##Scripts were written in shell, which call other scripts written in C, perl and python.###

Fully Automated SGMBL and WGMBL

To build probabilistic models for SGM and blast database,run following command on your terminal prompt:

usage: sh sgmmodel_blastdb.sh markov_order confid1 confid2 confid3

Similarly to build probabilistic models for whole genome and blast database together, run following command on your terminal prompt:

usage: sh wgmmodel_blastdb.sh m

Once models and blast database are ready, they can be use for all future analysis, without regenerating them.

Then you can run following script with any “metagenome file” (that you want to classify,test.fa here) and “prefix” for output files(output here) as input. The following script will also ask you to choose whether you want to use SGM or WGM for classifying reads. If WGM then which Markov model order. Also if user wants to do blast testfile or not.

usgae: sh autoclassify.sh test.fa output

INPUTS

---

input_file for building models: Genome Files in fasta format
keep all genomes files in a folder ”genomes” with name “accessionid.fna” format (For e.g.: NC_12345.fna)

Markov_order for building SGM: 0, 1, or 2. It is Markov model order.

confid1: Threshold parameter for the segmentation(takes values between 0 - 1)

confid2: Threshold parameter for the first-step clustering (grouping contiguous similar segments)(take values between 0 - 1)

confid3: Threshold parameter for the second-step clustering (grouping non-contiguous similar segments or clusters)(takes values between 0 -1)
you can tune/change segmentation and clustering parameters as well as markov model order at which you want to segment all genomes.

For e.g. : sh sgmmodel_blastdb.sh 2 0.90 0.95 0.99

where 2 is Markov model order, 0.90 is the threshold for segmentation, 0.95 is the threshold for first-step clustering (grouping contiguous similar segments), and 0.99 is the threshold for second-step clustering (grouping non-contiguous similar segments or clusters). These 4 parameters can be tuned by the user from command prompt.

Also BLAST+ tools must be installed on your computer. Again same genome as used for building models will be use to make BLAST DB.

OUTPUTS

All outputs will be stored in a folder Results.

For all the output files, if user choose SGM it will have sgm in output files and wgm.m if user choose wgm with markov order m.

Files with these suffix will contain final classification:

1. final.classify.txt : Classification with only SGM (or WGM).

2. formulabased.taxonomy.txt : Classification with only SGMBL using formula for combining BLAST and SGM (or WGM) prediction.

3. singularhit.taxonomy.txt : Classification with SGMBL (or WGMBL) by giving Blast priority.

To add new genomes to existing SGM and WGM models

Make a directory “addgenomes”. Keep all new genomes in this folder and also add to folder genomes that you want to add to existing SGM and WGM. Then run the following script for adding models to SGM and WGM directories repectively:

usage: sh addsgmmodel.sh markov_order confid1 confid2 confid3

usage: sh addwgmmodel.sh m

Users have to also update taxonomy.txt file manually for new genomes.

Examples: one test file test.fa is provided in folder with two genome sequences in genomes folder and 1 genome in addgenome folder.