HaMMLET – Supplementary material

John Wiedenhoeft, Eric Brugel, Alexander Schliep

August 5, 2015

The web supplement can be found at http://bioinformatics.rutgers.edu/Supplements/HaMMLET/. This document describes the scripts which can be used to reproduce those results. Please note that due to the use of random numbers both in HaMMLET and CBS, the results can vary slightly. To obtain the scripts, pull the biorxiv branch from the repository:

```
git clone https://github.com/wiedenhoeft/HaMMLET.git
cd HaMMLET
git checkout biorxiv
make .
cd evaluation
```

The scripts have been tested under Python 2.7.3 with the following packages:

- NumPy 1.8.2
- SciPy 0.13.3
- Pandas 0.13.1 (older versions might not work due to keyword changes!)
- PyYAML 3.10
- matplotlib 1.3.1
- multiprocessing 0.70a1
- dateutil 1.5

Simulated aCGH data

The directory willenbrock_fridlyand contains two scripts, as well as a HaMMLET model file.

```
Rscript getData.R
```

pulls the simulations from http://www.cbs.dtu.dk/~hanni/aCGH/, extracts all data tracks, creates separate files for each of them and place them into three subdirectories.

```
python run.py
```

will run CBS and HaMMLET on each of them, create plots for all results, and a boxplot of the F-scores. Notice that the vertical bars represent the true breakpoints.

High-density CGH array

The directory GSE23949 contains two scripts and an auxiliary file specifying chromosome sizes (for plotting).

```
        bash getData.sh
```

downloads the raw data for GSE23949 from the Gene Expression Omnibus, and prepares the BT474 cell line data for use with CBS and HaMMLET. Two files are created, `all_bt474.csv` contains the entire genome, `chr20.csv` contains only chromosome 20. CBS and HaMMLET are called by running

```
        python run.py
```

Note that CBS will take about 2 hours for the entire genome, depending on your platform. While HaMMLET is very fast, plotting relies on an external library (`matplotlib`) and may take a minute or two.

Speed and convergence effects of wavelet compression

The directory `simulations` contains a YAML file containing the parameters for the simulations. You need to have PyYAML installed (`https://pypi.python.org/pypi/PyYAML`). Running

```
        python yamlbatch.py test.yaml
```

will create 129,600 subdirectories (one for each parameter combination), and run the simulation and inference scripts on each of them. `test.yaml` contains the entry

```
        __nrThreads__: 36
```

which sets the number of threads to be used in parallel. Set this number according to your needs. After the simulations and inference is complete, run

```
        python simulationPlots.py test.yaml
```

in order to create the plots.

WARNING: This is a very extensive test. A total of 129,600 simulations will be created. Both CBS and HaMMLET will run twice on each of them: a simple run to benchmark the time, and an extended run in which ALL sampled state sequences will be output (more than 32 million values per run), in order to calculate the F-measures for each iteration. **Be prepared to spend several CPU months on this!** In total, about 180 GB of free disk space will be required.

Coriell, ATCC and breast carcinoma

The directory `snijders` contains two Python scripts and a file containing information about the database, cell line and accession number.

```
        python getData.py
```

downloads the relevant data (GSE16) from the Gene Expression Omnibus.

```
        python run.py
```

creates the input files and runs HaMMLET with plotting.