QSPEC/QPROT: Testing protein differential expression with false discovery rate estimation

Hyungwon Choi*

December 18, 2012

Abstract

QSPEC/QPROT implement a Bayesian hierarchical model for protein differential expression. QSPEC is for spectral count data and QPROT is for intensity data or any continuous transformation of spectral counts. Both QSPEC and QPROT allow group comparison for independent samples and paired samples. Significance analysis is performed by false discovery rate (FDR) analysis, which reports both local and global FDR for each protein.

1 Installation

Type 'make all' to install all components of the software. This software requires GNU Scientific Library for C language (any version is O.K.), freely downloadable from

`http://www.gnu.org/software/gsl/`

You will need to add the current installation directory to your shell login files such as `.cshrc` or `.bashrc` in order to run the command line at any location you want. One way to do this is to add the following lines to `.bashrc` in the home directory (∼/):

```
PATH=/home/hwchoi/software/qprot_v1.2.2/bin/:$PATH
```

2 Available programs

- qspec-param: group comparison using spectral counts with independent samples
- qspec-paired: group comparison using spectral counts with paired samples
- qprot-param: group comparison using intensity data or any continuous data with independent samples

*For troubleshooting, contact the author at hyung_won_choi@nuhs.edu.sg.

Protein	Length	0	0	0	1	1	1
ALB	609	2902	2749	2407	1603	1487	1499
MYH8	1937	718	752	724	2	1	1
MYH11	1972	627	583	490	94	98	90
KRT8	483	284	312	272	541	663	380
FN1	2355	179	151	169	600	527	472
LTF	710	187	134	150	20	20	18
KRT19	400	149	152	120	468	487	438
TUBA1C	449	137	128	143	221	220	156
KRT5	590	111	110	101	83	93	76
KRT9	623	103	94	92	108	118	104

Table 1: Top portion of a sample spectral count data matrix for QSPEC. In the case of QPROT, the second column "Length" should be removed from the input data.

- qprot-paired: group comparison using intensity data or any continuous data with paired samples
- getfdr: FDR calculation using the output from the four main programs

Note that there are parametric and nonparametric models available for each type of quantitative data. The computational load is much lighter for the MCMC sampler in the parametric models, and thus the computation is much quicker.

3 Input File Format

To use QSPEC/QPROT, the dataset must be prepared in a matrix data format. Header contains the group indicators, and the rest of the rows should contain the quantitative data for each protein.

Things to keep in mind when preparing the data:

- When the dataset has $\geq$6 samples, exclude proteins with just one non-zero count only.
- All non-observed counts/intensities should be filled in as zero.
- The column for 'length' should be used for spectral count data only (QSPEC). For continuous data, remove the column.
- Protein names should not contain any blank space.
- Remove all proteins with too many missing observations (e.g. proteins with single spectral count).

4 Fitting the models

Once the data is ready, the rest is straightforward. Use the following command lines to fit the models:

```
usage: qprot-param <matrixData> <nburnin> <niter> <normalize 0/1>
usage: qspec-param <matrixData> <nburnin> <niter> <normalize 0/1>
usage: qprot-paired <matrixData> <nburnin> <niter> <normalize 0/1>
usage: qspec-paired <matrixData> <nburnin> <niter> <normalize 0/1>
```

The recommended arguments are:

- nburnin: 2,000 or more
- niter: 10,000 or more
- normalized: 1 unless it is known that the protein concentrations are different across samples.
- numThreads: the number of threads for parallel computing, e.g. 4 for using 4 threads.

5 FDR estimation

To finalize the report, a separate module has to be run to calculate FDR at various thresholds of Z-statistic.

```
usage: getfdr <matrixData>
```

- matrixData: any output file from one of the four command lines above.

For instance, the command line

```
gouda$ qprot-param mydata 2000 10000 1
gouda$ getfdr mydata_qprot
```

generates the final output file mydata_qprot_fdr. The command line also generates a separate file reporting the density of null and alternative distributions. To plot the density information, the following R script can be run:

```
d = read.delim("mydata_qprot_fdr", header=T, as.is=T)$Zstat
tmp = read.delim("mydata_qprot_density", header=F)
hist(d, breaks=50, xlab="Zstat", main="mydata")
ff = 600  # scaling factor: need to be adjusted in each dataset
lines(tmp$V1, tmp$V4 * ff, col=3)
lines(tmp$V1, tmp$V2 * ff, col=4)
lines(tmp$V1, tmp$V3 * ff, col=2)
```