Billboard User's Manual

Release 0.2

Mateusz Patelak

September 11, 2008

Contents

1 Users manual

1.1 Installation

Prerequisites

In order to compile and run Billboard software, you should have **Java Development Kit** version 6 and the **Ant** build tool installed.

Compilation

In the root project folder (where the `build.xml` is located) run a command

```
ant deploy
```

This should build the application and place it in the `deploy` subfolder.

1.2 Setting Billboard up

After the compilation process you should have the following directory structure in your `deploy` folder.

- `Billboard.jar` - the application's jar file
- `lib` - library dependencies
- `res` - application's resources folder
- `log` - log files folder
- `log4j.properties` - log settings

All application parameters may be set in the `res/app.properties` file. You may also change these settings dynamically from the command prompt (see section 1.3). The available parameters are listed below.

- `pairsFile`
  denotes the path to the file containing pairs of names of the homologuous sequences; each row of this file should contain a tab-separated pair of names
- `defaultTFDir`
  denotes the path to a folder containing sets of transcription factors
- `tfSet`
  denotes the name of a set of transcription factors that should be used
- `sequencesDirectory`
  denotes the path to a folder containing sequences in *FASTA* file format
- `downloadSequences`
  indicates that the sequences should be downloaded directly from the *EnsEMBL* database; names listed in the `pairsFile` should match, in this case, *EnsEMBL* IDs; currently, only human, mouse and rat sequence IDs are supported
- `backgroundSequencesDirectory`
  denotes the path to a folder containing sequence files in *FASTA* format used for computing *CRM* hit scores; this directory should contain at least a 100 of sequences with no ambiguous nucleotides; the background sequences length should match the length of the sequences of interest
- `scoreParametersFile`
  denotes the file containing α, β and γ parameters as well as the length of the window and it's jump
- `upstreamLength` and `downstreamLength`
  if the `downloadSequences` parameter is set, these two parameters denote the values indicating the length of the sequences before and after the transcription start site (or before and after the gene start point)
- `thresholdType`
  indicates the threshold type for transcription factors' binding sites; the possible values are `balanced` and `absolute`
- `gui`
  indicates if the graphical user interface should be used
- `threads`
  denotes the number of the threads computing the alignments

1.3 Running the Billboard application

Enter the `deploy` subfolder and type

```
java -jar Billboard.jar
```

Depending on the length of sequences of interest and the number of threads (see section 1.2) you should allow an adequate amount of memory for the Java virtual machine to run. For example, for sequences of length of 20000 nucleotides, using only one thread, it is safe to allow 512 megabytes of memory. This can be done by using `-Xmx` virtual machine parameter (e.g. `java -Xmx512M -jar Billboard.jar`).

By default the application reads all of its parameters from the file `res/app.properties`. However, one may use a different file, by running the program with the `properties` parameter (e.g. `java -jar Billboard.jar properties=path/to/your/parameter/file`).

One may also overwrite some of the parameters from the parameter file using the command prompt. For example, if one would like to use another file containing the pairs of sequences, one would write `java -jar Billboard.jar pairsFile=path/to/your/pairs/file` in the command prompt.

1.4 Input formats

Transcription factor file format

A transcription factor file should have a `pfm` extension. The name of the TF is the file's name. The file contains a PSSM matrix. Subsequent rows of the file correspond to A, C, G and T nucleotides.

Scoring function parameters file format

Each row of this file contains tab-separated values of α, β, γ, window length and jump length.

Results file format

The results of computations are written in the *GFF* file format (http://www.sanger.ac.uk/Software/formats/GFF/) and placed in the `result.gff` file in the output directory. Each row contains information about one *CRM* that was found.

1.5 Using custom transcription factor sets

The `res/TF` directory contains some bundled transcription factor sets. One may add a custom factor set by adding a folder in the `defaultTFDir` path containing TF files in format described in section 1.4.