A Bayesian estimator of protein–protein association probabilities

Jason M. Gilmore¹, Deanna L. Auberry¹, Julia L. Sharp², Amanda M. White¹, Kevin K. Anderson¹ and Don S. Daly¹,∗
¹Pacific Northwest National Laboratory, Battelle Boulevard, Richland, WA 99352 and ²Department of Applied Economics and Statistics, Clemson University, Clemson, SC 29634, USA

Received on October 8, 2007; revised on May 15, 2008; accepted on May 15, 2008

Advance Access publication May 22, 2008

1 INTRODUCTION

Identifying associations between proteins is essential to the larger goal of inferring protein networks and their functions. Two closely related techniques for uncovering protein associations are the endogenous and exogenous protein affinity isolation assays closely related techniques for uncovering protein associations are between bait and prey protein pairs using data from multiple-bait, multiple-replicate, protein liquid chromatography tandem mass spectrometry LC–MS/MS affinity isolation experiments.

Availability: BEPro³ is public domain software, has been tested on Windows XP, Linux and Mac OS, and is freely available from http://www.pnl.gov/statistics/BEPro3.

Summary: The Bayesian Estimator of Protein–Protein Association Probabilities (BEPro³) is a software tool for estimating probabilities of protein–protein association between bait and prey protein pairs using data from multiple-bait, multiple-replicate, protein liquid chromatography tandem mass spectrometry LC–MS/MS affinity isolation experiments. BEPro³ is a software tool for estimating probabilities of protein–protein association between bait and prey protein pairs using data from multiple-bait, multiple-replicate, protein liquid chromatography tandem mass spectrometry LC–MS/MS affinity isolation experiments.

Certainty in a prey identity and, hence, in a prey-bait association is gained through replicate assays. The proportion of observations of a prey protein across replicates of bait is naïve estimate of the probability of association between that prey and bait. This is a naïve estimate in that it does not account for false positive and false negative identifications. Further, not all true prey–bait associations are of interest. For instance, a prey protein, such as ribosomal protein, that binds indiscriminately, or ‘ubiquitously,’ across the proteome of a prey protein across replicates of bait is naïve estimate of the probability of association between that prey and bait. This is a naïve estimate in that it does not account for false positive and false negative identifications. Further, not all true prey–bait associations are of interest. For instance, a prey protein, such as ribosomal protein, that binds indiscriminately, or ‘ubiquitously,’ across the proteome is scored using a weighted average of its Bayes’ Odds scores (Sharp et al., 2007). A prey with many large Bayes’ Odds across baits scores high ubiquity and a prey with few large Bayes’ Odds values receives a low ubiquity score. A BEPro 3 Bayes’ Odds calculation depends upon three parameters: the chance that a randomly chosen protein associates with another randomly chosen protein, and the false positive and false negative identification rates. For a large proteome of N proteins, our prior belief is that the probability that two randomly chosen proteins associate is small, say 1/(N − 1). The false positive and false negative identification rates may vary from prey-to-prey and bait-to-bait because one prey may be more easily observed by LC–MS than another, and because of analytical differences in the affinity isolation assay, or differences in sample concentrations submitted to LC–MS. For a prey with a statistically significant LRT, the algorithm estimates the prey true positive and false positive rates by segregating the observed frequencies of prey observation into high and low frequency classes. For a prey with a non-significant LRT, the true positive and false positive rates are estimated with the medians of the rates for those prey with statistically significant LRTs.

It is important to note that the interpretations of the LRT, Bayes’ Odds and ubiquity statistics depend upon the design of the affinity isolation experiment. An experiment featuring all bait proteins that knowingly associate should result in non-significant LRTs, high Bayes’ Odds across almost all bait proteins and high ubiquity scores for those bait proteins also observed as prey. Whereas, a second experiment featuring randomly selected bait proteins

ABSTRACT

1554
should result in statistically significant LRTs (i.e. non-uniform observation frequencies across baits), high Bayes' Odds for certain bait-prey combinations, and low ubiquity scores for all but those promiscuously sticky prey.

3 SENSITIVITY AND SPECIFICITY

The statistical sensitivity and specificity of the BEPro³ algorithm may be assessed using an endogenous affinity isolation experiment with known protein complexes that involved 75 LCMS injections of 2–10 replicates of 16 bait proteins with 9 baits having 4 replicates and 4 baits with 5 or more replicates (Sharp et al., 2007). The resulting 200 prey by 16 bait frequency matrix contained 43 prey that were observed in all but one injection, and 106 that were observed in 5 or fewer injections.

We estimate BEPro³ sensitivity with the true positive fraction of prey–bait associations, or the proportion of known prey–bait pairs (as identified by literature mining and previous affinity isolation experiments) that have a high Bayes' Odds. Similarly, we estimate BEPro³ specificity with one minus the false positive fraction of prey–bait associations, or one minus the proportion of prey–bait pairs not known to interact that have a high Bayes' Odds.

Assuming the preliminary identification of prey–bait interactors is true, 3,104 of 3,200 observed prey–bait associations, or 97%, fall in the 'not known interactors' category in the example affinity isolation experiment. The estimated BEPro³ specificity, with a Bayes' Odds cutoff of 0.5, is about 95% (Fig. 1A). The estimated sensitivity is about 50% (Fig. 1B). Alternatively, if we accept the BEPro³ identification of prey–bait interactors, then 165 prey–bait pairs with high Bayes' Odds in the 'not known interactors' category (Fig. 1A) deserve further investigation as potential protein interactors. Further, the 48 pairs with low Bayes' Odds in the 'known interactors' category (Fig. 1B) may provide guidance to improving the assay.

4 IMPLEMENTATION

BEPro³ was designed, developed and packaged to ensure a sound implementation of its sophisticated statistical algorithm, to facilitate easy, sensible usage, to ensure easy availability and to encourage modification. The core statistical routines are written in the R language (The R project for Statistical Computing. Vienna, Austria. http://www.r-project.org). A JAVA user interface (Sun Microsystems, Inc., Santa Clara, CA, USA. http://java.sun.com) facilitates data management and setting analysis parameters. BEPro³ returns tabular results and an HTML-annotated analysis summary as text compatible with Excel (Microsoft, Inc., Redmond, WA, USA), Cytoscape (Institute for Systems Biology, Seattle, WA, USA. http://www.cytoscape.org) or a file/internet browser. This software requires R version 2.2 and Java 1.5.0, or more recent versions. BEPro³, R and Java are free and open, allowing for unrestricted distribution under a general GNU license. The self-installing BEPro³ package includes a highly integrated user guide, example dataset and analysis, and supplementary documentation that detail the specifics of the algorithm and its implementation.

ACKNOWLEDGEMENTS

Funding: Funding was provided by the US Department of Energy Office of Advanced Scientific Computing Research under contract 47901 and by the Office of Biological and Environmental Research under contracts 41966 and 43930 with the Pacific Northwest and the Oak Ridge National Laboratories.

Conflict of Interest: none declared.

REFERENCES

