Documentation for mlest.c

Laura A. Salter

July 2, 2003

Description:

The program `mlest.c`calculates maximum likelihood estimates of the false negative random error rate, the false positive random error rate, and the prior probability two proteins co-occur in the same complex, ν, ϕ, and ρ respectively. These parameters are used to in modeling protein-protein interaction from high-throughput proteomic data sets and estimating the posterior probability two proteins co-occur in the same complex using the program `calc.post.c` which is available at www.unm.edu/∼compbio/software/Interaction_Assess/Posterior. See Gilchrist *et al.* (2003) for more details.

About the Program `mlest.c`:

Given one or more datasets consisting of protein-protein interaction data, this program will compute the maximum likelihood estimates (MLE's) of the parameters of Eq. 6 in the reference below. There are two random error rates (ν and ϕ) for each dataset, and a single parameter ρ for all datasets. The program uses the Nelder-Mead Simplex Algorithm to compute the estimates, as implemented in the GSL package (see http://sources.redhat.com/gsl/). GSL is free software distributed under the terms of the GNU General Public License (http://www.gnu.org/copyleft/gpl.html).

Using `mlest.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 datasets (N) and the total number of interactions (i). Next are a number of lines, each of which corresponds to one dataset and contains the number of interactions within that dataset. Following this are a series of lines

corresponding to each interaction, each of which contains the dataset number (1 to N), the number of trials within that interaction, the number of successes within that interaction, and the total number of interactions of this type. There are i such lines in the dataset.

Below is the input file used to compute the MLE's found in Table 1 of the paper cited above. There are 2 datasets, which contain a total of 126 interactions. There are 6 interactions in the first dataset, and 120 interactions in the second data set. The first interaction in the first dataset consists of no trials and no successes and occurred 1,821,186 times in that dataset. Data for all other interactions follow.

```
2 126
6
120
1        0        0        1821186
1        1        0        1119893
1        1        1        2452
1        2        0        171948
1        2        1        594
1        2        2        183
2        0        0        2013021
2        1        0        658915
2        1        1        1388
2        2        0        279672
2        2        1        978
2        2        2        97
2        3        0        94882
2        3        1        412
2        3        2        69
2        3        3        17
.
.
.
.
2        14       11       0
2        14       12       0
2        14       13       0
2        14       14       0
```

Output format:

All output will be written to the current window. After responding to the prompts for the name of the input file and a set of starting values for the parameters, the program will print information concerning each iteration of the algorithm. This information will include the

iteration number, the values of the parameters, the value of the negative of the log likelihood, and the simplex size.

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

```
        Program to find MLEs of the parameters nu, phi, and rho
               using the Nelder-Mead simplex algorithm


Please enter the name of the file containing your data: datafile


Input data were read as follows:

  There are 2 datasets with
    6 observations in dataset 1
    120 observations in dataset 2


Enter a starting value for nu in dataset 1: 0.3
Enter a starting value for phi in dataset 1: 0.01
Enter a starting value for nu in dataset 2: 0.5
Enter a starting value for phi in dataset 2: 0.01
Enter a starting value for rho: 0.01


 Iterations beginning .....

Iter. #    Nu-1      Phi-1      Nu-2      Phi-2      Rho     -Log Likelihood     Simplex size
    1    0.30000    0.01000    0.50000    0.01000    0.01000 f() = 79496.35867 ssize = 0.7242522
    2    0.30000    0.01000    0.50000    0.01000    0.01000 f() = 79496.35867 ssize = 0.6393544
    3    0.30000    0.01000    0.50000    0.01000    0.01000 f() = 79496.35867 ssize = 0.5548643
    4    0.30000    0.01000    0.50000    0.01000    0.01000 f() = 79496.35867 ssize = 0.4699353
    5    0.30000    0.01000    0.50000    0.01000    0.01000 f() = 79496.35867 ssize = 0.3830356
    6    0.30000    0.01000    0.50000    0.01000    0.01000 f() = 79496.35867 ssize = 0.3831607
    7    0.30000    0.01000    0.50000    0.01000    0.01000 f() = 79496.35867 ssize = 0.3460125
    8    0.30000    0.01000    0.50000    0.01000    0.01000 f() = 79496.35867 ssize = 0.3081447
    9    0.30000    0.01000    0.50000    0.01000    0.01000 f() = 79496.35867 ssize = 0.3127530
   10    0.30000    0.01000    0.50000    0.01000    0.01000 f() = 79496.35867 ssize = 0.2777755
   11    0.30000    0.01000    0.50000    0.01000    0.01000 f() = 79496.35867 ssize = 0.2576190
   12    0.30000    0.01000    0.50000    0.01000    0.01000 f() = 79496.35867 ssize = 0.2376798
   13    0.30000    0.01000    0.50000    0.01000    0.01000 f() = 79496.35867 ssize = 0.2219898
   14    0.30000    0.01000    0.50000    0.01000    0.01000 f() = 79496.35867 ssize = 0.2119714
.
.
.
.
  530    0.34515    0.00107    0.53879    0.00130    0.00188 f() = 50411.97319 ssize = 0.0000035
  531    0.34516    0.00107    0.53879    0.00130    0.00188 f() = 50411.97319 ssize = 0.0000035
  532    0.34516    0.00107    0.53879    0.00130    0.00188 f() = 50411.97319 ssize = 0.0000032
```

```
  533    0.34516    0.00107    0.53879    0.00130    0.00188 f() = 50411.97319 ssize = 0.0000026
  534    0.34516    0.00107    0.53879    0.00130    0.00188 f() = 50411.97319 ssize = 0.0000025
  535    0.34516    0.00107    0.53879    0.00130    0.00188 f() = 50411.97319 ssize = 0.0000025
  536    0.34516    0.00107    0.53879    0.00130    0.00188 f() = 50411.97319 ssize = 0.0000024
  537    0.34516    0.00107    0.53879    0.00130    0.00188 f() = 50411.97319 ssize = 0.0000020
  538    0.34516    0.00107    0.53879    0.00130    0.00188 f() = 50411.97319 ssize = 0.0000018
  539    0.34516    0.00107    0.53879    0.00130    0.00188 f() = 50411.97319 ssize = 0.0000019
  540    0.34515    0.00107    0.53879    0.00130    0.00188 f() = 50411.97319 ssize = 0.0000018
  541    0.34515    0.00107    0.53879    0.00130    0.00188 f() = 50411.97319 ssize = 0.0000017
  542    0.34515    0.00107    0.53879    0.00130    0.00188 f() = 50411.97319 ssize = 0.0000013
  543    0.34515    0.00107    0.53879    0.00130    0.00188 f() = 50411.97319 ssize = 0.0000013
  544    0.34515    0.00107    0.53879    0.00130    0.00188 f() = 50411.97319 ssize = 0.0000013
  545    0.34516    0.00107    0.53879    0.00130    0.00188 f() = 50411.97319 ssize = 0.0000010
converged to a local maximum at
  546    0.34516    0.00107    0.53879    0.00130    0.00188 f() = 50411.97319 ssize = 0.0000009


 Please note: Program should be run many times with varying starting points to detemine global
 maximum
```

Program notes:

This program makes use of the the GSL package (see http:// sources.redhat.com/ gsl/). GSL is free software distributed under the terms of the GNU General Public License (see http://www.gnu.org/copyleft/gpl.html). This means that you must have the GSL package installed on your system to compile and run the program, and you may need to change the include statements in the first few lines of mlest.c to reflect the location of the GSL package on your system. Alternatively, we have provided an executable version of the program for LINUX.

Please note that the Nelder-Mead Simplex Algorithm is guaranteed only to find a local maximum (which is a minimum of the function -log likelihood). Thus the program should be run many times from different starting points, and the parameters which give the smallest value of -log likelihood should be selected as the MLE's.

Reference

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