PSGfinder

Finding signal of positive selection using pairwise alignments

Description

PSGfinder scans pairwise alignment of homologous coding DNA sequence (CDS) for regions undergoing divergent positive selection. Candidate regions are defined according amino acid difference positions and their dN/dS ratio is estimated by an implementation of Yang and Nielsen's (2000) method, yn00 from PAML package.

Options

`-a, --alignment=DIR` Specifies a directory to look in for alignments. Each alignment file must be in FASTA format. There is only one pairwise alignment by file.

`-c, --clean-up` Masks putative non-homologous regions of the alignments before analysis. The cleaning method masks gappy regions and highly divergent regions. For that, each contiguous site region is filtered for its length, dS (estimated by yn00) and number of amino acid differences.

`-e, --estimations=FILE` Analyzes FILE which is the output of the program run with option `--no-test`.

`-f, --fdr` Applies a Bonferroni's style correction to the p-value according to an estimated number of independent tests. This number is calculated from the total number of analyzed windows weighted by the degree of overlap amongst windows.

`-l, --log-file=FILE` Writes the run parameter values, global statistics and the skipped alignments to FILE.

`-r, --data-range=X-Y` According the alphanumeric ascending order of the alignment file names, analyses alignment X^{th} to alignment Y^{th} (the first file name in the list is #1). You can also specify a single file #, or a range like `X-` or `-Y` to mean "from X to the end of the list" or "from the beginning of the list to Y", respectively.

`-x, --no-test`	Outputs yn00 estimated values of evolutionary rates for each gene and their regions and stops before re-calculating window dN/dS, calculating Fisher's exact p-value and filtering for significant results. Output table is slightly different from that of a complete run-through: it summaries yn00 output. Most importantly, dN/dS ratio is calculated by yn00, from each window dN and dS.
`-h, --help`	Display a short option descriptions.

Parameter file

Alignment cleaning algorithm parameters and windows parsing and filtering parameters can be defined through a control file named psgfinder.ctl that must be located in the directory of your analysis. Here is an example that you can use as a layout. If you do not specify a control file, these are the default values.

```
# alignment cleaning parameters
                        n = 30
               maximum dS = 1
                        k = 2
                        d = 4
                        q = 6

# windows parsing and filterng parameters
                 min. size = 4
  min. number of aa diff. = 3
                    dS max = -1
               S category = 36
                   lambda = -1
                    alpha = .05
```

Alignment cleaning

Since dN/dS estimation relies on the correctness of the analyzed nucleotide alignment, an alignment cleaning method has been implemented in PSGfinder, aiming to reduce unreliable regions of pairwise alignments. The alignment cleaning method proceeds in 3 steps. In the first place, it masks the regions that are shorter than a given length threshold (`n`). These regions are defined between gaps, as contiguous aligned sites. Secondly, it estimates the dS of the remaining regions, and masks those that exceeds a given threshold (`maximum dS`). Finally, it detects clusters of amino acid difference, and masks high density clusters. These clusters are formed given 3 parameters. The seeding parameters `k` defines the minimum number of contiguous amino acid difference to start a cluster. The density

parameters `d` defines the minimum distance between clusters to be aggregated. The threshold parameter `q` defines the minimum number of amino acid differences in a cluster to be masked.

Defining candidate regions and test for positive selection.

In order to identify CDS regions with a higher rate of amino acid changes than predicted by neutral evolution, PSGfinder analyses each alignment region that lays between two amino acid differences, with a minimum length (min. size) and containing a minimum number of amino acid changes (min. number of aa diff.; in addition to the ones that defines the region). Each region is analyzed with yn00 and its dN is compared to the whole gene dS which is a good proxy for the neutral evolution rate. An additional step of filtering discards region with unexpectedly high dS. This threshold is defined according a pool of small regions (with the same number of synonymous sites: `S category`). From this data, a Poisson distribution is fitted to predict expected the number of synonymous mutations. The dS threshold for all regions is defined at the 99% density of the fitted distribution. Alternatively, you can choose a dS threshold for regions using the parameter file, either by defining the threshold (`dS max`) or the lambda parameter of the fitted distribution (`lambda`). The discrepancy of dN and dS evolutionary rates is supported by a Fisher's exact test p-value. The p-value threshold is defined by parameter `alpha`. It can be adjusted for multiple testing using option `--fdr`.

Multiple testing p-value adjustment

The more analyzed regions, the higher chance to get false positives. To prevent this, the `--fdr` option allows to make the p-value threshold more stringent as the number of analyzed region increases. However, since analyzed regions often overlap and thus are not independent tests, correcting for the total number of analyzed region would be abusive. Instead, PSGfinder estimates a number of independent tests among overlapping regions given the amount of their overlap (Formula 1). The sum of these estimated numbers of independent tests is used to adjust the p-value threshold.

$$T_k = N_k \, x \, \frac{S_k}{\sum_{i=1}^{N} Sw_i}$$

Formula 1: For each group (k) of overlapping windows encompassing Sk sites with an a.a. mismatch and containing N_k windows with each Sw sites with an a.a. mismatch, the number of non-overlapping tests (T_k) is approximated.

Output format

The normal output of psgfinder.py is a tab delimited file listing analyzed alignment windows and their dN/dS, plus various evolutionary rate values estimated by yn00. The option `--header` allows to display a header line before the output. Here are the different columns of the output and their description:

`file name`	The file name in the alignment directory
`sequences`	Sequence names of the pairwise alignment
`window`	Window of analysis
`size`	Size of the window
`N`	Number of non-synonymous sites
`S`	Number of synonymous sites
`%gap`	Percentage of gap
`dN`	Window's dN
`dS`	Window's dS
`dN/dS(whole gene)`	Window's dN/Whole gene's dS
`p-value`	Chance of dN being greater dS according to a Fisher's exact test
`whole gene`	1 if the data refers to the whole gene alignment, 0 if it is a window

When using `--no-test` option, psgfinder.py returns yn00 estimated values for each candidate region, without searching for signal of positive selection. You can also combine this option with option `--header` to get the header line for this particular output. Here are the different columns and their description:

`file name`	The file name in the alignment directory
`sequences`	Sequence names of the pairwise alignment
`window`	Window of analysis
`size`	Size of the window
`N`	Number of non-synonymous sites
`S`	Number of synonymous sites
`%gap`	Percentage of gap
`t`	Estimated time value
`kappa`	Estimated kappa value
`dN`	Window's dN
`dN SE`	dN standard error

`dS`	Window's dS
`dS SE`	dS standard error
`omega`	Window's dN/dS
`whole gene`	1 if the data refers to the whole gene alignment, 0 if it is a window

You can analyze the data produced with the option `--no-test` using option `--estimation`.