

INSTRUCTIONS FOR BOOTSTRAPPERS GAMBIT

(Copyright 1999 J.A. Lake, All rights Reserved)Read_Me.doc

USEFUL METHODS FOR CALCULATING ROBUST TREES

It's not generally appreciated that molecular sequence analysis is a field in its infancy. Thus it is an inexact science, in which there are few analytical tools that are based on general mathematical principles. As a result many, perhaps most, phylogenetic trees reconstructed from molecular sequences are incorrect, because they make mathematical assumptions that are not met by the data being analyzed. Frequently these incorrect assumptions lead to long branch attraction.

Long Branch Attraction

Long branch attraction can be caused by one or more of three pitfalls of sequence analysis. For all three the effects are the same, In trees artifactually produced by long branch attraction, rapidly evolving sequences (represented by long branches on unrooted phylogenetic trees) will be placed with other rapidly evolving sequences, even if the sequences are only distantly related. In comparison with most problems in molecular biology, which can be solved by acquiring more data, long branch attractions are diabolical. When long branch attractions are present, if longer sequences are used, the incorrect solution will be even more strongly supported.

Specifically the mathematical steps in sequence analysis that produce this pitfall are; i. incorrect sequence alignments, caused by inadequate mathematical models and often related specifically to biases created by progressive alignment algorithms when they are used to align more than three taxa (organisms); ii. the failure to account properly for site to site variation (all sites within sequences can evolve at different rates, yet most algorithms assume they evolve at the same rate), and iii. unequal rate effects (the inability of most tree building algorithms to produce good phylogenetic trees when genes from different taxa in the tree evolve at different rates). Of the three pitfalls, alignment artifacts are potentially the most serious, because even if one solves the second and third problems, then misalignments can still produce incorrect trees. General algorithms are available for pitfalls two (site to site variation) and three (unequal rate effects) and are incorporated in the **Gambit** program, but none are available for the alignment problem. **Gambit** contains algorithms not significantly affected by site to site variation or by unequal rate effects. Specifically, paralinear (logdet) distances (1,2), is a truly additive method for determining distances between sequences. Since Paralinear distances is based on a very general Markov model, it is not significantly affected by unequal rate effects. Also, Pattern Filtering is a demonstrably optimal method for estimating the variation of rates at different sequence sites (3), and as such, is not significantly affected by site to site variation effects. Both of these methods are available in **Gambit**.

Tree Reconstruction

Determining globally optimal, multi-taxon phylogenetic trees is also computationally intensive because the number of possible trees increases rapidly with increasing taxa. (For four taxa, three unrooted trees must be compared; whereas for thirteen taxa, thirteen billion, 13,749,310,575 trees must be compared.) Given such large numbers it is difficult to search exhaustively more than 12-13 taxon trees even using the branch and bound algorithm (4). **Gambit** approaches this problem in a unique way. Once **Gambit** finds a solution (using heuristic methods), it uses the data to estimate the probability that a better solution exists (5). **Gambit** then accepts only solutions for which better solutions are unlikely (at either the 95% or 99% confidence levels). With these methods it is possible to calculate "best" trees in reasonable times for 15 - 30 taxa, depending upon the sequence data.

An additional difficulty found when constructing multiple taxon trees, is that many different optimality criteria are used for evaluating the "best" multi-taxon trees. For example, distance trees can be reconstructed by searching for local minima using least-squares criteria, or by the criterion of minimum distance, whereas parsimony methods minimize the number of nucleotide changes often using global searches (6). **Bootstrappers Gambit** is a multi-taxon tree reconstruction algorithm designed so that it can be used with most, if not all, phylogenetic methods. It uses a probability criterion as a common basis for comparing trees derived using diverse methods (7).

----- **The following paragraph may be skipped**

Bayesian and likelihood methods can assess the probabilities of trees and thus are useful for providing a common basis for reconstructing trees using different algorithms. Sinsheimer *et al.* (8) developed a method for calculating the probability of trees derived by evolutionary parsimony, but the calculations are complex for trees with more than five taxa. Felsenstein (9) has thoughtfully proposed that bootstrap replicates (10,11) might provide a good method of assessing the likelihood function in tree reconstruction. Both groups calculate the probability, $P(\text{tree}_j|S)$, that the j^{th} tree is correct given aligned sequences, S . These are complex calculations. **Bootstrappers Gambit** calculates something simpler - the probability, $P(H|S)$, that algorithm A applied to a sequence of infinite length (generated under the same model as S) would yield the j^{th} tree. Under a multinomial model (assuming a Jeffreys' prior on the underlying parameters) the integral for calculating $P(H|S)$ can be estimated by bootstrap replication. **Bootstrappers Gambit** combines this bootstrap with a multi-taxon algorithm that may be used in combination with any four-taxon method. **Bootstrappers** refers to the Bayesian-Jeffreys'-bootstrap method of estimating probabilities and **Gambit** or **dance** refers to the systematic search of trees based on their decomposition into four-taxon statements.

Bootstrappers Gambit combines various algorithms for phylogenetic analysis into a single package. The program is designed for personal computers and runs on the DOS operating system (a Windows 95/98 version is being tested). Among the phylogenetic reconstruction methods accommodated in **Gambit**, in addition to Paralinear distances are: Jukes-Cantor distances (12), Kimura two parameter distances (13), a 6 parameter distance method based on the evolutionary parsimony assumptions (Lake, unpublished), maximum parsimony (14), evolutionary parsimony

(15), and a symmetric transversion parsimony (16). Other algorithms, such as maximum likelihood are being added.

AN EXAMPLE OF GAMBIT

Gambit functions by decomposing multiple taxon trees into sets of four taxon statements as illustrated in the following Figure for a five-taxon tree (7). Five aligned sequences corresponding to taxa 1 through 5 are shown at the top of the figure. Three bootstrap replicates (obtained by sampling with replacement) from the original aligned sequences are shown at the top of the figure. For this example, maximum parsimony is used to analyze quartets of taxa. Distance methods are similar but use four-point equations to calculate winning quartet values (17). For four taxa (i, j, k, and l) three trees are possible (the E tree clusters i with j and k with l, the F tree clusters i with k and the G tree clusters i with l). For example, in the first column of replicate 1 the quartet represented by taxa 1, 2, 3 and 4 (denoted 1234), corresponds to the sequence pattern, AAAA. Since this pattern supports no tree, by parsimony, the result is indicated by a blank (-) in the table of quartet values for replicate 1. In the second column the sequences for quartet 1234 are TTCC. Parsimony interprets this pattern as support for the E tree (14) and an e is entered in the quartet value table. The most parsimonious four-taxon trees are then chosen by counting 'e's, 'f's and 'g' s at all sequence positions. The four-taxon trees supported at the most positions are entered into the quartet value table. (If two trees tie, then no tree is selected.) For replicate 1 the pattern of winning four-taxon tree values is EEEEE (quartets 1234, 1235, 1245, 1345, and 2345, respectively). This value pattern and the tree it uniquely supports are shown beneath each replicate. Some quartet value patterns are inconsistent with trees and may support non-tree graphs (15). For example the pattern from replicate 2, GEEFE, fits no tree. Details of Gambit, used to relate value patterns to trees, are described in the original paper (7).

The last step involves calculating the probability of each tree. The conditional probability that a particular tree would be supported with infinite data is given by the number of replicates supporting the tree divided by the total number of replicates supporting trees (for details see the Bayesian-Jeffreys'-Bootstrap Theorem in ref. 7). In this example, two trees corresponding to the EEEEE pattern and the GEFFF patter are present. The total number of trees is two, so that the probability of the EEEEE tree is estimated as 1/2 and the probability of the GEFFF tree is 1/2. Better estimates can be provided by taking more replicates. In practice, thousands, or even millions of replicates are calculated for a single set of trees, since for large multi-taxon trees most value patterns support no tree.

Taxa Original Sequence
 1 A T C G G T A C C G
 2 A T C G T G A G C G
 3 A C C C T G A A T G
 4 A C A T G G T G T G
 5 A C A C T C T A G G

Bootstrap Replicates

Taxa Replicate 1 Replicate 2 Replicate 3
 1 A T G G T A A C C C T T G G G T A A G G A T G G G G T C C
 G
 2 A T G G G A A G C C T T T T T G A A G G A T G G T T G G G
 G
 3 A C C C G A A A T T C C T T T G A A G G A C C C T T G A A
 G
 4 A C T T G T T G T T C C G G G G T T G G A C T T G G G G G
 G
 5 A C C C C T T A G G C C T T T C T T G G A C C C T T C A A
 G

Quartets Four Taxon tests
 1234 - e - - - - - e e e e g g g - - - - - e - - g g - - -
 -
 1235 - e e e - - - - - e e - - - - - - - - e e e - - - - -
 -
 1245 - e - - - e e - - - e e f f f - e e - - - e - - f f - - -
 -
 1345 - - - - - e e - - - - - f f f - e e - - - - - f f - - -
 -
 2345 - - - - - e e f - - - - - e e - - - - - f f
 -

1234 E G
 1235 E 1 3 4 E Pattern G
 1245 E \ | / E corresponds F 1 2 3
 1345 E \ | / F to no tree F \ | /
 2345 E 2 5 E F 4 5

Probability($\begin{matrix} 1 & 3 & 4 \\ & \backslash & | & / \\ & 2 & & 5 \end{matrix}$ | sequences) = (1 + 0 + 0) / (1 + 0 + 1) = .50

Probability($\begin{matrix} 1 & 2 & 3 \\ & \backslash & | & / \\ & 4 & & 5 \end{matrix}$ | sequences) = (0 + 0 + 1) / (1 + 0 + 1) = .50

Getting Started

Installing **Gambit** is simple. Move the executable named Gam95.xyz to your desktop, and rename the file Gam95.exe. A sample metazoan data set, LOPH1294.CUT (slightly modified from Halanych, Bacheller, Aguinaldo, Hillis, and Lake, Science, **267**, 1641-43, 1995) comes with the program, and also put this on the desk top.

Double click on the gambit icon. The program will start, and you will be queried as shown below. Just type in the location of your file (the reply is in shown in bold and should always be followed with a return),

```
What is the name of the input file? C:\...\loph1294.cut
```

(Note that upper or lower case will work for all queries in the program. Also if you have started **Gambit** from windows, you may need to type in the complete location of your input file, eg. C:\.....\loph1294.cut .

The program will next present a menu of sequences and a query that will look like (including your reply in bold):

```
1 Plumate N    5 Placopect    .....
2 Terebrata   6 Acanthopl    ....
3 Phoronis    7 Glycera a    ..
4 Glottidia   8 Artemia      .
```

```
How many taxa would you like to analyze? 8
```

Then the following prompt appears:

```
List the taxon numbers (Separate with commas) 2,3,4,7,8,9,10,14
```

(If you need to enter more than 10 taxa, then you should hit the return after every ten, or so, numbers, and after the last number). Also note that the order in which the numbers are entered is not important, since the program randomizes the order of the taxa.

Next the taxa will be listed. Make sure that they are listed correctly, and if not, the program can be terminated by typing <control>+c, at almost any time.

You are next queried (the reply is in bold as always):

```
How would you like to score gaped positions?
Select an option by typing its number
```

1. Exclude all positions with gaps
2. Exclude all positions with more than 50% gaps
3. Include all positions

```
1
```

and the computer will reply,

All positions with gaps are excluded

The next query will be,

Select data type by typing its number

1. Nucleic acid sequences (DNA or RNA)
2. Protein sequences in one letter code

1

and the computer will reply,

Nucleic acid sequences are being analyzed

You will next be asked to:

Select a site-to-site (STSV) criterion by typing the number

1. Classify sites by rates
2. Assume all sites are the same (i.i.d.)

1

The computer replies,

Sites classified by rate

Npositions = 1439

Enter Y (or N to plot the distances on the screen? **n**

As usual, the reply is not case sensitive. If you had replied y, then a crude plot of the rates for every 4th site would have been presented. From these rates a histogram of rate variation is calculated and presented on the next screen.

The cells are numbered from 0 to 60 corresponding to substitution rates from 0.0 to 1.0 substitutions/position (also labelled). From these you are asked to select the cells to be used for the analysis.

What starting and ending cells would you like? (Space with a comma) **0,16**

The histogram will appear again, this time with only distances labeled. At the bottom of the histogram will be a set of numbers below each of the cells selected for analysis. These are the bin numbers, and show which cells will be grouped for analysis. The notation appears as,

Bins 11233455667778888

Bin #1 corresponds to cells 0 and 1, bin #2 to cell 2, bin #3 to cells 3 and 4, etc. Each bin contains the appropriate number of sites to allow the reliable calculation of a tree. Then if the site to site variation (STSV) correction is subsequently requested, results from each set of bins will be analyzed

individually, and combined to reconstruct the tree. In this way STSV effects can be dramatically reduced.

In our example, the available sites are all evolving at reasonable rates. If, however, some sites had been evolving at rates higher than about .35 substitutions/site, one might want to exclude these from the analysis. Sometimes one might want to study the slower evolving sites separately, or to study faster evolving sites separately. **Gambit** gives us this flexibility.

The program then asks,

```
Gambit can perform the following tasks:
  1. Calculate trees and their probabilities
  2. -- Option not available --
  3. Calculate branch lengths
  4. Print nucleotide compositions
Type the number of the task:
1
```

In this case we have chosen to calculate the number of trees. Details of options 3 and 4 are provided by the prompts.

The program then replies,

```
Trees will be calculated
How many trees would you like to calculate? 100
```

Many times, when analyzing larger trees incorporating more taxa, it is worthwhile to specify initially a small number of trees, perhaps 1, in order to see how much time the calculations will take. Since **Gambit** will always prompt you for more trees following the end of the calculation, you can always increase the total number of trees later.

Next the program asks,

```
Select a nodal criterion by typing its number
----- Optimal solutions calculated below -----
  1. Probability of finding a better tree < 5%
  2. Probability of finding a better tree < 1%
1
```

Either option is fine, but the second option can take considerably longer, depending upon the data.

Next the computer returns the computed probability of a better tree and queries you about the randomizing the order of taxa.

```
Number of better solutions per tree < 0.499988E-01
How many random orderings of taxa (N < or = 100)?
(If N=0, taxa will be added in the order entered) 100
```

The computer then repeats your selection and asks you to select the phylogenetic algorithm.

```
The number of random orderings of taxa is:          100
Select an algorithm by typing the algorithm number
  1.    Paralinear/LogDet distances
  2.    Kimura two parameter distances
  3.    Jukes Cantor distances
  4.    Lake 6 parameter distances
  5.    Evolutionary Parsimony
  6.    Maximum Parsimony
  7.    Symmetric Transversion Parsimony
  8.    -- Not Available --
----- SITE TO SITE VARIATION -----
  9.    Paralinears for STSV
 10.    Kimura distances for STSV
 11.    Jukes Cantor distances for STSV
9
```

If no warnings have appeared (a warning would appear only if more than 29 bins were needed), then I recommend the site to site variation algorithms be used. These are available for Paralinear, Kimura, and Jukes Cantor distances. Also the site to site variation algorithms can frequently run MUCH FASTER than the same algorithms without site to site variation.

Following selection of the algorithm then a series of pairs of numbers appears on the screen as each tree is successfully discovered. The first number of the pair is the number of times that this particular tree topology has been found previously, and the second number is the percent of the requested number of trees that have been found.

After all of the requested trees have been calculated, **Gambit** asks,

```
Would you like to do additional bootstraps (Y or N)?  n
```

Then **Gambit** lists the number of bootstrap replicates that have been tried and the number of trees successfully found. It also lists the taxa in the order in which they were first selected and enquiries about the root of the tree,

```
NSUCCESS, NTRIES=          100          155
From the taxa listed below:
  1 Terebrata  2 ...          3 ...
  6 Eurypelma          7 Antedon          8 Anemonia
Type the taxon number of the outgroup for rooting the tree:  8
```

When this is finished, **Gambit** echoes back your reply, lists the number of clades found, and then asks how many trees you would like to display.

```
The tree is rooted in the branch to taxon          8
Number of Clades found
12345678
**.....          100
```

```
****.....      100
*****..       79
***.....       65
....**..       56
  etc.
```

To display:

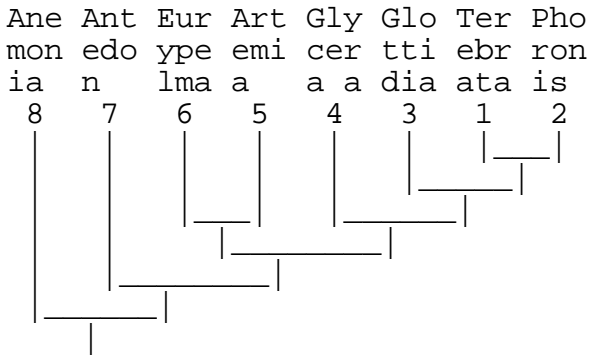
1. the most probable tree,
2. the 5 most probable trees,
3. the 10 most probable trees,
4. ALL the trees
5. NONE of the trees

Type the number of your choice, **2**

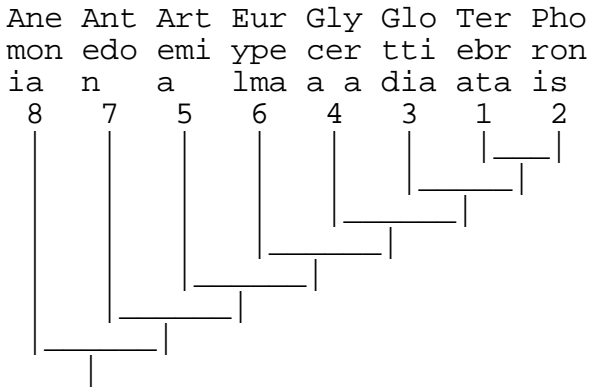
The computer then prints the various trees and their probabilities in the decreasing order of their probability,

The trees and their probabilities follow:
 Displaying the 5 most probable trees

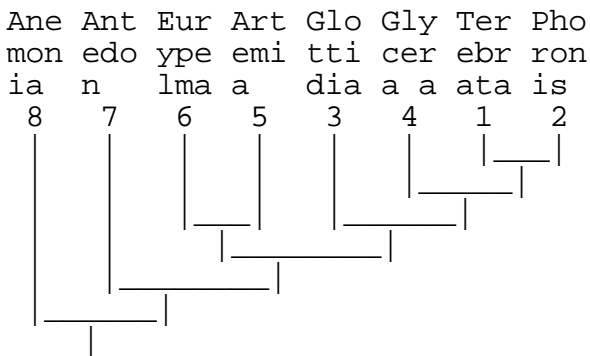
P = 32.0%



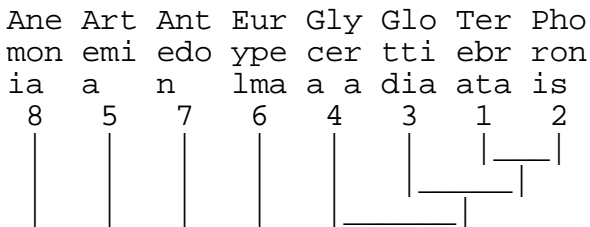
P = 20.0%

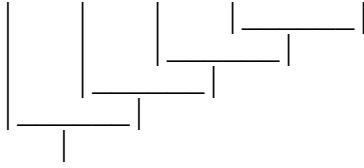


P = 14.0%



P = 11.0%





etc.

```

1 Terebrata      2 Phoronis      3 Glottidia      4 Glycera a      5
Artemia  6 Eurypelma  7 Antedon      8 Anemonia

```

After this the program prompts you if you want to end the session or asks if you want to go back to the histogram and continue the analysis, Type E to exit, or <enter> to return e

At this point, **Gambit** is finished and writes the results to the file **loph129.out**. Note the first name of the output file is truncated to no more than 7 letters, and the last name is changed to "out".

Making Data Files

In order to analyze your own data, you need to have aligned sequence files in the correct format. The file structures that are accepted by **Gambit** are described below.

Gambit currently recognizes three formats for input data. These are the GAMBIT format (to be described), the GCG (MSF) format (close the PHYLIP format), and the NEXUS (Paup, MacClade) format. It is not necessary to select the format type as these are identified by the program itself. In the future, PHYLIP and HENNIG86 formats will also be recognized.

The Gambit Format.

The Gambit format resembles that used for the GCG set of programs, *i.e.* Pileup, Lineup, Pretty Printouts, *etc.*, and that used for PAUP. Since it is not documented anywhere it is illustrated below. A glance at the file loph1294.cut, slightly modified from that used in (18), gives you an idea of the format of the data files. The first line contains the number of sequences, 17 in this case, and then the sequences follow in the next 17 lines. This is followed by a blank line and then by another 17 lines of sequence. In this example, the number of nucleotides has been totaled for each sequence and written to the right of the sequences **but these are ignored by the program**. Also the first 10 spaces of each line usually contain taxon names. These are required for the first set of 17 lines, but 10 blank spaces can be used if desired for all the subsequent sets of 17 lines.

```

17
Plumate N NNNCTGGTTG ATCCTGCCAG TAGTCATATG CTTGTCTCAA AGATTAAGCC 50
Terebrata ----- -AGTCATATG CTTGTCTCAA AGATTAAGCC 29
Phoronis ----- -AGTCATATG CTTGTCTCAA AGATTAAGCC 29
Glottidia ----- -AGTCATATG CTTGTCTCAA AGATTAAGCC 29
Placopect AACCTGGTTG ATCCTGCCAG TAGTCATATG CTTGTCTCAA AGATTAAGCC 50

```

Acanthopl	TACCTGGTTG	ATCCTGCCAG	TAGTCATATG	CTTGTCTCAA	AGATTAAGCC	50
Glycera a	TACCTGGTTG	ATCCTGCCAG	TAGTCATATG	CTTGTCTCAA	AGATTAAGCC	50
Artemia	TACCTGGTTG	ATCCTGCCAG	TAG-CATATG	CTTGTCTCAA	AGATTAAGCC	49
Eurypelma	TACCTGGTTG	ATCCTGCCAG	TAGTCATATG	CTTGTCTCAA	AGATTAAGCC	50
Antedon	-----	-----	-----	-----	-----	0
Lampetra	-ACCTGGTTG	ATCCTGCCAG	TAG-CATATG	CTTGTCTCAA	AGATTAAGCC	48
Branchios	--CCTGGTTG	ATCCTGCCAG	TAGTCATATG	CTTGTCTCAA	AGATTAAGCC	48
Tripedali	AACCTGGTTG	ATCCTGCCAG	TAGTCATATG	CTTGTCTCAA	AGATTAAGCC	50
Anemonia	TATCTGGTTG	ATCCTGCCAG	TAGTCATATG	CTTGTCTCAA	AGATTAAGCC	50
Plumate 0	-----	-----	TAGTCATATG	CTTGTCTCAA	AGATTAAGCC	30
Spacer 1	AACCTGGTTG	ATCCTGCCAG	TAGTCATATG	CTTGTCTCAA	AGATTAAGCC	50
Spacer 2	TACCTGGTTG	ATCCTGCCAG	TAG-CATATG	CTTGTCTCAA	AGATTAAGCC	49
Plumate N	ATGCATGTCT	AAGTACATAC	GTGAAACCGC	GAATGGCTCA	TTATATCAGT	100
Terebrata	ATGCATGTCT	AAGTACACAC	GTGAAACCGC	GAATGGCTCA	TTAAATCAGT	79
...						

The Nexus format.

Gambit also recognizes data sets written in the NEXUS format. When such a data set is used, it provides a message that a Nexus format is being read. **Gambit**, however, recognizes only a small subset of all possible NEXUS formats. In particular, taxon names must be on the "left" side of the sequences. The sequences should start less than 50 columns from the left margin. The file should contain only a single block (the data block) and all data in the block, except for the data matrix and its "matrix" header, will be ignored. The data matrix may be interleaved (as in the Gambit format shown in the third example below), either DNA, RNA, or one letter protein sequences may be used, uppercase or lowercase letters may be used, gaps may be coded as "-" or as any character not equal to {A,C,G,T,U,a,c,g,t,u} (or the 20 amino acid codes). The maximum length of a line may not exceed 510 characters. Once the sequence data are read, you will be queried for data just as in the example using the **Gambit** format. For example, the following NEXUS file conforms to these requirements:

```
#NEXUS

begin data;
  dimensions ntax=4 nchar=513;
  format datatype=rna gap=- interleave;
  matrix

      HumanTATCTGGTTG [1-10]
        Chimp   tatctggttg
      Gorilla   TATCTG-TTG
        15462   T--CTGGTTG

      HumanTATCTCCTTG [11-20]
        Chimp   tatctCGttg
      Gorilla   TATCTG-TTG
        15462   T--CTggTTG

end block;
```

The GCG format.

For completeness, an example of the GCG format is shown below. As with the NEXUS format, the additional information is ignored, and you will be queried by the program as if you were running a **Gambit** data set.

```
Onycho.Msf  MSF: 2222  Type: N  September 18, 1995  17:17  Check:
9825  ..
```

```
Name: Onyfin           Len: 2222  Check: 6653  Weight: 1.00
Name: Euripat          Len: 2222  Check: 4204  Weight: 1.00
Name: C12afor          Len: 2222  Check: 1250  Weight: 1.00
Name: O2gfor           Len: 2222  Check: 549   Weight: 1.00
Name: C12arev          Len: 2222  Check: 8618  Weight: 1.00
Name: O2grev           Len: 2222  Check: 8551  Weight: 1.00
```

```
//
```

```
1                                     50
Onyfin  TATCTGGTTG ATCCTGCTAG TAGTCATACG CTCGTCTCAA ACATTAAGCC
Euripat TATCTGGTTG ATCCTGCCAG TAGTCATACG CTCGTCTCAA AGATTAAGCC
C12afor TATCTGGTTG ATCCTGCTAG TAGTCATACN CTCGTCTCAA ACATTAAGCC
O2gfor  TATCTGGTTG ATCCTGCCAG TAGTCATACG CTCGTCTCAA AGATTAAGCC
```

```
51         60
Onyfin  ATGCACGTCT
Euripat ATGCACGTCT
C12afor ATGCACGTCT
O2gfor  ATGCACGTCT
```

A few additional details (for all three format types)

Blank columns, as illustrated above, are not necessary. The length of a line may vary considerably. (All characters between positions 11 and 510 will be read.) For nucleotide sequences, only the letters corresponding to nucleotides: a,c,g,t,u,A,C,G,T,U will be analyzed. For amino acid sequences only the letters a, A, c, C, etc. will be recognized as amino acids. All others, including numbers, spaces, any other letters not included in the set (including for nucleotides N, R, or Y, and symbols, like ? or -, etc.) will be treated like gaps (for the purpose of calculating whether or not a column is to be included in a calculation).

Currently the maximum length of sequence allowed is 32,768 and the maximum number of taxa is 101 (17 for the beta test version). Also note that this beta test version will expire January 15, 2000.

Program Swap

Program **Swap**, provided as **Swapc.exe**, is a useful practical tool that allows one to:

1. Reformat data sets going from the Gambit, GCG, and Nexus formats, to any of the other

two.

2. Change the order of taxa within a data set or delete selected taxa from the data set.

Swapc is simple to run. Simply type **Swapc**, and follow the instructions that appear on the screen. The screen instructions are similar to those for **Gamec** and are self explanatory.

I hope you enjoy using **Gambit**. I appreciate your suggestions, and will read them all. But since this is a minor part of our research program, it may not always be feasible to respond to your requests.

These are copyrighted programs, so please read the Legal Section below. If you can not accept the conditions, please destroy your copies of the programs, and notify me (Jim Lake; 232 Molecular Biology Institute; UCLA; Los Angeles, CA 90095.)

Jim Lake

Legal Matters

Disclaimer of Warranty

No warranties, express or implied, are made that the programs contained in this set are free of error, or are consistent with any particular standard of merchantability, that they will meet your requirements for any particular application, or that they will not damage your computer. The programs are accepted *AS IS*. They should not be relied on for solving a problem whose incorrect solution could result in injury to a person or loss of property. If you do use the programs in such a manner, it is at your own risk. The author disclaims all liability for direct or consequential damages resulting from use of the programs.

Licensing Information

Computer programs, like literary or artistic compositions are protected by copyright. Generally it is an infringement for you to copy into your computer a program from a copyrighted source, except with a license. (It also deprives the program's author of compensation for their creative work.)

I want to make it easy for you to use the programs. But to do so legally, you will need a license. You can get one as follows:

["Non-commercial beta users trial license"] If you are using this software solely for non-commercial purposes, then you may download one copy of the programs for your own personal use on one computer (one screen), provided you supply your name, address, and email address, and agree to the conditions of its use. You are not authorized to transfer or distribute a copy to any other computer or person. The trial license may be used only until January 15, 2001. Copies of the program are expected to be available for purchase late in the year 2000.

["Commercial beta users trial license"] If you are using this software for commercial purposes, you may license the programs for use on a single computer (one screen). To obtain the programs please mail your name, address, and a \$50 license fee (must accompany order) to: James Lake, 232 Molecular Biology Institute, UCLA, Los Angeles, CA 90095.

Reference number 7 should be cited in publications resulting from use of this program. References numbers 1 and 3, should also be cited if results were obtained using paralign distances or STSV corrections, respectively.

REFERENCES

1. Lake, J.A. (1994) Reconstructing evolutionary trees from DNA and protein sequences: Paralineal distances. *Proc. Natl. Acad. Sci. USA*, **91**, 1455-1459.
2. Lockhart, P.J., Steel, M.A., Hendy, M.D. & Penny, D. Recovering evolutionary trees under a more realistic model of sequence evolution. (1994) *Mol. Biol. Evol.*, **11**, 605-615; Steel, M.A (1994) *Appl. Math. Lett.*, **7**, 19-24.
3. Lake, J.A. (1998) Optimally recovering rate variation information from genomes and sequences: Pattern filtering. *Mol. Biol. Evol.* **15**, 1224-1231.
4. Hendy, M.D. & Penny, D. (1982).
5. Lake, J.A. (1999) Manuscript in preparation.
6. Hillis, D.M., Moritz, C. eds. (1990) *Molecular Systematics*, Sinauer Associates, Inc. Sunderland, MA, USA.
7. Lake, J.A. (1995) Calculating the probability of multitaxon evolutionary trees: Bootstrappers Gambit. *Proc. Natl. Acad. Sci. USA*, **92**, 9662-9666.
8. Sinsheimer, J.S., Lake, J.A., & Little, R.J.A. (1996) Bayesian hypothesis testing of four-taxon topologies using molecular sequence data. *Biometrics*, **52**, 193-210.
9. Felsenstein, J. (1992) *Genet. Res.*, **60**, 209-220.
10. Efron, B. (1979) *Ann. Statist.*, **7**, 1-26.
11. Felsenstein, J. (1985) *Evolution*, **39**, 783-791.
12. Jukes, T.H. & Cantor, C.R. In **Mammalian Protein Metabolism III.**, ed. H.N. Munro, pp. 21-132, Academic Press, N.Y., 1969.
13. Kimura, M. (1983) **The Neutral Theory**. Cambridge University Press, London.
14. Fitch, W. (1977) *Am. Nat.*, **111**, 223-257.
15. Lake, J.A. (1987) A rate-independent technique for analysis of nuclei acid sequences: Evolutionary parsimony. *Mol. Biol. Evol.*, **4**, 167-191.
16. Sinsheimer, J.S., Lake, J.A., & Little, R.A. (1997) Inference for phylogenies under a hybrid parsimony method: Evolutionary-symmetric transversion parsimony. *Biometrics*, **53**, 23-38.
17. Bandelt, H.-J. & Dress, A. (1986) *Adv. Appl. Math.*, **7**, 309-343.
18. Halanych, K.M., Bacheller, J.D., Aguinaldo, A.A., Hillis, D.M., & Lake, J.A. (1995) Evidence of 18S ribosomal DNA that the Lophophorates are Protostome animals. *Science*, **267**, 1641-1643.