Published in final edited form as:

Biometrics. 2011 September ; 67(3): 958–966. doi:10.1111/j.1541-0420.2010.01519.x.

Network-based auto-probit modeling for protein function prediction

Xiaoyu Jiang¹, David Gold², and Eric D. Kolaczyk^{3,*}

¹Boehringer Ingelheim Pharmaceuticals, Inc. 900 Ridgebury Road, Ridgefield CT 06877 USA

²Department of Biostatistics, School of Public Health and Health Professions University at Buffalo, the State University of New York, Buffalo NY 14214 USA

³Department of Mathematics and Statistics, Boston University 111 Cummington Street, Boston MA 02215 USA

Summary

Predicting the functional roles of proteins based on various genome-wide data, such as protein-protein association networks, has become a canonical problem in computational biology. Approaching this task as a binary classification problem, we develop a network-based extension of the spatial auto-probit model. In particular, we develop a hierarchical Bayesian probit-based framework for modeling binary network-indexed processes, with a latent multivariate conditional autoregressive (CAR) Gaussian process. The latter allows for the easy incorporation of protein-protein association network topologies – either binary or weighted – in modeling protein functional similarity. We use this framework to predict protein functions, for functions defined as terms in the Gene Ontology (GO) database, a popular rigorous vocabulary for biological functionality. Furthermore, we show how a natural extension of this framework can be used to model and correct for the high percentage of false negative labels in training data derived from GO, a serious short-coming endemic to biological databases of this type. Our method performance is evaluated and compared with standard algorithms on weighted yeast protein-protein association networks, extracted from a recently developed integrative database called STRING. Results show that our basic method is competitive with these other methods, and that the extended method – incorporating the uncertainty in negative labels among the training data – can yield nontrivial improvements in predictive accuracy.

Keywords

Auto-probit; Bayesian hierarchical model; Gene Ontology annotation uncertainty; MCMC algorithm; Protein function prediction; STRING

1. Introduction

Inferring the functional role of proteins is a primary task in biology, for purposes ranging from general knowledge to drug discovery and diagnostic development. Protein functions are commonly taken as terms from the Gene Ontology (GO) database, a controlled vocabulary which describes gene and gene product attributes in organisms

* kolaczyk@math.bu.edu .

Supplementary Materials

Web Figures referenced in Section 4 is available under the Paper Information link at the Biometrics website <http://www.tibs.org/biometrics>.

(www.geneontology.org). Proteins can have multiple functional annotations in GO as one protein can perform multiple biological roles. Protein-term annotations in GO must follow a *true-path rule*, which states that if a protein is categorized into a more specific functional class, it must also be categorized into all the less specific ancestor functional classes.

Protein-protein interaction (PPI) networks are one of the most commonly used sources of information for predicting protein functions (Lehne and Schlitt (2009)). PPI networks are routinely modeled by functional linkage graphs, with vertices corresponding to proteins and edges indicating interactions between a pair of proteins. For a given protein function, the corresponding binary variables describing protein annotation statuses can be thought of as constituting a binary stochastic process indexed on the PPI network. Protein function prediction is then viewed as a task of predicting binary labels at locations in the network where they are unknown, given the labels observed at other nearby locations. Classifiers are usually built for this purpose and various methodologies of this sort have been proposed, frequently based on the principle of “guilt-by-association”, a widely adopted principle in systems biology, supported by empirical experience. In the present context, “guilt-by-association” suggests that a protein is likely to share the functions with the majority of its neighbors. The principle has been primarily used for prediction of biological processes in GO, and in special cases for prediction of molecular function and cellular localization.

Note that protein function prediction has been extensively researched; a comprehensive summary of various methods is provided by Sharan, Ulitsky and Shamir (2007). Examples of methodologies based on the “guilt-by-association” principle include Nearest-Neighbor(NN) algorithm, methods introduced in Schwikowski, Uetz and Fields (2000); Hishigaki et al. (2001); Lanckriet et al. (2004); Chua, Sung and Wong (2006), and many, particularly Markov Random Field-based probabilistic approaches. For example, Letovsky and Kasif (2003) proposed a binomial model of local neighbor functional annotation probability combined with a Markov Random Field (MRF) propagation algorithm to assign functions. Deng et al. (2003) developed a Bayesian approach that uses Markov Random Field model on the protein neighborhood to infer predictive probabilities. Most recent work combines PPI data with other data sources, such as DNA binding motifs, protein localization, and gene expression data. For instance, Deng, Chen and Sun (2004) modified their MRF model to combine protein domain information with PPI. In Nariai, Kolaczyk and Kasif (2007), Jiang et al. (2008) and Jiang et al. (2008), the binomial model in Letovsky and Kasif (2003) was generalized to integrate the Gene Ontology hierarchical structure, gene microarray, protein motif and cell localization data.

In this paper, we pursue two avenues towards improvement in protein function prediction that differ notably from what has been considered in the literature to date. First, we wish to incorporate weighted – rather than binary – protein-protein association networks in a seamless fashion into a probabilistic framework. An example of a weighted PPI network, and one which we shall use later in the data analysis described in this paper, is that derived from the STRING database in Mering et al. (2005), which contains a combination of known and predicted protein-protein associations and corresponding scores. The scores express increased confidence when an association is supported by several types of evidence, which can be highly informative in inferring proteins' functional characteristics. Taking advantage of the scores in databases such as STRING has become a new challenge as well as an opportunity to improve function prediction accuracy. However, it is not obvious how this challenge can be met by simple adaptations of the “guilt-by-association” principle governing methods like those mentioned earlier.

Second, we wish to model and account for uncertainty in annotations in the Gene Ontology database, which is a critically important issue in functional genomics and proteomics. Most

methods using annotations in the GO database for training classifiers assume that a protein being or not being annotated with a function accurately reflects whether or not that protein truly has that function or not. However, while positive annotations – which traditionally have reflected experimentally confirmed protein functions – are generally reliable, negative annotations can be much less so. The reason for this disparity in reliability comes from the fact that negative annotation for a given GO term can reflect either a known lack of positive annotation (perhaps logically implied by certain positive annotations on other GO terms) or simply an absence of knowledge as to the protein status with respect to this term. This observation suggests that, instead of treating the task of protein function prediction as a binary classification problem, we actually have three classes - “having the function”, “not having the function” and “status unknown”. This observation does not appear to have been widely acknowledged in the literature. However, results in this paper show that acknowledging and, moreover, accounting for it appropriately can yield nontrivial improvements in predictive accuracy.

2. Bayesian hierarchical model for protein function prediction

2.1 Network-based auto-probit model

Suppose we have a collection of N proteins, $n \leq N$ of which have functional annotations in the Gene Ontology database, and $N - n \geq 0$ for which we wish to predict functionality. Let the binary variable y_{iG} denote the functional annotation of protein i for the GO term G . That is, $y_{iG} = 1$ if protein i is annotated with term G in the database; $y_{iG} = 0$ otherwise, for $i = 1, \dots, n$. Since protein functions are predicted separately, for each choice of G , we simplify y_{iG} to y_i in the following sections, to avoid making equations too busy.

Motivated by Weir and Pettitt (2000), who developed a spatial auto-probit model for lattice data, we find it useful in our own network-based context here to employ a latent Gaussian process in modeling the binary variables y_i . Specifically, we define each y_i through a continuous latent variable z_i , according to the sign of z_i , i.e.,

$$y_i = \begin{cases} 1, & \text{if } z_i \geq 0 \\ 0, & \text{if } z_i < 0 \end{cases} \quad (1)$$

for $i = 1, \dots, N$. The latent vector z is to be understood here simply as a continuous-valued representation of protein functional status, which will be convenient for purposes of model building and interpretation, as well as computational tractability. The values of the z_i are assumed unobserved. We assume that we observe only the functional status indicator y_i of n proteins $i = 1, \dots, n$ and take as our goal to predict the values y_i , $i = n + 1, \dots, N$ of the remaining $N - n$ proteins. For now we assume that the presence and absence of GO annotations are correct, i.e., Gaussian process z fully determines the value of y .

It is commonly assumed that proteins with the same or similar functions tend to interact more frequently than others. This assumption underlies, for example, the *local density enrichment assumption* in Letovsky and Kasif (2003). Therefore, we want to encode protein-protein association network structure into the model to aid in inferring the functional labels. Let A be the $N \times N$ adjacency matrix for a protein network. For a binary (i.e., unweighted) network, $a_{ij} = 1$ for interacting neighbors i and j , and 0 otherwise. For a weighted network, a_{ij} takes on the value of the weight for the edge $\{i, j\}$, with a weight of $a_{ij} = 0$ indicating no edge. For each i , let $d_i = \sum_{j \neq i} a_{ij}$ be the degree of protein i . If the network is binary, d_i simply counts the number of neighbors of protein i . Denote by $D = \text{diag}\{d_i\}$ the diagonal degree matrix.

The $N \times 1$ latent Gaussian process vector z is assumed to follow a multivariate normal distribution

$$z|\mu, \beta \sim \text{MVN}\left(\mu, \left(I - \beta D^{\frac{1}{2}} A D^{\frac{1}{2}}\right)^{-1}\right) \sim \text{MVN}\left(\mu, (I - \beta B)^{-1}\right), \quad (2)$$

where I is the identity matrix, μ is the location vector for proteins, and $B = D^{\frac{1}{2}} A D^{\frac{1}{2}}$. We constrain the value of β to ensure that the precision matrix $I - \beta B$ is positive definite. We restrict β to be non-negative based on *local density enrichment assumption*. Writing the

determinant of the precision matrix as $|I - \beta D^{\frac{1}{2}} A D^{\frac{1}{2}}| = \prod_{i=1}^N (1 - \beta \lambda_i)$, where λ_i is the i -th

eigenvalue of matrix $D^{\frac{1}{2}} A D^{\frac{1}{2}}$, we note that a sufficient condition for this determinant to be positive is that $\beta < \lambda_{\max}^{-1}$, where $\lambda_{\max} = \max_i \lambda_i$. Hence, we constrain β such that $0 < \beta < \lambda_{\max}^{-1}$, which is similar to Weir and Pettitt (2000) in the context of spatial auto-probit modeling. The parameter β measures the spatial dependence in probability between neighbors, of carrying the function G , over the global network topology. When $\beta = 0$, a protein's functional status is independent of its neighbors'; larger positive values of β are indicative of functional similarity between immediate graph neighbors.

To understand the manner in which this model incorporates the network topology, we note that the partial correlation coefficient ρ_{ij} between two neighbor proteins z_i and z_j takes the form $\rho_{ij} = \beta \sqrt{d_i d_j a_{ij}}$; 0 if they are not neighbors. This expression indicates increased influences of highly interactive proteins, in other words, proteins with large neighborhood is more likely to be consistent with the majority of its neighbors.

2.2 Prior and posterior distributions for the network auto-probit model

The joint posterior probability distribution is as follows

$$P(z, \mu, \beta | y) \propto P(y | z) \cdot P(z | \mu, \beta) \cdot P(\mu) \cdot P(\beta), \quad (3)$$

where $P(y | z) = \prod_{i=1}^n P(y_i | z_i) = 1$, since we assume for now that the GO annotation is correct and the sign of z_i fully determines the value of y_i . To fully specify this posterior, we need to equip the parameters μ and β with appropriate prior distributions.

We assign a conditional autoregressive (CAR) prior distribution with a hyperparameter τ^2 to μ which models the spatial dependence on the network and smooths individual μ_i locally. More specifically, we define a (singular) joint prior distribution on μ of the form

$$P(\mu | \tau^2) \propto \exp\left\{-\frac{1}{2\tau^2} \mu^T L \mu\right\} \propto \exp\left\{-\frac{1}{2} \sum_{ij} \frac{(\mu_i - \mu_j)^2}{\tau^2}\right\}. \quad (4)$$

where $L = D - A$ is the graph Laplacian matrix, which leads to the prior conditional mean of μ_i being a weighted average on protein i 's neighbors, i.e.,

$$\mu_i | \mu_1, \dots, \mu_{i-1}, \mu_{i+1}, \dots, \mu_n; \tau^2 N \left(\frac{\sum_j a_{ij} \mu_j}{d_i}, \frac{\tau^2}{d_i} \right). \quad (5)$$

To better understand the effect of the τ^2 , we derive the posterior distribution of μ , given the Gaussian process z and β

$$P(\mu | z, \beta) \propto P(z | \mu, \beta) \cdot P(\mu) \text{MVN} \left(z^T (I - \beta B) \left(I - \beta B + \frac{L}{\tau^2} \right)^{-1}, \left(I - \beta B + \frac{L}{\tau^2} \right)^{-1} \right). \quad (6)$$

The parameter τ^2 controls the extent to which the prior influences the posterior distribution and can be interpreted as the variance of the difference in expected latent characteristics for two neighbor proteins. Therefore, the choice of τ^2 is important in identifying the location vector μ . We discuss this issue further below.

Due to the constraint on β for obtaining a valid precision matrix, we apply a truncated normal prior to β , $\beta \sim TN(\beta_0, \sigma_\beta; 0, \beta_{\max})$, where β_0 and σ_β are the prior mean and standard deviation, respectively; $\beta_{\max}$ is the maximal possible value for β , as dictated by the largest eigenvalue of B in our implementation. We simply set β_0 as the midpoint of 0 and $\beta_{\max}$; σ_β is chosen to be small so that the truncated normal posterior distribution for β could be comfortably fit on the tight constraint.

2.3 Complexity of Smoothing Variances

Proper tuning of the prior distributions in this Bayesian auto-probit model is critical to obtaining useful posterior distributions of the location parameter μ and the neighborhood effect parameter β . Identification of the location vector μ in this over-parameterized model would be difficult without strong spatial smoothing. We develop an effective degree of freedom in analogy to the degrees of freedom for the smoother matrix in the smoothing splines (e.g., Hastie, Tibshirani and Friedman (2001)),

$$\rho(\tau^2) = \text{trace} \left[\left(V^T V + \frac{L}{\tau^2} \right)^{-1} V^T V \right] = \text{trace} \left[\left(I - \beta B + \frac{L}{\tau^2} \right)^{-1} (I - \beta B) \right], \quad (7)$$

where $L = D - A$, as defined earlier. The degree of freedom for μ is monotonely increasing and is confined between 0 and N . In applications, such as those shown later in this paper, we have found it useful to specify τ^2 so as to impose a fairly low number of degrees of freedom, forcing μ to be fairly smooth. Please see the supplementary materials for the derivation.

2.4 Markov chain Monte Carlo algorithm

Markov Chain Monte Carlo algorithm is used to draw samples from the joint posterior distribution. We update the individual z_i , μ_i , and the parameter β one at a time, which requires access to the fully conditional distributions.

The Gibbs sampler is used to update z_i , based on the conditional probability

$$P(z_i | z_{[-i]}, \mu, \beta, y) = \begin{cases} \frac{1}{\sqrt{2\pi}\Phi(\bar{z}_i)} \exp\left\{-\frac{1}{2}\left(z_i - \bar{z}_i\right)^2\right\}, & z_i \geq 0, \quad y_i = 1, \\ 0, & z_i < 0, \quad y_i = 1, \\ \frac{1}{\sqrt{2\pi}(1-\Phi(\bar{z}_i))} \exp\left\{-\frac{1}{2}\left(z_i - \bar{z}_i\right)^2\right\}, & z_i < 0, \quad y_i = 0, \\ 0, & z_i \geq 0, \quad y_i = 0, \end{cases} \quad (8)$$

where $z_{[-i]}$ is all of z except the i th element z_i , $\bar{z}_i = \mu_i + \beta \sum_{j \neq i} \sqrt{d_i d_j a_{ij}} (z_j - \mu_j)$, and Φ is the standard normal cumulative density function.

The Metropolis-Hasting algorithm is used to update μ_i and β . The proposal change from μ_i to a new μ'_i (or from β to a new β') is drawn from a normal distribution centered at the current value with a pre-defined standard deviation.

3. Incorporating uncertainty of functional annotation in Gene Ontology

In previous sections, we assume that the annotations from the Gene Ontology database reflect the actual protein functional status, that is, observing an annotation in the database ($y_i = 1$) means that the protein carries the function ($z_i \geq 0$), while not observing such an annotation ($y_i = 0$) means that the protein does not have that function ($z_i < 0$). Again, the Gaussian process z in our work is simply a modeling device, serving as a continuous-valued representation of protein functional status. However, in practice, just because a protein is not annotated with a particular function in a database (e.g., such as Gene Ontology) does not necessarily mean that it indeed is lacking that function. Instead, it is frequently the case that the functional status is currently unknown to science but listed as “un-annotated” in the absence of more accurate knowledge (Letovsky and Kasif (2003)). Therefore, since database annotations of functional status cannot be completely trusted, our hypothesis is that accounting for annotation uncertainty may improve predictive performance.

While we do not account for uncertainty in positive annotations, we acknowledge that positive annotations might be incorrect as well, a point we take up in Discussion. However, it is less likely where annotations are derived from experimentally confirmed evidence from literature. Since there is an arguably greater interest among scientists in discovering previously unknown functions of proteins, rather than confirming known function, we build up our framework under the assumption stated in the previous paragraph.

Therefore, we modify our model, writing

$$P(y | z, g) = \prod_{i=1}^n P(y_i | z_i, g), \quad \text{where } P(y_i = 0 | z_i, g) = \begin{cases} 1, & z_i < 0, \\ g, & z_i \geq 0. \end{cases} \quad (9)$$

This extended version of our network auto-probit model thus incorporates the probability of being “incorrectly un-annotated”. The device employed here is analogous to that used in Weir and Pettitt (2000), in modeling the spatial distribution of toads, to account for the fact that spatial regions for which no toads were observed are not necessarily devoid of toads.

The joint posterior distribution including g is given by

$$P(z, \mu, \beta, g | y) \propto P(y | z, g) \cdot P(z | \mu, \beta) \cdot P(\mu) \cdot P(\beta) \cdot P(g), \quad (10)$$

where we let $P(g)$ be the uniform prior distribution for g , $P(g) \sim U(0, 1)$, $P(y/z, g)$ is as described above, and all other terms are the same as mentioned earlier.

3.1 Markov Chain Monte Carlo algorithm with GO annotation uncertainty

When we consider the Gene Ontology annotation uncertainty and include the probability of being “incorrectly un-annotated”, it can be shown that the fully conditional distribution for updating individual z_i is different from before, being expressed as

$$P(z_i | z_{[-i]}, \mu, \beta, g, y) = \begin{cases} \frac{1}{\sqrt{2\pi}\Phi(\bar{z}_i)} \exp\left\{-\frac{1}{2}\left(z_i - \bar{z}_i\right)^2\right\}, & z_i \geq 0, \quad y_i = 1, \\ 0, & z_i < 0, \quad y_i = 1, \\ \frac{g}{\sqrt{2\pi}[1 - \Phi(\bar{z}_i) + g\Phi(\bar{z}_i)]} \exp\left\{-\frac{1}{2}\left(z_i - \bar{z}_i\right)^2\right\}, & z_i \geq 0, \quad y_i = 0, \\ \frac{1}{\sqrt{2\pi}[1 - \Phi(\bar{z}_i) + g\Phi(\bar{z}_i)]} \exp\left\{-\frac{1}{2}\left(z_i - \bar{z}_i\right)^2\right\}, & z_i < 0, \quad y_i = 0, \end{cases} \quad (11)$$

where $\bar{z}_i = \mu_i + \beta \sum_{j \in N_i} \sqrt{d_i d_j a_{ij}} (z_j - \mu_j)$, and Φ is the standard normal cumulative density function. The derivation of the conditional probability is included in the supplementary material. Gibbs sampler can be used to update g . It can be derived that the fully conditional distribution for g follows a beta distribution,

$$P(g | z, \mu, \beta, y) \propto P(y | z, g) \cdot P(g) \propto g^{N_{-+}} (1 - g)^{N_{++}}, \quad (12)$$

where $N_{-+} = \#\{i : y_i = 0, z_i \geq 0\}$, $N_{++} = \#\{i : y_i = 1, z_i \geq 0\}$.

In our algorithm, only the local neighborhood information is needed for updating for one protein at each iteration. Therefore, users do not need to store the entire network matrix in memory. For computational efficiency, one might want to employ modified methods that update at longer numbers of steps (e.g., every one pass through the network). In this case, the computations could be parallelized and substantial efficiency could be gained on a computer cluster. However, this is beyond our scope and is for future research to explore.

4. Results

4.1 Data

We have implemented our network-based auto-probit model on two yeast protein-protein association (sub)networks extracted from the STRING database, introduced in Mering et al. (2005). STRING contains known and predicted protein-protein associations, where “association” refers to both direct physical binding and indirect interaction such as participation in the same metabolic pathway or cellular process. Information for associations is obtained from seven evidence sources: database imports (PPI and pathway databases), high-throughput experiments, co-expression, homology based on phylogenetic co-occurrence, homology based on gene fusion events, homology based on conserved genomic neighborhood, and text mining.

STRING simplifies the access to protein-protein associations by providing a comprehensive collection of such associations for a large number of organisms. A score S is assigned to each interacting pair of proteins by bench-marking against the KEGG pathway (www.genome.jp/kegg/pathway.html). The score is calculated by $1 - S = \prod_i (1 - S_i)$, where i indicates the individual evidence type described at the end of the previous paragraph, and S_i

is the score from the i th source. As a result, STRING database provides users with weighted undirected protein-protein association networks.

For purposes of illustration, we extract two sub-networks of size $N = 211$ and 975 , respectively, from the overall protein-protein association network in yeast, a well-annotated model organism. In each case we choose genes annotated with a general term to define the sub-network, and then use a more specific descendent term as our annotation of interest (i.e., to define the label y_i). The smaller network is defined by those 211 genes annotated with the term *GO:0007154, cell communication (CC)* in the Gene Ontology database, as of November 2007, and will be referred to as the “CC” network in what follows. The target term of interest to us in this network is taken to be *GO:0007242, intracellular signaling cascade (ISC)*, a grand-child term of cell communication in the GO hierarchy. Of the 211 genes in this network, 120 are annotated with ISC (i.e., $y_i = 1$). The larger network is defined by those 975 genes annotated with the term *GO:0006996, organelle organization of biogenesis (OOB)*, and referred to as the “OOB” network in what follows. The target term of interest is taken to be *GO:0051276, chromosome organization and biogenesis (COB)*, one of the child terms of OOB. A total of 393 genes are annotated with COB. In the remainder of this section, the CC and OOB networks are used to explore the performance and behavior of our models with respect to model fitting and prediction. As an aside, we note that according to the GO annotation in November 2007, there are 226 and 1290 proteins annotated with cell communication and organelle organization and biogenesis, respectively. We only use the largest connected components, neglecting small connected neighborhoods and isolated proteins, therefore, the network sizes in this paper are smaller than the actual numbers of proteins.

4.2 Parameter estimation by the auto-probit model

Using the model in Section 2.2 and 2.3, we first take the observed annotation y as known for all N proteins, and examine the issue of fitting the model. The hyperparameters for the two networks are listed in Table 1. We discard the first 1000 iterations as burn-in and run 9000 iterations to get posterior samples. Convergence diagnostics by standard approaches indicate that all chains reach equilibrium. For trace plots and histograms of the posterior samples of β , please refer to Web Figures 1 and 2.

Estimates of the posterior means of some parameters are given in Tables 2 and 3, together with 95% credibility intervals. It can be seen that there is statistically significant positive spatial dependence on both networks. The small estimates for β are a result of the large eigenvalues $\lambda_{\max}$ (312.9047 and 4297.8 for CC and OOB networks, respectively), and hence a small $\beta_{\max}$ (0.0032 and 0.0002 for CC and OOB networks, respectively).

Additional model fitting results, including histograms of the posterior estimates of μ and the posterior predictive probabilities can be found in the supplementary material. These histograms are colored according to the presence and absence of the annotations. Note that the histograms for the two classes of proteins are well separated. Arguably this separation is primarily driven by the parameterization of our model, through the value of y_i at each node (i.e., recall equation (14)), however, the extent of this separation can be expected to be influenced at a secondary level by the values of y_i at neighboring nodes (an expectation easily confirmed in practice through simple toy networks, and by our study of stability reported below). That is, the posterior estimates of μ_i are affected by both the y_i observed and the network topology.

4.3 Prediction performance comparison by 10-fold cross-validation studies

As a sanity check, we compare the predictive performance of our network-based auto-probit model to two natural competitors. The first is the nearest-neighbor (NN) algorithm, which outputs the fraction of a protein's neighbors in the network possessing the annotation term in question. The second is the kernel logistic method in Zhu and Hastie (2005), which builds a Laplacian kernel matrix $L = D - A$ from the (weighted) network and produces predictive probabilities. Note that since NN does not use edge weights, we use the corresponding induced binary networks as input for that method. Accordingly, for the sake of comparison, we implement our network-based auto-probit model for both the weighted and binary versions of our networks.

We use the principle of 10-fold cross-validation to generate predictions from each of the four methods (i.e., NN on the binary network, kernel logistic method on the weighted network, and the auto-probit model on both binary and weighted networks). Specifically, at each randomly selected fold, we set the labels for 10% of the proteins as “unknown”, and use the labels from the remaining 90% of the proteins, as well as the network, to generate predictions of the unknown proteins. We note that, formally, the cross-validation implemented here is not the standard cross-validation used in typical supervised learning tasks. Rather, it represents a variant of semi-supervised learning, in that while we do not use the annotation status of proteins in each fold when predicting that status, we do make use of those proteins when incorporating the full network of protein-protein interactions into our model. See Liang, Mukherjee, and West (2007), for example, for a recent discussion on the implications on prediction of using such auxiliary information. Output from each method is compared to a predictive threshold, and true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) are calculated accordingly. The results are shown in Figure 1, summarized in the form of receiver-operating characteristic (ROC) curves.

On the CC network, the area under the curve (AUC) values for the auto-probit method with weighted STRING network, kernel logistic method, NN algorithm and auto-probit on binary STRING network are 0.8321, 0.8371, 0.8022, and 0.7817, respectively; on the OOB network these numbers are 0.8839, 0.8437, 0.8502, and 0.6397. A two-sided Wilcoxon signed rank test based on the ten folds (recommended by Demsar (2006) in this scenario) allows us to compare these AUCs formally. In comparing the auto-probit on weighted STRING to the same method on binary STRING, the p-values are 0.0640 and 0.0035, respectively, on the CC and OOB networks, indicating a reasonably strong degree of improvement from incorporating weights into this method. On the other hand, in comparing the auto-probit on weighted STRING to NN and kernel logistic methods, the p-values are 0.6232 and 0.7337, respectively, on the CC network; and 0.2640 and 0.4548, respectively, on the OOB network. Hence we cannot reject the null hypothesis that all three methods perform equally well on these data.

This last conclusion, however, is not surprising and is in fact reassuring. The similarity in performance of these rather different methods is most likely due, at least in part, to the fact that sequence similarity is the dominating information source for the protein-protein association. As pointed out in Peña-Castillo et al. (2008), it is hard to distinguish methods relying on protein-protein association data when this is the case. Therefore, we conclude that the results suggest that our proposed method of network-based auto-probit modeling is capable of performing on par with typical standard algorithms.

An important question is whether predictions from our model are stable with respect to natural variation in the training data. As an exploration of this issue, we perform a 100-trial simulation study on the CC network. A 10-fold cross-validation is conducted for each trial

and ROC curves are generated. The results demonstrate good reproducibility of predictive accuracy. Please refer to the supplementary material for details.

4.4 Analysis of the Gene Ontology annotation uncertainty

A distinguishing feature of our network-based auto-probit model for gene function prediction is the ease with which uncertainty information can be incorporated, which allows us, in particular, to explore the problem of GO annotation uncertainty, a key contribution of this paper.

We first fit the model introduced in Section 3 to the two networks CC and OOB, and conduct posterior inference on g , i.e., the probability that a protein is currently not annotated with the function of study but actually has the function. The hyperparameter setup here is the same as before. We run 9000 iterations after 1000 burn-in. Results from diagnostic methods indicate that both chains are converged. The histograms of the posterior estimates are shown in Figure 2. The posterior means of g are 0.1142 and 0.4186 for the CC network and the OOB network, respectively. Hence, we see that while in the CC network the rate of false negative annotations is estimated to be fairly small, that in the OOB network is estimated to be quite substantial – more than two out of every five.

Knowledge of these false negative rates may be in turn propagated through the process of posterior-based prediction to produce more accurate predictions based on currently “flawed” annotations. To illustrate, we perform a study implementing a variation of 10-fold cross-validation. Specifically, we use the “old” annotations for our two target functions (updated in June 2006) for training and the “new” annotations (updated in November 2007) for validation. That is, given the annotations in the 2006 GO database as training, for each of ten randomly selected folds, we set the labels of 10% of the proteins to “unknown”, analogous to the cross-validation study presented in Section 4.3, and the appropriate posterior predictive probabilities are computed. In validating our predictions, however, sensitivity and specificity are assessed using the labels in the 2007 GO database as a truth set. For each of the two target functions we train two models: (i) our original network-based auto-probit model (without g in the model), and (ii) the extended version of this model (with g in the model) accounting for annotation uncertainty.

Our intent in performing the study in this manner is to show that modeling the partially incorrect labeling in the 2006 data can, when evaluated against the more accurate 2007 data, be seen to improve predictive performance, as compared to not modeling such inaccuracies. Note that modeling the uncertainty in the negative annotations affects the entire model, and therefore in principle all predictions. Thus in presenting our results we show predictive performance evaluated on all proteins, not just those proteins with updated annotations. According to the GO database update in November 2007, there are 16 proteins in the CC network that were updated to have the annotation *intracellular signaling cascade*, i.e., compared to not having had the annotation in 2006; similarly, 38 proteins in the OOB network received the new annotation *chromosome organization and biogenesis* in 2007. Visualization and numerical assessment (see the supplementary materials) suggest that the 16 proteins updated in the CC network are scattered fairly uniformly throughout the network of protein-protein interactions. The 38 proteins updated in the OOB network show some modest (but not extreme) evidence of clustering. Thus the evaluation we present here does not seem likely to be unduly biased by some sort of advantageous placement of the updated proteins on the network.

We show the ROC curves from our model with and without g . For both networks, the auto-probit model with g outperforms the auto-probit model without g . For the CC network, the AUC's for our model with g and without g are 0.7529 and 0.6647, respectively. The p-value

for AUC comparison from a two-sided Wilcoxon signed rank test based on the ten folds is 0.0317, indicating a significant prediction improvement from incorporating the uncertainty in negative annotations. For the OOB network, the AUC's for the model with g and without g are 0.7260, 0.6616, respectively; and the p-value is 0.1602. These results demonstrate that our method is able to model and correct for missing annotations by producing improved predictions. While the improvement is modest in the OOB network, it is substantial in the CC network.

4.5 Additional results

We have conducted a number of additional numerical studies, largely aimed at evaluating the robustness and stability of our models and the predictions they produce to changes in various aspects.

We note that there is an important aspect of the overall modeling process implicit in the way that we have set up our prediction problems. That is, we have exploited our knowledge of the parent functions of the functions that we seek to predict. This decision is motivated by the fact that often when predicting a given protein some background knowledge may be available (either of a current but less specific GO annotation, for example, or of related biological information suggestive of such an annotation), allowing us to effectively restrict the prediction problem to a sub-network. If such knowledge is not available, then prediction would in principle need to be based on the a model fit to the full network.

We have explored the effect of this modeling issue on predicting our two terms, intracellular signal cascade and chromosome organization and biogenesis, in the CC and OOB networks, respectively. Specifically, we fit our model to the full PPI network in both cases, both with and without modeling for mis-annotations, and evaluate the performance on the CC and OOB sub-networks. The AUC under the corresponding ROC curves is given in Table 4 (the ROC curves themselves can be found in the Web Figures 6 and 7). We see that (a) the best results are achieved when we both model mis-annotations and restrict training to the respective sub-networks, but (b) whether modeling mis-annotations or not, on both datasets both methods perform similarly, and on par with the best results, when training on the full network. Since the sub-networks here are substantially smaller than the full PPI network, and the computational gains obtained by using smaller sub-networks will be only more critical for prediction in higher-level organisms (e.g., in human cell lines), it is clear that the modeling of mis-annotation proposed here has a fundamental role to play in successful prediction.

On a related note we point out that, given the hierarchical nature of GO, the choice of parent function can matter when deciding on the definition of a relevant sub-network. For example, for the function ISC, there are two non-trivial terms above it: *cell communication* (CC), which we have used in our work above, and *regulation of cellular process* (RCP). If we use RCP in place of CC and train our two models on the resulting sub-network, we again find that by modeling mis-annotation we outperform our original model. However, the performance of both models is noticeably worse than when using CC, with an AUC of 0.6498 and 0.5863, respectively. (Please see the supplementary material for ROC plots.) This change is possibly due to the fact that RCP defines a substantially larger sub-network than CC, and the proportion of proteins annotated with ISC in RCP is much smaller than in CC (i.e., 6.11% versus 56.87%). Together the above results suggest that some care is needed in choosing sub-networks upon which to fit our proposed models. A more detailed investigation of these issues is merited, but is beyond the scope of this paper.

It is also interesting to ask to what extent our model can predict false negative annotations. To obtain some insight into this question, we look at what percentage among the updated

proteins have their predictive probabilities increased by incorporating g into the model. We find that 12 out of the 16 proteins updated in the CC network (75%) and 23 out of the 38 proteins updated in the OOB network (61%) have increased probabilities, which is suggestive. See Web Figures 12 and 13 in the supplementary materials.

Finally, we note that the availability of the 2007 update data allows us to address not only questions related to predictions but also questions related to modeling fitting, such as stability of parameter estimates. For example, we compare the posterior mean value for each μ_i inferred under our model (without g) in 2006 and 2007. As described in the supplementary materials, we find that in the CC network very few of the estimated μ_i change sign from 2006 to 2007, that among those that do the majority correspond to the updated proteins; and that the vast majority of those that change sign and were not updated are immediate neighbors of updated proteins. These results suggest that the impact of changes in a value y_i on inference of parameters μ_i is relatively local. The results for the OOB network are qualitatively similar, though slightly less pronounced.

5. Discussion

This paper introduces a network-based fully Bayesian auto-probit model for protein function prediction. It takes protein-protein association networks as input and employs a latent Gaussian process z to encode proteins functional similarity. Using a hierarchical Bayesian model, we assign a conditional autoregressive (CAR) prior distribution with a single hyperparameter τ^2 to the location vector μ of the Gaussian process.

The incorporation of uncertainty in negative GO term annotations is shown here to yield substantial improvements in predictive accuracy. This observation has powerful implications, since the tendency toward emphasis on “positive results” in science, and the manner in which modern biological databases encode those results, means that this issue is not restricted to the Gene Ontology database alone, but rather is likely endemic to the area as a whole.

Although annotations that are derived from experimentally confirmed evidence are unlikely to be incorrect, it is still possible for positive annotations to be false. As genomic function prediction becomes more and more of a focus, many databases are starting to incorporate larger fractions of predicted annotations, where this issue is more common. As such, one may soon want to include the possibility of incorrect annotations in a model like ours, presumably accounting them in a similar fashion as what we propose here for the uncertainty in negative annotations.

In our work, protein annotations are predicted by conditioning on the protein-protein association network topology. However, it is known that protein-protein association data is traditionally noisy, in other words, the protein functional linkage graphs may lack of edges and contain false edges. One way to address this is to use more confident networks, such as those used in our paper extracted from the STRING database, rather than networks based on a single source of information. Another more sophisticated way is to account for the uncertainty in edges in the modeling process. Our network auto-probit model is structured in such a way that it should facilitate such modeling.

Note that our proposed model predicts protein function independently, without considering the relationship among functional terms. In fact, GO terms are structured as a directed acyclic graph (DAG), reflecting their ontological relationships with each other. Recent work has shown that some improvement in protein function prediction can be obtained by exploiting the structure among GO terms. See Jiang et al. (2008), for example, and

references therein. In principle, the network auto-probit model proposed here can be extended in an analogous manner.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank Brian Reich for helpful discussion at the start of this project. The authors also thank the reviewers for their comments and suggestions, which greatly improved the exposition of the manuscript. This work was supported in part by NIH grant R01 HG003367-01A1, NSF ITF 0428715, NSF DMS-0602204 (EMSW21-RTG, BioDynamics at Boston University), and ONR award N00014-06-1-0096.

References

- Chua HN, Sung WK, Wong L. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics*. 2006; 22:1623–1630. [PubMed: 16632496]
- Cressie, NAC. *Wiley Series in Probability and Statistics*. J Wiley; New York: c1991.
- Demsar J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*. 2006; 7:1–30.
- Deng M, Chen T, Sun F. An integrated analysis of protein function prediction. *Journal of Computational Biology*. 2004; 11:463–475. [PubMed: 15285902]
- Deng M, Zhang K, Mehta S, Chen T, Sun F. Prediction of protein function using protein-protein interaction data. *Journal of Computational Biology*. 2003; 10:947–960. [PubMed: 14980019]
- Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Statistical Science*. 1992; 7:457–472.
- Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning: data mining, inference, and prediction*. Springer. 2001
- Hishigaki H, Nakai K, Ono T, Tanigami A, Takagi T. Assessment of prediction accuracy of protein function from protein-protein interaction data. *Yeast*. 2001; 18:523–531. [PubMed: 11284008]
- Jiang, X.; Nariai, N.; Steffen, M.; Kasif, S.; Gold, D.; Kolaczyk, ED. Combining hierarchical inference in ontologies with heterogeneous data sources improves gene function prediction; *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine*; 2008b.
- Jiang X, Nariai N, Steffen M, Kasif S, Kolaczyk ED. Integration of relational and hierarchical network information for protein function prediction. *BMC Bioinformatics*. 2008a; 9:350. [PubMed: 18721473]
- Lanckriet GRC, Cristianini N, Jordan MI, Nobel WS. A statistical framework for genomic data fusion. *Bioinformatics*. 2004; 20:2626–2635. [PubMed: 15130933]
- Lehne B, Schlitt T. Protein-protein interaction databases: Keeping up with growing interactomes. *Human Genomics*. 2009; 3:291–297. [PubMed: 19403463]
- Letovsky S, Kasif S. Predicting protein function from protein-protein interaction data: a probabilistic approach. *Bioinformatics*. 2003; 19:i197–i204. [PubMed: 12855458]
- Liang F, Mukherjee S, West M. Understanding the use of unlabelled data in predictive modeling. *Statistical Science*. 2007; 22:189–205.
- Mering CV, et al. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research*. 2005; 33:D433–D437. [PubMed: 15608232]
- Nariai N, Kolaczyk ED, Kasif S. Probabilistic protein function prediction from heterogeneous genome-wide data. *PLoS ONE*. 2007; 2(3):e337. [PubMed: 17396164]
- Peña-Castillo, et al. A critical assessment of *Mus musculus* gene function prediction using integrated genomic evidence. *Genome Biology*. 2008; 9:S2.

- Schwikowski B, Uetz P, Fields S. A network of protein-protein interactions in yeast. *Nature Biotechnology*. 2000; 18:1257–1261.
- Sharan R, Ulitsky I, Shamir R. Network-based prediction of protein function. *Molecular Systems Biology*. 2007; 3:88. [PubMed: 17353930]
- Weir IS, Pettitt AN. Binary probability maps using a hidden conditional autoregressive Gaussian process with an application to finnish common toad data. *Applied Statistics*. 2000; 49:473–484.
- Zhu J, Hastie T. Kernel logistic regression and the import vector machine. *Journal of Computational and Graphical Statistics*. 2005; 14:185–205.

Figure 1.

Comparing the auto-probit model with the logistic kernel method and NN algorithm by ROC curves based on a 10-fold cross-validation. [Top]: predicting *intracellular signaling cascade* in the CC network; [Bottom]: predicting *chromosome organization and biogenesis* in the OOB network.

Figure 2. Histograms of the posterior estimates of g , the probability of being incorrectly un-annotated. [Top]: target function of *intracellular signaling cascade* in the CC network; [Bottom]: target function of *chromosome organization and biogenesis* in the OOB network.

Figure 3. Comparing the auto-probit model with and without annotation uncertainty by ROC curves based on 10-fold cross-validation. [Top]: predicting *intracellular signaling cascade* in the CC network; [Bottom]: predicting *chromosome organization and biogenesis* in the OOB network.

Table 1

Hyperparameter setup

	CC network	OOB network
τ^2	0.30	1.00
σ_μ in proposal	0.4	0.4
$\beta_{\min}$	0	0
$\beta_{\max}$	0.0032	0.0002
σ_β in proposal	0.0001	0.0001

Table 2

Estimated posterior means, Monte Carlo standard errors and 95% credibility intervals for the CC network with the model in Section 2

CC network			
Parameter	Degree	Estimate	95% credibility interval
β	–	0.0017	(0.0015, 0.0019)
$\mu_{YCR026C}$	1	−0.212	(−0.227, −0.197)
$\mu_{YCR038C}$	15	0.158	(0.153, 0.162)
μ_{GB035C}	25	0.082	(0.078, 0.086)
$\mu_{YMR037C}$	40	0.136	(0.132, 0.139)
$\mu_{YLR229C}$	58	0.157	(0.154, 0.160)

Table 3

Estimated posterior means, Monte Carlo standard errors and 95% credibility intervals for the OOB network with the model in Section 2

OOB network			
Parameter	Degree	Estimate	95% credibility interval
β	–	1.4764×10^{-4}	$(1.2608 \times 10^{-4}, 1.6921 \times 10^{-4})$
$\mu_{YBR172C}$	1	0.434	(0.374, 0.495)
$\mu_{YAL010C}$	15	−0.062	(−0.125, 0.000)
$\mu_{YLR068W}$	50	0.129	(0.072, 0.186)
$\mu_{YBL002W}$	100	−0.351	(−0.411, −0.292)
$\mu_{YLR175W}$	202	0.221	(0.161, 0.282)

Table 4

Summary of AUC for predictions on CC and OOB networks

	Predicting ISC on CC	Predicting COB on OOB
Train model on subnetwork and evaluate performance on subnetwork ^a	0.7539, 0.6641 ^b	0.7259, 0.6616
Train model on large network and evaluate performance on subnetwork	0.8061, 0.7766	0.6541, 0.6417

^aThis is the analysis in Section 4.4^bThe first number is AUC from the auto-probit model with g ; the second number is AUC from the auto-probit model without g .