Reproducing SBSP

The commands used to set up, reproduce, and graph results from the SBSP paper

Karl Gemayel

Mon 27 Apr 2020 10:37:05 EDT

Contents

1 Downloading and installing

1.1 Code

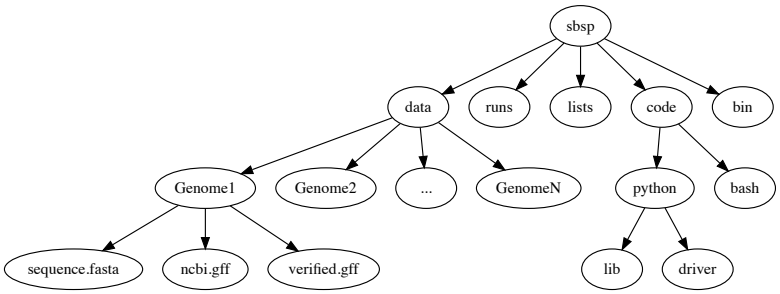
Downloading the code is fairly straightforward using `git`.

1.2 Data

We provide the databases for *Enterobacterales*, *Actinobacteria*, *Archaea*, and *FCB group*, and the sequence and label files for the genomes with verified starts: *E. coli*, *H. salinarum*, *N. pharaonis*, *M. tuberculosis*. We also provide the steps to create a data base with for any ancestor using data that can be downloaded from NCBI's website.

2 Code and data structure

After installing SBSP, you will have the following structure



2.1 Bin

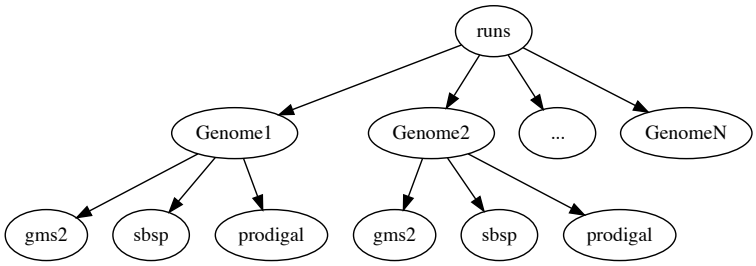
The python scripts can be located at

2.2 Data

The data directory contains all genomic raw information: mainly the sequence and labels files. If constructing databases from scratch, this directory will also include all genomes downloaded from NCBI.

2.3 Runs

For this analysis, all runs executed by SBSP, GMS2, and Prodigal will be put in subdirectories for each genome.



3 Running on verified genomes

SBSP takes as input:

- Query proteins: FASTA file
- Target protein database: Diamond database

It outputs:

- GFF file containing labels
- Multiple sequence alignment files for all queries
- details.csv: output file containing details of predictions

```
# List of genomes with verified genes
pf_list_verified=$lists/verified.list
pf_sbsp_conf=$config/sbsp_defaults.conf

toggle_pbs="--pf-conf-pbs $config/pbs_defaults.conf"     \
    # if PBS installed, set this option to empty: ""

$bin/run_sbsp_from_genome_list_py.sh --pf-genome-list    \
    ${pf_list_verified} --pf-conf-sbsp                   \
    ${pf_sbsp_conf} ${toggle_pbs}
```

4 GMS2 on metagenomes

4.1 Run GMS2 on genome fragments

```
$bin/analysis_on_genome_fragments_py.sh                    \
    --pf-genome-list $lists/verified.list --tools          \
    gms2 prodigal
```

5 Collecting Data

6 Tables and Graphs

6.1