Name

fetchMGs

Version

1.0

Software description

fetchMGs extracts the 40 single copy universal marker genes (decribed in Ciccarelli et al., Science, 2006 and Sorek et al., Science, 2007) from genomes and metagenomes in an easy and accurate manner. This is done by utilizing profile Hidden Markov Models (HMMs) trained on protein alignments of known members of the 40 MGs as well as calibrated cutoffs for each of the 40 MGs. Please note that these cutoffs are only accurate when using complete protein sequences as input files. The output of the program are the protein sequences of the identified proteins, as well as their nucleotide sequences, if the nucleotide sequences of all complete genes are given as an additional input.

Input

A fasta file with protein coding sequences, and optionally the nucleotide sequences of the proteins. If the DNA sequences are available, the corresponding nucleotide sequences of the proteins, are also extracted.

Output

The output of this software is saved within the specified output folder and consists of:

- 40 x COGxxxx.faa files (sequences of extracted proteins)
- 40 x COGxxxx.fna files (sequences of extracted genes)
- marker_genes_scores.table (protein <TAB> score <TAB> marker gene ID <TAB> genome identifier)
- temp (identifiers of proteins identified homologous to any marker gene)
- hmmResults (specific output files from HMMer3)

Synopsis

fetchMGs.pl -m|mode <*extraction*|*calibration*> [OPTIONS]

Extraction mode

./fetchMGs.pl [options] -m extraction <protein sequences>

Required options

<protein sequences>

Multi-FASTA file with protein sequences from which marker genes should be extracted

Further options

-o|outdir

Output directory; default = "output"

-b|bitscore

Path to bitscore cutoff file; Path to bitscore cutoff file; default = "$pathInWhichThisScriptResides/lib/MG_BitScoreCutoffs.[allhits|verybesthit].txt" (depending on -v option)

-l|library

Path to directory that contains hmm models; default = "$pathInWhichThisScriptResides/lib"

-p|protein_only

Set if nucleotide sequences file for <protein sequences> is not available

-d|dnaFastaFile

Multi-FASTA file with nucleotide sequences file for <protein sequences>; Not neccesary if protein and nucleotide fasta file have the same name except .faa and .fna suffixes

-v|verybesthit_only

Only extract the best hit to each OG from each genome. Recommended to use, if extracting sequences from reference genomes. Please do not use for metagenomes. If this option is set fasta identifiers should be in the form: taxID.geneID and, if needed, have " project_id=XXX" somewhere in the header

-c|og_used

Orthologous group id to be extracted; example: "COG0012"; default = "all"

-t|threads

Number of processors/threads to be used

-x|executables

Path to binaries used by this script. default = "" --> will search for variables in \$PATH Path to executables used by this script (hmmsearch; cdbfasta, cdbyank). default = "$pathInWhichThisScriptResides/bin" If set to "" will search for executables in \$PATH

Calibration mode

./fetchMGs.pl -m calibration <reference protein sequences> <true positives map>

Required options

<reference protein sequences>

Multi-FASTA file with protein sequences that include marker genes (true positives)

Further options

<true positives map>

Tab-delimited file with true positive protein identifiers and COG IDs

-o|outdir

Output directory; default = "output"

-b|bitscore

Path to bitscore cutoff file; Path to bitscore cutoff file; default = "$pathInWhichThisScriptResides/lib/MG_BitScoreCutoffs.uncalibrated.txt" (depending on -v option)

The other options for \'-m extraction\' can also be used here.

Software dependencies

The fetchMGs script requires the cdbyank, cdbfasta and HMMer3 executables. These software are (c) respecitve authors, and have been installed in the bin folder, within the fetchMGs folder. If these are incompatible with your system please install them yourself.

cdbyank, cdbfasta: http://sourceforge.net/projects/cdbfasta/

hmmsearch (HMMer3): http://hmmer.janelia.org/

Author

The **fetchMGs** package was developed by Shinichi Sunagawa and Daniel R Mende (Bork Group, EMBL) (http://www.bork.embl.de). External software used by the **fetchMGs** package are copyright respective authors.

Copyright