research articles Journal of proteome research

Statistically Inferring Protein–Protein Associations with Affinity Isolation LC–MS/MS Assays

Julia L. Sharp,*,† Kevin K. Anderson,‡ Gregory B. Hurst,§ Don S. Daly,‡ Dale A. Pelletier,§ William R. Cannon,‡ Deanna L. Auberry,‡ Denise D. Schmoyer,§ W. Hayes McDonald,§ Amanda M. White,‡ Brian S. Hooker,‡ Kristin D. Victry,‡ Michelle V. Buchanan,§ Vladimir Kery,‡ and H. Steven Wiley‡

Clemson University, 237 Barre Hall, Clemson, South Carolina 29634-0313, Pacific Northwest National Laboratory, P.O. Box 999, Richland, Washington 99352, and Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, Tennessee 37837

Received February 27, 2007

Affinity isolation of protein complexes followed by protein identification by LC–MS/MS is an increasingly popular approach for mapping protein interactions. However, systematic and random assay errors from multiple sources must be considered to confidently infer authentic protein–protein interactions. To address this issue, we developed a general, robust statistical method for inferring authentic interactions from protein prey-by-bait frequency tables using a binomial-based likelihood ratio test (LRT) coupled with Bayes' Odds estimation. We then applied our LRT-Bayes' algorithm experimentally using data from protein complexes isolated from *Rhodopseudomonas palustris*. Our algorithm, in conjunction with the experimental protocol, inferred with high confidence authentic interacting proteins from abundant, stable complexes, but few or no authentic interactions for lower-abundance complexes. The algorithm can discriminate against a background of prey proteins that are detected in association with a large number of baits as an artifact of the measurement. We conclude that the experimental protocol including the LRT-Bayes' algorithm produces results with high confidence but moderate sensitivity. We also found that Monte Carlo simulation is a feasible tool for checking modeling assumptions, estimating parameters, and evaluating the significance of results in protein association studies.

Keywords: protein–protein interaction • affinity isolation • LC–MS/MS • likelihood ratio test • Bayes' Odds

Introduction

Revealing protein networks and their function, such as multiprotein complexes responsible for DNA replication, transcription, and protein degradation; cell signaling carried out by kinase cascades in which protein–protein interactions mediate phosphorylation; and metabolite channeling in which small molecules are channeled between enzymes, is an important component of systems biology research. Several methods exist to characterize protein–protein interactions on a large scale, including classical isolation of protein complexes by sedimentation, cross-linking, surface plasmon resonance, imaging, two-hybrid analyses, and affinity purifications. Of these, only two-hybrid analysis [1–4] and affinity purification methods [5–7] are readily scalable to a genomic level. Two-hybrid analyses initially had problems with high false-positive identifications, but the methods have been refined such that the most pressing issue currently is reducing the false-negative rate. Affinity purification methods isolate an affinity-tagged "bait" protein that interacts as part of a protein complex with "prey" proteins.[8] These methods suffer from both false-positive identification of non-biologically relevant interactions and from failure to identify other authentic interactions (false negatives).[9] The recent development of affinity methods in which the bait protein is expressed from a plasmid[10] has not only resulted in the use of affinity methods for a wider range of organisms, but also readily allows for replicate experimental measurements in less genetically tractable organisms.

An affinity isolation assay uses an affinity-tagged bait protein to isolate associated prey proteins from a cellular lysate. Proteins are identified using liquid chromatography (LC) tandem mass spectrometry (MS/MS), and protein associations are inferred from the list of identified proteins.

In affinity isolation assays, systematic and random assay errors can arise for a number of reasons. First, the proteins could be incorrectly identified. In most affinity isolation assays, the affinity-isolated proteins are treated with a proteolytic enzyme such as trypsin, and the resulting tryptic peptides are identified by liquid chromatography separation followed by MS/MS. The error rates of this process can be controlled so

* To whom correspondence should be addressed. Phone: 864.656.3028. Fax: 864.656.5776. E-mail: julia.1.sharp@gmail.com.
† Montana State University.
‡ Pacific Northwest National Laboratory.
§ Oak Ridge National Laboratory.

Published on Web 08/11/2007
10.1021/pr0701106 CCC: $37.00
© 2007 American Chemical Society

that a low false-positive rate is obtained; however, this is at the expense of a high false-negative rate.[11–14] False negatives can be due to failure to detect certain tryptic peptides that are above or below the m/z scan range of the mass spectrometer, or are not efficiently ionized. Also, despite the use of the LC separation to provide temporal separation of peptides in line with the mass spectrometric detection, the mixture of peptides eluting from the LC at any given time may exceed the capacity of the scanning program of the mass spectrometer to obtain an MS–MS spectrum of every peptide. A high false-negative rate can also occur if there is more than one protein complex that can form with a selected bait, resulting in a heterogeneous population of protein interactors. In this case, it may be that one set of proteins is present in much lower concentrations (so-called "substoichiometric binding"). The amount of prey protein present may be undetectable because of dynamic range issues. Dynamic range may also be an issue with affinity isolation assays in which the bait protein is present at a much higher concentration than the targeted prey proteins.

Second, depending on the experimental protocol, the relative proportions of proteins present after the purification steps may not reflect the proportions present in the cellular environment. An incorrect proportion of proteins can either lead to a false identification by driving the thermodynamic equilibrium toward binding, or may lead to a missed identification by driving the equilibrium toward dissociation.

Furthermore, the cell lysate is a well-mixed solution of proteins in an affinity isolation. Within a cell, the local concentration of some proteins can be quite high because of cellular sorting of proteins, while others can be quite low, though cells and biological systems in general follow thermodynamic laws. Deviating from the native cellular environment may be a major source of error (one not addressed here) in affinity isolation assays that possibly limits the affinity isolation methods to the identification of binary protein–protein interactions. Accounting for these errors is necessary to confidently infer significant cellular protein–protein interactions.

False positive prey identifications can occur if prey proteins are isolated due to reasons other than *in vivo* interaction with the bait protein. Proteins that (for example) are abundant in the cell, or interact with the affinity resin itself rather than the bait protein, may survive the affinity isolation and be detected as preys across a large number of affinity isolation experiments, regardless of whether there is an authentic interaction with the bait.

Protein–protein interactions for *Escherichia coli*, and *Saccharomyces cerevisiae* have previously been studied using mass-spectrometry outcomes and validation databases.[5–7,15] Typically, statistical analyses performed on these associations has been only descriptive. Gilchrist et al., however, statistically analyzed the probability that two proteins associate in the same protein complex using a Bayes' Odds calculation with global estimates of prior detection probabilities.[16] Shen, Li, and Chen furthered Gilchrist's method by defining the rate of true association in a proteome and integrating prey–prey and prey–bait protein association rates into the model of all protein pairs.[17] The method proposed here estimates the Bayes' Odds for each prey–bait pair using individualized estimates of prior probabilities of detection by exploiting the structure of an experiment to parse prey proteins to make individualized estimates of prior probabilities of detection.

Prey protein identifications form lists from an affinity isolation LC–MS/MS assay that are cross-tabulated by bait protein to form an annotated prey-by-bait frequency matrix. The matrix summarizes the frequency at which each prey was observed with each bait. The frequency pattern across a given row (prey), or within a given column (bait), gives information about the interactions between a prey and the baits, or a bait and the prey, respectively. Both row and column frequency patterns also reflect assay errors. Fundamental to our estimate of the strength of each inference about a bait–prey interaction is an evaluation of row homogeneity. We propose a scheme to classify each prey protein represented in the frequency matrix as "nonuniform" or "uniform" corresponding to prey being exclusively detected with one or more baits (i.e., a non-homogeneous pattern of detects), or non-exclusively detected (i.e., a homogeneous pattern of detects or nondetects) across all baits in the set. We then use this classification to estimate individual prior probabilities of prey identification and posterior probabilities of bait–prey interaction. We also use this classification to screen for ubiquitous interactors and non-interactors. These uses, however, are highly dependent upon the baits selected and the design of the experiment.

Detection homogeneity is assessed, and uniformity assigned, using a likelihood ratio test (LRT) by comparing the likelihood of the data assuming a (uniform) homogeneous frequency distribution to the likelihood of the data under all alternative (nonuniform) non-homogeneous distributions. LRT modeling assumptions are investigated, and conservative parameter and p-value estimates are obtained using Monte Carlo simulation. The probability of an interaction between each bait–prey pair in the frequency matrix is then estimated by calculating its posterior Bayes' Odds of interaction from its observed frequency and prior probability estimates that depend upon the prey's uniformity and the bait's frequency distribution. For an experiment that includes baits of sufficient diversity, the algorithm can discriminate against artifacts in the form of preys that are apparently detected for reasons other than an *in vivo* interaction with the bait. The algorithm is illustrated with an endogenous affinity isolation experiment investigating a small group of hypothesized protein subnetworks of *Rhodopseudomonas palustris*, including RNA polymerase, ATP synthase, DNA polymerase, and the RNA degradosome (see Supporting Information).

Materials and Methods

In an affinity isolation LC–MS/MS assay, proteins are identified from lists of detected peptides and supporting data. For our example analysis, we apply the "two-peptide" rule with which we tentatively infer a protein is present in a sample if two or more of a protein's peptides are observed with SEQUEST[18] search results of XCorr $\geq$ 1.8 for singly charged ions, $\geq$2.5 for doubly charged ions, or $\geq$3.5 for triply charged ions, and DeltaCN $\geq$ 0.08.[19]

Lists of prey protein identifications from an affinity isolation LC–MS/MS assay are cross-tabulated by bait protein to form an annotated prey-by-bait frequency matrix summarizing the frequency at which each prey was observed with each bait. 'Uniform' prey exhibit a homogeneous frequency distribution of positive MS identifications across a collection's bait proteins. 'Nonuniform' prey have significantly higher, or lower, frequencies for one or more baits relative to the others. The homogeneity of the frequency distribution for each prey across the experiment is evaluated with a binomial-based Likelihood Ratio Test (LRT). A prey protein is flagged "nonuniform" or "uniform" by comparing the p-value of the LRT to a critical p-value

selected to control the false discovery rate. The Bayes' Odds of interaction is estimated for each bait−prey pair with individualized prior probability estimates based on the outcome of the prey's LRT.

Blind reliance on statistical theory, models, and results undermines confidence and may produce erroneous inferences about protein−protein interactions. To improve confidence and reduce modeling errors, Monte Carlo simulation is used in three ways. First, assumptions underpinning the LRT are investigated for the employed design of the affinity isolation experiment. Second, conservative parameter estimates are made for LRT significance tests. Third, the LRT p-values are estimated from simulated LRT null sampling distributions.

Classifying a Prey Protein by Its Detection Frequency Profile. Classifying a prey protein with respect to the uniformity of its detection frequency across baits relies on a statistical hypothesis test comparing one probability model defined by a null hypothesis about the homogeneity of prey detection across baits to another probability model defined by an alternative hypothesis encapsulating any non-homogeneous prey detection.

The biological null, or naive, hypothesis is that a given prey protein uniformly associates, meaning that the given prey associates with no baits or all baits equally. This translates into the statistical null hypothesis that the probability p_j of observing the given prey protein with the jth bait protein is the same probability p for all baits:

$$H_o\colon\ p_j = p\ \forall\ j$$

The number of times S_j that a particular prey protein is observed in n_j replicate samples of the jth bait protein can be described by a binomial probability model. The likelihood of the observed frequencies for B baits under the null hypothesis is the product of the prey's binomial probabilities:

$$L_o(p) = \prod_{j=1}^{B}\binom{n_j}{S_j}p^{S_j}(1-p)^{n_j-S_j}$$

where n_j is the number of replicate LC−MS/MS runs of the jth bait, and S_j is the number of times the prey is observed with the jth bait.

The biological alternative hypothesis is that the prey protein is a nonuniform associator. That is, the prey will associate with at least one bait exclusively over the others, or the prey will associate with almost all the baits excluding one or more. The statistical alternative hypothesis is that at least one of the probabilities p_j of observing the given prey is not the same as the probability p of observing the given prey with all j baits:

$$H_a\colon\quad \text{at least one}\quad p_j \neq p$$

Under the alternative hypothesis, the likelihood of the observed frequencies is

$$L_a(p_1, \ldots, p_B) = \prod_{j=1}^{B}\binom{n_j}{S_j}p_j^{S_j}(1-p_j)^{n_j-S_j}$$

where p_j is the expected proportion of observations a prey has with the jth bait. The alternative likelihood equation is very similar to the null likelihood, featuring p_j in the place of p.

The interpretation of both the null and alternative hypotheses depend upon the design of the affinity isolation experiment, and particularly, the choice of bait proteins.

The Likelihood Ratio Test. The LRT compares the likelihood of the observed frequencies for a given prey protein under the statistical null hypothesis to the likelihood of the observed frequencies under the statistical alternative hypothesis. If the ratio favors the null hypothesis, the biological inference is that a given prey protein is a uniform interactor. If the ratio favors the alternative hypothesis, we infer that the prey protein interacts nonuniformly. The LRT statistic is

$$\lambda = \frac{\sup\limits_{p} L_o(p)}{\sup\limits_{p_j} L_a(p_1, \ldots, p_B)} \tag{1}$$

where, in practice, the numerator or denominator is the appropriate maximum likelihood estimate and the (large sample) LRT statistic is employed:[20,21]

$$J_T = -2\ln\lambda \tag{2}$$

$$= 2\sum_{j=1}^{B}\ln[\hat{p}_j^{S_j}(1-\hat{p}_j)^{n_j-S_j}] - 2\sum_{j=1}^{B}\ln[\hat{p}^{S_j}(1-\hat{p})^{n_j-S_j}] \tag{3}$$

where $\hat{p}_j = S_j/n_j$ is the observed proportion for the jth bait and

$$\hat{p} = \left(\sum_{j=1}^{B}S_j\right)\Big/\left(\sum_{j=1}^{B}n_j\right)$$

is the observed proportion across all the baits for a given prey.

The statistical significance of the LRT statistic J_T is commonly evaluated assuming that the number of replicates per bait is large enough so that the sampling distribution of J_T under the null hypothesis follows approximately a χ^2 distribution, a questionable assumption given the low number of replicates, often three or less, most often found in a large-scale affinity isolation experiment. The validity of this approximation for a given affinity isolation experiment may be evaluated empirically using Monte Carlo (MC) simulation by comparing an MC-simulated J_T distribution to the appropriate χ^2_{B-1} distribution.

The MC approach closely approximates the sampling distribution of J_T for a given affinity isolation design (i.e., the numbers of baits and replicates of each) by simulating thousands of possible detection frequency profiles, computing J_T for each simulated profile under the null hypothesis, and then compiling the empirical distribution (histogram) from these J_T values. This MC simulation is repeated for every overall proportion p expected for the given experiment. If the sampling distribution of the LRT statistic under the given p and design of the affinity isolation experiment is well-approximated by a χ^2 distribution, then $C = (1-\alpha)100\%$ of the simulated J_T's should be smaller than the $\chi^2_{B-1,\alpha}$ critical value. If the theoretical χ^2_{B-1} distribution does not closely agree with the MC sampling distributions, the common reliance on the χ^2_{B-1} distribution to determine LRT significance is invalid (see Supporting Information: The large sample assumption and LRT significance).

Determining the Statistical Significance of the LRT. If the χ^2 assumption is not appropriate (which is likely, given few replicates per bait), the statistical significance of LRT (or p-value) may be estimated using a similar Monte Carlo simulation and a conservative estimate for the row proportion, $\hat{p}_{\text{bound}}$, that controls the false-positive rate and facilitates executing the LRT test when the overall proportion is zero or one.[22] The conservative value $\hat{p}_{\text{bound}}$ may be estimated in a separate Monte

Carlo exercise (see Supporting Information: Determining the statistical significance of the LRT).

As in the χ^2 assumption check, a MC approximation simulates thousands of realizations of the experiment, computes J_T for each realization, and compiles the sampling distribution from these values. Again, one realization is a set of observances simulated using the binomial distribution under the null hypothesis following the design of the affinity isolation experiment. In this case, however, the MC approximation is repeated for every observed overall proportion $\hat{p} = \hat{p}_{\text{bound}}$ for the experiment.

To estimate the LRT p-value for a given prey protein, the appropriate reference distribution simulation is chosen based on the prey's observed overall proportion $\hat{p}$. The p-value is then the proportion of simulated test statistics equal to or larger than the observed test statistic.

Controlling the Rate of False-Positive LRTs. When evaluating a test for a given prey, the p-value is compared to an appropriate false-positive rate (i.e., the rate at which the null hypothesis is rejected when it is in fact true), say $\alpha = 0.05$. If the p-value is larger than the chosen false-positive rate, the null hypothesis is not rejected, and the prey protein is deemed 'uniform'. If the p-value is smaller, the null hypothesis may be rejected, and the prey protein may be regarded as 'nonuniform'. Because many LRTs are performed simultaneously when evaluating an affinity isolation experiment, a greater number of prey may be identified as "nonuniform" by chance than called for by the chosen false-positive rate.[23] To reduce this incidence, we control this false discovery rate using a method by Benjamini and Hochberg[24] (see Supporting Information: Controlling the rate of false-positive LR tests).

Bayes' Odds of a Prey–Bait Interaction. If a prey does interact with a given bait, let the likelihood of a positive observation be a binomial distribution with parameter β. If a prey does not interact with a given bait, let the likelihood of a positive observation be a binomial distribution with parameter θ. The parameters β and θ are the true-positive and false-positive rates, respectively, of an affinity isolation assay for a given bait–prey pair.

Suppose there is knowledge prior to the experiment about whether a prey–bait association exists; this prior knowledge may be used in a Bayes' Rule application in a manner similar to Gilchrist et al.[16] Suppose the prior belief that prey i is associated with bait j is the probability π_{ij}. Then, the prior belief that prey i is not associated with bait j is the probability $1 - \pi_{ij}$.

By Bayes' Rule, the Bayes' Odds that prey i is associated with bait j is

$$P_{ij} = \left(1 + \frac{(1-\pi_{ij})\theta_{ij}^{S_{ij}}(1-\theta_{ij})^{n_j - S_{ij}}}{\pi_{ij}\beta_{ij}^{S_{ij}}(1-\beta_{ij})^{n_j - S_{ij}}}\right)^{-1}$$

where S_{ij} is the number of observances of prey i in n_j replicate affinity isolations of bait j as recorded in the prey-by-bait frequency matrix.

The Bayes' Odds (P_{ij}) quantifies the posterior confidence that prey i is associated with bait j. The Bayes' Odds here differs from the Bayesian probability of association proposed by Gilchrist et al.[16] in that the parameters β_{ij} and θ_{ij} may vary from prey-to-prey and bait-to-bait. These parameters may vary from prey-to-prey because one prey is more easily observed by LC–MS/MS than another. These parameters may vary from bait-to-bait because of analytical differences such as the difference of the affinity tag on function, associations of different bait proteins, or differences in sample concentrations submitted for LC–MS/MS. We attempt to exploit our more intimate knowledge of and involvement in the design and execution of an affinity isolation assay in hopes of improving inferences about probabilities of association under these conditions.

Estimating the Bayes' Odds Parameters. Consider the prior belief π_{ij} that prey i interacts with a randomly chosen bait j. Suppose the proteome is thought to have N proteins, and each of these proteins are thought to bind uniquely with only one other protein. If a protein is chosen at random from the N, the probability that the next randomly chosen protein will bind to it was $1/(N-1)$. Therefore, the prior belief that two proteins associate with each other is small; the estimated prior probability used for the Bayes' Odds was $\hat{\pi}_{ij} = 1/(N-1)$.

For each prey, an algorithm finds the largest difference between the estimated prey true-positive ($\hat{\beta}$) and false-positive ($\hat{\theta}$) rates (it is assumed that $\beta > \theta$). The algorithm assumes that the observed frequencies across the baits for a given prey can be segregated into two classes: low frequencies resulting from no association between some baits and the prey, and high frequencies resulting from associations between some baits and the prey. Because values of $\hat{\beta} = 1$ and $\hat{\theta} = 0$ will result in unintuitive values of the Bayes' Odds (e.g., $\hat{\theta} = 0$ can result in Bayes' Odds of one), the estimates are calculated using a Bayesian-based adjustment. The adjustment uses the fact that a binomial likelihood for observing S successes in n trials, combined with a beta prior distribution, $Beta(\gamma_1, \gamma_2)$, yields a beta posterior distribution $Beta(S + \gamma_1, n - S + \gamma_2)$, which has as its expected value

$$\frac{S + \gamma_1}{n + \gamma_1 + \gamma_2}$$

The adjusted parameter estimates will take this form using $\gamma_1 = \gamma_2 = 0.25$.

For a particular prey protein, or row in the prey-by-bait frequency matrix, the adjusted observed bait proportions are ordered from smallest to largest ($\hat{p}^*_{(1)} \leq \hat{p}^*_{(2)} \leq \ldots \leq \hat{p}^*_{(B)}$), where

$$\hat{p}^*_{(j)} = \frac{S_{(j)} + \gamma_1}{n_{(j)} + \gamma_1 + \gamma_2}$$

The ordered proportions $\hat{p}^*_{(j)}$ are divided into two groups, $\{\hat{p}^*_{(1)}, \ldots, \hat{p}^*_{(j)}\}$ and $\{\hat{p}^*_{(j+1)}, \ldots, \hat{p}^*_{(B)}\}$, from which θ is estimated with the data of the lower group and β is estimated with the data of the upper group. This is done for $j = 1, \ldots, B-1$. For example, for $j = 1$, $\hat{\theta}_{(1)} = \hat{p}^*_{(1)}$ and

$$\hat{\beta}_{(1)} = \frac{S_{(2)} + \ldots + S_{(B)} + \gamma_1}{n_{(2)} + \ldots + n_{(B)} + \gamma_1 + \gamma_2}$$

For the kth cut point, the estimates are

$$\hat{\theta}_{(k)} = \frac{S_{(1)} + \ldots + S_{(k)} + \gamma_1}{n_{(1)} + \ldots + n_{(k)} + \gamma_1 + \gamma_2}$$

and

$$\hat{\beta}_{(k)} = \frac{S_{(k+1)} + \ldots + S_{(B)} + \gamma_1}{n_{(k+1)} + \ldots + n_{(B)} + \gamma_1 + \gamma_2}$$

The parameter estimates of θ and β for a particular prey are $\hat{\theta}_k$ and $\hat{\beta}_k$ that have the largest difference between them (see Supporting Information: Estimating the Bayes' Odds parameters).

This estimation algorithm is also applied to each bait protein or column in the prey-by-bait frequency matrix. Here, the adjusted observed prey proportions are ordered from smallest to largest, and repeatedly divided into two groups. In this case, the $\hat{\theta}_k$ and $\hat{\beta}_k$ that have the largest difference between them are the estimates of θ and β for a particular bait. These estimates, along with the results from the likelihood ratio testing, are used to calculate the Bayes' Odds for each bait−prey pair.

For a prey in the "nonuniform" class, its prey-specific parameter estimates $\hat{\theta}$ and $\hat{\beta}$ are used to calculate the Bayes' Odds between that prey and each bait. For a "uniform" prey, the prey's estimated row proportion $\hat{p}_k$ becomes $\hat{\theta}$ and is combined with each bait's $\hat{\beta}$ to estimate a particular bait−prey pair's Bayes' Odds.

A Screen for Ubiquitous Interactors. A ubiquitous prey protein is one that interacts with many of the baits. A general formulation of a measure of ubiquity involves the prey's Bayes' Odds

$$U_i = \frac{1}{B}\sum_{j=1}^{B} f(P_{ij}) \tag{4}$$

where $f(x)$ is an increasing function that maps the interval [0, 1] into [0, 1]. The measure is large (near 1) if the prey has large Bayes' Odds for most of the baits considered and is small (near 0) if most of the Bayes' Odds are small.

Various choices of $f(x)$ are possible. For example, let $f(x) = 0$ for $x < a$, and $f(x) = 1$ for $x \geq a$. Then, for the given prey, U_i represents the fraction of baits that have Bayes' Odds greater than a. As another example, let

$$f(x) = \left(1 + c\left(\frac{1-x}{x}\right)^a\right)^{-1} \tag{5}$$

with user-defined parameters $a > 0$ and $c > 0$. This $f(x)$ produces an "S" curve; its steepness is controlled by a (bigger is steeper), while its position is controlled by c (a small value is to the left, a big value is to the right). Reasonable values are $a = c = 6$.

Interpretation of U_i depends on the number of baits, B, and their relationship. For example, if most or all of the baits in an analysis belong to a complex, then it would be expected that if a prey associates with one, it should associate with many of the baits. On the other hand, if the baits are fairly "independent," then the ubiquity measure has greater meaning. For some experimenters, the product of B and U_i, approximately the number of baits with high Bayes' Odds, may be more informative.

A Java/R Implementation. We have developed a self-installing, user-friendly implementation of this LRT-Bayes' algorithm. This statware (Bayesian Estimator of Protein−Protein Association Probabilities, or BEPro[3]) is an open-source, open-development coding in R (R Foundation for Statistical Computing, Vienna, Austria), a well-established statistical programming language, wrapped in a user interface based on Java (Sun Microsystems, Inc., Sunnyvale, CA). This allows users to know exactly the implemented statistical operations and facilitates modification or extension to improve the algorithm. The self-install package includes a tutorial featuring the discussed example and documentation. This statware can be freely obtained and used by Windows, Mac, and Linux users alike (http://www.pnl.gov/statistics/).

Illustrating the LRT-Bayes' Algorithm: An Example Affinity Isolation Experiment. Four model protein complexes that are conserved among multiple bacterial species and have been well-studied in other organisms were chosen to test subnetwork identification using an endogenous isolation method in *R. palustris* (see Supplementary Table 4 in Supporting Information). They are DNA-dependent RNA polymerase,[25] ATP synthase,[26] the RNA degradosome,[27] and DNA-dependent DNA polymerase III.[28] Candidate bait proteins were chosen from orthologues of known subnetwork members from *E. coli* that have also been reported in a large-scale affinity/mass spectrometry study.[7] The *E. coli* results were obtained from an organism for which the affinity-tagged proteins can be introduced into the chromosome;[7] in the present study, reagent-tagged proteins were expressed episomally from an introduced plasmid.

Ideally, the affinity isolation procedure would yield a sample containing a purified mixture of the proteins from a complex in which the bait protein participates as a protomer. However, protein−protein interactions exhibit a wide variety of physical and chemical characteristics that complicate the isolation of pure, intact complexes. Combined with uncertainties in the detection method, results of protein−protein interaction experiments are known to be fairly noisy, with significant false-positive and false-negative rates. Therefore, the expected outcome of a real experiment is the detection of the bait protein, some proteins that are authentic interactors with the bait, and some proteins that are not authentic interactors with the bait; missing from the result may be some proteins with which the bait interacts *in vivo*. For the four model complexes studied here, one expects each affinity isolation assay to identify one or more of the protein components expected to be associated with that complex, based on literature reports from other bacterial species.

Results and Discussion

The example endogenous affinity isolation experiment involved 75 LC−MS/MS injections spanning from 2 to 10 replicates of 16 bait proteins with 9 baits having 4 replicates and 4 baits with 5 or more replicates. The prey-by-bait frequency matrix contained 200 prey; 43 prey were observed in all but one injection, and 106 were observed in 5 or fewer injections. In total, 110 prey were tagged "nonuniform," of which 102 prey were observed with only a few baits, while the 8 others were observed with all but a few baits. Of the 90 "uniform" prey, 89 were observed 10 or fewer times randomly across injections, with 1 observed randomly 66 times across injections.

The two panels of Figure 1 summarize the results of the LRT-Bayes' Odds analysis of the four model complexes for the example 16 bait, 200 prey, 75 LC−MS/MS injection experiment. Panel A shows the distribution of 3104 prey-bait protein Bayes' Odds for prey that are not known members of a complex targeted by the 16 chosen baits; Supplementary Table 5 in Supporting Information lists the corresponding bait−prey pairs for which the LRT-Bayes' Odds is larger than or equal to 0.5. Panel B shows the distribution of 96 prey−bait protein Bayes' Odds for prey proteins that are known members of a targeted complex; Supplementary Table 6 in Supporting Information lists the bait−prey pairs summarized by panel B. The Supplementary Spreadsheet Example_RPal.xls in Supporting Information contains the prey-by-bait matrix and the results from the

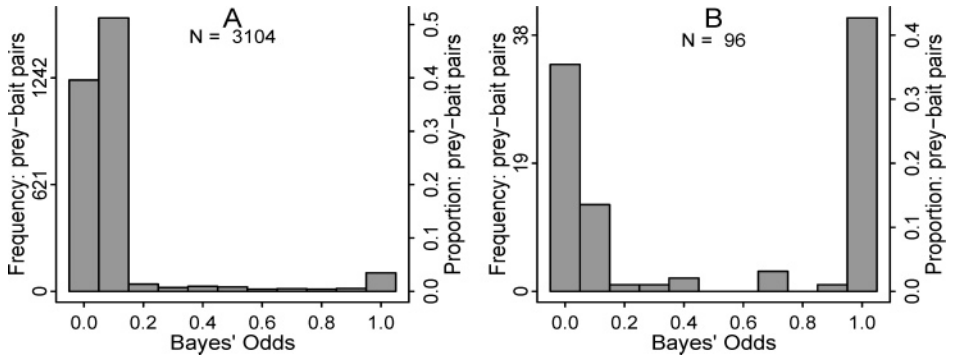

Figure 1. LRT-Bayes' Odds sensitivity and specificity. Empirical distributions of Bayes' Odds for prey proteins that are not known components of a complex targeted by the 16 chosen baits (A) and for those preys that are known members of a targeted complex (B). As expected, 97% (3104/3200) of all prey–bait associations fall in the "not known components" category given the design of the affinity isolation assay and experiment. Assuming the preliminary assessment of complex membership is true, the specificity of the LRT-Bayes' Odds classifier (using a Bayes' Odds cutoff of 0.5) is about 95% as displayed in panel A with a sensitivity of about 50% as shown in panel B. Believing the Bayes' Odds classification, 165 prey–bait pairs in the "not known components" category with high Bayes' Odds deserve further investigation as potential protein interactors. The 48 pairs with low Bayes' Odds in the "known components" category may provide guidance to improving the assay.

LRT-Bayes' Odds analysis. The example analysis also includes results from the Gilchrist et al. algorithm.[16]

From literature mining and previous affinity isolation experiments, the prey–bait pairs can be classified as members or non-members of a complex. Interactions defined as "known" were chosen from gene annotation and literature evidence (see Supplementary Table 4 in Supporting Information). These included any interactions, within a single complex, among components of the 4 model complexes mentioned above: for RNA polymerase, the core components (α, β, β', ω subunits) and several sigma factors (σ^{70}, σ^{54}, σ^{32}, σ^{E}, σ^{FecI}, and an ECF-type sigma factor, RPA4225); for ATP synthase, the cytosolic components (α, ϵ, β, γ, δ subunits) and membrane components (A, B, B′, C, I subunits); for DNA polymerase III, the α, β, ϵ, $\tau + \gamma$, δ, δ', χ subunits; and for the degradosome, pnp, rne, two homologues of rhlB (RPA3670 and RPA2477), a homologue to srmB (RPA2478), dead, and eno. Interactions between proteins not in this list, or between proteins in this list that are from different complexes, were defined as "unknown." While not perfect, this classification provided a useful reference point for evaluating the results of the algorithm.

The statistical sensitivity and specificity of the LRT-Bayes' Odds association classification can be judged using the Bayes' Odds calculations. Sensitivity is the proportion of prey–bait pairs where the pairs are known to be members of one of the complexes targeted, and the Bayes' Odds suggest that they are in fact known members of one of the targeted complexes. Specificity is the proportion of prey–bait pairs that are not members of the complex, and the Bayes' Odds reflect that they are not members. To judge the sensitivity and specificity of this prey–bait affinity isolation assay, consider a Bayes' Odds < 0.5 as suggesting that there is not an association between a prey–bait pair, whereas a Bayes' Odds ≥ 0.5 suggests an association between a prey–bait pair. Panel A in Figure 1 represents those prey–bait pairs that are not members of a complex. The prey–bait pairs that have Bayes' Odds < 0.5 represent the LRT-Bayes' Odds specificity, or the proportion of prey–bait pairs that are not known from the literature to associate that are identified as such by the Bayes' Odds. The specificity for the LRT-Bayes' Odds is approximately 0.95. Panel B in Figure 1 represents those prey–bait pairs that are members of a complex. The prey–bait pairs that have Bayes' Odds ≥ 0.5 in panel B represent the sensitivity of the LRT-Bayes' Odds. This affinity isolation experiment has a sensitivity of approximately 0.5.

The classification of members or non-members of a complex is done using previous experiments and literature. When the Bayes' Odds critical value of 0.5 is used, an assessment can be made about the validity of this classification. Whereas the specificity of the LRT-Bayes' Odds was assessed in panel A of Figure 1 using the prey–bait pairs with Bayes' Odds smaller than 0.5, the histogram also suggests that, 95% of the time, the classification is correctly identifying non-members, or that 5% of the time the classification of non-members is incorrect. These 5% of cases with Bayes' Odds ≥ 0.5 need to be scrutinized more closely as to their interactor relationship with the baits. Of the 153 interactions in Supplementary Table 5 in Supporting Information, 62 involve preys that associate with a large fraction of the baits used in this study ($U_i > 0.35$ for these preys). Aside from these suspect interactions, the other cases could represent novel interactions, could help identify problems with the experimental assay, or could be a result of the relatively small size of the experimental data set examined (see Supplementary Table 5 in Supporting Information). Panel B of Figure 1 suggests that about 50% of the prey–bait pairs with Bayes' Odds < 0.5 classified as members of a complex may require further investigation. Bait–prey pairs with low Bayes' Odds in panel B of Figure 1 include interactions between proteins that, while members of the same complex, would not be expected to interact, such as different sigma factors of RNA polymerase (see Supplementary Table 6 in Supporting Information).

The classification of members or non-members of prey–bait pairs to a complex from literature is of varying degrees of quality. A prey may have been assigned to the non-member class because there was no literature indicating it belonged to the network member class. The LRT-Bayes' Odds classification depends upon the specificity and sensitivity of the endogenous affinity isolation assay. For example, water-insoluble proteins would not be detected no matter their network membership classification. The LRT-Bayes' Odds classification tool appears to exhibit high specificity and acceptable sensitivity. Prey–bait pairs with large Bayes' Odds scores are most likely interactors, whereas prey–bait pairs with small Bayes' Odds are most likely

not interactors, though this inference is not strong. The LRT-Bayes' classifier may provide useful information when vetting a prior classification of prey−bait protein interactors.

The histograms for unknown and known complex members for the "nonuniform" and "uniform" likelihood ratio categories were similar, implying that the sensitivity and specificity of the LRT-Bayes' Odds are similar. The three components of the "core" RNA polymerase enzyme that were expressed as baits (rpoA, rpoB, and rpoC) yielded nonuniform interactions with high Bayes' Odds for all four core components (the three tagged proteins and rnpO), as well as the major sigma factor (σ^{70}). Two tagged core components also yielded high Bayes' Odds for interaction with the heat-shock sigma factor (σ^{32}). The two sigma factors, σ^{32} and σ^{E}, that were successfully expressed as baits yielded high Bayes' Odds for some, but not all, core components. The core components and sigma factors that were detected as preys were also scored as "nonuniform" interactors. In this case, high Bayes' Odds scores and nonuniform prey calls were obtained for known components of the RNA polymerase complex (see Supporting Information: Results and Discussion).

In addition to core components and sigma factors, other proteins were identified as interacting with the affinity-tagged baits. Heat shock proteins dnaK (RPA0333) and grpE (RPA0331) were identified when σ^{32} served as bait; these two proteins have also been observed in association with σ^{32} in *E. coli*.[29] Other interactors identified with high scores include RPA0060, a 14.5 kDa acidic conserved protein of unknown function. RPA0060 provides an example of the promise of the algorithm for identifying potentially novel protein−protein interactions. In subsequent experiments, RPA0060 was employed as the affinity-tagged bait, based on detection of RPA0060 as a prey for rpoA. In these experiments, rpoA was detected as a prey for the RPA0060 bait in each of two LC−MS/MS measurements performed on two separate cultures (data not shown). This "reciprocal" result strengthens the inference that RPA0060 is associated with the RNA polymerase complex. A similar protein (YacL) was identified as interacting with RNA polymerase in *E. coli*.[7] A third group of interacting proteins are also identified as nonuniform interactors with high Bayes' Odds but appear to have no known biological role in the complex under study: RPA4642 (fructose bisphospate aldolase) and RPA2953 (Hu DNA binding protein). These proteins exhibit relatively high values of U_i (0.70 and 0.93, respectively) and can thus be identified as likely artifactual preys that are not selectively isolated with baits in the RNA polymerase complex.

When members of the F1 portion of the ATP synthase complex were used as baits, nonuniform interactions were identified for all of the soluble F1 head piece members. For example, when RPA0177, the γ-subunit, was used as bait, nonuniform interactions for all members of the soluble F1 portion of ATP synthase were identified with high Bayes' Odds, including α, β, δ, and ϵ subunits. Additionally a number of interactions with proteins not previously reported in the literature were identified with high Bayes' Odds. As noted above for RNA polymerase, a number of these interactions involved prey proteins with relatively high values of U_i and are thus not likely to be authentic.

Inferences about protein−protein interactions that are derived from affinity-isolation experiments can be strengthened by performing "reciprocal" experiments, in which a prey protein identified in one experiment is employed as the bait in a subsequent experiment. Table 1 illustrates the LRT-Bayes' Odds scores for several protein−protein interactions from RNA polymerase and ATP synthase that were identified in reciprocal experiments. All of these interactions exhibit high Bayes' Odds for both bait−prey orientations, providing increased confidence that these measurements identified authentic biological protein−protein interactions.

Table 1. Reciprocally Tagged Protein LRT-Bayes' Odds

complex	protein 1	protein 2	LRT-Bayes' Odds: protein 1 bait/ protein 2 prey	LRT-Bayes' Odds: protein 2 bait/ protein 1 prey
RNA polymerase	RPA3226	RPA0367	0.984	1.000
	RPA3226	RPA3267	1.000	1.000
	RPA3226	RPA3268	0.990	1.000
	RPA3267	RPA0367	0.984	1.000
	RPA3267	RPA3268	0.990	1.000
ATP synthase	RPA0175	RPA0177	1.000	0.87
	RPA0176	RPA0177	1.000	0.976
	RPA0176	RPA0178	1.000	1.000

DNA polymerase III was chosen as a model for "low-abundance" species; native DNA polymerases are generally present in very low quantities in bacterial cells (10−20 copies per cell in *E. coli*[30]). Methods based on affinity isolation combined with mass spectrometric identification for detecting protein−protein interactions are known to be biased against low-abundance proteins.[9] In this study, when putative members of the DNA polymerase III complex or the RNA degradosome complex were used as baits, no obvious biologically relevant interactions were identified by manual inspection of the interaction data which was supported by the lack of nonuniform interactions identified with high Bayes' Odds scores.

Conclusions

The LRT-Bayes' method is an informative tool for appraising observations of protein association from protein affinity isolation LC−MS/MS experiments. The algorithm employs an objective statistical procedure, in contrast to often subjective, ad hoc appraisal methods. The example LRT-Bayes' analysis produced biologically sensible results that evoke more confidence than ad hoc analyses. Monte Carlo simulation proved a useful and doable alternative to check model assumptions, estimate model parameters, and judge significance. The LRT-based classification facilitated disparate follow-on analyses through segregation of the prey proteins.

Generally, prey proteins with small Bayes' Odds have relatively low frequencies, and for those in the "uniform" class, these frequencies are homogeneously distributed across all baits and are inferred to be false-positive observations. Prey with large Bayes' Odds are likely to be truly associated with one or more bait proteins or are ubiquitous proteins that associate non-discriminantly with the bait or affinity resin in the collection. If baits from a sufficient variety of complexes are included in the experimental design, these ubiquitous proteins can be flagged based on a weighted and normalized sum of their Bayes' Odds for interactions with all baits.

The reliance on large sample statistical theory is common but often does not apply. Powerful computing options are available to simulate the distributions associated with the data. Here, MC simulations were useful in examining the statistical assumptions, in generating data to find approximate p-values, and in determining which conservative bound on the row proportion should be used. For this LRT application, a comprehensive MC simulation showed that an LRT p-value, a

measure of LRT statistical significance, could deviate significantly from the true p-values when estimated by the common appeal to the χ^2 distribution. A conservative p-value estimate from a second simulation of the actual LRT statistics is more acceptable for affinity isolation studies with small numbers of replicates per bait.

Acknowledgment. This research was funded by the Genomics: GTL program and the U.S. 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. Pacific Northwest National Laboratory is a multi-program national laboratory operated by Battelle Memorial Institute for the U.S. Department of Energy under Contract DE-AC05-76RL01830. Oak Ridge National Laboratory is managed and operated by UT-Battelle, LLC, for the U.S. Department of Energy under contract DE-AC05-00OR22725.

Supporting Information Available: Supplementary Tables 1–6. Supplementary Figures 1–4. Supplementary Spreadsheets 1–3. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Fields, S.; Song, O. *Nature* **1989**, *340*, 245–246.
(2) Uetz, P., et al. *Nature* **2000**, *403*, 623–627.
(3) Ito, T.; Chiba, T.; Ozawa, R.; Yoshida, M.; Hattori, M.; Sakaki, Y. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 4569–4574.
(4) Miller, J.; Lo, R.; Ben-Hur, A.; Desmarais, C.; Stagljar, I.; Noble, W.; Fields, S. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 12123–12128.
(5) Gavin, A., et al. *Nature* **2002**, *415*, 141–147.
(6) Ho, Y., et al. *Nature* **2002**, *415*, 180–183.
(7) Butland, G.; Peregrin-Alvarez, J.; Li, J.; Yang, W.; Yang, X.; Canadien, V.; Starostine, A.; Richards, D.; Beattie, B.; Krogan, N.; Davey, M.; Parkinson, J.; Greenblatt, J.; Emili, A. *Nature* **2005**, *433*, 531–537.
(8) Dziembowski, A.; Seraphin, B. *FEBS Lett.* **2004**, *556*, 1–6.
(9) von Mering, C.; Krause, R.; Snel, B.; Cornell, M.; Oliver, S. G.; Fields, S.; Bork, P. *Nature* **2002**, *417*, 399–403.
(10) Markillie, L. M.; Lin, C.; Adkins, J. N.; Auberry, D. L.; Hill, E. A.; Hooker, B. S.; Moore, P. A.; Moore, R. J.; Shi, L.; Wiley, S. H.; Kery, V. *J. Proteome Res.* **2005**, *4*, 268–274.
(11) Havilio, M.; Haddad, Y.; Smilansky, Z. *Anal. Chem.* **2003**, *75*, 435–444.
(12) Kapp, E. A.; Schutz, F.; Connolly, L. M.; Chakel, J. A.; Meza, J. E.; Miller, C. A.; Fenyo, D.; Eng, J. K.; Adkins, J. N.; Omenn, G. S.; Simpson, R. J. *Proteomics* **2005**, *5*, 3475–3490.
(13) Cannon, W. R.; Jarman, K. H.; Webb-Robertson, B.-J.; Baxter, D. J.; Oehmen, C. S.; Jarman, K. D.; Heredia-Langner, A.; Auberry, K. J.; Anderson, D. C. *J. Proteome Res.* **2005**, *4*, 1687–1698.
(14) Sadygov, R.; Wohlschlegel, J.; Park, S. K.; Xu, T.; Yates, J. R., III. *Anal. Chem.* **2006**, *78*, 89–95.
(15) Krogan, N. J., et al. *Nature* **2006**, *440*, 637–643.
(16) Gilchrist, M.; Salter, L.; Wagner, A. *Bioinformatics* **2004**, *20*, 689–700.
(17) Shen, C.; Li, L.; Chen, J. Y. *Proteins: Struct., Funct., Bioinf.* **2006**, *64*, 436–443.
(18) Eng, J. K.; McCormack, A. L.; Yates, J. R., III. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976–989.
(19) Tabb, D. L.; McDonald, W. H.; Yates, J. R. *J. Proteome Res.* **2002**, *1*, 21–26.
(20) Sentilles, D. *A Bridge to Advanced Mathematics*; The Williams & Wilkins Company: Baltimore, MD, 1975.
(21) Casella, G.; Berger, R. *Statistical Inference*, 2nd. ed.; Duxbury Press: Pacific Grove, CA, 2001.
(22) Kendall, M. G.; Stuart, A. *The Advanced Theory of Statistics*; Vol. 2, 4th ed.; MacMillan Publishing Co., Inc.: New York, 1979; Chapter 20, pp 114–116.
(23) Miller, R. *Simultaneous Statistical Inference*, 2nd ed.; Springer: New York, 1981.
(24) Benjamini, Y.; Hochberg, Y. *J. R. Statist. Soc., Ser. B* **1995**, *57*, 289–300.
(25) Borukhov, S.; Nudler, E. *Curr. Opin. Microbiol.* **2003**, *6*, 93–100.
(26) Kersten, M. V.; Dunn, S. D.; Wise, J. G.; Vogel, P. D. *Biochemistry* **2000**, *4*, 3856–3860.
(27) Carpousis, A. J. *Biochem. Soc. Trans.* **2002**, *30*, 150–155.
(28) Stukenberg, P. T.; Odonnell, M. *J. Biol. Chem.* **1995**, *270*, 13384–13391.
(29) Gamer, J.; Bujard, H.; Bukau, B. *Cell* **1992**, *69*, 833–842.
(30) McHenry, C. S. *Annu. Rev. Biochem.* **1988**, *57*, 519–550.

PR0701106