No. 31

COMPUTER SCIENCE MONOGRAPHS

A Publication
of
The Institute of Statistical Mathematics

INTEROGATE 1.0. EXPLORATION AND TESTING OF STATIONARITY, REVERSIBILITY AND CLOCK-LIKENESS IN SEQUENCE DATA

by

Peter J. Waddell, Hiroshi Mine and Masami Hasegawa

March 2005

Computer Science Monographs

The series of Computer Science Monographs publishes the statistical software and the articles on computer applications originating at The Institute of Statistical Mathematics. All communications relating to this publication should be addressed to the Editorial Office, The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan.

The electronic version of this issue is downloadable at
http://www.ism.ac.jp/

INTEROGATE 1.0. Exploration and Testing of Stationarity, Reversibility and Clock-likeness in Sequence Data.

Peter J. Waddell[1,3*], Hiroshi Mine[1,4], and Masami Hasegawa[1,2].

[1] Institute of Mathematical Statistics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo, 106-8569, Japan
[2] Department of Biosystems Science, School of Advanced Studies, The Graduate University for Advanced Studies, Hayama, 240-0193, Japan
[3] Present Address, Laboratory of Biometry and Bioinformatics,Graduate School of Agriculture and Life Sciences, University of Tokyo, 1-1-1 Yayoi Bunkyo-ku, Tokyo 113-8657, Japan
[4] Present address, 3rd Research Dept. 307U Hitachi, Ltd., Systems Development Lab, Tokyo
[*] For correspondence: waddell@ab.a.u-tokyo.ac.jp

Abstract

INTEROGATE 1.0 is a set of C programs to facilitate the evaluation and editing of molecular sequence data. The package is intended to assist in phylogenetic analyses by giving the user important statistics on the data in conjunction with inferring an evolutionary tree. These statistics can assist in understanding which sites/models may be most suitable for analysis, and also which sequences may be significantly atypical. The user may then want to consider the results of analyses removing sites and/or particular sequences. Conversely, they can be used after inferring initial trees to ask what part, if any, of the tree may be incorrect. The programs are: Freqnuc, FreqAA and FreqCodon for analyzing the stationarity, reversibility and relative rate of nucleotide, amino acid and codon sequences, respectively; Capture, for estimating the proportion of invariant sites; Pcons, for assessing properties of constant sites; Exclcon, for removing constant sites from data; Perfect, for removing more rapidly evolving sites from the data or "site stripping"; Perfect2, for removing the same sites from matched amino acid and codon data. The manual includes a detailed worked example of these programs use. This includes sequences used to reconstruct the evolutionary origins of mammals, which gave unanticipated inferred relationships within the superorder Supraprimates.

INTRODUCTION

INTEROGATE 1.0 is a set of programs to facilitate the evaluation and editing of molecular sequence data. The package is intended to assist in phylogenetic analyses by giving the user important statistics on the data in conjunction with inferring an evolutionary tree. These statistics can assist in understanding which sites/models may be most suitable for analysis, and also which sequences may be significantly atypical. The user may then want to consider the results of analyses removing sites and/or particular sequences. Conversely, they can be used after inferring initial trees to ask what part, if any, of the tree may be incorrect.

A particular strength of the package is the large number of tests for assessing non-stationarity and non-reversibility in nucleotide, amino acid or codon data. These tests are more sensitive than the standard tests in packages such as PAUP (Swofford 2002), yet still computationally tractable for large data sets and/or large numbers of character states (e.g. amino acids or codons). Many of these tests and methods were introduced in Waddell et al. (1999b).

In reference to a to a known or estimated phylogeny, these tests may be summed across a set of nonintersecting paths in order to avoid the loss of independence usually associated with multiple pairwise tests on evolutionary data. This approach is inspired by the use of pairwise lnL statistics in providing a bound on the likelihood of a tree (Waddell 1995). The programs will then indicate whether, overall, there is statistically significant evidence that the data violate stationarity and reversibility assumptions. A further option allows a full set of "non-parametric" parsimony-based tests of whether two lineages have evolved at the same rate (Waddell et al.1999a). These are useful tests of whether there may really be a molecular clock. Unlike parametric likelihood ratio tests these tests do not rely upon specific assumptions about how characters evolve except that characters are independent.

Full tables of test output between all pairs of sequences can be used to diagnose which parts of the data are causing the most violation of the null hypothesis of stationarity or reversibility. To facilitate further analysis, the tables of pairwise test results (or distances between sequences) are also output as NEXUS files. These may be read directly into PAUP and the whole matrix of results visualized as a tree (essentially a hierarchical cluster analysis of multidimensional data). This tree may indicate those sequences that are similar to each other in their overall properties and those which are most different. An alternative to hierarchical clustering to look at structure in the data are multidimensional scaling methods, and the out distance matrices may also be read into analysis packages such as R with minimal effort (R development core team 2004).

INTEROGATE also contains tools to allow the researcher to focus on the more conservative parts of the data and/or remove invariant sites. Removal of invariant sites is often important when using likelihood or distance based methods (e.g. Lockhart et al.1996). These programs allow the user to do this taking into account a unique character (e.g., base) composition of invariant sites (Waddell and Steel 1997).

Another set of programs remove what appear to be the least conservative or most rapidly evolving sites. This is called site stripping (Waddell et al. 1999b). In particular, removing sites that vary within a particular group or groups may be beneficial because this tends to shorten all edges and in particular the lengths of the edges connecting a group of species to the rest of the tree. In a non-parametric way, this also removes what appear to be the fastest evolving sites in each group. Assuming that all sites have the same evolutionary rates in different groups, these are the sites most likely to be involved in long edge attraction (Felsenstein 1978, Hendy and Penny 1989) and repulsion (Waddell 1995) effects that may mislead phylogenetic inference. In a severe form, most useful on large data sets, the programs can quickly reduce the data to a set of characters linking major groups (e.g., mammalian orders) that show minimal evidence of change or homoplasy within groups. Hopefully, these are the most reliable characters for inferring the relationships between groups.

None of these methods are fool proof. For example, with the removal of variable sites the total amount of data is reduced so stochastic errors may increase. There is always the risk that much useful data has been thrown away and what remains is not without problems itself. In a sense, these "site stripping" methods are a form of character reweighting, where the prespecified groups represent non-nested partitions on the "real" tree. For example, they may non-nested clades. It is hoped that these methods are robust since they make minimal assumptions about what the true model and phylogeny are (c.f., iterative reweighting methods that often make strong assumptions about what the phylogeny is, based on an initial analysis of the data at hand).

Used effectively and together, these methods can give a much better feeling for how robust the data are and which parts of the tree may be incorrect. They were used extensively in the published and unpublished analyses that have become the basis of the modern phylogenetic classification of placental mammals (e.g., Waddell et al. 1999c, Waddell et al. 2001).

EXTRACTING AND COMPILING:

Source code and executables are supplied for Linux, Unix (including OSX for Macintosh) and Windows machines. The programs are all written in ANSI C. Source code comes tarred and zipped into one file. To extract, first unzip with the command,

```
gunzip filename.tar.Z
```

then untar with,

```
tar -xvf final.ver.tar
```

If you have any trouble with the executables supplied, the programs compile easily once you have cc or gcc installed. To compile each program in its respective folder:
i) for folders with files "matrix.c" and "matrix.h" type,

```
(g)cc programName.c matrix.c -lm -o programName
```

ii) for folders without both "matrix.c" and "matrix.h" type

```
(g)cc programName.c -lm -o programName
```

The programs should then run by calling them. In Linux either set the path to the executables, or by go to the directory they are in and type "./programname.exe". The "./" sets the path to the program to that of the current directory. To run under OSX, open a window with the "terminal" program and go to directory containing the executable. Then use the same commands mentioned above for Linux. For Windows, open the folder where the program resides and double click the console application. A DOS-console window opens. From this point onwards, the executions of all programs are similar for all platforms. Upon typing a files name, a brief description of what the program does and how it runs will appear.

LIST OF PROGRAMS:

This software suite includes:

(i) Freq-programs: These will perform a wide range of pairwise tests on a set of data. These tests will detect evidence of non-stationarity (changing base/amino acid/codon usage) or non-reversible evolution. Non-stationarity tests include those based on approximate X^2 or G^2 (=G) statistics (Waddell et al. 1999b), or an asymptotically more accurate sum of squares (SS) test (Tavaré 1986). For testing for the non-reversibility of true evolutionary model, there are asymptotically exact X^2 and G^2 test statistics (Waddell and Steel 1997). Importantly, these programs automatically replicate the same tests with grouping of cells (character states) to ensure that expected frequency counts are high enough to allow a reasonable approximation to a χ^2 distribution under the null hypothesis of stationarity or reversibility. There is an option to output a full set of "non-parametric" parsimony-based tests of whether pairs of sequences are evolving at the same rate (Waddell et al. 1999a). Output is also available in NEXUS format for matrixes of test results.

(a). Freqnuc.exe – nucleic sequence data

(b), Freqaa.exe – amino acid sequence data

(c), Freqcod.exe – codon sequence data

(ii) Capture.exe: Using either the capture-recapture method of Sidow et al. (1992) or that of Waddell et al. (1999b), this program estimates what proportion of sites are invariant along with the s.e. of this estimate. It uses either nucleotide or amino acid sequences (for the Waddell et al. test) or codon sequences (for the Sidow et al. test).

(iii) Pcons.exe: Makes a test of whether the frequencies of the varied and the unvaried characters are equal or not. Works with nucleotide, amino acid or codon sequence data.

(iv) Exclcon.exe: Will exclude a specified fraction of unvaried sites either in proportion to the frequencies of all characters or just the unvaried characters. Works with nucleotide, amino acid or codon sequence data.

(v) Perfect.exe: Removes the most rapidly evolving characters in the data. The user specifies a set of groups of sequences and any site showing variability in these groups is removed (nucleotide, amino acid or codon sequence data). There are options on how gaps and singletons are treated.

(vi) Perfect2.exe: Reads in an amino acid data set and it corresponding codon data set. Based on user specified groups of sequences, it will remove characters (amino acids or codons) that show variability in the amino acid data, and will do this for both data sets (i.e., if a specific amino acid site is to be removed, so will the corresponding codon). Thus, it is like Perfect.exe but it is useful when the user wants a matching amino acid and codon data set (for example to do perform matching nucleotide and amino acid analyses).

DESCRIPTION:

This topic is divided into two sections:

i) Input – provides descriptions of the types of input formats the programs accept.

ii) Execution, Output and Worked Example – provides descriptions of running all the programs in INTEROGATE1.0 package. This is done with a worked example of topical data. The format and interpretation of output from each program is also discussed.

INPUTS

Each program in this package accepts input in a standard format. They have also been modified to accept a simple NEXUS format. The standard format is a basic non-interleaved input format that will work with both PHYLIP and MOLPHY programs. The NEXUS format is used by various other phylogenetic software packages (PAUP4.0, MrBayes, etc). These two formats are further explained in the following section, and examples are supplied with the programs.

Standard format:

Each program accepts a standard input format that is compatible with PHYLIP (Felsenstein) and MOLPHY (Adachi and Hasegawa,1996). A standard input file starts with two numbers; the first indicating the number of sequences and the second indicating the length of each sequence. On a new line, the name of the sequence is given. This should not use spaces or special characters and be 10 or fewer characters if it is to be used further with PHYLIP programs (for MOLPHY name length may be at least 25 characters). There is then a new line and the first sequence is a single continuous line of amino acid or nucleic acid characters (see figure 1 below). Please ensure there the sequences are a continuous stream without line breaks, carriage returns or other special characters.

```
5 11
dog
FFINIISLIIP
seal
FMINIISLIIP
cat
FMINVLSLIIP
horse
FMINVLLLIVP
rhino
FTI--LLLVIP
```

Figure 1 Standard input data format. Sequences must be of equal lengths and are assumed to be aligned.

NEXUS format:

The programs in INTEROGATE1.0 will read in a simple NEXUS format as long as the file extension '.nex' is used (i.e., filename.nex). The following figure illustrates this format:

```
#NEXUS

Begin DATA;
      Dimensions ntax=42 nchar=344;
      Format datatype=PROTEIN gap=-;
      Matrix
      [                               1           11          21…]
Ovis_canadenensis        CWTFMHRKFSSAPCEVYSSRNTAMEWHPHTPSCDIC…
Okapia_johnstoni         CWSFMHRKFSSAPCEVYSPRNAAMEWHPHTPSCDIC…
Lagenorhynchus_obscurus  CWIFMHRRFSSAPCEAYSPRNATMEWSSHTTPCDIC…
Balaenoptera_physalus    CWSFMHRKFSSAPCEAYSPRNATMEWSSHTTSCDIC…
.
.
.
Sus_scrofa               CWSFMHRKFSSTPCEVYSPRNATMEWHPHTLNCDIC…

;
End;
```

Figure 2 Sample non-interleaved NEXUS input format

EXECUTION, OUTPUT and WORKED EXAMPLE

The use of these programs will be illustrated with an alignment of 42 sequences of the RAG1 gene from Waddell and Shelly (2003) covering all orders of placental mammals. This gene has proven very useful in evaluating the phylogeny of the placental mammals. Despite the sequenced exonic region being ~800 bp long, trees based on this data (e.g., figure 3, Waddell and Shelly 2003) are highly congruent with prior expectations (e.g. Waddell et al. 1999c, Murphy et al. 2001, Scally et al. 2001, Waddell et al. 2001). The afore mentioned tree of figure 3 is, however, less congruent within the superorder Supraprimates (made up of the clades Euarchonta + Glires). The taxa Primates (represented here by human, *Homo*, and tarsier, *Tarsier*) + Dermoptera (flying lemur, *Cynocephalus*) + Scandentia (tree shrew, *Tupaia*) = the clade Euarchonta , while Rodentia (mouse, *Mus*, porcupine, *Hystrix*, south American rodent, *Dolichotis*) + Lagomorpha (rabbit, *Sylvilagus*, pika, *Ochotona*) = Glires. While the internal structure of Supraprimates is unresolved in the sense that Bayesian posterior probability (*pp*) values for edges were less than 0.95 (assuming a specific model and an uninformative prior), this part of the tree is somewhat unusual. For example, the flying lemur is placed sister to the rodents and the tree shrew, lagomorphs and primates are not only together, but the tree shrew is sister to tarsier with a fairly high *pp* value. In particular, we are interested in diagnosing possible reasons why the prior expectations of the flying lemur sister to tree shrew or flying lemur sister to primates are not favored in these analyses.

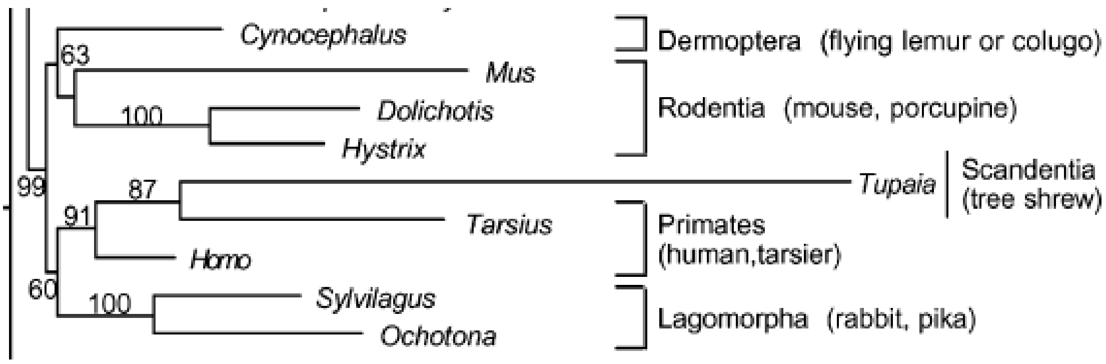

Figure 3. Relationships within Supraprimates based on a RAG2 alignment. Values are pp values under an invariant sites plus gamma HKY model. Reproduced from figure 1 of Waddell and Shelly (2003).

Freqnuc.exe

We begin with a diagnosis of base composition. Freqnuc.exe tests for stationarity of nucleotide composition, a reversible evolutionary process and clock-like evolution. It will do this via all possible pairwise tests and as a sum of user specified pairs. If these pairs are non-interesting on the true phylogeny, and if there is stationarity, then asymptotically each test result will be close to independent and the summed statistic will be close to a chi-square (χ^2) random variable with summed degrees of freedom (d.f.). Here, asymptotically means with very long sequences.

Taking our example data "ragCodon" we obtain the following output when the program is run (here ">" indicates a prompt for input):

```
>Freqnuc.exe
freqnuc Ver. 0.45
Please input the NUC file name > ragCodon
    1   295 R : unknown character, eliminated.
...
   41   236 Y : unknown character, eliminated.
Please input the OUTPUT file name > example1
  1 : Ovis_canadenensis     2 : Okapia_johnstoni      ...
 40 : Cyclopes_didactylus  41 : Choloepus_hoffmani   42 : Dasypus_novemcinctu
Would you like to specify outgroups for performing relative rate test
( atmost two! ) ?
(y/n) > y

If you specify more than two taxa, the program will only take first two!

please specify taxa numbers :
41 42
Would you like to specify more outgroups to exclude from the relative rate test?
(y/n) > y

please specify taxa numbers :
1 … 22 32 … 40

Which pairs of taxa indicate non-intersecting paths?
Or press Enter/Return for all pairs.
> 24 23 25 26 30 31 27 28
Do you want the output to be in Nexus file format?
(y/n) > y
Please wait...done.
```

The program reads in the specified file, in this case "ragCodon", which is a nucleotide file that starts at the first position of a codon and continues in multiples of three until the end. It then lists character states it excludes, in this case nucleotide ambiguity characters. It asks for an output filename, and in this case is given "example1". Species are then numbered according to input order and their names listed. If you would like to have the program run relative rate tests, then type "y" then specify up to two taxa as outgroups, here taxa 41 and 42. The program then asks if you would like to discard any other taxa from further analysis. In this example we are focusing on just Supraprimates, so we list all other taxa for exclusion. The user is then asked to indicate pairs of taxa indicating nominally non-intersecting paths in the true phylogeny. These are entered in pairs without punctuation, the pairs here being (23,24), (25, 26), (30,31), (27, 28). Having some expectation of the true phylogeny and edge lengths, this set of paths sum nearly all the edges in the subtree of Supraprimates sequences, leaving out the relatively short external edge to *Hystrix*. In this case, there is nearly zero chance these paths really intersect if our expectations of mammalian phylogeny are correct.

Finally, the program asks the user wants a NEXUS file of output to be produced in order to facilitate hierarchical clustering using PAUP.

The output as shown below, actual output being in courier font and discussion of the output is in Times font.

```
example1 output
freqnuc Ver. 0.45

                               Ungrouped                    Grouped

    Species                    Sym        Stat  St.All SS       SymG    df Sta2.5 df Sta5  df    SS5

  23 Cyncephalu           X2         58.8    96  22.77    57.1    58.4  5     96  3    96    3  57.1
  24 Tupaia_tan           G2         66.1 103.6  22.85            64.6  5  103.6  3 103.6    3
  25 Homo_sapie           X2         23.6  29.2   3.43    16.1    14.9  4   29.2  3  29.2    3  16.1
  26 Tarsius_sy           G2         28.5  30.1   3.43            15.6  4   30.1  3  30.1    3
  27 Mus_muscul           X2          5.9   2.2   0.38     1.6     5.9  6    2.2  3   2.2    3   1.6
  28 Dolichotis           G2          6.1   2.2   0.38             6.1  6    2.2  3   2.2    3
  30 Sylvilagus           X2         10.6     4   0.43     4.6     2.6  4      4  3     4    3   4.6
  31 Ochotona_s           G2         13.4   4.1   0.43             2.6  4    4.1  3   4.1    3
     Overall              X2         98.9 131.4  27.01    79.4    81.8 19  131.4 12 131.4   12  79.4
                          G2        114.1   140   27.1            88.9 19    140 12   140   12
```

The output begins with a table, like that above, that gives results for just the specified paths and the summed results. The sequence names and their numerical indices are listed in the leftmost columns. Reading the first line, results for tests without grouping cells are given first. To the right are the same tests using the automatic grouping of cells to have expected values of at least 2.5 or 5. The X^2 and G^2 test statistics are listed for each pair of sequences. The first number is 58.8. This is the X^2 result for an ungrouped test of symmetry of the pairwise divergence matrix, **F**, along the path between species 24 and 23. The number directly below it, 66.1, is the G^2 result for a test of symmetry of **F** (or model reversibility) for the same path. (See Waddell and Steel 1997 for further details of this test) The X^2 and G^2 results alternate down the table. The next column headed "Stat" lists results for the X^2 and G^2 results for the stationarity test (Waddell 1995, Waddell et al.1999b). The column "St.All" does likewise for a test including all sites (the "traditional" conservative test, e.g. like those in PAUP, e.g., see Swofford et al. 1996). The column SS is the sum of squares GLS test of stationarity (Taveré 1986), with value 57.1 for this first pair.

The next set of columns headed "Grouped" repeat the tests with grouping of cells. SymG repeats the earlier G^2 test of symmetry for grouped cells, so that expected values are at least 2.5. Grouping is expected to improve the convergence to the test statistics under the null hypothesis to a χ^2 distribution. The “df” columns indicates the degrees of freedom after grouping from this pairwise **F** matrix, in this case the number 5 appears, a drop from 6 degrees of freedom without grouping. The next four columns repeat the tests of stationarity with grouping so expected values are at least 2.5 and 5. The final column gives the SS5 results, that is the GLS test with expected values of at least 5 and the same degrees of freedom indicated in the previous “df” column. The most robust tests of stationarity are those in the SS5 column. These clearly show large values for the pairs *Tupaia* and *Cyncephalus* and also *Homo* and *Tarsius*.

The final two rows of the first table give the results of tests summed over the selected paths. The summed result clearly indicates non-stationarity. For example, with SS5 the summed statistic is 80.2 with 12 degrees of freedom. A χ^2_{12} distribution has its most extreme 1% of values greater than ~26.2 and the test statistic here is much larger. For symmetry or reversibility the summed grouped results are 82.8 (X^2) and 88.2 (G^2) with 19 degrees of freedom, while a χ^2_{19} distribution has its most extreme 1% of values greater than ~36.2.

More formally, to make a formal test of non-stationarity consider the following form. In this case it is a test of stationarity using the new non-intersecting summation method, grouping and the GLS test statistic,

Set α, the probability of a type1 error (rejecting Ho, when Ho is true), to 0.05

Ho: The sequences are stationary in base frequency

H1: The sequences are non-stationary

The test statistic is 80.2with 12 d.f., p <<0.05, therefore reject Ho in favor of H1, the sequences have undergone non-stationary evolution

Some of the test results will be closely correlated since non-stationarity coincides with non-symmetry of **F**. It is possible to have data that are non-reversible but observe stationary frequencies, but this requires unlikely conditions. Note also that symmetry of **F** arises in two cases, these being a reversible model, or a clock. If the data are close to clock-like it will be harder to detect a non-reversible model. If evolution is clock-like and the only non-stationarity is a non-equilibrium root-frequency, in which case all tips evolve towards the same expected value, this too will be hard to detect.

As is elaborated further below, with this type of test, the impact of alternative pairwise pathsets that may cross over on short internal edges should be considered. A contrived example here would be *Homo*/*Cyncephalus* , *Tupaia* /*Tarsius, Mus/Ochotona,* and *Dolichotis/Sylvilagus.* In this case the summed grouped SS5 test statistic is 40.6 with 12 d.f. and the summed grouped SymG value is 50.8 (X^2) and 55.9 (G^2) with 21 d.f. Both tests are still clearly significantly large compared with their matching χ^2 distributions and we have no reason to doubt that the data are in fact non-stationary. The reason for the drop in the size of the test statistic is that the pair *Tupaia* /*Tarsius* contains the two most deviant taxa with respect to base composition and they are deviating in the same general direction.

To summarize, the tests of stationarity use three statistics, G^2, X^2 and a Generalized Sum of Squares (GLS) to compare the base composition of sites that change between two species. The first two statistics may be close to $\chi^2{}_3$ distributed for a single pair (where the subscript indicates with 3 degrees of freedom), but do not converge to χ^2. They compare the sums of row and columns of a square matrix (ignoring just the diagonal). However, each value in this matrix will appear in both a row and a column, therefore row and column sums are not independent, but correlated. Failure to adjust for this correlation tends to make the observed values of G^2 and X^2 biased upwards. Therefore, the program also makes a GLS test of stationarity. This, with long sequences, will converge to χ^2 as this takes into account the aforementioned correlations (the test involves estimating and using the inverse of the variance covariance matrix of row and column sums).

The tests for a reversible evolutionary process are from Waddell and Steel (1997). They also use G^2 and X^2 statistics comparing the *ij*-th and *ji*-th entries of the square divergence matrix, **F**. These pairs of numbers are much closer to independent of each other and given independence of characters follow a multinomial distribution. With long sequences, these test statistics converge to χ^2 with degrees of freedom equal to $b(b\text{-}1)/2$ where b is the number of states (with 4 nucleotides this is 6 degrees of freedom).

In the output, the results according to the user specified pairs (or by default all pairs in the order of the input species) are shown first along with their sum. If these pairwise paths are non-intersecting on the true phylogeny, then the summed statistic will be close to a χ^2 distribution with degrees of freedom equal to the sum of the degrees of freedom of the component pairwise tests. Unless the true phylogeny is a star-like tree for an even number of sequences, it will be necessary to leave out some edges of the tree in constructing non-intersecting paths. Alternatively, if internal edges are short, then counting multiple pairs that cross them tends to have little effect on the overall result. A recommended mode of operation is to obtain a reasonable estimate of the tree then plot out one set of non-interesting paths that is near maximal in sum of lengths. Then, plot out another set of near maximal length paths that may cross each other along poorly resolved internal edges (hence may not cross if the inferred tree is incorrect). Run both sets and if they both reject the null hypothesis, the result is probably trustworthy. This can also be thought of as the "hand" version of a robust Bayesian method. This would calculate a minimal path on each optimal tree from bootstrap replicates, and average the test statistics across these trees.

When paths are specified, the matrices of pairwise output for testing stationarity and symmetry of **F** are for just these sets of sequences. Inspecting the table of results, the largest value within Supraprimates is clearly between *Cyncephalus* and *Tupaia*. This goes a long way to explaining why trees of this data do not support these two as sister taxa or recover Euarchonta.

The output file then contains a table like that below.

```
Nucleotids                  Adenine Cytosine  Guanine  Thymine
 23 Cyncephalu                 229      201      192      177
           proportion          29       25       24       22
                   X2          3.229    0.011    0.282    0.679
 24 Tupaia_tan                 122      185      167       89
           proportion          22       33       30       16
                   X2         32.560    1.054    5.294   35.984
 25 Homo_sapie                 219      203      202      176
           proportion          27       25       25       22
                   X2          1.200    0.061    0.031    0.557
 26 Tarsius_sy                 193      223      216      165
           proportion          24       28       27       21
                   X2          0.529    2.768    1.365    0.011
 27 Mus_muscul                 220      201      197      176
           proportion          28       25       25       22
                   X2          1.359    0.011    0.031    0.557
 28 Dolichotis                 217      201      209      171
           proportion          27       25       26       21
                   X2          0.913    0.011    0.452    0.129
 30 Sylvilagus                 208      192      207      192
           proportion          26       24       26       24
                   X2          0.105    0.282    0.282    3.947
 31 Ochotona_s                 219      190      206      185
           proportion          27       24       26       23
                   X2          1.200    0.452    0.212    2.085
 Overall                        1627     1596     1596     1331
                   X2           41.096    4.652    7.950   43.948
                   G2           46.361    4.586    8.218   51.058
 Expected                      203.38   199.50   199.50   166.38
```

This table gives base frequencies and the percentage of each base in each taxon along with an X^2 test statistic of the observed number against the expected number based on the mean composition of all taxa. This is like the test made in PAUP. These X^2 statistics give some idea how different each cell is from the mean. The summed numbers however have no known distribution (due to unaccounted for correlations) and should not be used as a formal test of stationarity. Notable results are that many sequences differ from the mean by substantial amounts, with *Tupaia* and *Tarsius* being enriched in G and C and depleted in A and T.

```
X-square test for symmetry df = 6

Species                     24        25        26        27        28        30        31

24 Tupaia_tana            58.8       6.5      26.2       3.1      11.2      11.7      11.9

23 Cyncephalus_volans  -              56      24.3      43.4      40.8      56.9        63

25 Homo_sapiens_publ   -       -              23.6       3.4        12        13      13.4

26 Tarsius_syrichta    -       -         -              14.6      12.9      18.3      18.8

30 Sylvilagus_sp       -       -         -         -               5.9       9.7       9.7

31 Ochotona_sp         -       -         -         -         -              10.7       4.6

27 Mus_musculus_publ   -       -         -         -         -         -              10.6

28 Dolichotis_patagonu  SX2Sym  = 595.0  Sdf = 168
```

Following this, there are many fully square tables (matrices) like that above, giving all possible pairwise statistics amongst the selected taxa (that is, all that were included in a pairwise path). If you want all taxa in the alignment listed, then hit return when asked to specify paths. The first such matrix is shown in full above. Each matrix begins with a description of the statistic calculated and how many degrees of freedom it has if this number is constant and it is a test statistic. If cells are grouped then an asterisk follows each statistic and the following number is the grouped degrees of freedom for just that pairwise statistic. Also provided are matrices of the Euclidean and Manhattan distance between the base composition of each pair of sequences. At the end of some tables is an overall sum of the statistic across the whole table and the summed degrees of freedom (e.g. SX2Sym = 620.2, Sdf = 168). Note, these sums do not take into account correlations and therefore should not be used for making overall tests.

Below is a list of all matrices that may be output.

```
 G-square test for symmetry df = 6

 X-square test for stationarity df = 3

 G-square test for stationarity df = 3

 X-square test for stationarity (all sites) df = 3

 Euclidian distance Matrix ( all sites )

Manhattan Distance Matrix ( all sites )

 Proportional Euclidian distance Matrix ( all sites )

Proportional Manhattan Distance Matrix ( all sites )

 G-square test for stationarity ( all sites ) df = 3
  Species                          23           25            26            30

 Test of base composition stationarity ( by GLS SS ) df = 3

 X-square test for symmetry ( grouped cells )

 G-square test for symmetry ( grouped cells )

 X-square test for stationarity ( grouped cells-2.5 )

 G-square test for stationarity ( grouped cells-2.5 )

 X-square test for stationarity ( grouped cells-5 )

 G-square test for stationarity ( grouped cells-5 )

 Test of base composition stationarity ( by GLS SS5 )
  Species                    23       25       26       30       31       27       28

 Relative Rate test using First Outgroup

 Relative Rate test using Second Outgroup

 Relative Rate for Liberal Case ( using both outgroups )
  OGspecies                        2            3             4             5

 Relative Rate test for Conservative Case ( using both outgroups )
  OGspecies

                        2            3             4             5
```

For the relative rate tests, all species other than those listed as “outgroups” or “outgroups to exclude” are included in the table. If a smaller table is desired, list any species to exclude as “outgroups to exclude.” Only the fist two species listed as outgroups are used to make these tests. The number for species *i* and *j* in these tables is an X^2 statistic following Waddell et al. (1999c). Two further tables are given which also feature tree independent tests of relative rate, but they use two outgroups and parsimony-like unambiguous reconstructions of characters. The liberal relative rate test counts a pattern such as CAAT (where the order is outgroups then ingroups) as indicating a change for the second ingroup, where as the conservative test does not count this and thus only counts patterns where all character states except one ingroup are the same.

At the very end of the output file, there are warnings advising how many entries in symmetricized **F** matrices were small before grouping. Following this is a list of how many entries in the expected π (here Pexp) base composition vector where small. Finally, a list of cases where the GLS SS test could not be performed on the raw F matrix since the variance covariance matrix could not be inverted (in none).

```
Warning.

 No. of entries in Fsym that are
  0  :   None
  0.5:   1( 23,25 )
  1-5:   2( 23,24 )   2( 23,25 )   3( 23,26 )   2( 23,27 )   4( 23,28 )
         4( 23,30 )   2( 23,31 )   2( 24,25 )   2( 24,26 )   1( 24,27 )
         2( 24,28 )   2( 24,30 )   2( 24,31 )   3( 25,26 )   2( 25,27 )
         3( 25,28 )   4( 25,30 )   3( 25,31 )   2( 26,27 )   2( 26,28 )
         3( 26,30 )   2( 26,31 )   2( 27,28 )   1( 27,30 )   2( 27,31 )
         4( 28,30 )   2( 28,31 )   3( 30,31 )

 No. of entries in Pexp that are
  0  :   None
  1  :   None
  2-5:   None

 Instances when SS cannot be calculated (as V cannot be inverted)
         None
```

Freqnuc.exe is one of three programs in the Freq suite of programs. It is used to test whether the nucleotide data indicate a non-reversible or non-stationary model. If so, then one should consider using the LogDet distances, and not just the general reversible distances (or, a non-reversible model for sequence evolution). Looking up the complete table of pairwise tests for homogeneity and stationarity in the output, one can often gauge which taxa seem most aberrant. A user may consider removing these to assess how they affect the overall tree.

In order to facilitate interpretation of matrices of output and identify taxa that in some way share similar properties, the matrices are also output in NEXUS format and can be read straight into PAUP. In order to work on one particular matrix all that need be done is to comment out (enclose) with square brackets the other results. Figure 4 shows the results of analyzing three of these matrices in PAUP. When making such analyses in PAUP, we prefer a consensus tree approach like that used in Waddell and Kishino (2000). We often find a consensus of NJ and OLS+ to be informative, but not overly detailed (e.g. small edges often collapse suggesting ambiguity in their representation). The consensus tree for Euclidean distances is shown in figure 4.The impression in figure 4a,which is based on Euclidean distances, is that base composition is generally similar in all species except *Tarsier* and *Tupaia* which are divergent in the same direction. Figure 4b based on GLS test distances is similar and confirms that only *Tupaia* and *Tarsier* deviate significantly. Note, this can be discerned visually by assuming the weighted tree reasonably represents the underlying pairwise distances and for any of these to correspond to test statistics significant at the 2% level, a distance should be greater than ~10, which with the two exceptions, these are not (an exact significance level to work with is somewhat arbitrary, as with most cases of multiple testing without a clear prior hypothesis). Figure 4c shows that the same general pattern emerges using as a distance the reversibility test statistic, and there is a hint of *Ocotona* diverging most from the others in its own direction. An alternative and *a priori* equally valid way to look at these output are with other multidimensional visualization techniques including multi-dimensional scaling (e.g., R core development team) and for nucleotides, the 3-axis plots of Waddell (1995).

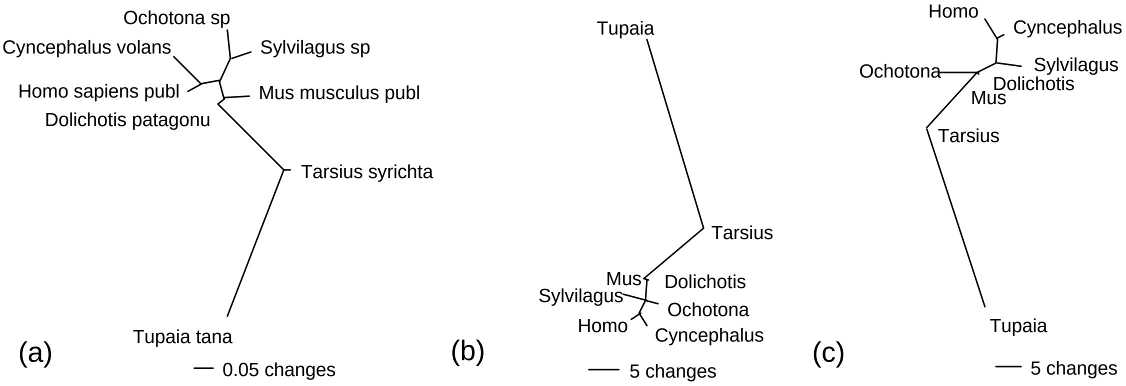

Figure 4. Tree based visualizations of distance matrices of (a) Euclidean base composition distances (b) GLS stationarity test distances (c) G^2 (Kullback-Liebler) reversibility test distances between all pairs of taxa. In figure (c) Mus and Dolichotis are located at the end of zero length edges that join the tree at the junction to Ochotona. Note the is different in each figure, and the orientation is chosen arbitrarily by PAUP when printing the trees. For figure (b) a distance of more than 10 has a p-value of ~0,02, for (c) a distance of 15 has a p-value of ~0.02. The trees are produced by taking a strict consensus tree of an NJ and on OLS+ tree for each data matrix. The edge lengths on this consensus tree are then estimated using OLS+ (see Waddell and Kishino 2000 for more details and examples). PAUP labels edge lengths as changes, but in this case these are actually distances.

Another aspect of the output are the relative rate tests. These test for clock-like evolution without any explicit assumption of the evolutionary tree. They are also model independent. This can be useful in that model-based tests can give the wrong answer even if the tree is correct. For the Supraprimates, using two divergent xenarthrans as the outgroups. the resulting table is obtained.

Relative Rate test for Conservative Case (using both outgroups)

Cyncephalus	0	223.4	1	149.1	23.4	140.3	138	132.6	11.6
Tupaia	223.4	0	248	14.4	148.5	17.1	19.2	19.4	165.8
Homo	1	248	0	176.3	34.7	162.4	159.3	158.3	19.2
Tarsius	149.1	14.4	176.3	0	77.2	4.9[10]	1.7	2.1	95.2
Mus	23.4	148.5	34.7	77.2	0	67.3	64.3	62.5	1.9
Dolichotis	140.3	17.1	162.4	4.9[10]	67.3	0	7.2[10]	5.2[10]	85.9
Hystrix	138	19.2	159.3	1.7	64.3	7.2[10]	0	1.1[100]	82.3
Sylvilagus	132.6	19.4	158.3	2.1	62.5	5.2[10]	1.1[100]	0	94.2
Ochotona	11.6	165.8	19.2	95.2	1.9	85.9	82.3	94.2	0

To return to our original question, why does *Cyncephalus* end up with the rodents and *Tupaia* within Primates? Taking everything together, it seems apparent that *Tupaia* groups with *Tarsier* in the estimated tree of figure 3 due to these two taxa showing accelerated rates of evolution and convergent base composition. This gives rise to a high *pp* value. Why *Cynocephallus* is not close to *Homo* or why Glires are not recovered are less immediately apparent. The proximity of the *Tarsier/Tupaia* group to *Homo* may be repelling *Cynocephallus*. There is no immediate explanation for why Glires is not recovered, or why lagomorphs end up with Primates. However, unlike the *Tupaia/Tarsier* grouping these groups do not have high *pp* values and it may simply be a lack of resolution. Indeed, rerunning the analysis with *Tupaia* and *Tarsier* excluded, suggests almost no resolution amongst the deep edges of this subset. That is, the biologically most reasonable tree (containing Rodentia, Glires and Primatomorpha = Primates + Dermoptera, here being *Homo, Tarsier* and *Cynocephalus*) is within 2 lnL units of the best tree, and a tree with Primatomorpha is within 0.1 lnL of the best tree. The BIC approximation suggests such trees will not have distinct *pp* values. One further point of interest from figure 4 is that for these data *Mus* does not seem particularly abnormal in base composition or evolutionary rate. This is in marked contrast to mtDNA sequences (e.g., Waddell et al. 1999b).

FreqAA.exe and FreqCodon.exe

Running these programs is basically the same as running Freqnuc. However the input data will need to be amino acid sequences using the standard one letter code or codon sequences starting with the first position and remaining in frame, respectively. Following the example data, the first part of the output of FreqAA is shown below. The string of zeros for the SS GLS test are due to the sparseness of the estimated variance covariance matrix, preventing it being inverted. In the last columns it can be seen considerable grouping was necessary, as would be expected, since there are 20 amino acids, only ~266 residues, and frequencies are unequal in nearly any sequence (e.g., tryptophan and cytosine tend to be rare). An interesting point is that the approximate test statistics for stationarity based on X^2 and G^2 have converged strongly to the more exact GLS values. This is expected, since with more cells the overall correlation of a row with a column tends to decrease. Consistent with this, the least grouped results are the most similar to GLS values. The overall result is that there is evidence of non-stationarity of amino acid frequencies (observed statistic 33.1, 0.05 > p-value > 0.01, based on χ^2_{21}) but no evidence of a non-reversible model. Inferring which species show strongest and significant divergence in amino acid composition is left as an exercise to the reader. The similarity of results with nucleotide frequencies suggests there may be a causal link, which would be made stronger and directed by confirming that third bases of codons moved in the same direction. Testing of codon frequencies using freqcod.exe reveals clear evidence of non-stationarity, but an insignificant result for non-reversibility.

Statistics for stationarity and reversibility of amino acid frequencies within Supraprimates.

```
                    Ungrouped                  Grouped
  Species              Sym    Stat  St.All SS    SymG    df     Sta2.5 df     Sta5  df     SS5
23 Cyncephalus      + X2   38.6 36.5  13.69     0     8.7     5   22.4      9  21.1      7    20
24 Tupaia             G2     52 43.5   13.9           9.7     5     24      9  22.6      7
25 Homo             + X2   29.3 25.6   5.04     0     0.5     2     10      5   8.4      4   7.6
26 Tarsius             G2      40.5     32  5.09           0.5      2  11.1      5   8.9     4
27 Mus              + X2   32.3 16.4   4.91     0     2.7     3    5.6      6   4.5      5   4.1
28 Dolichotis          G2      44.7   19.9  5.01           2.7      3   5.8      6   4.6     5
30 Sylvilagus       + X2     25   22   2.89     0     2.6     2    5.4      4   1.9      2   1.4
31 Ochotona            G2      34.7   29.4   2.9           2.6      2   5.6      4   1.9     2
                                100.
  Overall             X2  125.3    5  26.53     0    14.4    12   43.5     24    36     18    33
                                124.
                      G2  171.9    8   26.9          15.5    12   46.4     24    38     18
```

Statistics for stationarity and reversibility of codon frequencies within Supraprimates.

```
                      Ungrouped                  Grouped
  Species               Sym       Stat   St.All SS SymG     df     Sta2.5 df   Sta5    df    SS5
23 Cyncephalu      + X2         93   98.2  70.34  0       14     7   28.1  16    25.8    11  30.3
24 Tupaia_tan        G2      128.9  130.7  85.74        15.1     7   33.5  16    30.5    11
25 Homo_sapie      + X2       57.3   64.4  27.31  0      8.9     6   16.7  13    23.2     8  25.2
26 Tarsius_sy        G2       78.1   83.5  30.23         9.4     6   17.4  13    24.6     8
27 Mus_muscul      + X2       75.4   68.7  38.13  0      5.4     9   21.3  19    21.8    11  17.4
28 Dolichotis        G2      103.7   89.4  43.31         5.6     9   22.2  19    22.9    11
30 Sylvilagus      + X2       64.5     64  24.81  0      4.4     5   29.5  12    21.5     8  17.9
31 Ochotona_s        G2       88.9   84.6  27.06         4.5     5   34.7  12    22.7     8
  Overall            X2      290.3  295.3  160.6  0     32.8    27   95.5  60    92.2    38  90.7
                     G2      399.6  388.2 186.35        34.6    27  107.7  60   100.7    38
```

Capture.exe

This program will estimate how many sites are invariant using either the methods of Waddell et al. (1999b) for amino acid or nucleotide data, or the methods of Sidow et al. (1994) for codon data. For the former method, two groups need to be specified. At least one of these should be monophyletic.

i) At the first prompt the user is asked to enter an input file with its extension.
ii) Next, the user is asked to enter a name for the output file, where the results will be stored.
iii) The user is asked whether the data is codon sequence. Enter ‘y’ for yes or ‘n’ for no if the input data is an amino acid or nucleic acid data (and not a coding sequence).
iv) Then user will be asked to divide the taxa into two groups. List the index numbers of the taxa in the first group. Separate indices with a space.

For this example, Supraprimates was selected as one group and all other taxa constituted the other. The estimated proportion of invariant sites was ~11.4%. Using the codon data the estimate was ~12.8%. These numbers are within sampling error of each other. The number of amino acid positions that were totally constant (no "-" or "*" characters either) was 53.

```
capture Ver. 0.02
1: Ovis …
42: Dasyp
Group1 : 23 24 25 26 27 28 29 30 31
Group2 :  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 32 33 34 35 36
37 38 39 40 41 42
Result : V1 = 171 V2 = 154 V12 = 111
Capture-Recapture estimates
                Invariant sites                 Pinv
Standard        29.8 (s.e.=7.048)       0.11145 (s.e.=0.02640)
Unbiased        30.0 (s.e.=6.966)       0.11223 (s.e.=0.02609)

Total Sites (Codons)    : 801 (267)
Constant Sites (Codons) : 144 (48)
Result : V1 = 199 V2 = 174 V12 = 154
Capture-Recapture estimates
                Invariant codons                        Pinv
Standard        42.2 (s.e.=2.921)       0.15789 (s.e.=0.01094)
Unbiased        42.2 (s.e.=2.899)       0.15803 (s.e.=0.01086)
```

Pcons.exe

This program tests whether the character composition in the varied and the constant sites of the data are the same or not. The output begins with a table of the frequencies of each character among the constant sites (marked “Fc” in the output below). Cystine (CYS) is the highest at 14. PFc is the predicted proportion of sites that should be constant if the character state composition of all sites is the same (this is just the overall proportion of this character state times the number of constant sites). X2c is the X^2 statistic for the comparison of these two numbers (nominally 1 degree of freedom, but cells are not independent). Fv, PFv and X2v are the corresponding numbers among the variable sites. At the end of the table the total number of sites, C, is listed (here 344), then the number constant, Sc (here 53) and some other numbers for checking results. The test statistics are SX2c or the sum of X^2 for the constant sites (here 61.97) and SX2v the sum of X^2 for the variable sites (here 11.27). Under the hypothesis that the frequencies of these two types of character are the same in expectation, the sum of these two statistics (= 73.24) will converge asymptotically to a χ^2 value with degrees of freedom equal to the number of character states minus one (in this case 20 -1 = 19). The p-value under the null hypothesis is extremely low (< 0.0001). Care should be taken that expected values are in the range that gives reasonable convergence to a X^2 else the test result could be spurious. In this case expected values of constant sites are moderately even and between 1 and 4. Although there are a couple of constant sites with low expected frequencies, these do not contribute much to the overall statistic. The predominant contribution and reason for rejecting the null hypothesis are that cystine residues have a reasonably large expected value. It can be concluded the test result is not spurious and there really is a difference in the residue frequencies for these two classes of sites.

i) At the first prompt the user is asked to enter an input file with its extension. For example, file aatest.dat, user will input ‘aatest.dat’ and not just ‘aatest’.
ii) At the second prompt, user is asked whether to consider the symbol ‘-‘ as being a different state to others or to be effectively ignored. The user may judge these as absence of evidence (note: insertions or deletions do indicate

the region may have a higher than usual rate of evolution, so to be conservative, these would be treated as a different state).

iii) At the next prompt user is asked to enter a name for the output file, where the results will be stored.

iv) The user is asked to select sequences to be included in the calculations. Enter the respective taxa number from the list separated by a space.

v) Next the calculations are and results are printed on the screen and stored in the specified output file.

Output of Pcons with Rag amino acid data.

```
Sample :  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27
28 29 30 31 32 33 34 35 36 37 38 39 40 41 42
AA :   Ala   Arg   Asn   Asp   Cys   Gln   Glu   Gly   His   Ile
Fc :     2     8     3     5    15     3     5     5     9     1
PFc:   6.1   8.3   2.8   2.9   6.7   3.9   5.6   2.7   5.8   4.6
X2c:   2.8   0.0   0.0   1.5  10.2   0.2   0.1   2.0   1.8   2.8
Fv :  13.8  13.5   4.3   2.5   2.4   7.1   9.6   1.9   6.0  10.9
PFv:   9.7  13.2   4.5   4.6  10.7   6.2   9.0   4.3   9.2   7.3
X2v:   1.7   0.0   0.0   1.0   6.4   0.1   0.0   1.3   1.1   1.8

AA :   Leu   Lys   Met   Phe   Pro   Ser   Thr   Trp   Tyr   Val
Fc :    17     7     0     4     5     4     2     2     0     6
PFc:  10.1  10.1   1.9   3.6   5.1   9.1   3.9   0.8   1.6   6.7
X2c:   4.7   1.0   1.9   0.0   0.0   2.9   0.9   2.0   1.6   0.1
Fv :   9.1  19.2   4.8   5.4   8.1  19.6   8.1   0.0   4.2  11.4
PFv:  16.1  16.1   3.0   5.8   8.0  14.5   6.2   1.2   2.6  10.7
X2v:   3.0   0.6   1.2   0.0   0.0   1.8   0.6   1.2   1.0   0.0

C = 267   Sc = 103   Sv = 164   SFv = 162   SPFc= 102   SPFv = 163
SX2c = 36.4444  SX2v = 22.8888
```

Exclcon.exe

This program reads in a file, asks you if the data are amino acids, then excludes a certain number of constant sites in proportion to either the overall composition or the frequencies at the constant sites alone. In order to estimate how many of which type of site to exclude the program multiplies the number to be excluded by the proportion of that character state among the constant sites. For example, for ARG the calculation is 39 x 3.77/100 = 1.47. Normal rounding would see 1 such site excluded. However, the program rounds such that the total number of excluded sites is exactly 39, so the cut off for rounding may be slightly larger or smaller than 0.5 (in this case it was slightly smaller).

i) At the first prompt user is asked to enter an input file with its extension.

ii) At the next prompt user is asked to enter a name for the output file, where the results will be stored.

iii) This program treats '-' as a different state. This is indicated to the user during execution, by an onscreen warning. this means that a column of alone state except for one of these characters is not considered constant.

iv) At the next prompt user is asked whether the data is an amino acid sequence. Enter 'y' for yes or 'n' for no if the input data is a codon or nucleic acid data.

v) Subsequently the user is asked to choose the species that he/she wants to include in the output file. Select the preceding number of the taxa in order to choose the species as with other programs.

vi) The total number of sites and total number of constant sites are printed on the screen. User is now asked to enter the number of sites he/she wants to exclude from the data.
vii) The next option is whether the constant sites are removed in proportion to the total number of constant sites. The program "XXX" may be used to test for a significant difference in these compositions.
viii) The number of each type of character excluded is written to screen.

Below is output of Exclcon.exe for the ragaa.nex file after excluding 39 constant sites in proportion to the frequencies at the constant sites. “Pr” indicates the percentage of such amino acids among the constant sites, while “Ex” indicates how many of each type were actually excluded.

```
Total = 267, Constant = 53
How many constant sites do you want to exclude? > 30
Remove constant sites in proportions of constant sites? (y/n) y
Warning : number of sites adjusted.
Warning : number of sites adjusted.
Warning : number of sites adjusted.

 Constant = 53, Excluded = 30

 AA   Ala   Arg   Asn   Asp   Cys   Gln   Glu   Gly   His   Ile
 Pr  0.00  3.77  1.89  5.66 26.42  0.00  3.77  1.89  7.55  1.89
 Ex     0     1     1     2     8     0     1     1     2     1

 AA   Leu   Lys   Met   Phe   Pro   Ser   Thr   Trp   Tyr   Val
 Pr 15.09  5.66  0.00  5.66  7.55  5.66  1.89  1.89  0.00  3.77
 Ex     2     2     0     2     2     2     1     1     0     1
```

Perfect.exe

Perfect.exe allows the user to remove sites that show variability in specified groups. Selects sites that show evidence of slow rates and conservatism. This program allows the user to select a subset of the sites that are generally more slowly evolving and as such, should be more robust to deviations from the model used to reconstruct the tree. The basic idea is to remove sites within a monophyletic group that show evidence of change. It is these sites that are likely to evolve faster. Doing this, is also analogous to choosing to replace that monophyletic group's taxa with a set of just those sites for which an ancestral sequence can be built based on a unanimous consensus sequence.

i) At the first prompt user is asked to enter an input file with its extension.
ii) At the next prompt user is asked to enter a name for the output file.
iii) For the next two prompts, user is then asked whether to consider certain states as being different to others or to be effectively ignored. The first option is whether a single member of the group having a different state should exclude the site. The second is whether a single member of any of the selected groups having a unique state not found in any other member of any group should exclude the site. The final question is whether a gap character "-" should exclude a state.
iv) User is then asked to specify groups that should be monophyletic species or at least correspond to edges on the true unrooted tree. This is done by listing the

number of groups and then listing the members of each group. Finally the user is asked if they want to retain any species (and their sites) in the output if they are not included in the specified groups.

v) The program removes all sites showing variation in any of the defined groups. The sequences are then written to an output file with all sites with ‘?’ removed. Also generated is a file in which just the position of sites that do or do not show change in the monophyletic groups is recorded (e.g. sites with ‘??’). These may be copied directly into a NEXUS file to use as a character weight set (and so can be included or excluded at will).

There is no limit on the number of groups or the number of taxa within each group. There are a series of options allowing the user to refine their definition of variable. For example, should a singleton (a unique character state found in only one taxa at a specific site) result in the site being excluded? Should a gap result in it being excluded? Sites remaining after such a treatment tend to be closer to the ancestral sequences of each group defined. Assuming an i.i.d. model with unequal rates across sites, sites that are excluded will be biased towards the faster evolving sites. This approach has also be called site stripping (Waddell et al. 1999b). It allows the analyst to query the data in more depth, and its usefulness as a tool tends to be most limited by the users ability to devise appropriate hypotheses to test.

Consider, for example, our test data. The key question is should *Tupaia* be closer to *Tarsier* than *Tarsier* is to *Homo*? The evidence already suggests that *Tupaia* and *Tarsier* are GC rich. How might this probable attraction be mitigated, so we can then ask if there is any residual evidence that *Tupaia* is within Primates? For this data one way to do this is to remove characters that differ between *Tupaia* and *Homo*. Doing this produces something close to an ancestral sequence of these two taxa. If *Tupaia* is indeed a Primate or a clear sister to them, this treatment would not be expected to readily disrupt this (i.e., any characters that evolved on a common edge to these three species should still be present in the data). Further, when removing characters, one can even be a bit more liberal and allow either *Tupaia* or *Homo* to have singletons. Such sites may retain useful information in other parts of the tree, but do not afford much opportunity for convergence between *Tupaia* and *Tarsier* based purely on base composition. On the flip side, following this treatment, if *Tupaia* is only attracted to primates due to its base compositional similarity to *Tarsier* then it move any where in the tree. However, it is a long external edge and as such there might be some tendency to attract to other long edges. The next longest edge in this data is that to *Mus*. Removing all sites that vary between *Homo* and *Tariser* except those due to singleton states, then analyzing the data with the most general likelihood model in PAUP, the best tree places *Tupaia* sister to *Mus*. A tree following Waddell et al. (1999c) has a worse lnL of 1.9 units. Thus, there is no clear evidence that *Tupaia* is within Primates.

The results are consistent with no resolution within Supraprimates above the superordinal level. A final check of this may be made by retaining just those characters that show no change within the orders of Supraprimates represented by more than two species. All but one of the 16 retained characters define orders and this single character groups Primates, *Cynocephalus* plus Lagomorpha vs Rodentia and *Tupaia*. This helps to confirm the overall impression there is no apparent signal in this data relating orders within Supraprimates.

Perfect2.exe

A variant of **Perfect.exe** that reads in a matching amino acid and codon file. It evaluates which sites to keep and which to exclude based on the amino acid data, then excludes sites that are variable within certain groups of taxa from both the amino acid data set and then the corresponding codons in the codon data set. This is useful if you want to analyze the exact same data at the nucleotide and amino acid level.

i) At the first prompt user is asked to enter an input file of amino acid data.
ii) At the second prompt user is asked to enter an input file of codon data that matches species to species and codon site to amino acid the first file.
iii) At the next prompt user is asked to enter a name for the output file.
iv) For the next two prompts, user is then asked whether to consider certain states as being different to others or to be effectively ignored as in the program Perfect.exe.
v) User is then asked to specify groups of monophyletic species again as in Perfect.exe.

SUMMARY

In conclusion, the programs allow the user to become more familiar with their data and to plan analyses accordingly. Use of these programs was an important determinant in why the tree of Waddell et al. (1999a) was the first accurate tree of the relationships of the 18 orders of placental mammals. The programs should be useful in all situations where systematic error needs to be considered in addition to stochastic error affecting the estimated tree.

Acknowledgements: Thanks to JSPS Postdoctoral Fellowships (1997-1999) for supporting P.J.W. during the development of these programs. H.M. was supported by JSPS grant funds. Amit Patel worked on part of the program, as a student programmer, including initial coding of NEXUS formats and relative rate tests (see Waddell et al., 2004).

REFERENCES

Adachi, J., and M. Hasegawa. (1996). MOLPHY version 2.3 (manual), The Institute of Statistical Genetics, Tokyo, Japan.

Felsenstein, J. (1993). PHYLIP (Phylogeny Inference Package) and manual, version 3.5c. Department of Genetics, University of Washington, Seattle.

Murphy, W.J., ED. Eizirik, S.J. O'Brien, O. Madsen, M. Scally, C.J. Douady, E. Teeling, O.A. Ryder, M.J. Stanhope, W.W. de Jong, and M.S. Springer. (2001). Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science. 294: 2348-2351.

R Development Core Team (2004). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3, URL http://www.R-project.org.

Scally, M., O. Madsen, C. J. Douady, W.W. de Jong, M.J. Stanhope and M.S. Springer. (2001). Molecular evidence for the major clades of placental mammals. J. Mamm. Evol. 8: 239-277.

Sidow, A., T. Nguyen, and T.P. Speed. (1992). Estimating the fraction of invariable codons with a capture-recapture method. J. Mol. Evol. 35: 253-260.

Swofford, D.L., G.J. Olsen, P.J. Waddell, and D.M. Hillis. (1996). Phylogenetic Inference. In: "Molecular Systematics, second edition" (ed. D. M. Hillis and C. Moritz), pp 450-572. Sinauer Assoc, Sunderland, Mass.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. Lectures on Mathematics in the Life Sciences 17: 57-86.

Waddell, P.J. (1995). Statistical methods of phylogenetic analysis, including Hadamard conjugations, LogDet transforms, and maximum likelihood. PhD Thesis. Massey University, New Zealand.

Waddell, P.J., Y. Cao, M. Hasegawa, and D.P. Mindell. (1999a). Assessing the Cretaceous superordinal divergence times within birds and placental mammals using whole mitochondrial protein sequences and an extended statistical framework. Systematic Biology 48: 119-137.

Waddell, P. J., Y. Cao, J. Hauf, and M. Hasegawa. (1999b). Using novel phylogenetic methods to evaluate mammalian mtDNA, including AA invariant sites-LogDet plus site stripping, to detect internal conflicts in the data, with special reference to the position of hedgehog, armadillo, and elephant. Systematic Biology 48: 31-53.

Waddell , P.J., and H. Kishino. (2000). Cluster Inference Methods and Graphical Models evaluated on NCI60 Microarray Gene Expression Data. Genome Informatics Series 11: 129-141.

Waddell, P.J., N. Okada, and Hasegawa (1999c). Towards resolving the interordinal relationships of placental mammals. Systematic Biology 48: 1-5.

Waddell, P.J., H. Kishino, and R. Ota. (2001). A phylogenetic foundation for comparative mammalian genomics. Genome Informatics Series 12: 141-154.

Waddell, P. J., H. Mine, A. Patel, and M. Hasegawa. (2004). INTEROGATE 1.0. Exploration and Testing of Stationarity, Reversibility and Clock-likeness in Sequence Data. Institute of Statistical Mathematics Research Memorandum 929, ISM, Minato-ku, Tokyo.

Waddell, P.J., and Shelley, S. (2003). Evaluating placental inter-ordinal phylogenies with novel sequences including RAG1, gamma-fibrinogen, ND6, and mt-tRNA, plus MCMC-driven nucleotide, amino acid, and codon models. Molecular Phylogenetics and Evolution 28: 197-224.

Waddell, P.J., and M.A. Steel. (1997). General time reversible distances with unequal rates across sites: Mixing G and inverse Gaussian distributions with invariant sites. Mol. Phyl. Evol. 8: 398-414.

The Institute of Statistical Mathematics

4-6-7 Minami-Azabu, Minato-ku
Tokyo 106-8569, Japan