Web-based tools for Bioinformatics; A (free) introduction to (freely available) NCBI, MUSC and World-wide.

When and Where---Wednesdays 1pm-2pm Room 438 Library Admin Building Beginning September 10, 2003.

Overview


Introduction/Scope

You should always bear in mind that life begets life and that not surprisingly, offspring tend to resemble their parents. The fossil record is replete with forms that do not exist and those whose resemble current living species. The bewildering plethora of forms provokes a trend in the human mind to organize this diversity in systematic ways. Descriptions of the surrounding world stratigraphy of fossils and more recently genome composition have contributed to the various ways to organize this information. The central working tool for this analysis is evolutionary theory. The observer reviews the available evidence and from this infers the order and plausible relationships between the pieces of the biological world. One of the most compelling of problems in the field of biology is the effort to explain the diversity and distribution of life on earth.

- **Phylogeny**
  
  The historical relationships among lineages of organisms or their parts (e.g., genes).(Hillis, Noritz & Mable, 1996)

  The tree-like evolutionary history of a set of taxa. (Chapter 17)(Gusfield, 1997)

- **Phylogenetics**
  
  The reconstruction of the evolutionary history of a group of taxa or genes.(Li & Graur, 1991)

- **Phylogenetic Tree**
A graphic representation of the phylogeny of a group of taxa. (Li & Graur, 1991) or genes. (Li & Graur, 1991)

- **Phylogeography**

  The study of biogeography as revealed by a comparison of estimated phylogenies of populations or species with their geographic distributions. (Hillis, Noritz & Mable, 1996)

- **Phylogram**

  A tree that depicts inferred historical relationships among entities. Differs from a cladogram in that the branches are drawn proportional to the amount of inferred character change. (Hillis, Noritz & Mable, 1996)

NCBI has a series of Science Primers. Here's link to the systematics and phylogenetics section. This is worth the time to pick up some terms and get the general picture.

**Wood Hole Marine Biology Lab** offers a two week **Workshop on Molecular Evolution** July 25-Aug 3 2004 ($1500 Application deadline March 1, 2004).

Uses of these methods include:

- Mitochondria, efficiency, heat production, haplotypes and climate


- Epidemiologists use phylogenetic methods to understand the development of pandemics, patterns of disease transmission, and development of antimicrobial resistance or pathogenicity:


- Conservation biologists may use these techniques to determine which populations are in greatest need of protection, and other questions of population structure:


http://people.musc.edu/~hazards/WebBioInformatics/PhylogeneticTools.htm 1/14/2004
Pharmaceutical researchers may use phylogenetic methods to determine which species are most closely related to other medicinal species, thus perhaps sharing their medicinal qualities:


Phylogenetic methods may be used to solve crimes, test purity of products, and determine whether endangered species have been smuggled or mislabeled:


Current phylogenetic methods generally employ (in order of increasing computational intensity)

- Distance methods
- Parsimony methods
- Maximum Likelihood Methods

Each of these methods, in turn, have a number of different modes and nuances which add complication and sophistication to the methods. Whether you are beginning for the first time or a seasoned veteran, the core of these methods is that they begin with a multiple alignment. Changes which are observed at a given position in an alignment may be modeled by a wide variety of techniques, none of which presume anything about the validity of the alignment. Invest your time and intellect on the alignment first. When you are satisfied, then proceed with the phylogenetic methods. This is a long way to get at the apt phrase "garbage in-garbage out!"

**A worked Example**

In order to perform some phylogenetic analyses, you need a candidate gene (or genes or even genomes) from a number of species or individuals or experimental conditions

- See the links in the Sample Questions/Data section to retrieve the example dataset.
- Mount your Homeroom.
- Retrieve and save the snapping shrimp data into it's own subdirectory in your Homeroom space.

Notice if you use my version of the shrimp data that I have modified the original FASTA files to reduce the ID of the sequences to just the gene ID number. This keeps the number of characters in the name to 10 or so. Why? Because the PHYLIP programs that we currently have installed with the BCR limit the name to 10 characters and you don't want the geneID number trimmed or several of the sequences will then have the same name--BAD IDEA.

The snapping shrimp data will come as a FASTA format file. We will then perform some multiple alignment operations to achieve two ends. First we want to see if we can reproduce the Williams alignment independently and secondly, by digesting the data through CLUSTALW we can produce
output in a PHYLIP format which can be used directly for phylogenetics.

FASTA files are readable by CLUSTALW and T_COFFEE (and can converted to GCG format very easily). Login to a BCR machine and change to the directory where you placed the shrimp FASTA files. T_coffee is a more sophisticated alignment program than clustalw but in its basic mode it's easy to use. Type:

```
t_coffee -infile your-sequence-file-name
```

To start clustalw simply type "**clustalw**" then select option 1.

```
Alpheus_50.tfa
erg2 4% clustalw
```

```
***********************************************************************
****** CLUSTAL W (1.73) Multiple Sequence Alignments *******
***********************************************************************

1. Sequence Input From Disc
2. Multiple Alignments
3. Profile / Structure Alignments
4. Phylogenetic trees

S. Executs a system command
H. HELP
X. EXIT (leave program)

Your choice: 1
```

Notice that the choice 1 is highlighted, that the program sees "Pearson" aka FASTA format and then prints the names and lengths of the individual sequences.
Your choice: 1

Sequences should all be in 1 file.

7 formats accepted:
NBRF/FIR, EMBL/SwissProt, Pearson (Fasts), GDE, Clustal, GCG/MSF, RSF.

Enter the name of the sequence file: Alpheus_50.tfa

Sequence format is Pearson
Sequences assumed to be DNA

Sequence 1: 13386623 401 bp
Sequence 2: 13386657 401 bp
Sequence 3: 13386691 401 bp
Sequence 4: 13386613 401 bp
Sequence 5: 13386649 401 bp
Sequence 6: 13386683 401 bp
Sequence 7: 13386641 401 bp
Sequence 8: 13386675 401 bp
Sequence 9: 13386709 401 bp
Sequence 10: 13386633 401 bp
Sequence 11: 13386667 401 bp
Sequence 12: 13386701 401 bp
Sequence 13: 13386625 401 bp
Sequence 14: 13386659 401 bp
Sequence 15: 13386693 401 bp
Sequence 16: 13386617 401 bp
Sequence 17: 13386651 401 bp
Sequence 18: 13386685 401 bp
Sequence 19: 13386643 401 bp
Sequence 20: 13386677 401 bp
Sequence 21: 13386711 401 bp
Sequence 22: 13386635 401 bp
Sequence 23: 13386669 401 bp
Sequence 24: 13386703 401 bp
Sequence 25: 13386627 401 bp
Sequence 26: 13386661 401 bp
Sequence 27: 13386695 401 bp
Sequence 28: 13386619 401 bp
Sequence 29: 13386653 401 bp

At the end of the listed input sequences you have a new menu. Select 2 to perform multiple
alignments.

The next menu allows you to perform an alignment; choice 1, OR to set up output format options; choice 9.
Option 4 in the next window allows you to toggle (turn on) the PHYLIP output format. Entering 4 will request CLUSTALW to produce the alignment in PHYLIP as well as its native format. Note as well the option 9. This allows you to write the current alignment to another format. Suppose you had Williams' original multiple alignment, you might want to use his alignment to see if you could reproduce his inferences given his alignment but you might want the alignment in PHYLIP format. You would enter 9 and the input alignment would be written to a different format without any readjustment by the clustalw program. In our example the Alpheus elongation factors are in FASTA format, and not aligned. We will align them and write them to PHYLIP format.
1. Do complete multiple alignment now (Slow/Accurate)
2. Produce guide tree file only
3. Do alignment using old guide tree file
4. Toggle Slow/Fast pairwise alignments = SLOW
5. Pairwise alignment parameters
6. Multiple alignment parameters
7. Reset gaps before alignment? = OFF
8. Toggle screen display = ON
9. Output format options

S. Execute a system command
H. HELP
or press [RETURN] to go back to main menu

Your choice: 9

******** Format of Alignment Output ********

1. Toggle CLUSTAL format output = ON
2. Toggle NBRF/PIR format output = OFF
3. Toggle GCG/MSF format output = OFF
4. Toggle PHYLIP format output = OFF
5. Toggle GDE format output = OFF
6. Toggle GDE output case = LOWER
7. Toggle CLUSTALW sequence numbers = OFF
8. Toggle output order = ALIGNED

9. Create alignment output file(s) now?

0. Toggle parameter output = OFF

H. HELP

Enter number (or [RETURN] to exit): 1

Hitting return send you back to the previous menu and you select 1 to perform the alignment. CLUSTALW will write the alignment as its own format and as a PHYLIP format.

Regardless of where or how you obtain an alignment, engage your brain and examine the alignment as if it is DANGEROUS. Be suspicious, fiddle with some parameters, try another alignment program. ALL of the phylogenetic inferences DEPEND on this alignment. CLUSTALW and most other methods available today do NOT create an OPTIMAL alignment. Instead they rely on heuristic tricks which demonstrably produce an alignment that is CLOSE TO OPTIMAL but there is no guarantee of optimality.
In addition to thinking hard about your data and the alignment, you can perform a series of mathematical tricks to assess the reliability of your alignment. Generally this is done by a procedure known as BOOTSTRAPPING. This procedure essentially re-samples your input data in a clever way, leaving something out each time. If you perform say 1000 bootstrap cycles, then the number of times a cluster of taxa appear together in the same clade is a measure of the reliability of the data.

When the alignment finishes CLUSTAW displays on screen the result and then returns you to the alignment menu. Hit return to go to the main menu. Select 4 to perform a quick and dirty bootstrapping operation on the current data.

Press [RETURN] to continue:

****** MULTIPLE ALIGNMENT MENU ******

1. Do complete multiple alignment now (Slow/Accurate)
2. Produce guide tree file only
3. Do alignment using old guide tree file
4. Toggle Slow/Fast pairwise alignments = SLOW
5. Pairwise alignment parameters
6. Multiple alignment parameters
7. Reset gaps before alignment? = OFF
8. Toggle screen display = ON
9. Output format options

S. Execute a system command
H. HELP
or press [RETURN] to go back to main menu

Your choice:

******************************************************************************
****** CLUSTAL W (1.73) Multiple Sequence Alignments ******
******************************************************************************

1. Sequence Input From Disc
2. Multiple Alignments
3. Profile / Structure Alignments
4. Phylogenetic trees

S. Execute a system command
H. HELP
X. EXIT (leave program)

Your choice: 4

Then select option 5 Bootstrapping. You are prompted to make some choices and then the
operation begins. You can watch the progress. Upon completion hit return and exit the program.

S. Execute a system command
H. HELP
or press [RETURN] to go back to main menu

Your choice:

****************************************************************************
***** CLUSTAL W (1.73) Multiple Sequence Alignments *****
****************************************************************************

1. Sequence Input From Disc
2. Multiple Alignments
3. Profile / Structure Alignments
4. Phylogenetic trees

S. Execute a system command
H. HELP
X. EXIT (leave program)

Your choice: 4

***** PHYLOGENETIC TREE MENU *****

1. Input an alignment
2. Exclude positions with gaps? = OFF
3. Correct for multiple substitutions? = OFF
4. Draw tree now
5. Bootstrap tree
6. Output format options

S. Execute a system command
H. HELP
or press [RETURN] to go back to main menu

Your choice: 5

Here's a view of the Boostraping options
Here's the first page of the phylip "interleaved" output format. It shows the number of taxa (50) and the length of the alignment. Note the names are the bare geneID numbers.
Here for the sake of comparison is a section of the CLUSTALW format output file. Note the Asterices which suggest conservation of bases at particular positions.
At this point we have at last produced a multiple alignment in a format suitable for phylogenetic analysis but what about the bootstrapping we asked CLUSTALW to perform. To examine this its useful to get a treeviewing program. A nice free one is TreeView. Here is our result viewed with treeview. Note the Trichotomy. The taxa are listed on right and adjacent to various nodes we see numbers. These numbers represent the number of times per 1000 trials that all the taxa to the right were in the same cluster.
With our alignment in the phylip format we are almost ready to start the analyses. We need to understand one more trick for using UNIX PHYLIP. All the programs look around in the current working directory for a file called literally "infile". Therefore we need to copy the phylip format file to "infile". To do this type:

```
cp Alpheus_50.phy infile
```

That should do the trick.

To perform a DNA Distance calculation type

```
dnadist
```

You will see a standard PHYLIP start menu like this:

```
Nucleic acid sequence Distance Matrix program, version 3.572c

Settings for this run:
D Distance (Kimura, Jin/NcI, ML, J-C)? Kimura 2-parameter
T Transition/transversion ratio? 2.0
C One category of substitution rates? Yes
L Form of distance matrix? Square
M Analyze multiple data sets? No
I Input sequences interleaved? Yes
0 Terminal type (IBM PC, VT52, ANSI)? ANSI
1 Print out the data at start of run Yes
2 Print indications of progress of run Yes

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

You can toggle the various options to tailor your distance calculation or type "y" to accept them and run the program. In a manner analogous to the input file process, the output is written to a generic file called "outfile". YOU WILL NEED TO COPY THE CONTENTS TO ANOTHER FILE OR PRESERVE THEM!!!
The meaning of the numbers is dependent on the model used but basically some normalized value is written such as the average number of mutations (differences) per residue or per 100 residues. 13386677 is identical to itself but rather different from the next to the last entry.

A parsimony search seeks to find the arrangement of the sequences which produces the smallest overall number of changes. This makes intuitive sense but there are plausible circumstances where different arrangements can be equally parsimonious. To perform type

```
dnapars
```
DNA parsimony algorithm, version 3.572c

Setting for this run:

U Search for best tree? Yes
J Randomize input order of sequences? No, use input order
O Outgroup root? No, use as outgroup species
T Use Threshold parsimony? No, use ordinary parsimony
M Analyze multiple data sets? No
I Input sequences interleaved? Yes
0 Terminal type (IBM PC, VT52, ANSI)? 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 No
5 Print sequences at all nodes of tree No
6 Write out trees onto tree file? Yes

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

There is another menu which you may customize, then press "Y" to execute. This time both an out file and tree file are produced which will need to be saved. It's not really apparent given the size of this data set but parsimony methods are slower than distance methods.

Here's a TreeView view of the FIRST of 96 trees found by DNAPARS which are equally parsimonious; that is they have the same length or the it takes an equal number of changes to produce each of the 96. Note that you can scroll through the 96 trees with the arrow button in the upper left.
There is a maximum likelihood program called "dnaml" which may be run in the manner of all PHYLIP
programs. dnaml in particular is rather slow. A related program called fastdnaml is also available. It uses the same input file format but you need to tell it where to write the output. For instance

```
fastDNAml <infile >Alpheus_50.fastDNAmlout &
```

The "&" runs the program in "background". You want to run all maximum likelihood programs in background because the calculations are seriously lengthy. Even the fast method takes noticeably longer to run. Here's the output tree from fastDNAml running on the Alpheus_50 data with default options.
In addition to the PHYLIP programs and fastDNAml, MUSC users can run a number of molecular evolution programs as listed here. PAML and MOLPHY as well as the PUZZLE program. Here's the output from a PUZZLE run on the Alpheus_50 data.
Well, maybe the UNIX command line is not for you. If you really cannot bear command line issues then check out the free program MEGA (http://www.megasoftware.net/).
MEGA is a very nice program with a number of capabilities and a nice tree editor function.

PAUP for MAC/WINDOWS/UNIX is not free but it is inexpensive and worth the cost.

Joe Felsenstein (author of PHYLIP) maintains a nice web site listing for lots of phylogenetic programs.

Here's a link to Sinauer Associates Systematics for several phylogenetics related titles including a link to PAUP purchase and the new book by Felsenstein. Amazon.com has many of these also. For some evolutionary biology titles try this Sinauer page.

Sample Questions/Data

Here's the ENTREZ page. Queries from this omnibus page also poll PopSet.

From the PopSet page follow the link to the snapping shrimp (Genus Alpheus) and retrieve the Alpheus Elongation Factor sequences in FASTA format (see the Links menu on the right or download directly here). This data is discussed in the following article by Williams et al (note the link to the full text). You will have to align them via PAUP, PHYLIP, GCG Pileup, CLUSTALW or t_coffee etc to begin a phylogenetic analysis.

Here's a link to a data set of fruitfly hexokinase coding regions from a PopSet entry derived from a study by Duvernell et al. You will have to align them via PAUP, PHYLIP, GCG Pileup, CLUSTALW or t_coffee etc to begin a phylogenetic analysis. This is a very different dataset from the snapping shrimp elongation dataset. Why?