ParaSel

Version 1.0

The ParaSel method is a Maximum Likelihood method for testing and detecting site-specific parallel selection. Given a multiple sequence alignment (MSA) and a phylogenetic tree, the program determines whether or not directional selection has operated along the phylogeny and led to the fixation of parallel substitutions (nucleotides) or replacements (amino-acids). For more detailts on the model and its use see Stern et al. "The Evolutionary Pathway to Virulence of an RNA Virus", Cell 2017.

Manual

Contents

Download and Installation

Source code (C++) is available at https://github.com/sternadi/parasel

Compiling Parasel

1. In order to unzip and untar the files please type:
   ```
   tar -xzvf parasel.tar.gz
   ```
   This will create the following directories:
   ```
   libs/phylogeny
   programs/directionalSelection
   ```
2. In some operating systems, you may use the makefiles to compile the program. If this does not work, skip to item 3.
   Make sure you are in the directory where you unzipped the files, and type:
   ```
   cd libs/phylogeny
   ```
   In order to run the Makefile, type:
   ```
   make
   ```
   Now, type:
   ```
   cd ../../programs/directionalSelection
   ```
   to get to the directionalSelection directory. Type:
   ```
   make
   ```
   in order to run the Makefile.
   This will result in an executable file called *directionalSelection* which will reside in the *programs/directionalSelection* directory.
3. In some systems the makefiles will not be operable. Thus, follow step 1 and compile directly using g++:
   a. Make sure you are in the directory where you unzipped the files.
   b. Type:
   ```
   mv libs/phylogeny/* programs/directionalSelection/
   ```
   c. cd to the directionalSelection library:
   ```
   cd programs/directionalSelection
   ```
   d. To compile, type
   ```
   g++ -O3 -o directionalSelection *.cpp
   ```
   This will result in an executable file called *directionalSelection* which will reside in the directory where you ran the g++ command.

If there are any problems with the compilations (occasionally, with old versions of g++) - please email sternadi@post.tau.ac.il and I'll try to help. To modify the code, or use parts of it for other purposes, permission is requested. Please note that the use of the program is for academic use only.

Running ParaSel

The aim of ParaSel is to detect directional selection in which one allele is preferred (selected for) and hence the probability of fixation of such an allele is elevated. Such a case is expected to lead to parallel fixation events along the phylogeny. Hence, we use the terms directional selection and parallel fixation, or parallel selection, interchangeably in this document.

In order to infer parallel selection for a certain dataset, we recommend performing the following stages:

When the ancestral sequence is known, the method has the most power:

1. Perform a statistical test to infer whether significant directional selection is operating on the dataset at hand. To this end compare the Akaike Information Crierion (AIC) between a null model and ParaSel (see below). If this test is in support of directional selection, proceed to the next stage.
2. Infer sites under directional selection based on their posterior probability (see below).

To run the program you must supply a parameters text file. Simply type in the command line:

```
parasel parameters_file_name
```

A basic example parameters file is available at the ParaSel webpage (`parasel.params`). See below how to use it.

For more complex options see the `parasel.allOptions.params` file, also available at the ParaSel webpage.

1. Performing a statistical test to assess significance of directional selection

In order to assess statistical significance of directional selection, the program must be run twice: once with a directional selection enabling model and once with a null model (which does not enable directional selection). Then the likelihood values of the two runs need to be used to calculate AIC (or AICc) scores (see here).

Thus, run the program twice, with the following parameters files (make sure to use the same phylogenetic model name `_modelName` in both runs):

1. For the directional selection model use the file: `parasel.params.` The number of parameters *k* in this model is 6.

2. For the null model to run, use the file: `null.params`. The number of parameters *k* in this model is 4.

The likelihood of the data given each model will be in the results file, which can then be converted into AIC scores, and used to assess whether the directional selection model better fits the data at hand. AIC is necessary since none of the asymptotic methods are applicable for using the χ^2 approximation for a likelihood ratio test, typically used when comparing two phylogenetic models. This pathology occurs since *S* is only estimated in the alternative model and not in the null model.

2. Inferring sites under directional selection

The results file of the directional selection model from stage 1 will contain all the information required for determining which sites have experienced rate shifts. The file will look like this:

```
#dirSel Results File
#=========================================================
#Parameters are:
#Log-likelihood: -41231.6
#S= 10
#Prob (S)= 0.00735633
#alpha=0.27
#K=11
#beta=0.000207357
#tau=0.757477
#q=0
#Rate categories are
#       0.000243454
#       0.00611082
#       0.0367097
#       0.126226
#       0.330245
#       0.75329
#       1.6675
#       5.07968
#(Parameters are ML estimates)
#Results of analysis
#Displayed on sequence XXX
#==========================================================================================
#POS(QUERY_SEQ)        POS(ALN)       RESIDUE POSTERIOR PROBABILITY OF DIRECTIONAL
SELECTION
                      A       C       G       T
1       1       G       1.3e-16 8.6e-06 0.0035  1e-05   1
2       2       G       1.3e-16 8.6e-06 0.0035  1e-05   1
3       3       G       2.1e-08 0.00081 3.7e-14 0.00078 1
4       4       T       0.95    6.6e-15 6.5e-06 4.1e-08 0.048
```

The top of the results file reports the values of various parameters of the model. Th bottom part presents the posterior probability that each site experienced directional selection for a specific character. The last column represents no directional selection. Thus, in the example above, the first three sites are *not* inferred to be under directional selection. On the other hand site 4 has a posterior probability of 0.95 for undergoing directional selection for “A”.

More options and instructions

You may more complex options aslisted below. In order to specify a root sequence (as in Stern et al. Cell 2017), you will need to add the root sequence to both the alignment and to the tree. Within the tree, this sequence will have a branch length of zero. Next, specify this sequence as the query sequence (`inQuerySeq`) and invoke `useQueryFreqsAtRoot.`

The basic options are:

	Name	Description	Default	Remarks
Input	_inSeqFile	Input aligned sequence file	Obligatory	Use full path. Formats accepted are: Fasta, Clustal, Phylip, Mase
	_inTreeFile	Input user tree in Newick file.	NJ tree	Use full path
	_inQuerySeq	Name of query sequence	1st in the sequences file	This determines the numbering of the sequence for which results are displayed.
Output	_outResFile	Results output file	Obligatory	Use full path
	_logFile	Log file name		Use full path
	_outTreeFile	Output tree file		Will report tree with optimized branch lengths, based either on tau (see below) or on individual branch

				length optimization (see below _bblOpt)

The more complex options are:

Name	Description	Default	Remarks
_modelName	{hky (HKY),jtt (JTT), rev (REV - for mitochondrial genomes), day (DAY), HIVb, HIVw, aajc (JC amino acids)}	hky	
_numOfCategoriesForRateDistr	Number of categories for among site rate variation, discretized gamma distribution	4	Integer
_useQueryFreqsAtRoot	The root sequence is determined to be that of the query sequence specified	0=false	0 or 1
_doMutationMapping	Map where mutations (nucleotide substitutions or amino-acid replacements) occurred along the phylogeny, using joint posterior probabilities at each node of the tree (see Stern et al. Cell 2017 for more details). If invoked, a file with suffix .map is created listing sites and mutations, and their posterior probability.	0=false	0 or 1
_verboseLevel	Verbose level for log file	5	Integer
_isNull	Use a null model by invoking ProbS = 0	0=false	0 or 1
_bblOpt	Perform branch length	1=true	0 or 1

	optimization on each branch. Due to computational intensity this option is not recommended for large sequence alignments		
_fixedS	Fix S value; S is not optimized using ML and is set at _initS	0=false	0 or 1
_initS	Initial value for S	10	real
_fixedProbS	Probability of site undergoing directional selection (P_{DS}) is fixed	0=false	0 or 1
_initProbS	Initial value for P_{DS}	0.01	real
_fixedKappa	Kappa (parameter for HKY model for Ts/Tv ratio) is fixed	0=false	0 or 1
_initKappa	Initial value for kappa	2	real
_fixedAlpha	Alpha (shape parameter for gamma distribution) is fixed	0=false	0 or 1
_initAlpha	Initial value for alpha	1	real
_fixedBeta	Beta (relaxation factor for tree leaves rates) is fixed	1=true	0 or 1
_initBeta	Initial value for beta	0	real
_fixedTau	Tau (inflation/deflation factor for all tree branch lengths) is fixed	1=true	0 or 1
_initTau	Initial value for tau	1	real