OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

# Bayesian Inference for Genomic Data Integration Reduces Misclassification Rate in Predicting Protein-Protein Interactions

**Chuanhua Xing[1]*, David B. Dunson[2]**

**1** Department of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina, United States of America, **2** Department of Statistical Science, Duke University, Durham, North Carolina, United States of America

## Abstract

Protein-protein interactions (PPIs) are essential to most fundamental cellular processes. There has been increasing interest in reconstructing PPIs networks. However, several critical difficulties exist in obtaining reliable predictions. Noticeably, false positive rates can be as high as >80%. Error correction from each generating source can be both time-consuming and inefficient due to the difficulty of covering the errors from multiple levels of data processing procedures within a single test. We propose a novel Bayesian integration method, deemed nonparametric Bayes ensemble learning (NBEL), to lower the misclassification rate (both false positives and negatives) through automatically up-weighting data sources that are most informative, while down-weighting less informative and biased sources. Extensive studies indicate that NBEL is significantly more robust than the classic naïve Bayes to unreliable, error-prone and contaminated data. On a large human data set our NBEL approach predicts many more PPIs than naïve Bayes. This suggests that previous studies may have large numbers of not only false positives but also false negatives. The validation on two human PPIs datasets having high quality supports our observations. Our experiments demonstrate that it is feasible to predict high-throughput PPIs computationally with substantially reduced false positives and false negatives. The ability of predicting large numbers of PPIs both reliably and automatically may inspire people to use computational approaches to correct data errors in general, and may speed up PPIs prediction with high quality. Such a reliable prediction may provide a solid platform to other studies such as protein functions prediction and roles of PPIs in disease susceptibility.

**Citation:** Xing C, Dunson DB (2011) Bayesian Inference for Genomic Data Integration Reduces Misclassification Rate in Predicting Protein-Protein Interactions. PLoS Comput Biol 7(7): e1002110. doi:10.1371/journal.pcbi.1002110

**Editor:** Christos A. Ouzounis, The Centre for Research and Technology, Hellas, Greece

**Received** January 3, 2011; **Accepted** May 17, 2011; **Published** July 28, 2011



**Funding:** This work was partly support by Award Number R01ES017436 from the National Institute of Environmental Health Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Environmental Health Sciences or the National Institutes of Health. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

* E-mail: chuanhua.xing@gmail.com

## Introduction

Protein interactions play important roles in most fundamental cellular processes. There has been increasing interest in reconstructing the interactome of a cell as large-scale data become available [1]. The improved knowledge of protein-protein interactions (PPIs) assists in detecting the susceptibility to human complex diseases [2–3] and then in discovery of new drugs and pharmaceuticals [4–8]. A variety of high-throughput experimental approaches have been developed to identify sets of interacting proteins, including yeast two-hybrid (Y2H) screening and mass spectrometry methods. However, these approaches are known to suffer from high false positives [9–12] and also high false negatives. A wide variety of computational approaches have been proposed for predicting PPIs. Some are based on data mining from published literature [13–17]. Please refer to [18] for a more complete review. The other studies are based on the amino acid sequences combined with additional information, such as co-expression patterns, phylogenetic distributions of orthologous groups, co-evolution patterns, the order of genes in the genome, gene fusion and fission events, and synthetic lethality of gene knockouts [19–27]. For reviews, refer to Bork et al. 2004, Shoemaker and Panchenko 2007, and Valencia and Pazos 2002 [28–30]. In this article, we focus on integrating the information from disparate data sources for the prediction of protein-protein interactions.

Genomic data integration has become popular in recent years with the intention of improving the power in predicting PPIs, as more disparate PPIs data are available. Several such methods have been recently developed, including decision trees [31–33], support vector machines (SVMs) [34], Bayesian models [22,35–38,39–44] and other considerations such as improving gold standard negative (GSN) set [45]. Among them, Bayes models have provided the most widely used paradigm for probabilistically integrating diverse data types. To calculate the score for each protein pair in each data source, the protein pairs are typically divided into subgroups based on features. One then calculates the likelihood ratio for the protein pairs in each feature subset by evaluating the ratio of the proportion of protein pairs in gold positive data set and the proportion in gold negative data set. Gold positive (negative) set is a dataset that includes protein pairs that are highly believed to be interacting (non-interacting). Naïve Bayes then multiplies directly the scores from multiple data sources for predicting whether a protein pair is interacting or not.

## Author Summary

Protein interactions are the basic units in almost all biological processes. It is thus vitally important to reconstruct protein-protein interactions (PPIs) before we can fully understand biological processes. However, critical difficulties exist. Particularly the rate of wrongly predicting PPIs to be true (false positive rate) is extremely high in PPIs prediction. The traditional approaches of error correction from each generating source can be both time-consuming and inefficient. We propose a method that can substantially reduce false positive rates by emphasizing information from more reliable data sources, and de-emphasizing less reliable sources. We indicate that it is indeed the case from our extensive studies. Our predictions also suggest that large numbers of not only false positives but also false negatives may exist in previous studies, as validated by two human PPIs datasets having high quality. The ability to predict large numbers of PPIs both reliably and automatically may inspire people to use computational approaches to correct data errors in general, and speed up PPIs prediction with high quality. Reliable prediction from our method may benefit other studies involving such as protein function prediction and roles of PPIs in disease susceptibility.


A protein pair is defined as interacting by observing a $>1$ posterior odds ratio, after multiplying prior odds to the likelihood ratio. Lee et al. 2004 and Lu et al. 2005 [36–37] integrated diverse functional genomics to reconstruct a functional gene network for Saccharomyces cerevisiae. Their results are comparable in accuracy to small-scale interaction assays with an increased true positive rate. Rhodes et al. 2005 [38] employed a naïve Bayes model to combine four data sources, ortholog interactions, co-expression, shared biological function, and enriched domain pairs. With a careful selection of prior information, their naïve Bayes model predicts nearly 40,000 protein-protein interactions in humans. They reported a false positive rate of 50%, though Hart et al. 2006 [39] later estimates this rate to be 85%. Also using a naïve Bayes approach, Scott and Barton, 2007 [40] predicted 37,606 human PPIs, with an estimated false positive rate as high as 76%.

As reviewed above, the predictions of PPIs still suffer a rather high false positive rate, which can be as high as $>80\%$. In addition, current PPIs prediction is far from complete with yeast $\sim 50\%$ and human only $\sim 10\%$ identified [1,39]. Hence, it is of critical importance to effectively reduce the false positive rate for a more reliable and complete prediction of large numbers of PPIs. Some of the errors may result from inaccurate data collection and error-prone data sources, though it is reasonable to assume that such data are in the minority and the majority properly reflects the evidence of interactions. To reduce the misclassification rate, it is necessary for the method to be robust to biased and non-informative data sources. The popular naïve Bayes model flexibly integrates the interaction information in a probabilistic way, compared with other data integration methods. However, the direct multiplication of likelihood ratio scores may not be able to effectively handle the effects of errors including missing interactions, sampling biases, and false positives [1]. Such errors can lead to completely wrong predictions even if they may come only from one single data source. It is therefore critically important to develop a novel algorithm that is able to effectively minimize the effects of the errors in data and therefore reduce the misclassification rate for obtaining a reliable prediction of PPIs.

We propose a nonparametric Bayes latent class discriminant analysis approach, which we refer to as nonparametric Bayes ensemble learning (NBEL) due to the ability to flexibly ensemble information about the presence of PPIs across different data sources. The goal of NBEL is to lower the false positive rate through automatically up-weighting the data sources that are most informative about a PPI, while down-weighting less informative and biased sources. None of existing integration methods, as far as we know, is able to flexibly weight the data sources for optimally capturing the information of PPIs. Bader et al. 2004 [46] weighted their positive and negative training examples inversely according to their fraction of the training set to favor 0.5 as the prior dividing threshold. InPrePPI [47] used a naïve Bayesian fashion to integrate multiple data sources by multiplying a weight, which is approximately estimated for each data source. However, the contributions of data sources can be different for the different protein pairs due to their different biological functions. NBEL learns the distributions of the likelihood ratios (LRs) for interacting and non-interacting protein pairs within each data source. If the distribution of the LRs for interacting and non-interacting pairs is not well separated for a particular data source, then that source will be down-weighted automatically in calculating the posterior probability of a PPI. In this manner, NBEL does not equally weight the different data sources, but instead learns the weights adaptively in a probabilistic manner. NBEL is thus more robust than classic naïve Bayes to unreliable, error-prone and contaminated data, and our extensive studies indicate this is indeed the case. On a large human data set our NBEL approach predicts many more PPIs than naïve Bayes, which suggests that large numbers of not only false positives but also false negatives may exist in previous studies. The validation on two experimental datasets having high quality supports our observation.

## Results

We conducted extensive simulation studies to evaluate and validate the performance of the proposed NBEL method. We compared the results with two methods, naïve Bayes and logistic regression. We then tested our approach on human genomic data sets. We finally validated the performance of NBEL via two experimental human PPIs data having high quality.

### Simulation Studies

The goal of our simulations is to assess the performance of our NBEL algorithm compared with two popular methods, naïve Bayes and logistic regression, in cases in which the interaction status is known. Current genomic integration approaches usually evaluate their prediction by comparing with the protein pairs in gold positive and negative datasets. However, only using gold positive and negative datasets may be misleading, as such data sets do not represent random samples of the entire set of human PPIs. In addition, the standards of selecting interacting protein pairs from each data source are rather ad hoc, and there is no known interacting information available for evaluation. One can verify the prediction using a small portion of experimental PPIs, but it is obviously not enough for evaluating large amount of computationally predicted PPIs. Hence, we also extensively tested on simulation data in which the truth is known.

We set up the simulations with 4 data sources. For the types of methods we are proposing, the performance of NBEL should improve as more data sources become available. We consider 5000 total protein pairs. We set the status of the first 1250 protein pairs as interacting, with the remaining 3750 non-interacting. We generated the simulated data using the model expressed by

Equation (2) in the Methods section (The parameters used are summarized in Table 1 in Text S1). We chose the distributions to allow a varying degree of separation in the interacting and non-interacting distributions for the different data sources. As discussed in detail in the Methods section, the more separated the distributions are, the more informative the data source is about a PPI. With a high degree of separation in which the two component distributions have minimal overlap, misclassification errors will be low for any reasonable method, so our focus is on the more realistic case in which there is substantial overlap.

**Tests on uncontaminated data.** We applied our NBEL approach and compared the performance with the naïve Bayes and logistic regression under different simulation scenarios, with the first case assuming uncontaminated data. Uncontaminated data refers to the data that are simulated error free. We calculated the estimated posterior probability for an interacting protein pair by averaging its conditional probabilities over MCMC iterations after burn-in. We then predicted that there is an interaction in protein pair $i$ if the estimated posterior probability is above a threshold. As noted in the Methods section, a 0–1 loss function results in an optimal threshold of 0.5, with this choice minimizing the Bayes risk defined as the posterior expectation of the overall misclassification rate obtained by weighting false positives and negatives equally.

The histogram of the estimated posterior probabilities for an example simulation is shown in Figure 1. There is a clear bimodal distribution with most of the interacting pairs having values close to one and most of the non-interacting pairs having values close to zero. The optimal 0.5 threshold separates them well. To compare with naïve Bayes, we directly multiplied the likelihood ratios (LRs) from the different data sources to obtain a final score for a protein pair. We then estimate a threshold that maximally separates the two modes in the histogram of all the final scores (we call this estimated threshold as *the alternative threshold* in short later on in this paper). To assess the impact of threshold choice on the performance and build a direct connection for comparing with naïve Bayes, we also evaluated the performance using the alternative threshold in applying our NBEL method. We chose 0.5 as the threshold for logistic regression, which typically produces very close results to using *the alternative threshold* based on our observations.

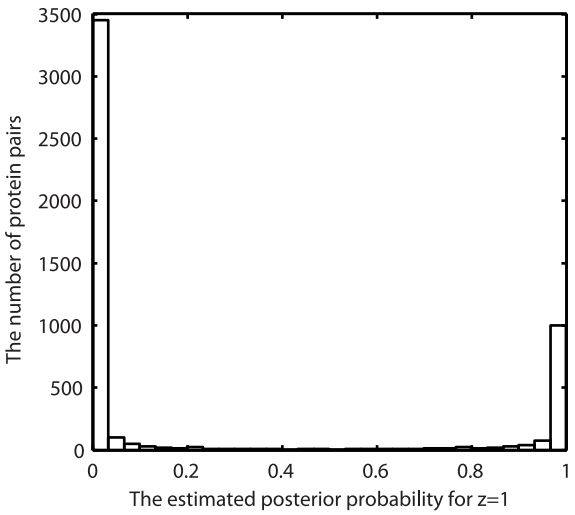


**Figure 1. The histogram of the estimated posterior probabilities of interacting protein pairs from NBEL algorithm.** This is from an example simulation using our NBEL. We can observe a clear bimodal pattern with almost all of the interacting pairs having posterior probabilities close to one and almost all of the non-interacting pairs having posterior probabilities close to zero.
doi:10.1371/journal.pcbi.1002110.g001

We analyzed 50 simulated data sets using three methods. Our NBEL method had lower misclassification rates (misclassification rate is defined as the average of the false positive (FP) rate and the false negative rate (FN)) than both naïve Bayes and logistic regression in all 50 simulated data sets, with the averaged misclassification rates 1.99% for NBEL using the threshold 0.5, 2.25% for NBEL using the alternative threshold, 7.57% using the alternative threshold for naïve Bayes, and 5.85% for logistic regression using the threshold 0.5. We can observe that the NBEL misclassification rates using the two thresholds are very close, and both are much better than naïve Bayes with an average difference of 5.58% and logistic regression with an average difference of 3.86%. Given the ideal case that the data sources are uncontaminated, such misclassification rate reduction from 7.57% or 5.85% to 1.99% can be a remarkable improvement especially when there is thousands and millions of PPIs data in the real data tests.

**Tests on contaminated data.** PPIs data in the real world however involves a large portion of false positives with possible varying false positive rates as we discussed in the Introduction section, although we expect that the situation can be better as the research goes on. We therefore simulated a series of data involving varying levels of contaminated data to examine the performance of our NBEL in reducing misclassification. We carried on the tests by repeating the same procedure as above but inducing errors to the data, in which a randomly-selected proportion of the protein pairs had their interaction status reversed. We created five sets of contaminated data with different levels of errors, and tested the performances of our NBEL algorithm on them. For the first data set, we randomly picked 25 out of 1250 interacting protein pairs for each data source, and reversed their status into non-interacting. We then randomly picked 75 out of 3750 non-interacting protein pairs for each data source, and, similarly, reversed their status into interacting. The induced error rate is slightly greater than 7% over all data sources, with the majority of errors occurring in fewer than two out of four data sources. The error rate is measured as the average of the induced false positive and false negative rates. We generated the remaining contaminated data sets by multiplying the number of protein pairs with scores appropriately reversed by 2, 4, 8, and 16 times of that for the first contaminated data set. Data were otherwise simulated and analyzed exactly as in *Uncontaminated Data*. The generated data are summarized in Table 1.

We applied three methods to each data set. This procedure was repeated on 50 independently generated data sets. The averaged misclassification rates, together with the ones for uncontaminated data, are summarized and plotted in Figure 2 (A). Similarly to uncontaminated data, our NBEL algorithm using either threshold has much lower misclassification rate than both naïve Bayes and logistic regression. Meanwhile, the misclassification rate reduction tends to be larger when the induced error rate in the data is higher, with a rather remarkable rate reduction of $>22\%$ from both naïve Bayes and logistic regression when the error rate in the data is as high as 46.95% in *Contaminated data IV*. The averaged standard deviation of FP and FN for 6 datasets varies from 0.0063 to 0.0158 for our NBEL, from 0.0078 to 0.0095 for logistic regression, and is high for naïve Bayes varying from 0.03 to 0.05. This suggests that our NBEL has a very strong function of error-correction, especially when the proportion of errors in the data is higher. This makes sense in that NBEL algorithm is designed to

**Table 1.** Summary of the contaminated data sets.

| | $L_1$ (out of 1250) | $L_2$ (out of 3750) | Induced Error Rate |
|---|---|---|---|
| Contaminated Data I | 96 | 288 | 7.68% |
| Contaminated Data II | 182 | 551 | 14.63% |
| Contaminated Data III | 342 | 1031 | 27.43% |
| Contaminated Data IV | 582 | 1775 | 46.95% |
| Contaminated Data V | 894 | 2670 | 71.36% |

In table 1, $L_1$ represents the number of interacting protein pairs that are reversed, and $L_2$ represents the number of non-interacting protein pairs that are reversed. We set the status for 1250 out of 5000 protein pairs as interacting, and 3750 out of 5000 protein pairs as non-interacting.
doi:10.1371/journal.pcbi.1002110.t001

flexibly integrate multiple data sources by up-weighting the more informative but down-weighting the less informative and biased sources in calculating the posterior probability of a PPI. Such a weight adjustment procedure is carried through by examining how well the learnt distributions of interacting and non-interacting protein pairs are separated within each data source. NBEL is therefore able to minimize the effects of the problematic data source that may be the results of missing data, sampling bias, false positive, or simple data entry errors, while maximize the information from the authentic interacting PPIs. In contrast, both naïve Bayes and logistic regression are barely functional in correcting data errors, with logistic regression slightly better and the misclassification rates for naïve Bayes close to the given error rates in all data sets. This may explain why the false positive rate is so high in the previous PPIs predictions. NBEL algorithm therefore provides a more powerful tool to integrate multiple data sources for a better prediction of PPIs. This property can be critically important, especially when current data for PPIs prediction include heavy data errors.

However, when the data error rates are extremely high as illustrated as the last points in Figure 2(A), we can observe that the misclassification rate is close to non-informative random rate 50%. These overlapped points correspond to *Contaminated data V*, having the data error rate 71.36%. This random non-informative prediction can be expected because the ability to accurately detect PPIs intuitively requires the majority of the data sources to be informative with error rate less than 50%. However, performance can be improved to some extent in the presence of large amounts of contamination by eliciting prior information as to whether a protein pair interacts or not from the literature. To assess this, we repeated the above tests with a fixed prior. This prior pre-assigns a probability weight to a protein pairs as to how possible it can be interact or not. However, such exhaustive prior information is difficult or impossible to obtain for all protein pairs. We therefore simply pre-assign a weight that represents the weight of interacting or the proportion of interacting protein pairs in the whole dataset. We randomly choose such a prior that is close to the known proportion of 1/4 in this study. We then predicted the interacting protein pairs from the posterior analysis of the MCMC procedure. The results are summarized in Figure 2 (B).

We can observe from Figure 2 (B) that the performance is similar to the one using the unknown prior in the presence of lower contamination when the error rate is less than 50%. The standard deviations for 6 datasets also have the similar pattern to the tests using unknown prior but slightly smaller values. However, when the contamination is rather high as in *Contaminated data V*, the elicited prior leads to much better performance in having a noticeable reduction of ~7% in misclassification rate. In addition to further supporting the previous conclusions, the elicited prior of being interacting or not may provide a realistic approach for genomic integration of PPIs data, especially when data includes rather high false positives and/or false negatives.

**Tests as the number of data sources increases.** As the number of data sources increases, NBEL will have more evidence to predict whether a protein pair interacts or not. We varied the number of data sources to observe the influence on the misclassification rate. We used the first contaminated data, and apply NBEL and naïve Bayes when the number of data sources $p$ is 4, 6, 8, 10, 12, 14, and 16, respectively. The simulation for every p is repeated 50 times. The averaged misclassification rates for the three approaches are plotted in Figure 3 (A). We can observe that the misclassification rates for our NBEL method using both the thresholds are much smaller than the ones for both naïve Bayes and logistic regression, with the misclassification rate for logistic regression obviously smaller than the one for naïve Bayes. As the number of data sources increases, the misclassification rates for NBEL and logistic regression reduce substantially. However, the misclassification rate for logistic regression is obviously higher than our NBEL, while naïve Bayes keeps a level of 8%~10%. To observe whether the misclassification rate can be reduced to such a low rate when the contamination in data is high, we repeated the above test but using contaminated data set III, and the comparison of misclassification rates among three methods are plotted in Figure 3 (B). We can see that naïve Bayes has a certain level of error correction when the misclassification rate is rather high, as the number of data sources increases. However, it stops decreasing when it reaches a level of 8%~10%, while NBEL and logistic regression decrease further. From Figure 3, logistic regression also produces a smaller misclassification rate as the number of data sources increases to be a large number such as 14 or greater. However, the number of reliable data sources to integrate in real data tests is usually not that large. While our NBEL is able to quickly reduce misclassification rate close to zero when the number of data sources increases to be 6 or 8. These tests supported our previous tests that our NBEL provides a more practical tool in predicting reliable PPIs from error-prone data, and learns from additional sources of informative data.

**Receiver operating characteristic.** In this part, we show the performance for all methods using a receiver operating characteristic (ROC) curve, which is the plot of the true positive (TP) rate versus false positive (FP) rate. We observed the ROC curves for all 6 sets of data in Table 1, with and without the known prior information of interaction. We illustrate our observations of ROC curves using the contaminated data III in Figure 4, which has error rate equal to 27.43% with the unknown prior information of interaction. From Figure 4, we can observe that our NBEL has a better performance than logistic regression, and logistic regression has a better performance than naïve Bayes. This is consistent with the observations in the previous tests. When the error rate in a dataset is lower, the curves are closer to the left top corner; when the error rate is higher, the curves are closer to the diagonal line which is TP rate equal to FP rate. The curves using the known prior information of interaction are very close to the ones using the unknown prior. However, when the error rate in a dataset is extremely high, for example in contaminated data V, the ROC curve using the known prior gives more reasonable results than the one using the unknown prior. This confirmed our observation indicated in Figure 2.

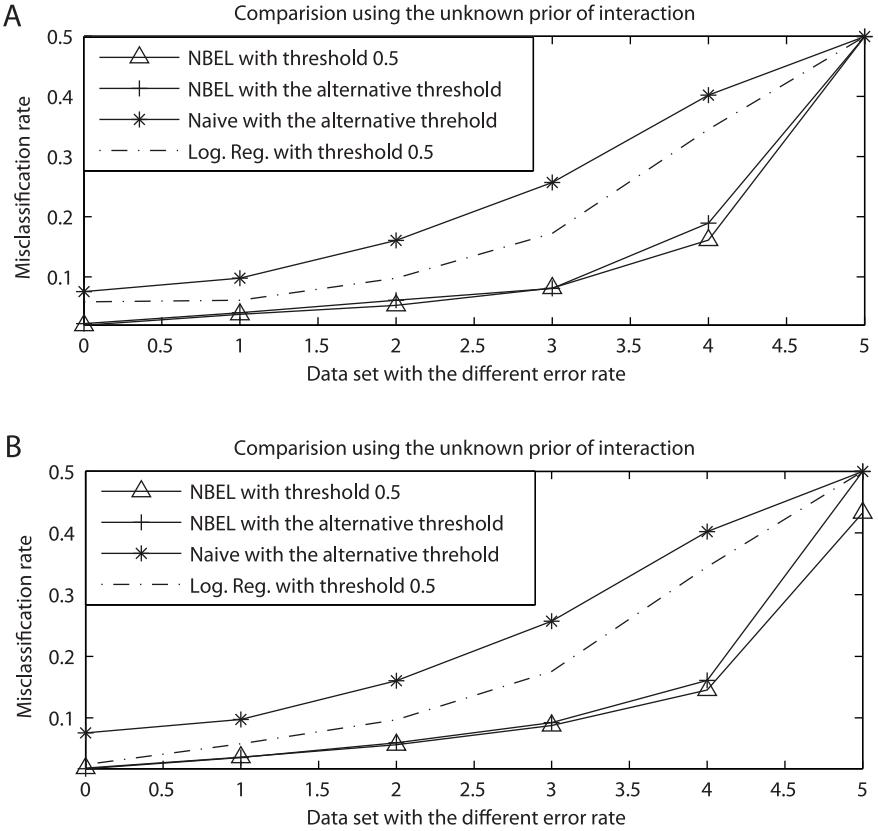


**Figure 2. Comparison with Naïve Bayes and logistic regression when the data sets have the different induced error rates.** The data set at x axis 0 represents noncontaminated data, and the data sets from x axis 1 to 5 represent contaminated data set I to V, with the induced error rates varying from 7.68% to 71.36%. y axis represents misclassification rate. A) is the comparison using the unknown prior of interaction. B) is the comparison using the known prior of interaction.
doi:10.1371/journal.pcbi.1002110.g002

## Tests on Human Data Sets

Rhodes et al. 2005 [38] collected human protein pairs from four data sources, ortholog from model organism interactome data (ortholog), genome-wide gene expression data (coexpression), protein domain data (domain), and biological functional annotation data (bio-function). Scott et al. 2007 collected more data sources in addition to the major ones in Rhodes et al. (2005) including coexpression, ortholog, domain, subcellular localication, post-translational modification co-occurrence, and protein intrinsic disorder [40]. We chose to test our approach on the data from Scott et al. (2007). The protein pairs collected from each data source are believed to be indicative of the possible interacting protein pairs, and they are measured by likelihood ratios (LRs). The protein pairs collected in each data source are firstly divided into the different feature states. The LR is then calculated for the protein pairs within that feature state by calculating the ratio of the proportion of protein pairs in the gold positive dataset to the proportion in the gold negative dataset. We chose to test on 79,441 protein pairs that have the product of LRs from all data sources greater than 100. Please review Rhodes et al. 2005 [38] and Scott et al. 2007 [40] for the principle of data collection.

We applied all the methods to integrate the scores of LRs from all the data sources of the collected human data for predicting PPIs. We tested the logistic regression model on LRs, as tested in Qi et al. (2006) [48]. We used the overlapped data with gold positive (GP) dataset and gold negative (GN) dataset to train the parameters for logistic regression model (Please review Text S1 for more information about GP and GN datasets). We predicted 39,334 PPIs using our NBEL algorithm, 16,234 PPIs using logistic regression, and 37,606 PPIs using naïve Bayes. The elucidated prior proportion of interaction for our NBEL is set as 0.5. The prior proportion of interaction was close to the empirical proportion by dividing the predictions from the naïve Bayes to the total number of protein pairs, 37,606/79,441 = 0.4734. Using a beta hyperprior can lead to an unrealistically high estimated proportion of PPIs. This is reasonable as current datasets for PPIs prediction are known to include many false positives, with rate varying from 50% to 85% [38–40]. As we analyzed in simulation studies, the extremely high proportion of errors in data may lead to non-informative prediction of a random probability of 0.5. An elucidated prior for the proportion of interactions however may alleviate the situation with a noticeable misclassification rate reduction.

Naïve Bayes requires a prior odds ratio, which is usually estimated by averaging the interactions per protein in the gold positive dataset. However, this value may be underestimated, since we do not know all the true interactions even in a small subset of proteins [38–40]. As discussed in Scott and Barton, 2007 [40], the prior odds ratio can change from 1/370 to 1/1093 across the different datasets. We picked prior odds ratio 1/400 for naïve Bayes as Scott and Barton 2007 and close to 1/381 in Rhodes et al. 2005 [38].

The number of PPIs predicted from NBEL, 39,334, however, is larger than 37,606 from naïve Bayes and 16,234 from logistic regression. We further analyzed the number of distinct proteins and the distinct interactions for the identified interacting protein pairs using three methods and their overlaps, as summarized in Figure 5. It appears that most of the unique proteins and protein

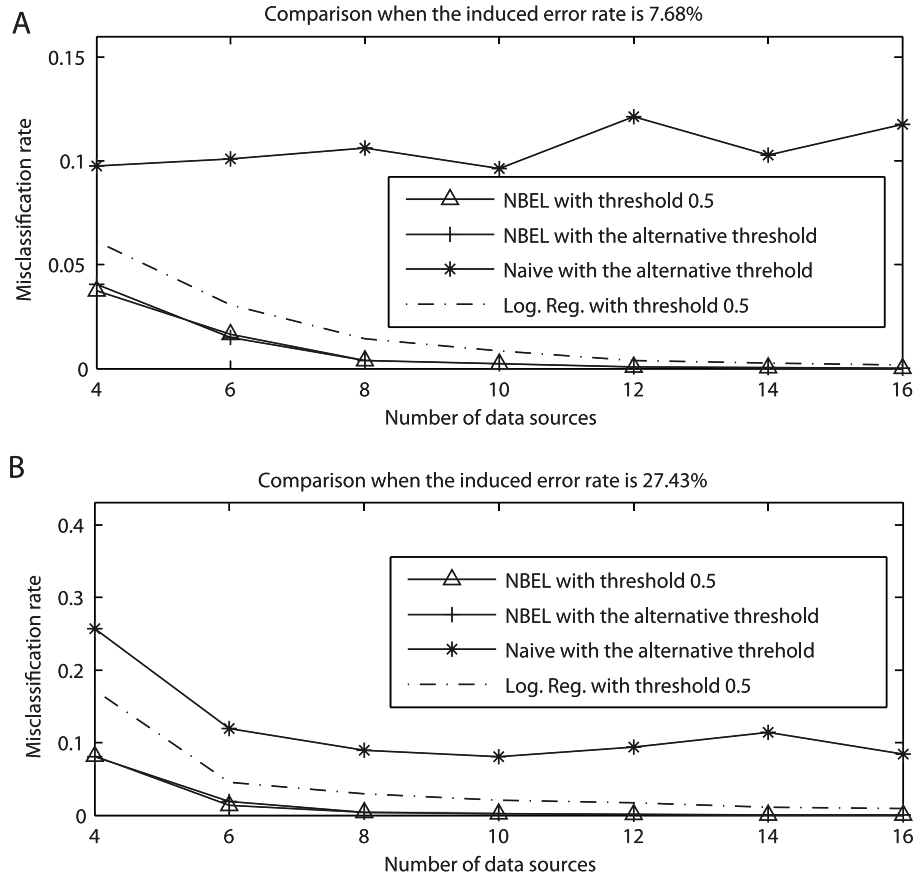


**Figure 3. Comparison with Naïve Bayes and logistic regression when the number of data sources increases.** X axis represents the number of data sources, and y axis represents misclassification rate. A) used contaminated data set I with the induced error rate 7.68%, and B) used contaminated data set III with the induced error rate 27.43%.
doi:10.1371/journal.pcbi.1002110.g003

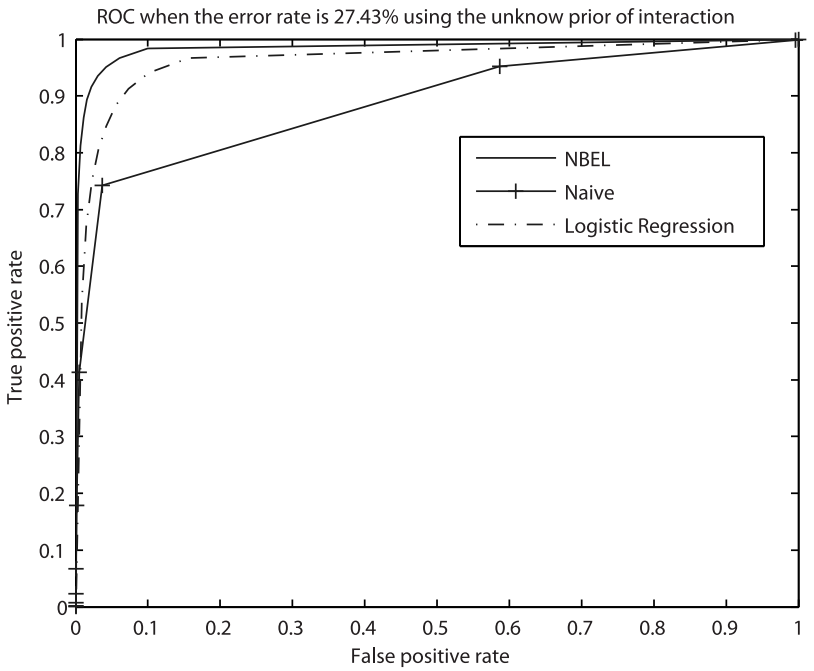


**Figure 4. Receiver Operating Characteristic (ROC) curves for our NBEL algorithm, naïve Bayes, and logistic regression.** We illustrate ROC curves using the data having error rate of 27.43% without the prior interaction information.
doi:10.1371/journal.pcbi.1002110.g004

pairs predicted by logistic regression are also predicted by NBEL and naïve Bayes. However, we observed a more reliable performance for NBEL and logistic regression than naïve Bayes from simulation studies, we suggest being skeptical of the protein pairs that are predicted by naïve Bayes but not as much those by NBEL and logistic regression. We can also observe that many more unique proteins and protein pairs are predicted by NBEL. This may be again the result of the function of error-correction from NBEL, as discussed in detail in the Methods section. We therefore expect a more reliable prediction using NBEL than naïve Bayes. A larger number of predicted PPIs may suggest that the previous estimations may not only have a large false positive (FP) rate [1], but also may have a large false negative (FN) rate. This also suggests the necessity of considering both the FP and FN rates for PPIs predictions.

We validated the above analysis by testing on another two human PPIs datasets with high quality. Mammalian protein-protein interaction database (MIPS) [49] manually curates high-quality experimental PPI data from the scientific literature, and includes only data from individually performed experiments that are believed to have the most reliable evidence from physical interactions. We downloaded 355 human PPIs with 423 proteins from MIPS. After eliminating the protein pairs with undesignated IDs and the ones with IDs mapping problems, we had 351 protein pairs and 420 proteins. We then compared 351 protein pairs with the human data set we collected, and found 46 protein pairs that are overlapping between two datasets. Among which, we had 26 interacting proteins falling into the set that are predicted by NBEL, but had 23 by naïve Bayes and only 11 by logistic regression. Further analysis indicates that the predictions from NBEL include all the ones from naïve Bayes and logistic regression. This observation coincides to what is observed applying NBEL to our collected human dataset. Our NBEL algorithm predicted an additional portion of protein pairs that are missed by naïve Bayes and logistic regression. This indicates that the predictions using naïve Bayes and logistic regression may have a rather large number of false negatives so that a large portion of interacting protein pairs are missed and predicted as non-interacting.

We tested on another large dataset, HomoMINT [50] having 38,414 PPIs. The data together with the ones from MIPS data are summarized in Table 2, and the test results are summarized in Table 3. In Table 3, $N_1$ indicates the total number of protein pairs in a database that are overlapped with our collected 79,441 human PPIs; $N_2$ indicates the number of protein pairs in $N_1$ that are predicted by NBEL; $N_3$ indicates the number of protein pairs in $N_1$ that are predicted by naïve Bayes; $N_4$ indicates the number of protein pairs in $N_1$ that are predicted by logistic regression. We measured true positive (TP) by calculating the proportion of protein pairs in $N_1$ that are predicted by either naïve Bayes or NBEL. Thus, $TP=\frac{N_2}{N_1}$ for NBEL, $TP=\frac{N_3}{N_1}$ for naïve Bayes, and $TP=\frac{N_4}{N_1}$ for logistic regression. False negative (FN) is simply $1-TP$.

We observe that the analysis on the second dataset has a similar pattern to that observed in the first experimental data from MIPS. The analyses from all datasets have a high true positive rate (a low false negative rate) from NBEL and a low true positive rate (a high false negative rate) from naïve Bayes and logistic regression. The overlapped predictions between three methods occupy most of the predictions from naïve Bayes and logistic regression but only a small portion from NBEL, which is consistent with the our previous analysis on our whole human data as shown in Figure 5. Again, the analysis using naïve Bayes and logistic regression missed a large portion of interacting protein pairs in having not only a large false positive rate [1] but also a large false negative rate.

## Discussion

The emergence of large-scale data has made it popular to study protein-protein interactions (PPIs) in recent years. However, one of the major issues is that a rather high proportion of false positives and negatives exist in current predictions. Data errors may occur from every data source and every stage of data collection and processing procedure. The usual approach to reduce the data errors is to minimize them from their generating source. However, such an approach can be extremely time-consuming and inefficient. Particularly, information may change as we improve our understanding in the underlying biological mechanism. A breakthrough to significantly reduce the misclassification rate is demanded for a reliable prediction of PPIs.

We proposed a nonparametric Bayes ensemble learning (NBEL) algorithm to integrate the multiple genomic data for obtaining a more powerful prediction of PPIs. Instead of the direct multiplication of scores from all data sources in naïve Bayes, our NBEL algorithm learns the distributions of interacting and non-interacting proteins within each data sources, and then automatically up-weights the informative and down-weights the less informative data sources. NBEL therefore has the function of error-correction which leads to a significant lower misclassification rate in predicting PPIs. We tested our NBEL algorithm on extensive simulations with various input data error rates varying from 0% to >70%, which mimic a rather high false positive rate >70% that is reported in previous PPIs predictions. Our simulation results indicated that our NBEL algorithm has a much lower misclassification rate, with the rate reduction varying from 7% to 25% from naïve Bayes and logistic regression. This suggests that NBEL is significantly more robust than naïve Bayes and logistic regression to highly contaminated data. Such a function becomes stronger as the number of data sources increases. Our tests on a large human data set indicate that NBEL predicts a larger number of PPIs than naïve Bayes and logistic regression, which are validated using two reliable experimental PPIs data. This indicates that rather high not only FP rate but also FN rate may exist in previous studies. This also suggests the importance of evaluating both the FP and FN rates in PPIs prediction.

We successfully demonstrated the feasibility of predicting high-throughput PPIs computationally, with substantially reduced false positives and false negatives. Our work may inspire people to utilize computational approaches to correct data errors for any problem in the field of computational biology that needs predictions from multiple data sources. The ability of predicting large numbers of PPIs both reliably and automatically may speed up PPIs prediction. Such a reliable prediction may provide a solid platform to other related studies. One example is the study of protein functions prediction since the group of protein pairs that tend to interact with each other may have similar functions. Another example is the study of roles of PPIs in disease susceptibility as the dynamic changes of PPIs may relate to disease causality.

There are still future works left for obtaining more complete and reliable inferences of PPIs. Current estimates of PPIs have a very low coverage [1]. The set of known interactions is even less representative of the whole network since the subset of interactions is by no means random. The analysis also showed that there is little overlap between the high-throughput datasets [1]. Paradoxically, some attempts to increase data quality, for example, multiple validations, make these biases more severe [1]. Although Lu et al. 2005 [37] indicated no appreciable dependence between any possible pairs of data sources for yeast. Information sharing does exist in the different levels among data sources for human. For example, it is believed that the interacting protein pairs sharing the same biological process may also

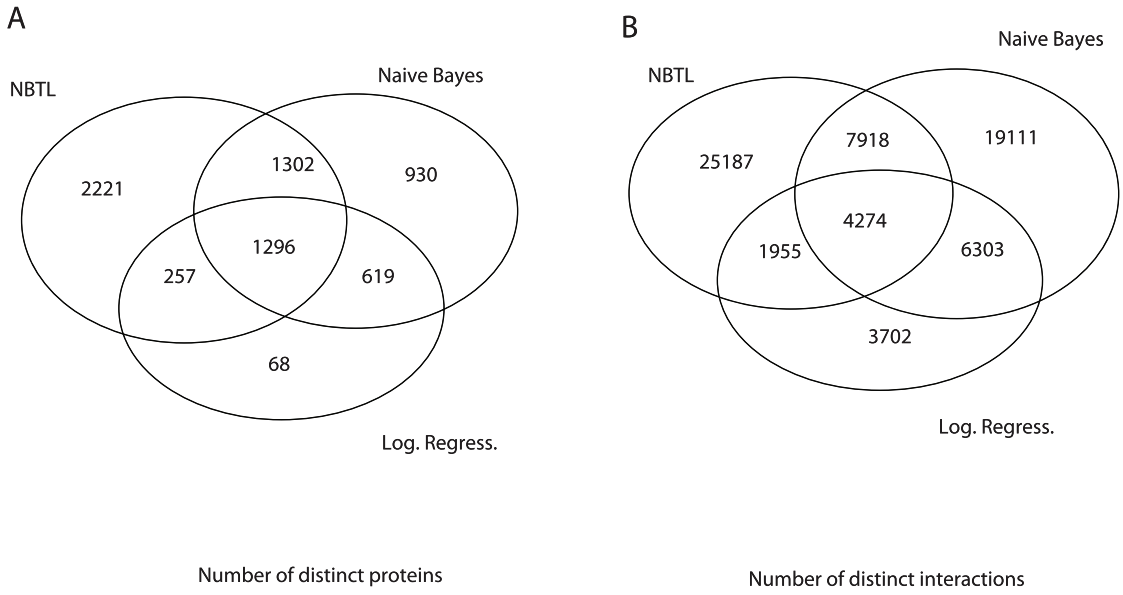


**Figure 5. Prediction comparison among our NBEL algorithm, naïve Bayes, and logistic regression.** A) listed he number of distinct and the overlapped proteins between two methods. B) listed the number of distinct and overlapped interactions among three methods.
doi:10.1371/journal.pcbi.1002110.g005

have physical associations between the enriched domains. Redundant information thus exists among the data sources of the biological functional annotation data and the protein domain data [38]. This invalidates our assumption of conditional independence in the different data sources given the unknown PPI status. Although some manual procedure as a semi-naïve Bayes is proposed in current works [22,35–38,40–41] to reduce such dependency, dependency exists more or less among any two of the disparate data sources. An effective integration method releasing the restriction of the conditional dependence is therefore demanded. Furthermore, since the network of PPIs is essentially time-evolving, an approach that is able to model the PPIs dynamically is desirable.

## Methods

In this section, we describe our NBEL method to integrate the likelihood ratios (scores) from the disparate data sources for the prediction of PPIs.

Let $\mathbf{Y}$ denote an $n \times p$ matrix, with rows corresponding to different protein pairs and columns to different types of scores from different data sources, with high values of the scores providing evidence of an interaction between the proteins. Typical analyses of protein interaction networks are based on one type of data, but here we propose a nonparametric Bayes latent class discriminant analysis approach for combining information from different data sources. We refer to this as ensemble learning following terminology in the machine learning literature. Let $y_{ij}$ denote the score in row $i$ and column $j$ of matrix $\mathbf{Y}$ and let $z_i = 1$ if the $i_{th}$ pair interacts with $z_i = 0$ otherwise.

Our nonparametric Bayes ensemble learning (NBEL) model assumes that

$$(y_{ij}|z_i = 0) \sim f_{0j};\ (y_{ij}|z_i = 1) \sim f_{1j}, \tag{1}$$

where $f_{0j}$ is the unknown distribution of the $j_{th}$ score across protein pairs that do not interact, and $f_{1j}$ is the unknown distribution of the $j_{th}$ score across protein pairs that do interact, for $j = 1,...,p$. For identifiability, we assume that $f_{0j} < f_{1j}$, denoting that $f_{0j}$ is stochastically less than $f_{1j}$. Following a Bayesian approach, we place priors on the unknown distributions $f = \{f_{0j}, f_{1j}, j = 1,...,p\}$.

In particular, we characterize each distribution using an infinite mixture model with

$$f_{z_i j}(y) = \sum_{h=1}^{\infty} \pi_h g(y; \Theta_{hz_i j}, \tau_h^{-1}), \tag{2}$$

where g(.) is a parametric kernel (e.g., Gaussian), $\pi_h$ is a mixture weight on component $h$, $\tau_h$ is a precision parameter specific to mixture component $h$, and $\Theta_{hz_i j}$ are location parameters specific to mixture

**Table 2.** Human protein-protein interactions datasets with high quality for validating our NBEL algorithm.

| Databases | Number of proteins | Number of protein pairs | Online Websites |
|---|---|---|---|
| From MIPS [49] | 420 | 351 | http://mips.helmholtz-muenchen.de/proj/ppi/ |
| From HomoMint [50] | 38,414 | 8,030 | http://mint.bio.uniroma2.it/HomoMINT/Welcome.do |
| Combined dataset | 38,834 | 8,381 | |

We overlapped protein pairs from each database in Table 2 with the whole collected human PPIs dataset that we tested on, and then compared the predictions out of the overlapped protein pairs for validating the performance of our NBEL algorithm.
doi:10.1371/journal.pcbi.1002110.t002

**Table 3.** Validation by comparing NBEL algorithm with naïve Bayes via two human datasets.

| | | Our NBEL | | | Naïve Bayes | | | Logistic Regression | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $N_1$ | $N_2$ | TP | FN | $N_3$ | TP | FN | $N_4$ | TP | FN |
| From MIPS [49] | 46 | 26 | 56.52% | 43.48% | 23 | 50.00% | 50.00% | 11 | 23.91% | 76.09% |
| From HomoMint [50] | 1688 | 1235 | 73.16% | 26.84% | 1005 | 59.54% | 40.46% | 484 | 28.67% | 71.33% |

In table 3, $N_1$ indicates the total number of protein pairs in a database that are overlapped with our collected 79,441 human PPIs; $N_2$ indicates the number of protein pairs in $N_1$ that are predicted by NBEL; $N_3$ indicates the number of protein pairs in $N_1$ that are predicted by naïve Bayes; $N_4$ indicates the number of protein pairs in $N_1$ that are predicted by logistic regression.

doi:10.1371/journal.pcbi.1002110.t003

component $h$, interaction status $z$, and score type $j$. It is well known that mixtures are extremely flexible. By allowing the kernel locations for each component to vary flexibly with interaction status and score type, we obtain a highly flexible model. The stochastic ordering restriction can be enforced by restricting $\Theta_{h0j} < \Theta_{h1j}$ for all $h, j$.

Dunson and Peddada 2008 [51] propose a restricted dependent Dirichlet process (rDDP) prior for modeling of unkown stochastically ordered distributions of the form shown in (2). However, they do not consider the case in which the stochastic ordering is over latent groups or cases in which data are available from different data sources. Conditionally on the data $\mathbf{Y}$ and the distributions $f$, the posterior probability of an interaction in pair $i$ is

$$\Pr(z_i = 1 \mid Y, f) = \frac{\psi_i \prod_{j=1}^{p} f_{1j}(y_{ij})}{\psi_i \prod_{j=1}^{p} f_{1j}(y_{ij}) + (1-\psi_i) \prod_{j=1}^{p} f_{0j}(y_{ij})} \qquad (3)$$

where $\psi_i$ is the prior probability of an interaction in pair $i$. This prior probability can be set to 0.5 to be uninformative, or one can incorporate available information outside of that included in the score $y_{i1},...,y_{ip}$ in the choice of $\psi_i$. Expression (3) describes that the information, such as the sharing and dependence among protein pairs, is borrowed via the normal mixture model and integrated for predicting protein-protein integrations.

The information can be transferred across the different protein pairs within columns (data sources). The distributions for interacting protein pairs and non-interacting protein pairs are learnt via the normal mixture model in expression (2). If only one data source were available ($p = 1$), there would be no ability to predict the interaction status latent variables $\{z_i\}$ and separately estimate the interacting and non-interacting score distributions without labeled data in which $z_i$ was known without error for a training subset. However, when repeated scores are available ($p > 1$), we obtain identifiability through the dependence structure in the multiple scores. In particular, the model will automatically interpret multiple scores that are high as evidence that the pair is more likely to be interacting. Essentially, the shared dependence on the latent class $z_i$ induces dependence in the multiple scores $y_{i1},...,y_{ip}$, allowing us to nonparametrically identify the different score densities under the stochastic ordering restriction. If a particular score (say score $j = 3$) tends to be unreliable, then it will have relatively low correlation with the other scores marginalizing out the latent $z_i$s, and hence the separation between $f_{0j}$ and $f_{1j}$ will be small. This small separation and low correlation will automatically lead to unreliable data sources being down-weighted and potentially even effectively excluded. This type of flexible adaptive weighting should substantially improve misclassification rates, and hence reduce false positives. This will be assessed through simulation studies in Section Results.

To complete a Bayesian specification of the model, we choose $g(y; \Theta, \tau) = N(y; \Theta, \tau^{-1})$, the univariate Gaussian distribution centered on $\Theta$ with precision $\tau$. In addition, following an rDDP specification (Dunson and Peddada 2008 [51]), we let

$$\begin{aligned}
&\psi_i = \psi \sim beta(a_\psi, b_\psi),\ i = 1,...n,\\
&\pi_h = V_j \prod_{l<h} (1 - V_l); \quad V_h \sim beta(1, \alpha),\\
&\Theta_{h0j} \sim N(\mu_j, \gamma_j^{-1}),\\
&\Delta_{hj} \sim N_+(0, \kappa_j^{-1}),\\
&\tau_{hj}^{-1} \sim Ga(\alpha_\tau, \beta_\tau),\ h = 1,...T,
\end{aligned}$$

where $\Delta_{hj} = \Theta_{h1j} - \Theta_{h0j}$, $N_+$ denotes a normal distribution truncazted below by zero, and $Ga(\alpha_\tau, \beta_\tau)$ denotes the gamma distribution. Letting $\psi_i = \psi$ for simplicity, $\psi$ represents the prior probability that a random selected protein pair is interacting. By choosing a beta hyperprior on $\psi$, we let the data inform about the proportion of interacting pairs. Normalizing the scores prior to analysis within each column of Y, we recommend the following default hyperparameter values, $a_\psi = b_\psi = 1$, $\alpha = 1$, $\mu_j = 0$, $\gamma_j = 1, \kappa_j = 1$, $\alpha_\tau = \beta_\tau = 1$.

We propose a blocked Gibbs sampler to estimate the posterior probabilities of unknowns (Ishwaran and James 2001 [52]) (Please find the details from Text S1). Our focus is on inference on the protein interactions based on the marginal posterior probabilities of $z_i = 1\,(i = 1,...n)$, which can be calculated using a Rao-Blackwellized approach. In particular, discarding a burn-in to allow convergence, we average the conditional posterior probabilities $\Pr(z_i = 1 \mid -)$ for each $i$ across a large number of MCMC iterations. Under 0–1 loss, the Bayes optimal classification rule sets $\hat{z}_i = 1(\hat{\psi}_i > 0.5)$ where $\hat{\psi}_i$ is the estimated posterior probability of $z_i = 1$. We recommend collecting 5,000 iterations, with the first 1,000 iterations discarded as a default.

## Supporting Information

**Text S1 The text file includes the parameters used to generate the simulated datasets, posterior computation, and the description of Gold Standard datasets.**
(PDF)

## Acknowledgments

We would like particularly to thank Scott MS and Barton GJ for kindly providing their human data sets [40] for us to test on.

## Author Contributions

Conceived and designed the experiments: CX DBD. Performed the experiments: CX. Analyzed the data: CX DBD. Contributed reagents/materials/analysis tools: CX. Wrote the paper: CX DBD.

## References


1. Hakes L, Pinney JW, Robertson DL, Lvell SC (2008) Protein-protein interaction networks and biology – what's the connection? Nat Biotechnol 26: 69–72.
2. Sham P (2001) Shifting paradigms in gene-mapping methodology for complex traits. Pharmacogenomics 2: 195–202.
3. Taylor IW, Linding R, Warde-Farley D, Liu Y, Pesquita C, et al. (2009) Dynamic modularity in protein interaction networks predicts breast cancer outcome. Nat Biotechnol 27: 199–204.
4. Butcher EC, Berg EL, Kunkel EJ (2004) Systems biology in drug discovery. Nat Biotechnol 22: 1253–1259.
5. Dolma S, Lessnick SL, Hahn WC, Stockwell BR (2003) Identification of genotype-selective antitumor agents using synthetic lethal chemical screening in engineered human tumor cells. Cancer Cell 3: 285–296.
6. Hood L, Perlmutter RM (2004) The impact of systems approaches on biological problems in drug discovery. Nat Biotechnol 22: 1215–1217.
7. Hopkins AL (2007) Network pharmacology. Nat Biotechnol 25: 1110–1111.
8. Yildirim MA, Goh KI, Cusick ME, Barabási AL, Vidal M (2007) Drug-target network. Nat Biotechnol 25: 1119–1126.
9. Mrowka R, Patzak A, Herzel H (2001) Is there a bias in proteome research? Genome Res 11: 1971–1973.
10. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, et al. (2002) Comparative assessment of large-scale data sets of proteinprotein interactions. Nature 417: 399–403.
11. Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, et al. (2008) High Quality Binary Protein Interaction Map of the Yeast Interactome Network. Science 322: 104–110.
12. Huang H, Bader JS (2009) Precision and recall estimates for two-hybrid screens. Bioinformatics, 25: 372–378.
13. Jensen LJ, Saric J, Bork P (2006) Literature mining for the biologist: from information retrieval to biological discovery. Nat Rev Genet 7: 119–129.
14. Huang M, Zhu X, Hao Y, Payan DG, Qu K, et al. (2004) Discovering patterns to extract protein-protein interactions from full texts. Bioinformatics 20: 3604–3612.
15. Saric J, Jensen LJ, Ouzounova R, Rojas I, Bork P (2006) Extraction of regulatory gene/protein networks from Medline. Bioinformatics 22: 645–650.
16. Kim S, Shin S-Y, Lee I-H, Kim S-J, Sriram R, et al. (2008) PIE: an online prediction system for protein-protein interactions from text. Nucleic Acids Res 36: W411–W415.
17. Malik R, Franke L, Siebes A (2006) Combination of text-mining algorithms increases the performance. Bioinformatics 22: 2151–2157.
18. Chowdhary R, Zhang J, Liu JS (2009) Bayesian inference of protein-protein interactions from biological literature. Bioinformatics 25: 1536–1542.
19. Aloy P, Böttcher B, Ceulemans H, Leutwein C, Mellwig C, et al. (2004) Structure-based assembly of protein complexes in yeast. Science 302: 2026–2029.
20. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA (1999) Protein interaction maps for complete genomes based on gene fusion events. Nature 402: 86–90.
21. Goh CS, Cohen FE (2002) Co-evolutionary analysis reveals insights into protein-protein interactions. J Mol Biol 324: 177–179.
22. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, et al. (2003) A Bayesian networks approach for predicting protein–protein interactions from genomic data. Science 302: 449–53.
23. Lu L, Arakaki AK, Lu H, Skolnick J (2003) Multimeric threading-based prediction of protein-protein interactions on a genomic scale: Application to the Saccharomyces cerevisiae proteome. Genome Res 13: 1146–1154.
24. Marcotte EM, Xenarios I, Eisenberg D (2001) Mining literature for protein-protein interactions. Bioinformatics 17: 357363.
25. Pazos F, Valencia A (2001) Similarity of phylogenetic trees as indicator of protein-protein interaction. Protein Eng 14: 609–614.
26. Ramani AK, Marcotte EM (2003) Exploiting the coevolution of interacting proteins to discover interaction specificity. J Mol Biol 327: 273–284.
27. Tsoka S, Ouzounis CA (2000) Prediction of protein interactions: Metabolic enzymes are frequently involved in gene fusion. Nature Genetics 26: 141–142.
28. Bork P, Jensen LJ, von Mering C, Ramani AK, Lee I, et al. (2004) Protein interaction networks from yeast to human. Curr Opin Struct Biol 14: 292–299.
29. Shoemaker BA, Panchenko AR (2007) Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. PLoS Comput Biol 3: e43.
30. Valencia A, Pazos F (2002) Computational methods for the prediction of protein interactions. Curr Opin Struct Biol 12: 368–373.
31. Lin N, Wu B, Jansen R, Gerstein M, Zhao H (2004) Information assessment on predicting protein protein interactions. BMC Bioinformatics 5: 154.
32. Zhang LV, Wong SL, King OD, Roth FP (2004) Predicting co-complexed protein pairs using genomic and proteomic data integration. BMC Bioinformatics 5: 38.
33. Qi Y, Klein-Seetharaman J, Bar-Joseph Z (2007) A mixture of feature experts approach for protein-protein interaction prediction. BMC Bioinformatics 8: S6.
34. Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, et al. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci U S A 97: 262–7.
35. Date SV, Stoeckert CJ, Jr. (2006) Computational modeling of the Plasmodium falciparum interactome reveals protein function on a genome-wide scale. Genome Res 16: 542–549.
36. Lee I, Date SV, Adai AT, Marcotte EM (2004) A probabilistic functional network of yeast genes. Science 306: 1555–1558.
37. Lu LJ, Xia Y, Paccanaro A, Yu H, Gerstein M (2005) Assessing the limits of genomic data integration for predicting protein networks. Genome Res 15: 945–953.
38. Rhodes DR, Tomlins SA, Varambally S, Mahavisno V, Barrette T, et al. (2005) Probabilistic model of the human protein-protein interaction network. Nat Biotechnol 23: 951–959.
39. Hart GT, Ramani AK, Marcotte EM (2006) How complete are current yeast and human protein-interaction networks? Genome Biol 7: 120.
40. Scott MS, Barton GJ (2007) Probabilistic prediction and ranking of human protein-protein interactions. BMC Bioinformatics 8: 239.
41. Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D (2003) A Bayesian framework for combining heterogeneous data sources for gene function prediction (in S. cerevisiae). Proc Natl Acad Sci USA 100: 8348–8353.
42. von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, et al. (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res 33: D433–D437.
43. Browne F, Wang H, Zheng H, Azuaje F (2010) A knowledge-driven probabilistic framework for the prediction of protein-protein interaction networks. Comput Biol Med 40: 306–317.
44. Elefsinioti A, Ackermann M, Beyer A (2009) Accounting for redundancy when integrating gene interaction databases. PLoS One 4: e7492.
45. Wu CC, Asgharzadeh S, Triche TJ, D'Argenio DZ (2010) Prediction of human functional genetic networks from heterogeneous data using RVM-based ensemble learning. Bioinformatics 26: 807–13.
46. Bader JS, Chaudhuri A, Rothberg JM, Chant J (2004) Gaining confidence in high-throughput protein interaction networks. Nat Biotechnol 22: 78–85.
47. Sun J, Sun Y, Ding G, Liu Q, Wang C, et al. (2007) InPrePPI: an integrated evaluation method based on genomic context for predicting protein-protein interactions in prokaryotic genomes. BMC Bioinformatics 8: 414.
48. Qi Y, Bar-Joseph Z, Klein-Seetharaman J (2006) Evaluation of Different Biological Data and Computational Classification Methods for Use in Protein Interaction Prediction. Proteins 63: 490–500.
49. Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I, et al. (2005) The MIPS mammalian protein-protein interaction database. Bioinformatics 21: 832–834.
50. Persico M, Ceol A, Gavrila C, Hoffmann R, Florio A, et al. (2005) HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms. BMC Bioinformatics 6 Suppl 4: S21.
51. Dunson DB, Peddada SD (2008) Bayesian nonparametric inference on stochastic ordering. Biometrika 95: 859–874.
52. Ishwaran H, James LF (2001) Gibbs sampling methods for stick-breaking priors. J Am Statist Assoc 96: 16173.