy, sometimes hard, sometimes impossible. Note that scientists are not professional software developers, and do not have software design teams, error checkers, etc. Often they don't even have much formal training in programming!

2. Programs with nice-ish menus, buttons to click, etc., are rare. Users who are beginning to use programs like the menus, etc., since they are easier to figure out initially. However, in the long run, menu-based programs aren't useful for tasks that are more complex than a once-off analysis. Much serious phylogenetics work requires processing a large collection of trees, or conducting the same analysis while varying a bunch of different options, e.g. to assess how confident you are that your result is independent of the specific choices you made.

Command-line programs are (a) much easier for scientists to write, and (b) are much easier for users to automate through scripts. Thus we will learn some basic scripting throughout this course. But basically, scripting is just like typing commands into a command-line, but instead saving all those commands to a text file, and having the computer run the commands for you.

3. But we are getting ahead of ourselves. The *first* thing you want to do when running a new program is making sure you can get it to work at all. Typically programs will have an example data file (in the correct format and everything) and either a script or a set of basic commands. The example data and scripts will often be included in the program download, and the basic commands will be on a help page for the program.

4. After you get the basics working, then you start to learn how the program works by opening up the input files, scripts, and output files in a text editor and seeing what the commands were, what the formats were, etc. You can then start to try getting your own data into the right format, try running the default analysis, and then start adding other commands and analyses that the program will perform.

5. As you add or change commands, you save your script files. At the end, you should be able to run your analysis again, e.g. years later when you have some new data, by just running the script. Re-doing all of this by clicking menus would be (a) impossible to remember and (b) tedious. Also, journals often require that you upload your data and scripts as part of the supplementary material, so that others may replicate your work, which is what science is about.

6. When you get really good, you will learn Perl or Python (Python is better) and use it to gather data from online databases, process it, get it into the correct format, pipe it into on analysis program, take the output and pipe it into another program, summarize the results, repeat the whole process 1000 times for a thousand different input datasets, etc. This is called "building an analysis pipeline." It might not be necessary for e.g. just doing a phylogeny of a single group, but for any task involving repeated analysis of different datasets, or large datasets (e.g. genomes), it is very useful.

This concludes Nick's General Philosophy of Using Scientific Programs. Let's use TNT as an example.

Getting TNT to run the example file

1. With your file/folder viewer, navigate to your TNT directory. You should see the actual TNT program (e.g. *tnt.command* on Macs is the command-line program) and some other files. *example.tnt* is the example data file. *aquickie.run* is the example script.
2. Open a Terminal window (Mac) or Command Line window (Windows). Navigate to your TNT directory.
    a. Macs: use "pwd" to see what directory you are in, "cd directoryname" to change directory, "ls" to list files.
    b. PCs: use "dir" to see what directory you are in and list files, "cd directoryname" to change directory
3. Run TNT and process the example file. (directions here: http://www.zmuc.dk/public/phylogeny/TNT/)
    a. Macs: type "./tnt.command" to start TNT. Then type "proc example.tnt ; aquickie ; [enter]"
    b. Windows: just type at the command line: "tnt example.tnt ; aquickie ; [enter]"

Question #1: open in a text editor the example file, the script file, and the output files.

- Via email, give a brief description of what is in each.
- The script file is kind of complicated, so just look up what these specific commands do: qnelsen, resample, export (list of commands is here: http://tnt.insectmuseum.org/index.php/Commands)

Getting TNT to run a draft file of Nick's

On the website you will find *dq_v2b.tnt*, a file Nick generated from a Mesquite data matrix. It is actually a character matrix based on a misquote of Charles Darwin which is circulating on the internet and elsewhere. Last fall, some URAPs (including class member Jason) tracked down hundreds of versions of the quote and coded them up. Now we are analyzing them to try to find the shortest tree explaining the data.

1. Load the file with:
proc dq_v2b.tnt;

2. Try:
Xmult;

Question 2: How many trees did you get, and how long were they?

3. Try:

hold 1000 ;
mult 100=tbr ;

hold 1000;
mult 1000=tbr;

Question 3: How many trees, how long, and what was the difference between the two runs, if any?

4. Try:
chomo 1;

Question 4: Look up this command. What does it do?

5. Try some other parameters, try to find the shortest tree possible. When you are satisfied (particularly if you find a tree lower than 553 steps), write down your parameters, and export the trees to a nexus file:

taxname=;
export – yourname_dq.trees;

Then open up the NEXUS file in Mesquite. File→Export→ Phylip trees, then open that Newick file in Dendroscope.

Question 5: Send me the parameters from your best search, and a screen image of one of the trees.

Getting your own data into TNT

If there is time left in lab, work on your own NEXUS file, try to export your character matrix to TNT format and get TNT to read it ("proc yourfilename") and build some trees.