Phast Indel User Guide

Introduction

PHylogenetic **A**nalysis with **S**pace/**T**ime Models, “Phast is a software package for comparative and evolutionary genomics. It consists of about half a dozen major programs, plus more than a dozen utilities for manipulating sequence alignments, phylogenetic trees, and genomic annotations.” [Phast Website]. This version of Phast has been extended with a DNA model of evolution that includes insertion and deletion events by using a 5x5 conditional probability matrix with a slight modification for gap to gap transitions [Rivas & Eddy 2008]. This model is dubbed F84e, it is an extended version of the original Felsenstein F84 model [Felsenstein & Church 86].

This documents assumes the user is familiar with fitting models with PhyloFit and computing conservation scores with PhyloP. The user should have also read the original documentation of PhyloFit and PhyloP.

Feature Summary

This version has three major extensions:

1. PhyloFit F84 & F84e Models: Fit a matrix using the F84 or the F84e model which includes rate estimations for insertion and deletion events.
2. PhyloFit Multi: Originally, PhyloFit takes a single alignment file and newick tree and fits the specified model by maximum likelihood. This extension allows PhyloFit to take two directories containing an equal number of newick files and alignment files and fit with the specified model.
3. PhyloP F84 & F84e: Given an alignment and model file, PhyloP will calculate conservation scores as P-scores for the given alignment file. F84e computes conservation scores for insertion and deletion events.

Extensions to PhyloFit:

F84 and F84E Models

PhyloFit was extended to include the F84 and F84e models. The F84 model is a straightforward substitution only model using Felsentein's pruning algorithm for calculating likelihood. The F84e model uses an extended pruning algorithm and performs various extra steps in the calculation of likelihood.

To use the F84 or F84e model you can use the PhyloFit command line options *--subst-mod* to specify *F84* or *F84e*:

```
./phyloFit --tree tree.newick msa.fa --subst-mod F84E -G -O branches
```

```
./phyloFit --tree tree.newick msa.fa --subst-mod F84 -O branches
```

The F84e model requires the user to add the additional command line options: *-G* using “gaps as bases”. Both models need *-O branches* “keep branch lengths constant”. Therefore the branch lengths of the tree must be specified in the tree file. An error will be printed if either flag is omitted.

The output is a *mod* file containing the background frequencies, estimated rate matrix, and phylogenetic tree. Additionally an *infoX* file is created. This file contains information about the individual parameters of the rate matrix and F84e model. The *infoX* file is required as input to PhyloP.

Additional Flags

Several additional flags have also been added to PhyloFit. These options were originally implemented to compare results with Dnaml-erate version of Dnaml from the Phylip package.

--*dnaMlNormalize:* With this flag the tree branches are scaled by a factor of the fractional rate of change.

--*reroot*: Peforms a midpoint rerooting on the tree. The rerooted tree is used for the fitting but the output prints the original input tree.

--*originalF84E:* Fits the F84e model exactly as described by Rivas & Eddy. This is not the default and instead we use a slightly modified version to allow computation of conservation scores for indel events. This option is not recommended when the overall goal is to compute conservation scores with PhyloP.

Multi

PhyloFit has been extended to allow multiple trees and sequence alignments to be fitted at once. Originally, PhyloFit takes a single alignment and newick tree to fit a model of DNA evolution. With this extension a user specifies a folder of alignments and a folder of newick trees to fit the data set.

Rate matrix parameters are shared among all the alignments and the likelihood is computed per alignment. All the likelihoods are added together, this value is optimized to obtain the most likely model to describe the entire data set. Two separate directories are required:

1. A *newick* directory containing *n* newick files with the extension ".newick".
2. An *alignment* directory containing *n fasta* files with the extension ".fa".

The number of newick trees and alignment files must be equal, the names for each newick/alignment pair must be equal except for the extension e.g. *firstFile.newick* must have a matching *firstFile.fa* file. The name of each specie in the tree must match the species in the alignment (this true for the original Phast as well).

The flag -*M* is used to specify that multiple files should be fitted at once. -*M* must be followed by the *newick* directory and the *alignment* directory. The *-O branches* option for keeping branch length constant is required. Sample command:

```
./phyloFit -M NewickFiles/ FaFiles/ -O branches
```

The slowest part of PhyloFit is estimating the matrix parameters. Fitting multiple alignments at once will take even more computation time, this process is still moderately fast and even fitting thousands of files at once should take less than ten minutes on most modern computer.

This option also works with other models and flags e.g.

```
./phyloFit -M NewickFiles/ FaFiles/ -O branches --subst-mod F84E -G
```

The output is written to a new folder (if it does not exist) named *phyloFitResults* in the current working directory. This folder will contain *n mod* files, one for each newick/alignment pair used to fit the model. Each files will have the same name as it's respective newick/alignment file with the extension ".mod". The rate matrix and background frequencies for all models will be the same but each file will contain it's alignment log-likelihood and respective newick tree. If the F84e or F84 model is used for fitting this folder will also contain a *infoX* file per newick/alignment pair, all *infoX* files are identical.

Extension to PhyloP:

PhyloP was extended to allow the user to compute conservation and acceleration scores for insertion and deletion events from a fitted F84e model. Given an alignment, PhyloP will calculate the likelihood for the null model and alternate model and compute conservation scores. For each site, the alternate model is created by fitting over a parameter that scales insertion and deletion rates. The null and alternate model are compared with a likelihood ratio test to determine significance per site.

PhyloFit is first used to create a F84e *mod* file and *infoX* file for a given alignment and phylogenetic tree by using the command line option *--subst-mod F84E.* Given a *mod* file, an alignment and an *infoX* file we can calculate conservation scores for insertion and deletion events with the following command:

```
./phyloP --refidx 0 --wig-scores --mode CONACC --method LRT data.mod alignment.fa -x
data.infoX
```

As seen in the command above PhyloP requires several additional command line options:

--*refidx 0*: Specifies specie for coordinate frame of alignment. Zero indicates to use entire alignment's coordinates, this is needed so gaps are not skipped.

--*wig-scores*: Output conservation scores in fixed-step wig format.

--*mode CONACC*: Uses positive values to indicate conservation and negative values to indicate acceleration for conservation scores.

--*method LRT*: Uses likelihood ratio test to compute conservation scores.

-*x:* Used to specify the name of the *infoX* file for the F84 or F84E model.

PhyloP will print the conservation scores in fixed-step wig format to standard output. With F84 or F84e model a file named *computedLikelihoods.txt* will also be created. This file contains three columns where each row represents the *nth* column in the alignment. Every row contains the computed log likelihoods for the null model and alternate model and the estimated scale parameter. This information is useful for calculating p-values and inferring information from the scaling parameter.

Sample *computedLikelihoods.txt* file:

```
-6.369980 -5.010082 10.285876
-7.092342 -5.725826 10.350118
-6.894692 -5.538195 10.225085
-7.334239 -5.979709 10.188605
-8.395074 -7.042011 10.158101
…
```

References

[1] http://compgen.cshl.edu/phast/index.php "PHAST: Home." PHAST: Home. N.p., n.d. Web. 13 Jan. 2016.

[2] Felsenstein, Joseph, and Gary A Churchill. "A Hidden Markov Model approach to variation among sites in rate of evolution." Molecular Biology and Evolution 13.1 (1996): 93-104.

[3] Rivas, Elena, and Sean R Eddy. "Probabilistic phylogenetic inference with insertions and deletions." PLoS Comput Biol 4.9 (2008): e1000172.