PHYML user's guide

Introduction

PHYML is a software implementing a new method for building phylogenies from sequences using maximum likelihood. The executables can be downloaded at: http://www.lirmm.fr/~guindon/phyml.html. Homologuous sequence data sets can be analyzed under several models of nucleotide or amino acid substitution. A discrete-gamma model (Yang, 1994) is also implemented to accommodate rate variation among sites and invariable sites can be taken into account too. PHYML has been compared to several other softwares using extensive simulations. Results indicate that its topological accuracy is at least as high as that of fastDNAml, while being much faster.

Files in the package

- `exe/` contains executable files for Linux, SunOS, Mac OSX and Windows systems.
- `example/` contains an example to test the software: `seq` contains five DNA data sets in PHYLIP sequential format with 60 taxa and 500 bp sequences; `seq_phyml_tree` shows the trees that you should obtain with default parameter values, except the **I** option that has to be switched to sequential format and the **S** option that must be switched to Yes to indicate multiple data sets; `seq_phyml_out` is the statistics (e.g. tree likelihood) file that you should get with `seq`.
- Source code is available on request (s.guindon@auckland.ac.nz)

Input sequence file

The input sequence file is a standard PHYLIP file of aligned sequences in interleaved or sequential format. It should look like this:

```
5 60
Tax1          CCATCTCACGGTCGGTACGATACACCTGCTTTTGGCAG
Tax2          CCATCTCACGGTCAGTAAGATACACCTGCTTTTGGCGG
Tax3          CCATCTCCCGCTCAGTAAGATACCCCTGCTGTTGGCGG
Tax4          TCATCTCATGGTCAATAAGATACTCCTGCTTTTGGCGG
Tax5          CCATCTCACGGTCGGTAAGATACACCTGCTTTTGGCGG

GAAATGGTCAATATTACAAGGT
GAAATGGTCAACATTAAAAGAT
GAAATCGTCAATATTAAAAGGT
GAAATGGTCAATCTTAAAAGGT
GAAATGGTCAATATTAAAAGGT
```

in interleaved format. The same data set in sequential format:

```
5 60
Tax1          CCATCTCACGGTCGGTACGATACACCTGCTTTTGGCAGGAAATGGTCAATATTACAAGGT
Tax2          CCATCTCACGGTCAGTAAGATACACCTGCTTTTGGCGGGAAATGGTCAACATTAAAAGAT
Tax3          CCATCTCCCGCTCAGTAAGATACCCCTGCTGTTGGCGGGAAATCGTCAATATTAAAAGGT
Tax4          TCATCTCATGGTCAATAAGATACTCCTGCTTTTGGCGGGAAATGGTCAATCTTAAAAGGT
Tax5          CCATCTCACGGTCGGTAAGATACACCTGCTTTTGGCGGGAAATGGTCAATATTAAAAGGT
```

The maximum number of characters in species name MUST not exceed 50. Blanks within the species name are NOT allowed. However, blanks (one or more) MUST appear at the end of each species name.

In a sequence, three special characters '.', '-', and '?' may be used: a dot '.' means the same character as in the first sequence, a dash '-' means an alignment gap and a question mark '?' means an undetermined nucleotide. Sites at which one or more sequences involve '-' are NOT excluded from the analysis. Therefore, gaps are treated as unknown character (like '?') on the grounds that "we don't know what would be there if something were there" (J. Felsenstein, PHYLIP documentation). Finally, standard ambiguity characters for nucleotides are accepted (Table 1).

Table 1

Character	Nucleotide
M	A or C
R	A or G
W	A or T
S	C or G
Y	C or T
K	G or T
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
N	A or C or T or G

Multiple data sets are allowed (option **S**), e.g. to perform bootstrap analysis using SEQBOOT (from the PHYLIP package). In this case, the data sets are given one after the other, in the formats above explained. For example (with three data sets):

```
5 60
Tax1          CCATCTCACGGTCGGTACGATACACCTGCTTTTGGCAGGAAATGGTCAATATTACAAGGT
Tax2          CCATCTCACGGTCAGTAAGATACACCTGCTTTTGGCGGGAAATGGTCAACATTAAAAGAT
Tax3          CCATCTCCCGCTCAGTAAGATACCCCTGCTGTTGGCGGGAAATCGTCAATATTAAAAGGT
Tax4          TCATCTCATGGTCAATAAGATACTCCTGCTTTTGGCGGGAAATGGTCAATCTTAAAAGGT
Tax5          CCATCTCACGGTCGGTAAGATACACCTGCTTTTGGCGGGAAATGGTCAATATTAAAAGGT

5 60
Tax1          CCATCTCACGGTCGGTACGATACACCTGCTTTTGGCAGGAAATGGTCAATATTACAAGGT
Tax2          CCATCTCACGGTCAGTAAGATACACCTGCTTTTGGCGGGAAATGGTCAACATTAAAAGAT
Tax3          CCATCTCCCGCTCAGTAAGATACCCCTGCTGTTGGCGGGAAATCGTCAATATTAAAAGGT
Tax4          TCATCTCATGGTCAATAAGATACTCCTGCTTTTGGCGGGAAATGGTCAATCTTAAAAGGT
Tax5          CCATCTCACGGTCGGTAAGATACACCTGCTTTTGGCGGGAAATGGTCAATATTAAAAGGT

5 60
```

```
Tax1          CCATCTCACGGTCGGTACGATACACCTGCTTTTGGCAGGAAATGGTCAATATTACAAGGT
Tax2          CCATCTCACGGTCAGTAAGATACACCTGCTTTTGGCGGGAAATGGTCAACATTAAAAGAT
Tax3          CCATCTCCCGCTCAGTAAGATACCCCTGCTGTTGGCGGGAAATCGTCAATATTAAAAGGT
Tax4          TCATCTCATGGTCAATAAGATACTCCTGCTTTTGGCGGGAAATGGTCAATCTTAAAAGGT
Tax5          CCATCTCACGGTCGGTAAGATACACCTGCTTTTGGCGGGAAATGGTCAATATTAAAAGGT
```

User supplied tree

A tree input file can also be supplied by the user as starting phylogeny to be refined by PHYML. This tree must be written in the standard parenthesis representation (NEWICK format) with branch lengths, and must be unrooted. Labels on branches (such as bootstrap proportions) are supported. Therefore, a tree with four taxa named A, B, C, and D with a bootstrap value equals to 90 on its internal branch, should look like this:

```
(A:0.02,B:0.004,(C:0.1,D:0.04)90:0.05);
```

Output files

After each run of the program the inferred phylogenies are stored in `your_seqfile_phyml_tree` . Trees are written in NEWICK format and can be plotted using a standard tree viewer program. Various descriptive statistics are listed in `your_seqfile_phyml_out` which should be self-explanatory. Notably this file contains the likelihood of the inferred phylogenies and the computing times.

Options

The PHYML program has a PHYLIP-like interface:

```
 - PHYML -

Settings for this run:

  D                                             Data type (DNA/AA)  DNA
  U                                   Input tree (BIONJ/user tree)  BIONJ
  M  Model of nucleotides substitution (JC69/K2P/F81/HKY/F84/TN93)  HKY
  V               Proportion of invariable sites (fixed/estimated)  fixed (p-invar = 0.00)
  T                                  Ts/tv ratio (fixed/estimated)  fixed (ts/tv = 4.00)
  R                    One category of substitution rates (yes/no)  yes
  S                                     Analyze multiple data sets  no
  I                    Input sequences interleaved (or sequential)  interleaved

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

The user can change the default parameter values of the program by typing the option characters.

- **D**: Data type. The default choice is to analyze DNA sequences
- **U**: Input tree. This tree is used as the starting tree to be refined by the algorithm. The default is to use BIONJ tree. The user can also supply a tree in NEWICK format (see above)
- **M**: substitution model. For DNA sequences, the default choice is HKY (Hasegawa et al., 1985). This model is analogous to K2P (Kimura, 1980), but allows for different base frequencies. The other models in PHYML are JC69 (Jukes and Cantor, 1969), K2P, F81 (Felsenstein, 1981), F84 (Felsenstein, 1989) and TN93 (Tamura and Nei, 1993). The rates matrices of these models are given in Swofford et al. (1996). Dayhoff (1978), JTT (Jones, Taylor and Thornton, 1992), and mtREV (as implemented in Yang's PAML) are the models that can be used when analyzing protein sequences.

- **V**: proportion of invariable sites. The default is to consider that the data set does not contains invariable sites (p-invar=0.0). However, the user can fix this proportion to any value in the 0.0-1.0 range. This parameter can also be adjusted so as to maximise the likelihood of the phylogeny. The latter makes the program slower.

- **T**: the user is prompted to fix the transition/transversion ratio (as used in all model except JC69 and F81), or to use the value that maximizes the likelihood of the phylogeny. The latter makes the program slower as it requires the optimization of the ts/tv parameter. The default value is 4.0. The definition of the transition/transversion ratio is the same as in PAML (Yang, 1994). In PHYLIP, the "transition/transversion rate ratio" is used instead. 4.0 in PHYML roughly corresponds to 2.0 in PHYLIP.

- **R**: use of the discrete-gamma model. The default is to use a unique substitution rate for each site. A discrete-gamma distribution can be used to account for variable rates among sites. The number of categories that define this distribution is user supplied. The higher this number, the better is the goodness-of-fit regarding the continuous distribution (but see option **C**).

- **C**: number of substitution rates categories. The default is to use four categories. The likelihood of the phylogeny at one site is averaged over four conditional likelihoods corresponding to four rates. The computation of the likelihood is therefore four times slower than with a unique rate. Number of categories less than four or higher than eight are not recommended. In the first case, the discrete distribution is a poor approximation of the continuous one. In the second case, the computational burden becomes high and an higher number of categories likely not enhances the accuracy of phylogeny estimation.

- **A**: the shape of a gamma distribution is defined by a numerical parameter. The higher its value, the lower the variation of substitution rates among sites. The default value is 1.0. It corresponds to a moderate variation. Values less than say 0.7 correspond to high variations. Values between 0.7 and 1.5 corresponds to moderate variations. Higher values correspond to low variations. This value can be fixed by the user. It can also be adjusted to maximize the likelihood of the phylogeny.

- **S**: is used in cases of multiple data sets. When switched to Yes, the user is prompted for the number of data sets to be analyzed.

- **I**: the input sequences can be either in interleaved (default) or sequential format (see above).

References

- Z. Yang (1994) *J. Mol. Evol.* **39**, 306-14.
- S. Ota & W.-H. Li (2001) *Mol. Biol. Evol.* **18**, 1983-1992.
- N. Saitou & M. Nei (1987) *Mol. Biol. Evol.* **4**(4), 406-425.
- W. Bruno, N. D. Socci, & A. L. Halpern (2000) *Mol. Biol. Evol.* **17**, 189-197.
- J. Felsenstein (1989) *Cladistics* **5**, 164-166.
- G. J. Olsen, H. Matsuda, R. Hagstrom, & R. Overbeek (1994) *CABIOS* **10**, 41-48.
- N. Goldman (1993) *J. Mol. Evol.* **36**, 182-198.
- M. Kimura (1980) *J. Mol. Evol.* **16**, 111-120.
- T. H. Jukes & C. R. Cantor (1969) in *Mammalian Protein Metabolism*, ed. H. N. Munro. (Academic Press, New York) Vol. III, pp. 21-132.
- M. Hasegawa, H. Kishino, & T. Yano (1985) *J. Mol. Evol.* **22**, 160-174.
- J. Felsenstein (1981) *J. Mol. Evol.* **17**, 368-376.
- David L. Swofford, Gary J. Olsen, Peter J. Waddel, & David M. Hillis (1996) in *Molecular Systematics*, eds. David M. Hillis, Craig Moritz, & Barbara K. Mable. (Sinauer Associates, Inc., Sunderland, Massachusetts, USA).
- K. Tamura & M. Nei (1993) *Mol. Biol. Evol.* **10**, 512-526.
- Dayhoff, M. O., R. M. Schwartz, and B. C. Orcutt. (1978). A model of evolutionary change in proteins. In: Dayhoff, M. O. (ed.) Atlas of Protein Sequence Structur, Vol. 5, Suppl. 3. National Biomedical Research Foundation, Washington DC,

pp. 345-352.

- Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. *CABIOS* **8**: 275-282.

Contacts

For any question about the algorithm, the options of the program or the source code, you can contact Stéphane Guindon (s.guindon@auckland.ac.nz, Thomas Building. Auckland University. New Zealand) or Olivier Gascuel (gascuel@lirmm.fr, LIRMM, 161 rue Ada, Montpellier, FRANCE, (33) 04 67 41 85 47)