SAINTq: scoring protein-protein interactions in affinity purification - mass spectrometry experiments with fragment or peptide intensity data

Guo Ci Teo and Hyungwon Choi

February 27, 2016

1 Compilation

Run "make" to compile the software. g++ version 4.4 or above is required to compile the source code. When the compilation finishes, the executable will appear in the "bin" directory.

2 Input files

Two input files will be used by the program for the analysis. The first is a table of intensity measurements, and the second is a parameter file for the input data information and optional parameters for inclusion criteria in scoring. Example input files can be found in the "examples" directory (MSPLIT/PeakView data from the SAINTq manuscript, Teo *et al*, 2015).

2.1 Parameter file

In the parameter file, the following information should be provided by the user.

- input_filename: the name of the file containing intensity data
- normalize_control: normalize control intensities by multiplying a constant to all control intensities so that the average observed test intensities is equal to the average observed control intensities. "true" or "false".
- input_level: the type of intensity data (valid entries are "protein", "peptide" or "fragment" and they are case sensitive).

- protein_colname: the header in the column of protein names in the intensity table
- pep_colname: the header of the column of peptide names (valid for fragment or peptide intensities only).
- frag_colname: the header of the column of fragment names (valid for fragment intensity data only).
- compress_n_ctrl: the number of control baits used in calculations, with priority for baits with greater intensities. Setting this number to a large number makes the program use all available control data (recommended in cases with at most several controls).
- compress_n_rep: the number of test bait replicates used for scoring, with priority given to the baits with higher probability scores. If this number is greater than or equal to the number of available replicates, then the scores will use the data from all replicates. Otherwise, the highest scoring replicate scores will be averaged to yield the final probability score.
- best_prop_pep: the proportion of peptides to be used for protein score calculation. Default is 0.5 (50%).
- min_n_pep: a minimum number of peptide intensities to be used for protein score calculation. This sets a lower bound on the number of peptides used when the "best_prop_pep" parameter selects too few peptides. Default is 3.
- best_prop_frag: the proportion of fragments to be used for peptide score calculation. Default is 0.5 (50%).
- min_n_frag: a minimum number of fragment intensities to be used for peptide score calculation. This sets a lower bound on the number of fragments used when the "best_prop_frag" parameter selects too few fragments. Default is 3.

Note that compress_n_ctrl, compress_n_rep, min_n_pep, best_prop_pep, min_n_frag and best_prop_frag control the number of intensity measurements utilized for the scoring, with priority for higher intensities. The program will not execute when unused options are set. For example, a protein level data should not have pep_colname and a peptide level data should not have frag_colname. Option lines can be commented out by putting a pound symbol (#) in the beginning of the line.

2.2 Intensity table file

This file should be a tab-separated values file, with the first three lines describing the bait and experimental information. The third line is a row of column names containing names of protein, peptide, fragment columns and the bait replicate IP names. Directly above the bait IP names, in the second line, are the corresponding bait names. The first line indicates whether each bait is a test protein (T) or control (C). Missing intensities can be represented with "0" or an empty character. Bait replicate IP names of the same bait names must be placed next to one another.

3 Running the program

In a command line prompt, execute the program with the parameter filename:

```
$ ./saintq input_parameter_file
```

4 Output file format

The output file is a tab-separated value file, with each row corresponding to a bait-prey interaction pair.

- The first two columns "Bait" and "Prey" are bait and prey proteins, respectively.
- #Rep: the number of bait replicates.
- #Pep: the number of peptides in each prey protein. This appears only for peptide and fragment level inputs.
- #Frag: the numbers of fragments available for each peptide in each prey protein. The numbers are separated with a bar "|" character. This appears only for fragment level inputs.
- AvgP: probability score for a bait-prey interaction.
- BFDR: the Bayesian false discovery rate (FDR).

5 Copyrights