1 NAME

t2prhd - a program to generate pairwise repeat homology diagrams

2 SYNOPSIS

t2prhd -m *HMM_file* [**-v -p -s -t -x -i** *default|omit|filter|filter-both|showall* **-d** *output_directory* **-w** *color_gradient_parameter* **-o** *PhyML_parameters* **-e** *hmmsearch_options* **-r** *width*x*height* **-c** *repeat_CSS_color* **-c2** *repeat_CSS_color2*] input_file1.fas input_file2.fas

3 DESCRIPTION

This is a program to generate pairwise repeat homology diagrams from two sequences and a profile HMM. The sequences can be DNA or amino acid sequences and the profile HMM must be the same type. The basic workflow of the program is as follows:

*** Reads two sequence files in Fasta format and a HMM file in HMMER format.**

*** Identifies repeats using the hmmsearch program from the HMMER package.**

*** Combines all repeats in one Fasta file.**

*** Checks if the number of total detected repeats is at least 4 and aborts if not. The reason for this is that the homology relations are detected by using an unrooted phylogenetic tree and such trees with less than 4 leaves can have only one possible topology.**

*** Aligns repeats using the hmmalign program and converts the alignment in Fasta format using Bio::AlignIO. It converts the alignment into sequential PHYLIP format with sequence names mapped to indices.**

*** Calculates a repeat phylogenetic tree: a Neighbor Joining (NJ) tree using CLUSTAL W or a Maximum Likelihood (ML) tree using PhyML. The default parameters for PhyML are '1 i 1 0 WAG e 4 e BIONJ y y' (WAG+Gamma+I, 4 gamma categories, gamma parameter and proportion of invariable sites estimated by ML, BIONJ starting tree) for amino acid and '0 i 1 0 HKY e e 4 e BIONJ y y' (HKY+Gamma+I, 4 gamma categories, gamma and Kappa parameters and proportion of invariable sites estimated by ML, BIONJ starting tree) for DNA data. The default parameter strings can be changed in the "PhyML settings" section of the script or can be overridden by the -o switch.**

*** Parses the tree using Bio::TreeIO searching for sister leaf nodes. They are regarded as unambiguously identified homologues.**

*** Marks the sister leaf nodes which are parts of a larger clade of perfectly identical leafs as spurious, as inside these clades the branching order is arbitrary. The handling of connections corresponding to these spurious sister leaf nodes can be controled through the -i option.**

*** Generates the diagram in SVG (Scalable Vector Graphics) format using XML::Writer.**

*** Generates a LaTeX file with the diagram if the -t switch is present.**

In the diagrams repeats are represented as rectangles (colored in red by default). The identified orthology relations are represented by blue lines, whereas the identified paralogy relations by brown arcs.

The colour intensity of the connecting lines is calculated as $(1 - \frac{patristic\_distance}{total\_tree\_length})^w$, so the intensity is a function of the patristic distance between the two repeats divided by the total tree length. The default value of the color gradient parameter **w** is 1 (linear color scale) but it can be adjusted by the **-w** switch. The distances between sister nodes are

generally low and they lie in a narrow interval close to zero, so by powering by **w** the colour scale can be tuned to accommodate to the span of distance range.

The input sequence files must be in Fasta format. If the sequence name is in the *Species.Gene.Accession number* format, it is parsed and the *Accession number* is discarded. Only alphanumeric characters and '.' can be used in the sequence names.

4 REQUIREMENTS

The program uses the *BioPerl* modules (needs version 1.4.0 or higher) to manipulate sequences and alignments and to parse *hmmsearch* results and trees. The *XML::Writer* module is used for SVG generation. You must also have the following binaries installed in your path:

*** hmmsearch and hmmalign from the HMMER package (`http://hmmer.janelia.org`) (version 2.3.2 or higher). User's Guide: `ftp://selab.janelia.org/pub/software/hmmer/CURRENT/Userguide.pdf`**

*** CLUSTAL W for NJ tree calculation (`ftp://ftp.ebi.ac.uk/pub/software/unix/clustalw/`, tested with version 1.83).**

*** PhyML for ML tree calculation (optional, `http://atgc.lirmm.fr/phyml/`, tested with version 2.4.4). If you run the script on other platform than Linux, you must change the name of the PhyML executable stored in the $PHYML_EXE variable.**

The generated SVG file can be viewed by **Firefox** 1.5 or higher and can be rasterized (and also viewed) using the **Batik** toolkit (`http://xmlgraphics.apache.org/batik/`) as an example.

The (optionally) generated LaTeX file must be run trough **pdflatex**, the *pgf/TikZ*, *xcolor*, *fancyhdr* and *fullpage* packages are required. The size of the diagram can be adjusted by modifying the scale parameter of the TikZ picture in the generated LaTeX file.

5 OPTIONS

-h - prints the help message and exits.

-m *HMM_file* - The name of the *HMMER* HMM file used for repeat identification.

-d *output_directory* - The name of the directory used to store the output files (optional, the "OUTDIR" directory is created and used if not specified).

-w *color_gradient_parameter* - Parameter used to tune the color intensity scale. It must be an integer.

-r *width*x*height* - The size of the generated SVG image in pixels, the default being 600x950.

-c *CSS_color* - The color of repeats, affects only the SVG output. Make sure that you use a valid CSS color (`http://www.w3schools.com/css/css_colornames.asp`) or the repeats will not be displayed.

-c2 *CSS_color* - The second repeat color, used with the -x switch, affects only the SVG output. Make sure that you use a valid CSS color (`http://www.w3schools.com/css/css_colornames.asp`) or the repeats won't be displayed.

-p - Use PhyML to calculate the phylogenetic tree. You must have the executable denoted by $PHYML_EXE in path.

-o *PhyML_parameters* - Overrides the default parameter string passed on to PhyML. You must specify a full parameter string. The syntax is described at `http://atgc.lirmm.fr/phyml/`.

-e *hmmsearch_options* - Extra options passed on to hmmsearch. Using this switch you can set for example the E-value cutoffs (-e '-domE 0.0001').

-s - Suppress colour legend and scale bar.

-i - Controls the handling of spurious connections. The possible options are:

default - filter spurious external connections, do not display internal connections at all

omit - show all external connections, do not display internal connections at all

filter - show all external connections, filter spurious internal connections

showall - show all connections

filter-both - do not display spurious internal and external connections

-x - Detect runs of perfectly identical repeats. If this switch is on, then the consecutive perfectly identical repeats will have the same color. If two consecutive repeats differ then the script switches back and forth between the colors specified by -c and -c2.

-t - Outputs the diagram in LaTeX/TikZ format besides SVG.

-v - Verbose output mode.

6 EXAMPLE

```
t2prhd -m fn3.hmm txh.fas txm.fas
t2prhd -x -c 'magenta' -c2 'green' -m fn3.hmm txh.fas txm.fas
t2prhd -x -i filter-both -m fn3.hmm txh.fas txm.fas
t2prhd -t -w 3 -m fn3.hmm -c 'magenta' -d MyOutdir txh.fas txm.fas
t2prhd -p -o '1 i 1 0 Dayhoff e 4 e BIONJ y y' -t -w 10 -m fn3.hmm \
-r 600x2000 -c '\#FF00FF' -d MyOutdir txh.fas txm.fas
```

7 FILES

The following files will be placed in the output directory:

*** the modified copies of the input files in Fasta format (.fas)**

*** the hmmsearch result files (.hmmres)**

*** the identified repeats in Fasta format (_reps.fas)**

*** all repeats combined in one Fasta file (Seq1_vs_Seq2.fas)**

*** the aligned repeats in MSF (.msf) and Fasta (_aln.fas) format**

*** if you use CLUSTAL W: the NJ tree calculated by CLUSTAL W (.ph)**

*** the SVG file containing the diagram (.svg)**

*** the CLUSTAL W and hmmalign log files (.log)**

*** the LaTeX file containing the diagram (.tex) - written if the -t switch is present**

*** the repeat alignment in sequential PHYLIP format (.phy) with the names replaced by indices**

*** if PhyML is used for tree calculation, the standard PhyML output files and a log file (PhyML.log) will be created. A tree file with the real repeat names will also be created (.nwk).**

8 KNOWN BUGS

*** CLUSTAL W (used to build the NJ tree) has a limit on the sequence name length. The repeats are identified by the sequence name and the appended starting and ending positions, so sequence names must be kept short to have enough space for the positional information. If the sequence name is too long, it will be truncated and the resulting tree cannot be parsed. If this happens, the script will die with an error message. Increasing the MAXNAMES parameter in *clustalw.h* should solve this problem.**

*** TikZ fails to draw some internal connections (arcs) and throws a "Dimension too large error" message. The respective connection is ignored by LaTeX, but the others are not affected. LaTeX can also fail to process diagrams produced for very long sequences.**

*** If the patristic distance between the sister leaf nodes is very small, and/or the value of the w parameter is large, the calculated colour intensity can become too small. The small numbers expressed with the exponential notation are not handled by TikZ and SVG, so the script will fail. Specify a smaller value by the -w switch to solve this issue.**

9 LIMITATIONS

The script places no restrictions on the length of the sequences, but limits imposed by the external applications are still present. As they can change from version to version it is advisable to consult their manuals.

10 AUTHOR

Botond Sipos (sbotond (at) gmail.com)

11 COPYRIGHT