# Tool Guide: Bayesian Estimator of Protein-Protein Association Probabilities (BEPro$^3$)

JM Gilmore, DL Auberry, AM White, JL Sharp, KK Anderson and DS Daly
Pacific Northwest National Laboratory

## 1.0 Introduction.

This document provides guidance in the use of the Bayesian Estimator of Protein-Protein Association Probabilities (BEPro$^3$) to estimate probabilities of protein-protein association using protein identifications from bait-prey affinity isolation LC-MS experiments.

BEPro$^3$ features two algorithms that differ in the estimation of the prior probabilities of both true and false positive identifications. One algorithm features local prior probability estimates that are specific to a particular bait and prey protein pair. The other algorithm uses global prior probability estimates that apply to all bait and prey pairs. For details about the local estimation algorithm, see:

> Sharp, J. L., Anderson, K. K., Hurst, G. B., Daly, D. S., Pelletier, D. A., Cannon, W. R., Auberry, D. L., Schmoyer, D. D., McDonald, W. H., White, A. M., Hooker, B. S., Victry, K. D., Buchanan, M. V., Kery, V. and Wiley, H. S. (2007). "Statistically Inferring Protein-Protein Associations With Affinity Isolation LC-MS/MS assays," *Journal of Proteome Research*, 6: p. 3788-3795

For details about the global estimation algorithm, see:

> Gilchrist, MA and Salter, LA and Wagner, A. 2004. "A statistical framework for combining and interpreting proteomic datasets." *Bioinformatics*, 20(5):689-700.
>
> *The user should note that our implementation of the global estimator is based on our understanding of the approach outlined in the Gilchrist article. We apologize if our global estimator unintentionally strays from what Gilchrist, et al, intended.*

This guide has six sections following this introduction. Section 2 briefly describes the underpinnings of the BEPro$^3$ software. Sections 3, 4 and 5 describe the requirements of the input files, the function of the user interface and the content of the output, respectively. Section 6 provides an example analysis to illustrate how BEPro$^3$ works. Section 7 describes some common problems and repairs.

## 1.1 Contents.

**Section 1:** Introduction.

**Section 2:** The BEPro$^3$ Software.

**Section 3:** Input data files.

**Section 4**: User interface.

**Section 5:** Output files.

**Section 6:** An example analysis.

**Section 7:** Troubleshooting examples

*All documentation related to BEPro$^3$ may be found in the "BEPro3\Documentation" directory.*

## 2.0 The BEPro$^3$ software.

The Bayesian Estimator of Protein-Protein Association Probabilities (BEPro$^3$) software is a self-installing package containing the BEPro$^3$ executable, this guidance document, the companion journal article, and example input files with their corresponding output.

BEPro$^3$ uses the statistical programming environment R (The R Project for Statistical Computing, Vienna, Austria. http://www.r-project.org) to implement the featured statistical algorithms. R contains an extensive and sophisticated set of vetted statistical tools that allow custom statistical algorithms to be quickly coded and tested. The R statistical routines have been wrapped in a graphical user interface written in Java (Sun Microsystems, Inc. Sunnyvale CA) to relieve the user from having to learn the R language. This combination of R and Java, which are both free and open-source software, allows this tool to be distributed without restrictions under a GNU license. The BEPro$^3$ package (self-installing software, user guide and example dataset) may be obtained from the PNNL statistics group (http://www.pnl.gov/statistics/) at http://www.pnl.gov/statistics/BEPro3/.

### 3.0 Input Data.

BEPro[3] requires experiment data and a set of analysis parameter inputs. Data can be uploaded in one of two ways: 1) as a prey protein by bait protein table of association scores in a comma separated value (*.csv) file with a sample pedigree .csv file, or 2) as a long format matrix. Each of the two formats is described below.

1. **Data Crosstab and Pedigree Inputs**
   - **Prey protein by bait protein table of association scores**
     - This table is derived from the LC-MS/MS proteomic output of an affinity isolation experiment featuring two or more replicates of two or more bait proteins
     - Format: comma separated value (CSV) file
       - The file can be created in MS Excel by choosing 'Save As', then choosing 'CSV' in the 'Save As' type field.
     - The first column of the table must contain prey IDs
     - The first row of the table must contain unique sample IDs
     - Excluding the first column and row, each cell of the table must contain a protein association score for the given prey-bait protein pair
       - Examples include: Number of identified prey peptides, the estimated prey protein abundance, and total SEQUEST XCorr (Thermo Finnigan, Waltham, MA)
       - Any score that is an indicator of protein presence or absence via comparison to a user-defined critical value is acceptable.
     - **DO NOT** leave empty cells in the table. Use '0', or another "absence" equivalent score
     - For an example, see "BEPro3\Example\Data\RPal_data.csv" (distributed with this BEPro3 software)

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | ORF | S1 | S2 | S3 | S4 | S5 | S6 |
| 2 | RPA0002 | 0 | 0 | 0 | 0 | 0 | |
| 3 | RPA0009 | 0 | 0 | 0 | 0 | 0 | |
| 4 | RPA0016 | 0 | 0 | 0 | 0 | 1 | |
| 5 | RPA0018 | 0 | 0 | 0 | 1 | 0 | |
| 6 | RPA0023 | 0 | 0 | 0 | 0 | 0 | |
| 7 | RPA0028 | 0 | 0 | 0 | 0 | 0 | |
| 8 | RPA0029 | 0 | 0 | 0 | 0 | 0 | |
| 9 | RPA0030 | 0 | 0 | 0 | 0 | 0 | |
| 10 | RPA0035 | 0 | 0 | 0 | 0 | 0 | |
| 11 | RPA0037 | 0 | 0 | 0 | 0 | 0 | |
| 12 | RPA0038 | 0 | 0 | 0 | 0 | 0 | |
| 13 | RPA0040 | 0 | 0 | 0 | 0 | 0 | |
| 14 | RPA0041 | 0 | 0 | 0 | 0 | 0 | |

   - **Sample pedigree table**
     - This table contains one record for each sample (column) featured in the prey by bait table of protein associations
     - Format: comma separated value (CSV) file

- The file can be created in MS Excel by choosing 'Save As', then choosing 'CSV' in the 'Save As' type field.

- The fields of the sample pedigree table summarize the history of each sample. Two columns are currently required, though more may be added in future versions of BEPro[3]
  - "sample ID" and "bait ID" column must contain unique identifiers for the sample and bait pairs. The names of these columns are not important but must be entered into the GUI.
  - Default names for these columns are "Sample.ID" and "Bait.ID"
  - Entries in the sample ID column must match exactly the header of its corresponding column in the prey by bait table of protein association scores
- For an example, see "BEPro3\Example\Data\RPal_pedigree.csv"

| | A | B | C |
|---|---|---|---|
| 1 | Sample.ID | Bait.ID | |
| 2 | S1 | RPA3226 | |
| 3 | S2 | RPA3226 | |
| 4 | S3 | RPA3226 | |
| 5 | S4 | RPA0176 | |
| 6 | S5 | RPA3226 | |
| 7 | S6 | RPA0176 | |
| 8 | S7 | RPA0176 | |
| 9 | S8 | RPA0176 | |
| 10 | S9 | RPA3226 | |
| 11 | S10 | RPA0175 | |
| 12 | S11 | RPA0175 | |
| 13 | S12 | RPA0177 | |
| 14 | S13 | RPA0367 | |
| 15 | S14 | RPA4225 | |

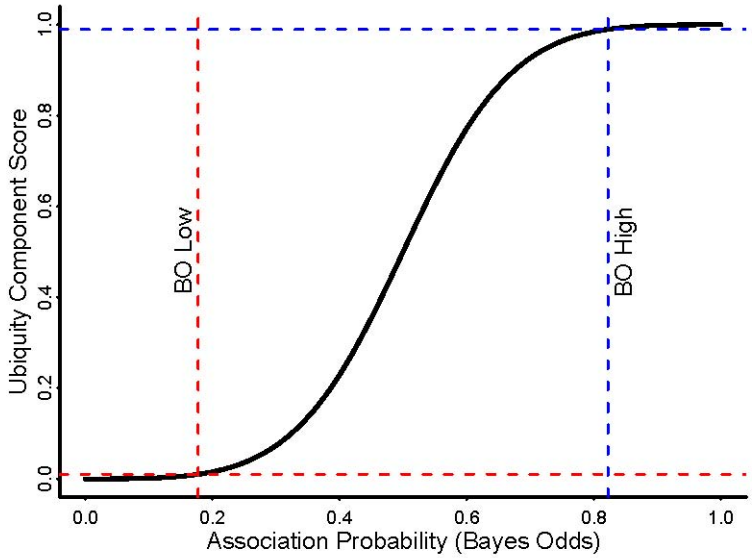
2. **Long Format Input**
   - Instead of a prey-by-bait table of association scores and a pedigree file, a single *.csv file may be uploaded with the following column headers in the following order from left to right
     - SampleID
       - The "SampleID" column must contain sample IDs that would have appeared in the pedigree file sample ID column described above
     - BaitID
       - The "BaitID" column must contain bait IDs as they would have appeared in the pedigree file bait ID column described above
     - PreyID
       - The "PreyID" column must contain the prey protein IDs as they would have appeared in the first column of the prey by bait protein association scores table described above
     - Observation
       - The "Observation" column must contain the protein association scores as they would have appeared in the prey by bait protein association table described above

- **NOTE:** The Gilchrist Analysis does not accept long format input and will not be run when data is provided by this method

A list of analysis parameter values for keyboard input is described below.

- **Maximum number of observable prey protein**
  - **NOTE:** This number may be smaller, perhaps much smaller, than the size of the proteome because not all proteins can be observed using a particular analysis.
  - The default value for this parameter is the number of unique prey proteins appearing in the input data
- **Number of prey protein in the population**
  - The default value for this parameter is double the number of observable prey protein
- **Minimum prey/bait score indicating association**
  - This value is used to make a preliminary determination of prey protein LC-MS/MS detection or non-detection.
  - Protein association scores in the input greater than or equal to this value signify positive detections. Any values less than this cutoff signify non-detection.
    - **NOTE:** If all protein association values of the input data are greater than or equal to this cutoff an error will result. This would indicate that all prey proteins are associated with all bait proteins under all conditions. The analysis cannot be run without some values above this threshold.
- **Number of Monte Carlo simulations for LRT p-value estimate**
  - This parameter underpins the estimate of the p-value used in the screening likelihood ratio test featured in the local Bayesian estimator.
  - We have found a default value of 10,000 simulations to be adequate
- **False discovery rate**
  - Controls the false discovery rate of the analysis with respect to determining prey/bait pair association.
  - The default value is 0.05.
- **Cytoscape probability threshold**
  - A threshold posterior probability of prey-bait protein association for use in filtering the Cytoscape (Institute for Systems Biology et al, Seattle WA. http://www.cytoscape.org/) compatible network results file
  - The default value is 0.05
- **Two parameters related to the calculation of ubiquity scores**
  - Ubiquity scores are weighted averages of posterior probabilities of association between bait and prey proteins. They are calculated using a sigmoid curve similar to a step-function.
  - The first ubiquity parameter is the **Bayes' Odds Lower bound ($BO_L$)**. When the sigmoid is drawn, this Bayes' Odds value will correspond to a component ubiquity score of 0.01. Bayes' Odds this low or lower are providing an essentially zero weight for ubiquity. This horizontal line is shown on the graph in red, as is its vertical intercept of the sigmoid.

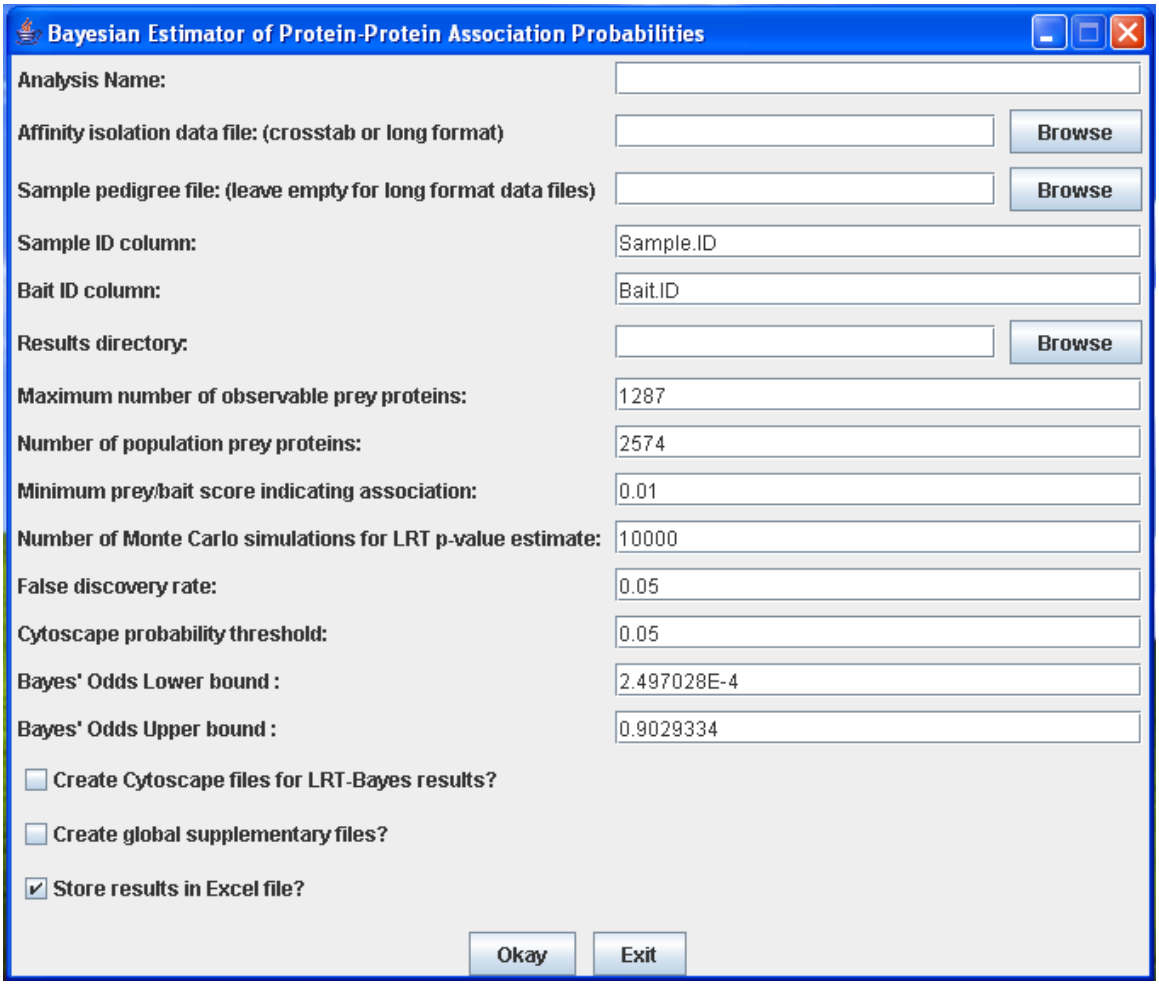
- The second ubiquity parameter is the **Bayes' Odds Upper bound ($BO_H$)**. When the sigmoid is drawn, this Bayes' Odds value will correspond to a component ubiquity score of 0.99. Bayes' Odds this high or higher are given a weight of approximately one for ubiquity. This horizontal line is shown on the graph in blue, as is its vertical intercept of the sigmoid.




  - **NOTE:** The shape of the ubiquity sigmoid determines the component ubiquity scores for Bayes' Odds between $BO_L$ and $BO_H$
  - **NOTE:** For nearly equivalent $BO_L$ and $BO_H$, the sigmoid is approximately a step function
- Default values for $BO_L$ and $BO_H$ are calculated using the "calculatePosteriors()" function using worst case estimates for pi (the overall probability of association between two random protein pairs) and the experimental false-positive and true-positive rates of protein association determination. These extremes are 0.05, 0.01, and 0.95, respectively. For $BO_L$, a rate of one observation in four trials is used. For $BO_H$, a rate of two observation in three trials is used
  - **NOTE:** These observation/trial rates are reasonable cutoffs for low confidence and high confidence results in our example dataset; however, these should be adjusted to reflect the different expectations of different datasets.

## 4.0 User Interface

To run BEPro[3], double-click the "BEPro[3]" shortcut on your desktop. The following window will appear:

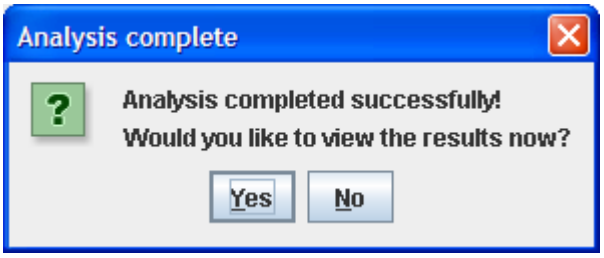



Enter the following information into this window:

- **Analysis Name**:
  - a unique name for this analysis which will be used to name output files
- **Prey protein by bait sample file**:
  - the *.csv file containing the annotated prey-by-bait protein association matrix
- **Sample pedigree file**:
  - the *.csv or *.txt file containing the sample pedigree data
- **Sample ID column**:
  - The exact name of the sample ID column in the pedigree file
  - **NOTE:** The default is Sample.ID

- **Bait ID column**:
  - The exact name of the bait ID column in the pedigree file
  - **NOTE:** The default is Bait.ID
- **Results directory**:
  - The directory in which to write output files
- **Maximum number of observable prey proteins** (see input data section 3.0)
- **Number of population prey proteins** (see input data section 3.0)
- **Minimum prey/bait score indicating association** (see input data section 3.0)
- **Number of Monte Carlo simulations for LRT p-value estimate** (see input data section 3.0)
- **False Discovery Rate** (see input data section 3.0)
- **Cytoscape probability threshold** (see input data section 3.0)
- **Bayes Odds Lower Bound ($BO_L$)** (see input data section 3.0)
- **Bayes Odds Upper Bound ($BO_H$)**(see input data section 3.0)
- **Store results in Excel file** (see output data section 5.0)

Once these values have been entered, click **Okay** to begin the analysis. During the analysis, a small window will open and remain static. When the analysis is successful and complete, a message will appear asking if you wish to view the results. The analysis may take several minutes depending on the size of the data files and the number of simulations run. While running the BEPro$^3$ GUI may go blank. This does not indicate a problem with the analysis.




If you select **Yes**, an HTML page will be opened in your web browser displaying a summary of the results. This HTML page is saved in the results directory.

For common errors, see troubleshooting section 7.0 at the end of this document.

## 5.0 Output Files

BEPro[3] creates several output files including an HTML summary of the analysis, a table of prey-bait protein frequencies of detection with supplemental statistics, and two tables of prey-bait protein probabilities of association with supplemental statistics (one each for the local and global Bayesian estimates). BEPro[3] also creates a .txt file suited for Cytoscape (Institute for Systems Biology. Seattle, WA. http://www.cytoscape.org). MS-Windows users may also have the *.csv tables written to one MS-Excel spreadsheet featuring one *.csv table per worksheet. These files are described in detail below.

1. HTML Analysis Summary
   <AnalysisName>_BEPro3_Analysis_Summary.html
   The analysis summary contains information about 1) the input data files, 2) the parameters used in the analysis, and 3) the location of the output files. The summary also includes estimates of the random protein association rate for a given proteome, and estimates of parameters related to the global Bayesian analysis including the false positive, true positive and false negative rates.
2. Frequency of Detection Table
   <AnalysisName>_<Date>_PreyBait_Frequencies.csv
   This file contains a table that lists prey proteins in the rows and bait proteins in the columns. The cells of the table indicate the number of replicates for each prey-bait pair for which the prey was observed, i.e. had an association score above threshold. It also shows the LRT decision between "Non-Uniform" and "Uniform" made for each prey, as well as the p-values, LRT statistics, row sums and row proportions of hits across all baits.
3. Posterior Probability Table
   <AnalysisName>_<Date>_LRT-Bayes_Posterior_Probabilities.csv
   This file is in the same format as the 'frequency of detection table' but the values of the cells are replaced with the Posterior Probabilities of Protein-Protein Association based on the LRT-Bayes method using local priors (Sharp *et al*, 2007). Additional information columns and rows include prey and bait specific ubiquity counts and percentages.
4. Supplementary Files for Global Analysis
   When using the crosstab/pedigree file input method, you may select an option to run a supplementary global analysis. The files generated include an additional Posterior Probability Table based on the global estimation of priors (Gilchrist *et al*, 2004) as well as other diagnostic and summary files such as TS.csv.
5. Cytoscape file
   <AnalysisName>_NetworkResults.txt
   This *.txt file contains a Bayes' Odds, a ubiquity score, a p-value, and other information for each prey-bait pair. This file is formatted for easy upload and visualization using Cytoscape.

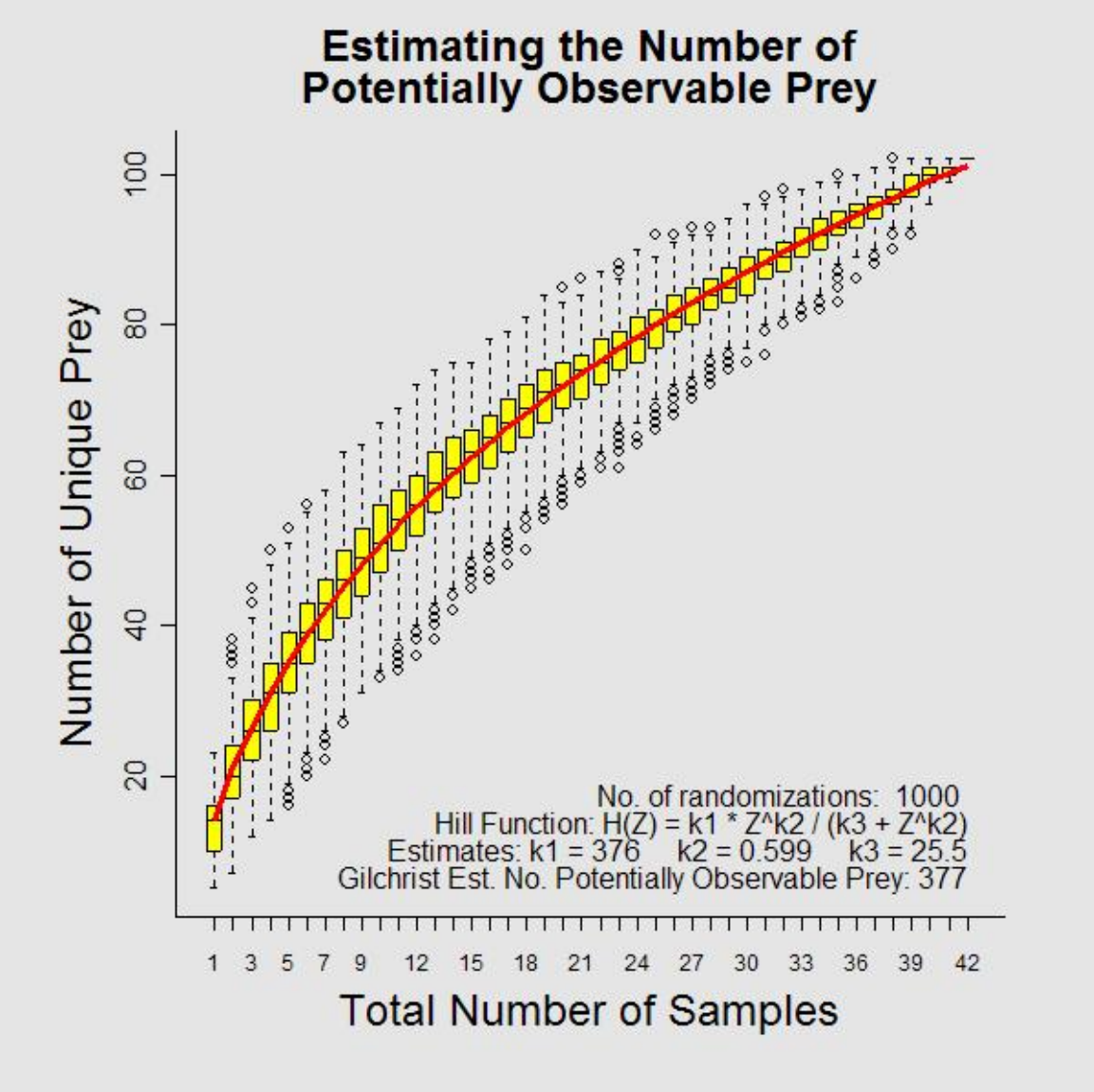
## 6.0 An Example.

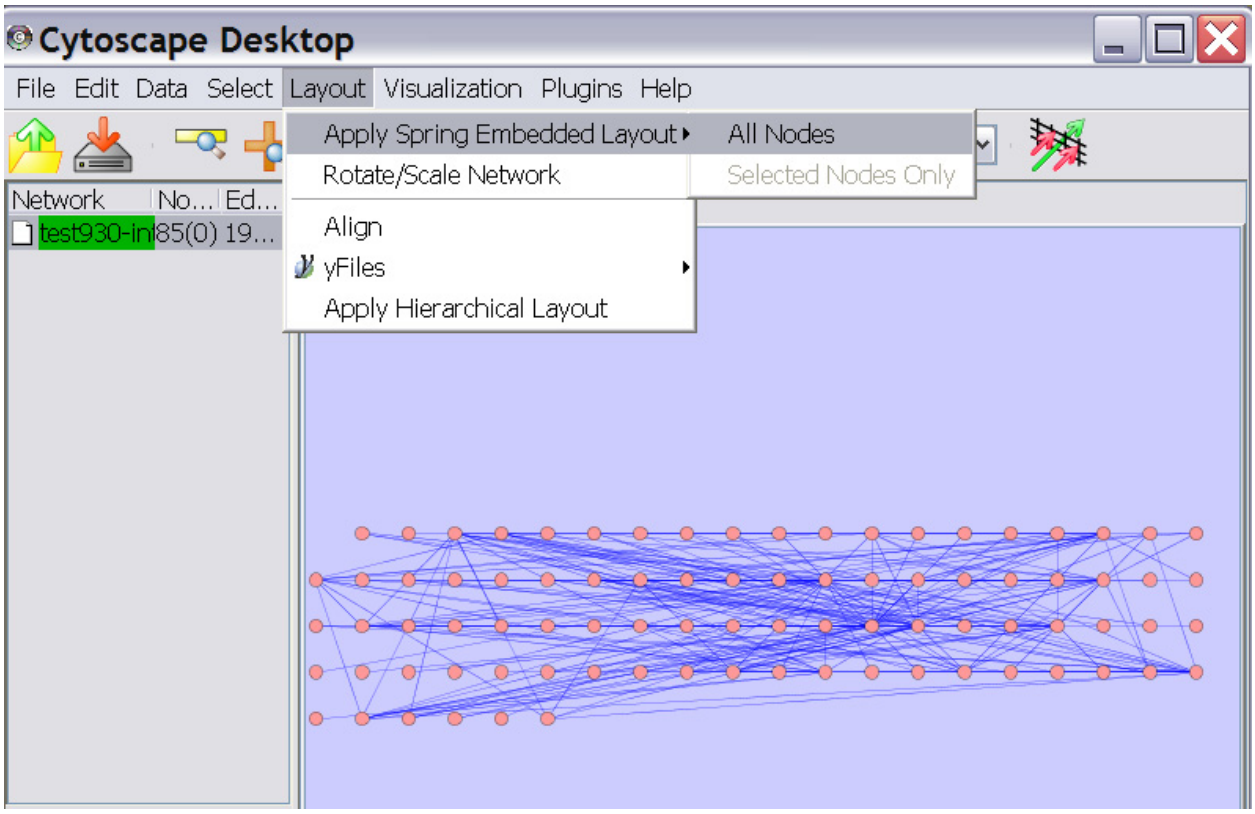
The objectives of this example are to illustrate the use of the software and to provide the user with results to verify the operation of the user's BEPro$^3$ installation. Any discussion of the biological significance of the results is secondary.

The example dataset contains MS-MS SEQUEST results on 70 affinity isolation samples of prey protein mixtures from 16 unique protein baits. The data file RPal_data.csv includes rows for 1287 prey proteins that have been observed over the course of numerous experiments. Each entry is the number of a prey's peptides observed in the given bait sample. Many of these rows, however, contain only zero entries because the corresponding prey proteins were not observed in this particular set of 70 samples.

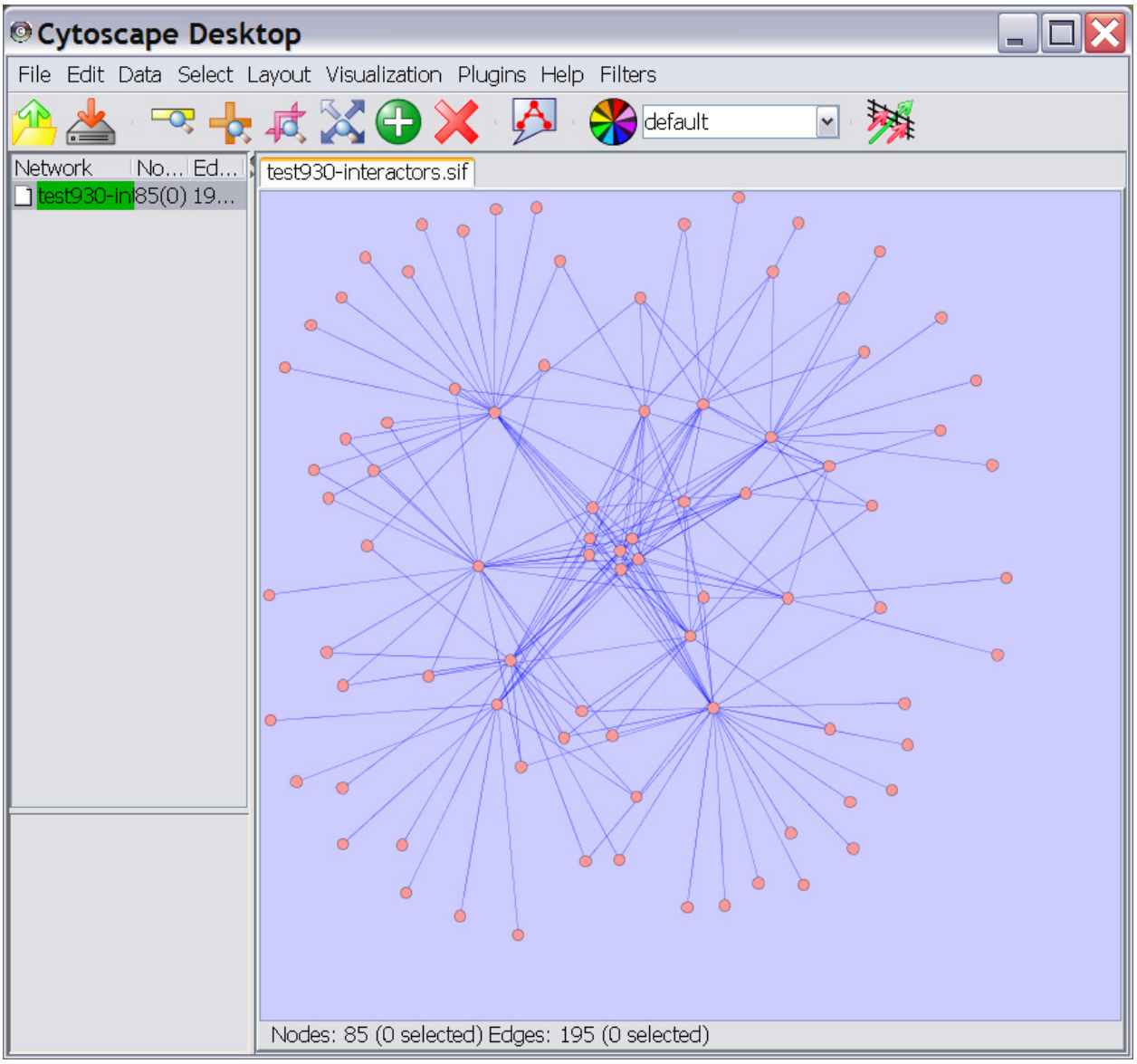
To run the illustration and obtain the verifying results, start BEPro$^3$ by clicking the BEPro$^3$ icon and then enter the necessary parameters. The example dataset can be found in the "\BEPro3\Example" directory. The prey-bait matrix ("RPal_data.csv") and pedigree file (Rpal_pedigree.csv") can be found in the "\BEPro3\Example\Data" Directory. For the example dataset, the remaining parameter values should be entered as shown in GUI figure (Section 4.0).

The various output files, and in particular, the HTML summary file, provide a reasonable summary of the analysis and results. The reader is encouraged to review the output files to obtain a better grasp of the results and compare these to the example results in order to verify the BEPro$^3$ installation. Example results may be found in the directory "\BEPro3\Example\Results\". The following image is one of the files produced and displays one way the global Bayesian algorithm estimates the number of potentially observable prey. Easily generated Cytoscape displays of the example results are also included.


Estimating the Number of Potentially Observable Prey
Number of Unique Prey
20
40
60
80
100
1
3
5
7
9
12
15
18
21
24
27
30
33
36
39
42
Total Number of Samples
No. of randomizations: 1000
Hill Function: H(Z) = k1 * Z^k2 / (k3 + Z^k2)
Estimates: k1 = 376 k2 = 0.599 k3 = 25.5
Gilchrist Est. No. Potentially Observable Prey: 377




The default layout is not very useful for this type of data. Change the layout to Spring Embedded under the Layout Menu.



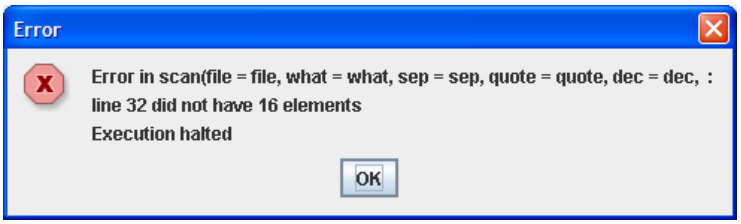

The Spring Embedded Layout is much clearer and makes the task of identifying prey proteins associated with multiple baits easier.

## 7.0 Troubleshooting common errors.

In testing this application, several common errors occurred, so it may be helpful to know the reason why, and how to fix them.

1. 

   a. Why: This occurs when two sample IDs are identical.
   b. Solution: Change the redundant sample name so all are unique.

2. 

   a. Why: This occurs when one or more cells in the data matrix are left blank.
   b. Solution: Fill each blank cell with a “below detection” value like zero.