Documentation for calc.post.c

Laura A. Salter

July 2, 2003

Description:

The program `calc.post.c`calculates of the posterior probability two proteins co-occur in the same complex given a set of trials, t, and successes, s, in one or more high-throughput proteomic data sets. The program also requires a false negative random error rate, ν, a false positive random error rate, ϕ, and a prior probability, ρ. See Gilchrist *et al.* (2003) for more details on this approach. The parameters ν, and ϕ, are specific to a particular dataset while ρ is universal. These parameters can be estimated using the program `mlest.c` which is available at www.unm.edu/∼compbio/software/Interaction_Assess/Estimate.

About the Program `calc.post.c`:

Given one or more datasets consisting of protein-protein interaction data, this program computes the posterior probability that a pair of proteins is part of the same complex. It is assumed that the false-positive and false-negative errors rates are known or have been previously estimated for the experimental technique used to generate each dataset.

Using `calc.post.c`

Format for input data:

All data should be placed in a single file in the format described below. The file can have any name shorter than 20 characters, and the user will be prompted to enter the name of the file upon running the program.

The first line of the input file contains the number of interactions (i) under consideration. The second line contains the number of datasets (N)to be analyzed. The third line contains the value of the parameter ρ. Next there will be N lines, each of which contains the ν and ϕ value for each dataset. These must be in the same order as the information from each dataset below. Finally, there will be i lines, each of which contains information on a single interaction from all datasets. The format for each line is the number of the protein

interaction (1 through i), the number of trials in dataset 1, the number of successes in dataset 1, the number of trials in dataset 2, the number of successes in dataset 2, etc.

Below is a portion of the input file used to generate Table 4b in the paper cited above. There are 120 interactions, 2 datasets, $\rho = 0.00188$, the error rates for the first dataset (HMS-PCI) are $\nu = 0.539$ and $\phi = 0.0013$, the error rates for the second dataset (TAP) are $\nu = 0.346$ and $\phi = 0.00107$, and the first interaction has no trials or successes in the HMS-PCI dataset and 2 trials with 1 success in the TAP dataset. Data for all other interactions follow.

```
120
2
0.00188
0.539 0.00130
0.346 0.00107
1 0 0 2 1
2 1 0 2 1
3 1 1 2 1
4 2 0 2 1
5 2 1 2 1
6 2 2 2 1
7 3 0 2 1
8 3 1 2 1
9 3 2 2 1
10 3 3 2 1
11 4 0 2 1
12 4 1 2 1
13 4 2 2 1
14 4 3 2 1
15 4 4 2 1
.
.
.
.
119 14 13 2 1
120 14 14 2 1
```

Output format:

The posterior probabilities for each interaction will be written to an output file called yourfilename.results, where yourfilename is the name you have given to the file containing the input data. The first few lines in the output file will contain the information you have entered

concerning the number of interactions and datasets, as well as the information about the parameters. Next, there will be a table which lists the interaction number, the number of trials and successes for each dataset, and the posterior probability of each interaction given the data and parameters.

A sample output file corresponding to the input file above is:

```
The number of interactions is 120
The number of datasets is 2
The value of rho is 0.001880

For dataset 1, nu is 0.539000 and phi is 0.001300
For dataset 2, nu is 0.346000 and phi is 0.001070

Interaction  (t,s)  (t,s)   Post. Prob.
     1       (0,0)  (2,1)   0.285080
     2       (1,0)  (2,1)   0.177097
     3       (1,1)  (2,1)   0.992978
     4       (2,0)  (2,1)   0.104063
     5       (2,1)  (2,1)   0.987066
     6       (2,2)  (2,1)   0.999980
     7       (3,0)  (2,1)   0.058988
     8       (3,1)  (2,1)   0.976297
     9       (3,2)  (2,1)   0.999963
     10       (3,3)  (2,1)   1.000000
     11       (4,0)  (2,1)   0.032725
     12       (4,1)  (2,1)   0.956951
     13       (4,2)  (2,1)   0.999932
     14       (4,3)  (2,1)   1.000000
     15       (4,4)  (2,1)   1.000000
.
.
.
.
     119      (14,13)  (2,1)   1.000000
     120      (14,14)  (2,1)   1.000000
```

Reference

Gilchrist, M.A., L.A. Salter, and A. Wagner. 2003. A statistical framework for interpreting high-throughput proteomic datasets, submitted to *Bioinformatics*.