

# BIOINFORMATICS

Electronic edition <http://bioinformatics.oupjournals.org>

VOLUME 20  
NUMBER 8  
22 May 2004  
PAGES 1205–1213

DOI: 10.1093/bioinformatics/bth061

## ***Probabilistic inference of molecular networks from noisy data sources***

*Ivan Iossifov<sup>1,2,\*</sup>, Michael Krauthammer<sup>1,2</sup>, Carol Friedman<sup>1</sup>, Vasileios Hatzivassiloglou<sup>3</sup>, Joel S. Bader<sup>5</sup>, Kevin P. White<sup>4</sup> and Andrey Rzhetsky<sup>1,2</sup>*

<sup>1</sup>Department of Medical Informatics, <sup>2</sup>Columbia Genome Center, Columbia University, New York, NY 10032, USA, <sup>3</sup>Department of Computer Science, Columbia University, New York, NY 10027, USA,

<sup>4</sup>Department of Genetics, Yale University School of Medicine, New Haven, CT 06520, USA and

<sup>5</sup>CuraGen Corporation, New Haven, CT 06511, USA

Received on July 29, 2003; accepted on September 3, 2003

Advance Access publication February 10, 2004

---

\*To whom correspondence should be addressed.



GO BACK

CLOSE FILE

# Abstract

**Summary:** *Information on molecular networks, such as networks of interacting proteins, comes from diverse sources that contain remarkable differences in distribution and quantity of errors. Here, we introduce a probabilistic model useful for predicting protein interactions from heterogeneous data sources. The model describes stochastic generation of protein–protein interaction networks with real-world properties, as well as generation of two heterogeneous sources of protein-interaction information: research results automatically extracted from the literature and yeast two-hybrid experiments. Based on the domain composition of proteins, we use the model to predict protein interactions for pairs of proteins for which no experimental data are available. We further explore the prediction limits, given experimental data that cover only part of the underlying protein networks. This approach can be extended naturally to include other types of biological data sources.*

**Contact:** [iossifov@dbmi.columbia.edu](mailto:iossifov@dbmi.columbia.edu)

Abstract

Introduction

Probabilistic Model

Network inference...

Three parts of the...

Results and Discussion

Acknowledgements

References



GO BACK

CLOSE FILE

# Introduction

The past five decades of molecular biology have brought an incredible wealth of high-quality information about the molecular machinery of living cells. Although the driving force behind this knowledge acquisition has been the satisfaction of individual research questions, an underutilized product is a mass of data that could be used to generate testable predictions in much the same way as is done in physics. However, by and large, these data are locked within the literature, and extracting them is a challenging process that introduces noise associated with misinterpretation of results by text-mining algorithms or by data curators. In addition to the vast biological literature, recent developments in genomics have resulted in another type of data that comprise hundreds or thousands of measurements of gene function, such as gene-expression levels or protein-protein interactions, under a given set of experimental conditions. These functional genomic datasets tend to be consolidated and therefore easily accessible to analysis, but they also tend to have a high level of noise associated with spurious or irrelevant measurements.

The amount of molecular data from these two sources is immense and is growing rapidly. A critical mass of these data should allow generation of testable models of molecular networks as the combined evidence helps to separate the relevant biological signals from the underlying noisy data. Current progress toward automated generation of molecular networks is limited by the rate of information processing and interpretation rather than

*Abstract*

**Introduction**

*Probabilistic Model*

*Network inference...*

*Three parts of the...*

*Results and Discussion*

*Acknowledgements*

*References*



**GO BACK**

**CLOSE FILE**

by the rate of accumulation of new information; hence, development of predictive mathematical models appears especially important.

There are many types of molecular interactions that are routinely reported in the literature or are analyzed using high-throughput methods. To test our methods for extracting these data and building molecular networks, we have chosen to focus on interactions between proteins. In this paper, we consider the problem of predicting interactions between pairs of novel proteins with known sequences, given a set of experimentally determined interactions. Previous studies have introduced a framework for predicting protein–protein interactions (Marcotte *et al.*, 1999; Sprinzak and Margalit, 2001; Gomez *et al.*, 2001; Bock and Gough, 2001; Gomez and Rzhetsky, 2002; Deng *et al.*, 2002; Tong *et al.*, 2002; Goldberg and Roth, 2003) but have only led to only moderate success, in part because they focused on a single type of experimental data. Here, we overcome this limitation by assuming that the available experimental data are heterogeneous, arising from more than one source and exhibiting different error patterns. Furthermore, we present a method for integrating diverse data types. We consider here data arising from large-scale yeast two-hybrid experiments and from automated analysis of numerous research articles by information extraction systems such as GeneWays, developed by our group (Koike and Rzhetsky, 2000; Rzhetsky *et al.*, 2000; Friedman *et al.*, 2001; Hatzivassiloglou *et al.*, 2001; Krauthammer *et al.*, 2002; Fig. 1).

**Abstract**

**Introduction**

**Probabilistic Model**

**Network inference...**

**Three parts of the...**

**Results and Discussion**

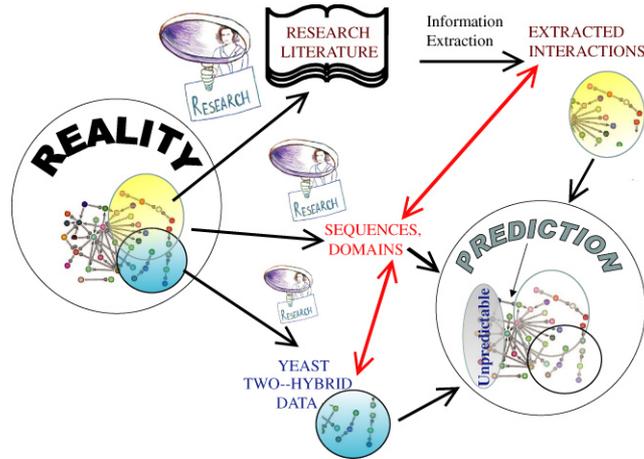
**Acknowledgements**

**References**



**GO BACK**

**CLOSE FILE**



**Fig. 1.** Main model diagram.

We present a mathematical framework for prediction of protein–protein interactions in the real world from published statements about such interactions in the research literature and from observed interactions in yeast two-hybrid experiments. Our framework allows heterogeneous data-production processes and different types of error during each process. Domain composition of proteins is used as a common point of reference for the different data types. Using prior observations, automatically extracted from the literature and derived from yeast two-hybrid data, our model allows estimation of the confidence in a given set of predicted

**Abstract**

**Introduction**

**Probabilistic Model**

**Network inference...**

**Three parts of the...**

**Results and Discussion**

**Acknowledgements**

**References**



**GO BACK**

**CLOSE FILE**

or experimentally determined interactions. Additionally, confidence estimates can be calculated for protein interactions that have no available experimental data.

We offer evidence for the power and utility of this approach by first constructing a plausible, simulated model of real-world protein–protein interactions proteome-wide. These real-world networks are composed of potential interactions that can occur between proteins. We generate these real-world networks stochastically from well-described distributions of domain compositions of real proteomes and protein–protein interaction networks. We then use this model as the basis for simulating experimental results from yeast two-hybrid data and results extracted from journal articles. Simulating the real-world protein–protein interactions and the processes that generate observable evidence allows perfect measurement of whether our predictions are correct by matching the predictions to the simulated real-world network. Further, as research articles and yeast two-hybrid data are usually incomplete and cover only partially overlapping sets of interactions between proteins, we are interested in exploring the prediction limits, given these fractional data. Using simulated data, we can study the influence of incomplete datasets by changing the coverage of the observable experimental results.

In addition to predicting protein–protein interactions from multiple data sources, our model also generates valuable information on the quality of the observed datasets. It does so by making a joint assignment of

*Abstract*

*Introduction*

*Probabilistic Model*

*Network inference...*

*Three parts of the...*

*Results and Discussion*

*Acknowledgements*

*References*



**GO BACK**

**CLOSE FILE**

labels ‘true’ and ‘false’ to interactions observed in two-hybrid experiments and of labels ‘correctly extracted, true,’ ‘correctly extracted, false’ and ‘incorrectly extracted’ to actions derived by text-mining systems (such as GeneWays) from research articles. Therefore, our model is useful for filtering noise from the observed data.

This paper describes the model in its entirety, but we emphasize the model’s ability to generate novel protein–protein interaction predictions and explore its prediction limits, given fractional or incomplete datasets.

**Abstract**

**Introduction**

**Probabilistic Model**

**Network inference...**

**Three parts of the...**

**Results and Discussion**

**Acknowledgements**

**References**



**GO BACK**

**CLOSE FILE**

# Probabilistic Model

Our model includes three major stochastic components: proteome-wide simulation of potential protein–protein interactions, simulation of statements published in the research literature and simulation of yeast two-hybrid data.

## Protein–protein interaction reality

Following an established view in structural biology, we consider each protein as a collection of protein domains; the linear order of such domains within a protein is not important for the purpose of this study, and so each protein is treated as a collection of unordered domains. A domain is defined as a portion of a polypeptide chain that is identified and described by a human expert, or by a computer program based on expert annotation. Domains often correspond to spatially compact structures; the same domain (although with variation at the amino-acid level) may occur in multiple proteins within the same organism.

Relatively recently, it became apparent that the domain compositions of real proteomes and real protein–protein interaction networks have highly non-random properties. In the context of our model, the following five distributions appear important:

- (A) The proportion of proteins having exactly  $k$  interactions with other proteins (including self-interactions).

[Abstract](#)

[Introduction](#)

[Probabilistic Model](#)

[Network inference...](#)

[Three parts of the...](#)

[Results and Discussion](#)

[Acknowledgements](#)

[References](#)



GO BACK

CLOSE FILE

- (B) The proportion of domain types per proteome having exactly  $k$  copies.
- (C) The proportion of proteins having exactly  $k$  domain copies (of any type).
- (D) The proportion of proteins having exactly  $k$  domain types (regardless of the number of copies of each domain type).
- (E) The proportion of domain types having exactly  $k$  interactions with other (or the same) domain types.

It appears ([Barabasi and Albert, 1999](#); [Albert and Barabasi, 2000](#); [Jeong \*et al.\*, 2000](#); [Albert \*et al.\*, 2000](#); [Jeong \*et al.\*, 2001](#); [Rzhetsky and Gomez, 2001](#); [Bader and Hogue, 2002](#); [Koonin \*et al.\*, 2002](#); [Park \*et al.\*, 2001](#)) that distributions A, B, C and E are Zeta-distributions [the Zeta- or Zipf–Estoup distribution is a discrete counterpart of Pareto distribution ([Johnson and Kotz, 1969](#))], whereas distribution D is an exponential distribution. The choice of distributions A, B, C and E but not D significantly affects the outcome of modeling. Therefore, in our simulations of reality, we generated protein networks with Zeta-distributions for distributions A–E (details of the simulation are given in the following section).

In our model, we assume that domain interactions specify protein interactions in a deterministic way: every pair of domains either is interacting or is not interacting (nothing in between); if two proteins contain at least one pair of interacting domains that belong to different proteins, the two

**Abstract**

**Introduction**

**Probabilistic Model**

**Network inference...**

**Three parts of the...**

**Results and Discussion**

**Acknowledgements**

**References**



**GO BACK**

**CLOSE FILE**

proteins also interact (Deng *et al.*, 2002). However, because of omissions, incorrect statements and errors in data analysis, the interacting proteins may appear as non-interacting in two-hybrid experiments and in scientific publications. We have also developed a probabilistic model for relating domain–domain interactions with protein–protein interactions (Gomez *et al.*, 2001; Gomez and Rzhetsky, 2002).

### Generating the simulated reality

We first generated protein–protein interaction networks by creating protein domain families that consist of all copies in the genome for each domain type. We then simulated interactions between domains and combined domains into multidomain proteins.

Denoting the size of the domain-type universe with  $N_D$ , and the parameter of a Zeta-distribution with  $\Gamma_D$ , we draw  $N_D$  samples from the Zeta-distribution, which gives us the set of numbers of copies for  $N_D$ -domain-type families. Next, to generate interactions between domain pairs and to combine domains into multidomain proteins, we used variations of the stochastic model introduced by Barabasi *et al.* (2000). We began by selecting a random pair of domain types, and then we formed an interaction between them. Next, we continued sampling random (previously unused) domain types, each time connecting them to existing domain types such that the probability of attaching a new domain type to an old domain type with  $k$  interactions is proportional to  $k^r$ , where  $r$  is a positive parameter. Further, once domain-type interactions were completely

Abstract

Introduction

Probabilistic Model

Network inference...

Three parts of the...

Results and Discussion

Acknowledgements

References



GO BACK

CLOSE FILE

defined, we generated multidomain proteins from domain families. In this case, we started with two randomly selected domain copies that represent two single-domain proteins. We then continued by randomly selecting a pair of unused domain copies of either the same domain type or of distinct domain types (either interacting or not). We used one copy of the pair to create a new one-domain protein, while we concatenated the second copy to one of the existing proteins such that the probability of adding a new domain to a protein with  $k$  domain copies (of any type) was proportional to  $k^q$ , where  $q$  is a positive parameter. The process stopped once all domain copies were used.

Analysis of simulated data showed that the statistical properties of a protein–protein interaction network generated in this manner are indeed close to the expected properties.

### **Generating statements in the literature**

We use the simulated real-world network described in the previous section to generate simulated research results on protein interactions published in scientific articles. In our model, each published result on a particular protein–protein interaction is defined as a ‘statement’. We assume two types of statements: ‘true’ statements—statements that agree with the real-world network—and ‘false’ statements that disagree with the real-world network. Further, each statement is either ‘positive’ (‘protein A activates protein B’) or ‘negative’ (‘protein A does not activate protein B’).

*Abstract*

*Introduction*

*Probabilistic Model*

*Network inference...*

*Three parts of the...*

*Results and Discussion*

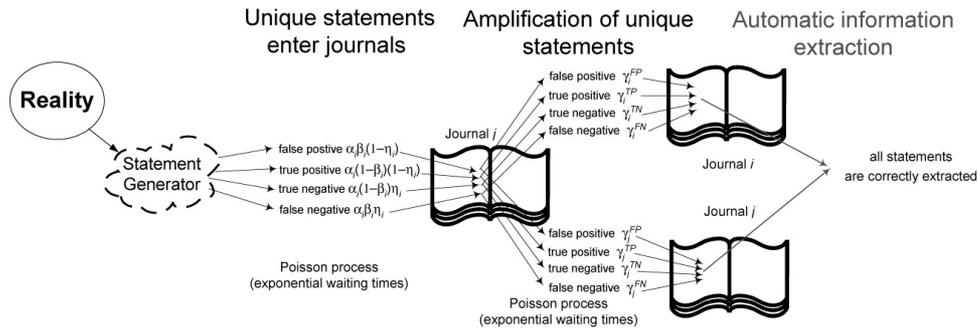
*Acknowledgements*

*References*



**GO BACK**

**CLOSE FILE**



**Fig. 2.** Stochastic model of the generation and propagation of statements about molecular interactions through journal articles.

We began the stochastic generation of literature statements by sampling interactions from the simulated real-world network of protein interactions (Fig. 2) by a noisy truth generator. Each scientific journal in our model has an associated parameter,  $\alpha_i$ , that represents the rate of a Poisson process per time unit that supplies the  $i$ -th journal with unique statements (statements not published previously) about interactions or non-interactions of proteins. The total number of unique statements published by all journals during time interval  $T$  is a random number sampled from a Poisson distribution with parameter  $\Lambda T$ , where  $\Lambda$  is the sum of the  $\alpha_i$ s over all journals. After the total number of unique statements is determined, the statements are distributed among journals—each statement has a probability,  $\alpha_i/\Lambda$ , of being published in the  $i$ -th journal. After the total number of unique

- Abstract**
- Introduction**
- Probabilistic Model**
- Network inference ...**
- Three parts of the ...**
- Results and Discussion**
- Acknowledgements**
- References**



⏪
⏩

◀
▶

GO BACK

CLOSE FILE

statements per journal has been decided, the noisy truth generator classifies each statement into one of four categories: false positive, true positive, true negative or false negative, with probabilities  $\beta_i(1 - \eta_i)$ ,  $(1 - \beta_i)(1 - \eta_i)$ ,  $(1 - \beta_i)\eta_i$  and  $\beta_i\eta_i$ , respectively (Fig. 2). For each statement, the noisy truth generator samples an appropriate unknown interaction from the simulated real-world protein-interaction network. For example, if the required statement is to be false positive, the noisy truth generator picks a random unknown negative interaction, converts the latter into a positive interaction (hence, the statement becomes false) and supplies the interaction to the  $i$ -th journal. Parameter  $\beta_i$  represents the expected proportion of unique statements that are false in the  $i$ -th journal; parameter  $\eta_i$  represents the expected proportion of unique statements that are negative in the  $i$ -th journal. Both parameters can vary from journal to journal, representing differences in journal quality (low  $\beta_i$ ) and bias toward positive findings (low  $\eta_i$ ).

After each unique statement is published for the first time, it becomes subject to amplification (re-publishing of the original statement by fellow researchers), which is a separate Poisson process whose rate is different for each of the four types of unique statements: these rates are  $\gamma_i^{\text{FP}}$ ,  $\gamma_i^{\text{TP}}$ ,  $\gamma_i^{\text{TN}}$  and  $\gamma_i^{\text{FN}}$ , for false-positive, true-positive, true-negative and false-negative unique statements, respectively, published in the  $i$ -th journal.

Unlike our previous stochastic model of research literature (Krauthammer *et al.*, 2002), the new model does not assume that the supply of true

**Abstract**

**Introduction**

**Probabilistic Model**

**Network inference...**

**Three parts of the...**

**Results and Discussion**

**Acknowledgements**

**References**



**GO BACK**

**CLOSE FILE**

and false statements is infinite. Additionally, in the present model, the real world has a finite size in terms of the number of proteins and interactions between them. We also allow the past history of published statements to affect the pattern of sampling of statements in the present. Moreover, in the current version of the model, we assume that the sampling is uniform over all unknown interacting protein pairs (positive interactions) and over all unknown non-interacting protein pairs (negative interactions), but the probability of sampling a positive interaction is generally greater than the probability of sampling a negative interaction. The latter assumption is based on the observation that it is much harder to publish a negative interaction ('protein A does not interact with protein B') than to publish a positive one. A key assumption of this model is that the publication patterns for true and false statements are different. In other words, we can derive the trustworthiness of actions by studying how a particular action has been published over time.

### Yeast two-hybrid data

Stochastic generation of yeast two-hybrid data (Davy *et al.*, 2001; Bader and Hogue, 2002; Gietz and Woods, 2002) was accomplished by sampling from the simulated real-world interactions and then by performing simulated experiments for the sampled interactions. We implemented two versions of the sampling from the simulated real-world interactions: the simplest version used a random selection of unknown interactions one

**Abstract**

**Introduction**

**Probabilistic Model**

**Network inference...**

**Three parts of the...**

**Results and Discussion**

**Acknowledgements**

**References**



**GO BACK**

**CLOSE FILE**

by one, regardless of the type of the sampled interaction. A more complex version allowed random selection of a set of proteins and testing of all interactions between them, and it included the possibility of multiple experiments describing each interaction.

Our stochastic model of simulated yeast two-hybrid experiments had just two parameters,  $\rho_N$  and  $\rho_P$ , which represented the probability of error, given either negative or positive interactions sampled from the real-world network simulation. For each interaction sampled, we generated experimental data, drawing from binomial distributions with parameter  $\rho_N$  for negative interactions, and parameter  $\rho_P$  for positive interactions.

**Abstract**

**Introduction**

**Probabilistic Model**

**Network inference...**

**Three parts of the...**

**Results and Discussion**

**Acknowledgements**

**References**



**GO BACK**

**CLOSE FILE**

## Network inference and parameter estimation

With the generative stochastic model defined, we computed the probability of data, given the model and model parameters (the likelihood). Our data were represented by three components: redundant statements about protein–protein interactions automatically extracted from the literature, sets of positive and negative protein interactions generated by the yeast two-hybrid experiments, and sequences of all proteins from the species under analysis. In this study, we chose not to estimate the parameters of generating the real-world protein–protein network, instead treating these parameter values as known. Therefore, the joint likelihood ( $L$ ) in our application is a product of four components: the probability of label assignment (‘correctly extracted, true’, ‘correctly extracted, false’ and ‘incorrectly extracted’) in the text-derived data given the current network ( $L_{\text{Literature}}$ ), the probability of label assignment (‘false’ and ‘true’) in the yeast two-hybrid data given the current network ( $L_{\text{Y2H}}$ ), the probability of the current network given protein domain composition (either zero or one under the current model) and domain interaction matrix, and the probability of the protein–protein interaction network topology ( $L_{\Psi}$ ), i.e.

$$L = L_{\text{Literature}} \times L_{\text{Y2H}} \times L_{\Psi}.$$

[Abstract](#)

[Introduction](#)

[Probabilistic Model](#)

[Network inference...](#)

[Three parts of the...](#)

[Results and Discussion](#)

[Acknowledgements](#)

[References](#)



GO BACK

CLOSE FILE

The probability of label assignment in the text-derived data, given the current network ( $L_{\text{Literature}}$ ), is calculated as follows:

$$\begin{aligned}
 L_{\text{Literature}} = & \prod_i f_{\text{Poiss}} \left( m_i^{\text{unique}} \mid \alpha_i T_i^{\text{unique}} \right) \\
 & \times b \left( n_i^F \mid n_i^F + n_i^T, \beta_i \right) \\
 & \times b \left( n_i^{\text{negative}} \mid n_i^{\text{negative}} + n_i^{\text{positive}}, \eta_i \right) \\
 & \times f_{\text{Poiss}} \left( n_i^{\text{FP}} \mid \gamma_i^{\text{FP}} T_i^{\text{FP}} \right) \times f_{\text{Poiss}} \left( n_i^{\text{TP}} \mid \gamma_i^{\text{TP}} T_i^{\text{TP}} \right) \\
 & \times f_{\text{Poiss}} \left( n_i^{\text{TN}} \mid \gamma_i^{\text{TN}} T_i^{\text{TN}} \right) \times f_{\text{Poiss}} \left( n_i^{\text{FN}} \mid \gamma_i^{\text{FN}} T_i^{\text{FN}} \right),
 \end{aligned}$$

where subscript  $i$  refers to the  $i$ -th journal;  $m_i^{\text{unique}}$  and  $T_i^{\text{unique}}$  are the total numbers of unique statements and the time during which these statements were accumulated;  $n_i^F$  and  $n_i^T$  correspond to the observed numbers of false and true unique statements;  $n_i^{\text{negative}}$  and  $n_i^{\text{positive}}$  are the observed numbers of positive and negative unique statements; pair  $(n_i^{\text{FP}}, T_i^{\text{FP}})$  represents the observed number of amplified false-positive statements and corresponding amplification time, and pairs  $(n_i^{\text{TP}}, T_i^{\text{TP}})$ ,  $(n_i^{\text{TN}}, T_i^{\text{TN}})$  and  $(n_i^{\text{FN}}, T_i^{\text{FN}})$  represent analogous quantities for true positive, true negative and false negative amplified statements, respectively; and  $b(x \mid n, p) = \binom{n}{x} p^x (1 - p)^{n-x}$  is the probability density function of the binomial distribution and

- [Abstract](#)
- [Introduction](#)
- [Probabilistic Model](#)
- [Network inference...](#)
- [Three parts of the...](#)
- [Results and Discussion](#)
- [Acknowledgements](#)
- [References](#)



◀

▶

◀

▶

GO BACK

CLOSE FILE

$f_{\text{Pois}}(x|\lambda) = (\lambda^x/x!)e^{-\lambda}$  is the probability density function of the Poisson distribution.

The likelihood of label assignments in the yeast two-hybrid data, given the current network ( $L_{\text{Y2H}}$ ), was calculated as follows:

$$L_{\text{Y2H}} = b(m_{\text{FN}}|m_{\text{FN}} + m_{\text{TP}}, \rho_P) \\ \times b(m_{\text{FP}}|m_{\text{FP}} + m_{\text{TN}}, \rho_N),$$

where  $m_{\text{FN}}$ ,  $m_{\text{TP}}$ ,  $m_{\text{FP}}$  and  $m_{\text{TN}}$  are the observed numbers of yeast two-hybrid data points that are negative and labeled ‘false’, positive and labeled ‘true’, positive and labeled ‘false’ and negative and labeled ‘true’, respectively.

Finally, the likelihood of the protein–protein interaction network topology ( $L_{\Psi}$ ) is

$$L_{\Psi} = \binom{N_P}{k_0 k_1 \cdots k_{N_P+1}} \prod_{i=0}^{N_P+1} z_0(i|\Gamma_P, p_0)^{k_i},$$

where  $N_P$  is the number of proteins in the network;  $k_x$  is the number of proteins with degree  $x$  (the maximum degree of a protein is  $N_P + 1$ ; it is achieved when the protein interacts with itself and all the other proteins);

- [Abstract](#)
- [Introduction](#)
- [Probabilistic Model](#)
- [Network inference...](#)
- [Three parts of the...](#)
- [Results and Discussion](#)
- [Acknowledgements](#)
- [References](#)



◀

▶

◀

▶

GO BACK

CLOSE FILE

and  $z_0(n|\gamma, p_0)$  is the probability density function of a modified Zeta-distribution with added probability for 0 equal to  $p_0$  and slope  $\Gamma_P$ —i.e.

$$z_0(x|\Gamma_P, p_0) = \begin{cases} p_0 & \text{if } x = 0 \\ cx^{-\gamma} & \text{if } x = 1, 2, \dots \end{cases}$$

$$c = \frac{1 - p_0}{\sum_{i=1}^{\infty} x^{-\gamma}}.$$

We assumed an uninformative prior distribution over parameter values and interaction assignment, and we inferred the posterior distribution of parameter values and network topologies (i.e. the probability of network and parameter values, given the data).

### Inferring the posterior distribution

For the statistical inference, we used a version of the Markov chain Monte Carlo (MCMC) technique (Gilks *et al.*, 1996). The essence of the MCMC in our case was a random walk through the discrete space of all possible protein–protein interaction networks and through the continuous space of admissible values of model parameter. We continued the walk for a large number (millions) of cycles of full update for all model parameters and for the network topology. After each full update cycle, we recorded the current values of parameters and current network topology. We estimated the posterior distributions of network edges and parameters directly from the frequencies of visiting corresponding states in the recorded MCMC run.

**Abstract**

**Introduction**

**Probabilistic Model**

**Network inference...**

**Three parts of the...**

**Results and Discussion**

**Acknowledgements**

**References**



**GO BACK**

**CLOSE FILE**

The MCMC random walk started with a randomly chosen set of parameter values and an arbitrary network (together we will refer to them as a state of the system,  $X$ ). Next, a potential change for the state (the proposal state,  $Y$ ) was generated and accepted with probability  $A$ , which was calculated in the following way (Hastings, 1970; Metropolis *et al.*, 1953):

$$A(X, Y) = \min \left[ 1, \frac{L(Y)q(X|Y)}{L(X)q(Y|X)} \right],$$

where  $L(X)$  and  $L(Y)$  are likelihood values for states  $X$  and  $Y$ , respectively, and  $q(X|Y)$  and  $q(Y|X)$  are conditional probabilities of proposing state  $X$  being in state  $Y$  and vice versa, respectively; distributions  $q(X|Y)$  and  $q(Y|X)$  are referred to as proposal distributions. In our application, we updated the parameters and network topology in a stepwise fashion, such that  $X$  differed from  $Y$  either by the value of a single parameter or by a single-edge change in the protein–protein interaction network topology. Since the update order does not affect the outcome, all the parameters and the network topology were updated in alphabetical order of the corresponding symbols. Further, an analysis of the model showed that every given set of domain interactions completely determined the protein interaction network, which in turn determined the set of labels for the observed data points (both literature and yeast two-hybrid); given a fixed set of data labels, however, all parameter values were independent. We analytically derived full conditional distributions for each parameter, given fixed labels

**Abstract**

**Introduction**

**Probabilistic Model**

**Network inference...**

**Three parts of the...**

**Results and Discussion**

**Acknowledgements**

**References**



**GO BACK**

**CLOSE FILE**

of the data points. Parameter values were updated by directly sampling from the corresponding conditional distribution, always accepting the proposed parameter value [Gibbs sampling, see [Gilks \*et al.\* \(1996\)](#)]. Indeed, for the  $i$ -th journal, we have

$$\alpha_i \sim \text{Gamma}\left(m_i^{\text{unique}} + 1, \frac{1}{T_i^{\text{unique}}}\right),$$

$$\beta_i \sim \text{Beta}\left(n_i^F + 1, n_i^T + 1\right),$$

$$\eta_i \sim \text{Beta}\left(n_i^{\text{negative}} + 1, n_i^{\text{positive}} + 1\right),$$

$$\gamma_i^{\text{FP}} \sim \text{Gamma}\left(n_i^{\text{FP}} + 1, \frac{1}{T_i^{\text{FP}}}\right),$$

$$\gamma_i^{\text{TP}} \sim \text{Gamma}\left(n_i^{\text{TP}} + 1, \frac{1}{T_i^{\text{TP}}}\right),$$

$$\gamma_i^{\text{TN}} \sim \text{Gamma}\left(n_i^{\text{TN}} + 1, \frac{1}{T_i^{\text{TN}}}\right),$$

$$\gamma_i^{\text{FN}} \sim \text{Gamma}\left(n_i^{\text{FN}} + 1, \frac{1}{T_i^{\text{FN}}}\right),$$

**Abstract**

**Introduction**

**Probabilistic Model**

**Network inference...**

**Three parts of the...**

**Results and Discussion**

**Acknowledgements**

**References**



**GO BACK**

**CLOSE FILE**

where  $\sim$  stands for ‘follows distribution’, and the notations  $\text{Beta}(x, y)$ , and  $\text{Gamma}(z, w)$  represent the probability density functions of the univariate beta and gamma distributions with parameters  $(x, y)$  and  $(z, w)$ , respectively ([Johnson and Kotz, 1970](#)).

In the case of multiple scientific journals in the model, the update cycles ran through all journals and for every parameter of every journal.

Similarly, parameters related to yeast two-hybrid data were sampled from the following full conditional distributions:

$$\begin{aligned}\rho_P &\sim \text{Beta}(m_{FN} + 1, m_{TP} + 1) \\ \rho_N &\sim \text{Beta}(m_{FP} + 1, m_{TN} + 1).\end{aligned}$$

Therefore, the full MCMC and the computation of the Metropolis–Hastings acceptance probability in our case was required only for updating the network topology in the model. We found that network updating was the most difficult component of the MCMC implementation. When implemented in the simplest fashion, where a random pair of domains was selected and the sign of the interaction between them was reversed, it led to abysmally slow convergence of the MCMC simulation. We therefore implemented an alternative strategy for updating domain type interactions. First, by sampling from a trinomial distribution, we determined whether an existing edge should be moved or deleted or whether a new edge should be added. In the case of interaction addition or deletion, we selected and reversed a random negative or positive interaction. In cases where a

[Abstract](#)

[Introduction](#)

[Probabilistic Model](#)

[Network inference...](#)

[Three parts of the...](#)

[Results and Discussion](#)

[Acknowledgements](#)

[References](#)



GO BACK

CLOSE FILE

translocation of a positive interaction (which is synonymous to moving an edge) was required, we started by selecting a random domain type in the network. If this domain type had no interactions, the network-update iteration ended. If the selected domain type did have interactions, we selected and removed a random interaction, while we randomly reconnected the original domain type to a new domain type such that the probability of reconnecting to a domain type with exactly  $k$  interactions was proportional to  $k^x$ , where  $x$  is a positive parameter. (We also allowed for reconnecting to domain types with zero connectivity, with a small positive probability.) The protein-interaction network was obtained deterministically from the domain-type interaction network. This version of network updating led to surprisingly fast convergence of the MCMC process. This faster MCMC convergence seems to be due to the tendency of the proposed network-update method to generate random networks that have Zeta-distributed connectivity of proteins.

**Abstract**

**Introduction**

**Probabilistic Model**

**Network inference...**

**Three parts of the...**

**Results and Discussion**

**Acknowledgements**

**References**



**GO BACK**

**CLOSE FILE**

## Three parts of the protein-interaction space

We can decompose the whole space of protein–protein interactions into three disjoint portions (Fig. 1): observable subspace [i.e. interactions that are directly observable (with errors)], subspace of predictable or deducible interactions and a subspace of interactions that are unpredictable under our model (in the absence of additional data). It is clear what the first subspace is, but the second and third subspaces require explanation. Under our model, we can correctly predict an interaction between a pair of proteins, given a domain-type interaction matrix deduced without errors from the observed data, in the two following cases: the first case is when two proteins under consideration have at least one pair of domain types present in different proteins that are known to interact (then we predict a positive interaction). Another predictable case is when two proteins contain only domain types that are known to lack interaction (then we predict a negative interaction). If two proteins contain no known interacting domain types and we lack information for at least one domain-type pair for domain types in distinct proteins, interaction cannot be predicted under our model. In reality, the domain-type interaction matrix is estimated statistically rather than observed directly, and so only a portion of interactions in the predictable subspace will be predicted correctly. Every domain-type interaction matrix with values 1 (positive interaction between domain types),  $-1$  (negative interaction between domain types), and 0 (no information

*Abstract*

*Introduction*

*Probabilistic Model*

*Network inference...*

**Three parts of the...**

*Results and Discussion*

*Acknowledgements*

*References*



GO BACK

CLOSE FILE

about the interaction) unambiguously partitions protein–protein interactions into three subspaces, according to predictive rules just discussed for a set of proteins with known domain composition.

**Figure 3** shows the dynamics of the absolute and relative size of the predictable area as a function of the number of visible interactions computed for a repeatedly simulated protein universe of 83 proteins. As we would expect, the absolute size of the predictable area is small when the observed data are few; it peaks at the point when about one-third of all interactions is known and then decreases again as the absolute size of the unobserved area (predictable plus unpredictable areas) decreases to zero. The proportion of the unobserved area that is predictable grows steadily with the growth of the visible area, almost to the point that nearly all interactions are known (**Fig. 3**).

In our evaluation of the stochastic model described in the following section, we measured, based on a simulated set of real-world protein interactions, the efficiency of our model for the two tasks described in the present and previous sections: cleansing noisy data of experimental errors introduced by the two-hybrid experiments and transferred to the literature, and predicting interactions between proteins not directly seen to interact in the real world, as discussed earlier in this section.

*Abstract*

*Introduction*

*Probabilistic Model*

*Network inference...*

**Three parts of the...**

*Results and Discussion*

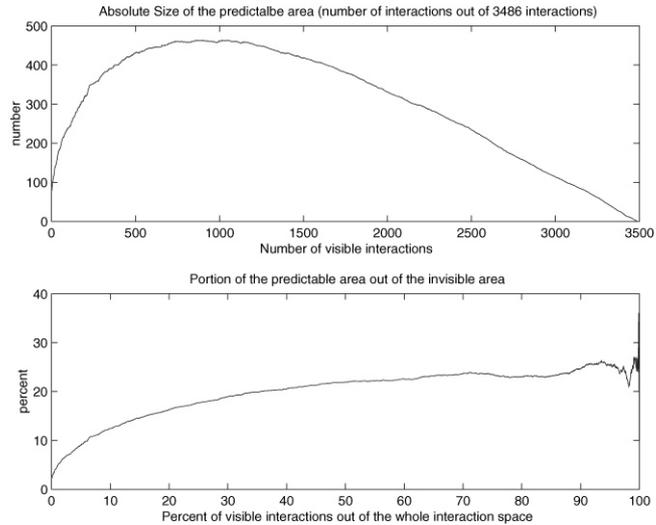
*Acknowledgements*

*References*



GO BACK

CLOSE FILE



**Fig. 3.** Predictable area curve. **(A)** The curve shows how many new interactions can be predicted after we have sampled a certain number of observable interactions. **(B)** The same data as in **(A)** are shown but are expressed as a percentage of the number of predictable interactions out of the number of unobservable (predictable plus unpredictable) interactions as a function of the percentage of observable interactions out of the total number of possible interactions.

- [Abstract](#)
- [Introduction](#)
- [Probabilistic Model](#)
- [Network inference...](#)
- [Three parts of the...](#)
- [Results and Discussion](#)
- [Acknowledgements](#)
- [References](#)



⏪
⏩

◀
▶

GO BACK

CLOSE FILE

## Results and Discussion

To test the correctness of our inference approach and its implementation we used computer-generated data, following our model exactly. Our model of a real-world network contained just 83 proteins with 100 domain types. The scientific literature was limited to a single journal that was ‘published’ over a period of 2 years, with preference given to positive statements about protein interactions: only 30% of published statements were negative, and only 1% of all published statements were false. (We currently have no information about the proportion of false statements in the research literature, but we suspect that it is  $>1\%$ .) Our simulated journal published protein-interaction information with a Poisson rate of 0.18 statements per day; true-positive, true-negative, false-positive and false-negative statements were amplified at rates of 0.1, 0.05, 0.02 and 0.01 statements per day, respectively. Our simulated yeast two-hybrid dataset contained errors at a rate of 15% for positive interactions and 15% for negative interactions. We chose a relatively small set of simulated proteins and domains to reduce the time required for analysis. We ran an MCMC simulation for five million iterations that took  $\sim 5$  h of single-processor time on an IBM Regatta computer. [Figure 4](#) shows the set of simulated interactions among the 83 proteins of the simulated protein–protein interaction network, and the predicted interactions after the MCMC algorithm was run.

*Abstract*

*Introduction*

*Probabilistic Model*

*Network inference...*

*Three parts of the...*

**Results and Discussion**

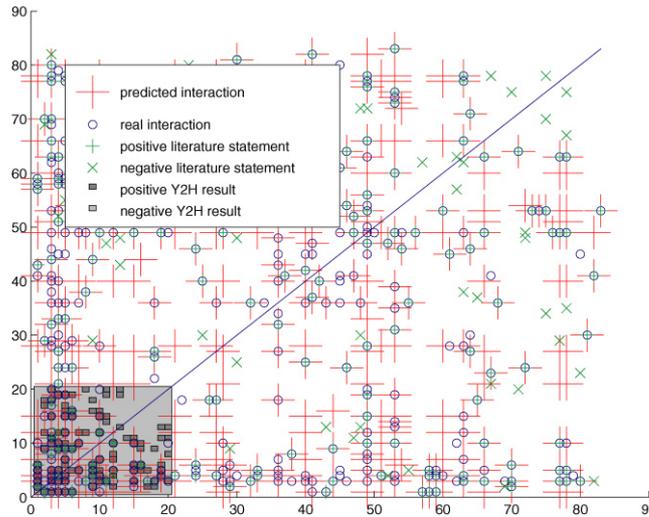
*Acknowledgements*

*References*



GO BACK

CLOSE FILE



**Fig. 4.** Real and predicted protein interactions for 83 proteins of a simulated protein–protein interaction network. Also shown: yeast two-hybrid and literature coverage of the same set of interactions.

After estimating the posterior probabilities of individual interactions, we evaluated the quality of the predictions both with respect to the whole set of interactions in our simulated real-world network and with respect to the predictable part of that set. Recall from the previous section that we can automatically recover more than 500 interactions not directly observed (based on domain information) once our system has seen 1000 interactions, some of which are incorrect (Fig. 3).

- Abstract**
- Introduction**
- Probabilistic Model**
- Network inference ...**
- Three parts of the ...**
- Results and Discussion**
- Acknowledgements**
- References**



Navigation controls:

- ◀ ▶
- ◀ ▶
- GO BACK**
- CLOSE FILE**

To measure the accuracy of our system's predictions, we ranked all predicted interactions such that the most likely positive interactions were situated at the beginning of the list, while the most certain negative interactions were at the end of the list. With this ranking, we were able to set a likelihood or rank threshold, treating the predictions above the threshold as positive interactions and the predictions below the threshold as negative interactions. This threshold allowed us to measure the specificity and sensitivity of our method by comparing its predictions with the known but not directly observable simulated real-world network interactions. Sensitivity (or recall) is defined as the percentage of true positives among true positives plus false negatives; specificity is the percentage of true negatives among true negatives plus false positives. Further, we varied the rank threshold, from the very beginning of the list, where specificity is 1 and sensitivity is 0, to the very end of the list, at which point sensitivity grows to 1 and specificity drops to 0. By varying the threshold, we computed the receiver-operator characteristic (ROC) curve, which plots sensitivity against specificity at different threshold values. We calculated an ROC score equal to 0.96 when we took into account all the interactions and an ROC score equal to 0.99 when we used only the interactions in the predictable area. These values indicate that our prediction method works very well—a powerless method has an ROC score close to 0.5.

The optimum cutoff between the positive and negative interactions is naturally the one that gives a number of positive predictions that is close

[Abstract](#)

[Introduction](#)

[Probabilistic Model](#)

[Network inference...](#)

[Three parts of the...](#)

**[Results and Discussion](#)**

[Acknowledgements](#)

[References](#)



**GO BACK**

**CLOSE FILE**

to the expected number of edges in a protein–protein interaction network. Since, we assumed that the protein–protein interaction network had a Zeta-distributed connectivity and we assumed the parameter of the Zeta-distribution to be known, we could easily compute the optimum number of positive predictions for a network of any given size. Using the optimum cutoff between false positive and negative predictions gave us a sensitivity of 0.782 and a specificity of 0.979 for the whole set of unobserved interactions in our simulated real-world network, and a sensitivity of 1 with a specificity 0.991 for the predictable part of that set.

We expect that the results would become less impressive if we introduced more noise into the simulation and into the information-extraction system; i.e., sensitivity and specificity would be expected to decrease as we increased the error rate in the simulated yeast two-hybrid system, raised the proportion of false statements published in the research literature and admitted imperfect information extraction by our automated system. Nevertheless, the volume of real data currently available is tremendous (and continues growing), and future implementation of our method will determine whether a large sample size will enable reliable parameter estimation even if the noise level is very high. For example, we estimated that currently there are at least one million full-text articles available on-line that may contain information about molecular interactions.

The performance of any protein-interaction prediction method depends to a large extent not only on the method's own merits but also on the

*Abstract*

*Introduction*

*Probabilistic Model*

*Network inference...*

*Three parts of the...*

**Results and Discussion**

*Acknowledgements*

*References*



**GO BACK**

**CLOSE FILE**

real-world properties of the networks that it is meant to simulate. For example, we can think of a world with only a few protein domain types occurring in proteins in roughly uniform fashion; these few domain types are recombined to produce a large number of proteins. If protein interactions are completely defined by interaction of domain types, as is assumed in our model, protein-interaction prediction methods would be extremely successful in this hypothetical universe. On the opposite extreme is an unlikely world where new types of protein domains evolve every time there is a need for new interacting proteins, which leads to an enormous number of domain types, each with a single domain copy. Obviously, assuming that protein interaction can be predicted only through knowledge of domain types, any method designed for interaction prediction would fail miserably in such a universe. In organisms found in nature, the real universe appears to lie somewhere in between these two extremes: the total number of domain types appears to be rather large (we currently know thousands of domain types) and it is clear that a lot of additional rare domain types are about to be described ([Bertone \*et al.\*, 2001](#)). Nevertheless, the frequencies of occurrence of these types in a proteome are extremely far from uniform, and so there are a few domain types with thousands of domain copies per proteome and a large number of domain types with a single copy per proteome. This simple consideration indicates that we should be able to derive a theoretical upper bound of performance of a ‘perfect’ interaction-prediction method.

**Abstract**

**Introduction**

**Probabilistic Model**

**Network inference...**

**Three parts of the...**

**Results and Discussion**

**Acknowledgements**

**References**



**GO BACK**

**CLOSE FILE**

Application of our model to real datasets will require a significant amount of computation which is a challenge on its own.

As a preliminary evaluation of a real dataset, we examined a yeast two-hybrid dataset generated for a subset of *Drosophila* proteins by CuraGen, Inc. (Giot *et al.*, 2003). This dataset comprises more than 20 000 experimental interactions among about 7000 proteins. Our current GeneWays database contains more than 1.5 million unique interactions extracted from 120 000 full research articles. These numerous relations between substances represent multiple organisms and are of various types (e.g. ‘activate’, ‘phosphorylate’ and ‘bind’). Out of 1.5 million interactions, approximately 10 000 interactions of type ‘bind’ among 2000 proteins can be unambiguously assigned to *Drosophila melanogaster*. Out of these 2000 proteins, about 1200 are common with the *Drosophila* yeast two-hybrid dataset. According to the current estimate, there are approximately 18 000 proteins in the *Drosophila* proteome (Adams *et al.*, 2000). Although the number of known (‘visible’) interactions seems small compared with the total number of all possible interactions  $[18\,000 * (18\,000 + 1)/2] \approx 162$  millions, we believe that the following assumption is reasonable. (Let us call a protein visible if there is at least one interaction or a lack of interaction reported for it.) We assume that every non-reported interaction between two ‘visible’ proteins is negative. The number of visible interactions is around 25 million, which is 15% of the whole interaction

**Abstract**

**Introduction**

**Probabilistic Model**

**Network inference...**

**Three parts of the...**

**Results and Discussion**

**Acknowledgements**

**References**



**GO BACK**

**CLOSE FILE**

space (162 millions). Therefore, using the predictable area curve (Fig. 3), we estimate that we might be able to predict the existence or absence of interactions for about 10% of the pairs of ‘invisible’ proteins—i.e. for about 13.7 million pairs.

*Abstract*

*Introduction*

*Probabilistic Model*

*Network inference...*

*Three parts of the...*

***Results and Discussion***

*Acknowledgements*

*References*



**GO BACK**

**CLOSE FILE**

## Acknowledgements

This study was supported by grants from the National Institute of Health (HG00045) to K.P.W. and from the National Science Foundation (EIA-0121687), National Institutes of Health (GM61372) and US Department of Energy (DE-FG02-01ER25500) to A.R.

*Abstract*

*Introduction*

*Probabilistic Model*

*Network inference...*

*Three parts of the...*

*Results and Discussion*

**Acknowledgements**

*References*



GO BACK

CLOSE FILE

## References

- Adams,M.D., Celniker,S.E., Holt,R.A., Evans,C.A., Gocayne,J.D., Amanatides,P.G., Scherer,S.E., Li,P.W., Hoskins,R.A., Galle,R.F. *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **287**, 2185–2195.
- Albert,R. and Barabasi,A.L. (2000) Topology of evolving networks: local events and universality. *Phys. Rev. Lett.*, **85**, 5234–5237. [MEDLINE Abstract](#)
- Albert,R., Jeong,H. and Barabasi,A.L. (2000) Error and attack tolerance of complex networks. *Nature*, **406**, 378–382. [MEDLINE Abstract](#)
- Bader,G.D. and Hogue,C.W. (2002) Analyzing yeast protein–protein interaction data obtained from different sources. *Nat. Biotechnol.* **20**, 991–997. [MEDLINE Abstract](#)
- Barabasi,A.L. and Albert,R. (1999) Emergence of scaling in random networks. *Science*, **286**, 509–512. [MEDLINE Abstract](#)
- Barabasi,A., Albert,R. and Jeong,H. (2000) Scale-free characteristics of random networks: the topology of the world-wide web. *Physica A*, **281**, 69–77.
- Bertone,P., Kluger,Y., Lan,N., Zheng,D., Christendat,D., Yee,A., Edwards,A.M., Arrowsmith,C.H., Montelione,G.T. and Gerstein,M. (2001) Spine: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. *Nucleic Acids Res.*, **29**, 2884–2898. [MEDLINE Abstract](#)

[Abstract](#)

[Introduction](#)

[Probabilistic Model](#)

[Network inference...](#)

[Three parts of the...](#)

[Results and Discussion](#)

[Acknowledgements](#)

[References](#)



GO BACK

CLOSE FILE

- Bock,J.R. and Gough,D.A. (2001) Predicting protein–protein interactions from primary structure. *Bioinformatics*, **17**, 455–460. [MEDLINE Abstract](#)
- Davy,A., Bello,P., Thierry-Mieg,N., Vaglio,P., Hitti,J., Doucette-Stamm,L., Thierry-Mieg,D., Reboul,J., Boulton,S., Walhout,A.J., Coux,O. and Vidal,M. (2001) A protein–protein interaction map of the *Caenorhabditis elegans* 26S proteasome. *EMBO Rep.*, **2**, 821–828. [MEDLINE Abstract](#)
- Deng,M., Mehta,S., Sun,F. and Chen,T. (2002) Inferring domain–domain interactions from protein–protein interactions. *Genome Res.*, **12**, 1540–1598. [MEDLINE Abstract](#)
- Friedman,C., Kra,P., Yu,H., Krauthammer,M. and Rzhetsky,A. (2001) Genies: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, **17**(Suppl. 1), S74–S82. [MEDLINE Abstract](#)
- Gietz,R.D. and Woods,R.A. (2002) Screening for protein–protein interactions in the yeast two-hybrid system. *Methods Mol. Biol.*, **185**, 471–486. [MEDLINE Abstract](#)
- Giot,L., Bader,J.S., Brouwer,C., Chaudhuri,A., Kuang,B., Li,Y., Hao,Y.L., Ooi,C.E., Godwin,B., Vitols,E. *et al.* (2003) A protein interaction map of *Drosophila melanogaster*. *Science*, 1727–1736, Epub 2003 Nov 06. [MEDLINE Abstract](#)

[Abstract](#)

[Introduction](#)

[Probabilistic Model](#)

[Network inference...](#)

[Three parts of the...](#)

[Results and Discussion](#)

[Acknowledgements](#)

[References](#)



GO BACK

CLOSE FILE

- Gilks, W., Richardson, S. and Spiegelhalter, D. (eds) (1996) *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC, New York.
- Goldberg, D.S. and Roth, F.P. (2003) Assessing experimentally derived interactions in a small world. *Proc. Natl Acad. Sci., USA*, **100**, 4372–4376. [MEDLINE Abstract](#)
- Gomez, S. and Rzhetsky, A. (2002) Towards prediction of complete protein–protein interaction networks. *Pac. Symp. Biocomput.*, 413–424. [MEDLINE Abstract](#)
- Gomez, S.M., Lo, S.H. and Rzhetsky, A. (2001) Probabilistic prediction of unknown metabolic and signal-transduction networks. *Genetics*, **159**, 1291–1298. [MEDLINE Abstract](#)
- Hastings, W. (1970) Monte carlo sampling methods using markov chains and their applications. *Biometrika*, **57**, 97–109.
- Hatzivassiloglou, V., Duboue, P.A. and Rzhetsky, A. (2001) Disambiguating proteins, genes, and RNA in text: a machine learning approach. *Bioinformatics*, **17**(Suppl. 1), S97–S106. [MEDLINE Abstract](#)
- Jeong, H., Mason, S.P., Barabasi, A.L. and Oltvai, Z.N. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42. [MEDLINE Abstract](#)
- Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. and Barabasi, A.L. (2000) The large-scale organization of metabolic networks. *Nature*, **407**, 651–654. [MEDLINE Abstract](#)

[Abstract](#)

[Introduction](#)

[Probabilistic Model](#)

[Network inference...](#)

[Three parts of the...](#)

[Results and Discussion](#)

[Acknowledgements](#)

[References](#)



GO BACK

CLOSE FILE

- Johnson,N.L. and Kotz,S. (1969) *Discrete Distributions. Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics.* Wiley, New York. pp. 319–323.
- Johnson,N.L. and Kotz,S. (1970) *Continuous Univariate Distributions. Houghton Mifflin Series in Statistics.* Houghton Mifflin, New York.
- Koike,T. and Rzhetsky,A. (2000) A graphic editor for analyzing signal-transduction pathways. *Gene*, **259**, 235–244. [MEDLINE Abstract](#)
- Koonin,E.V., Wolf,Y.I. and Karev,G.P. (2002) The structure of the protein universe and genome evolution. *Nature*, **420**, 218–223. [MEDLINE Abstract](#)
- Krauthammer,M., Kra,P., Iossifov,I., Gomez,S.M., Hripsak,G., Hatzivassiloglou,V., Friedman,C. and Rzhetsky,A. (2002) Of truth and pathways: chasing bits of information through myriads of articles. *Bioinformatics*, **18**(Suppl. 1), S249–S257. [MEDLINE Abstract](#)
- Marcotte,E.M., Pellegrini,M., Ng,H.L., Rice,D.W., Yeates,T.O. and Eisenberg,D. (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science*, **285**, 751–753. [MEDLINE Abstract](#)
- Metropolis,S., Rosenbluth,A., Rosenbluth,M., Teller,A. and Teller,E. (1953) Equation of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087–1092.

[Abstract](#)

[Introduction](#)

[Probabilistic Model](#)

[Network inference...](#)

[Three parts of the...](#)

[Results and Discussion](#)

[Acknowledgements](#)

[References](#)



GO BACK

CLOSE FILE

Park,J., Lappe,M. and Teichmann,A. (2001) Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. *J. Mol. Biol.*, **307**, 929–938.

[MEDLINE Abstract](#)

Rzhetsky,A. and Gomez,S.M. (2001) Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome. *Bioinformatics*, **17**, 988–996.

[MEDLINE Abstract](#)

Rzhetsky,A., Koike,T., Kalachikov,S., Gomez,S., Krauthammer,M., Kaplan,S., Kra,P., Russo,J. and Friedman,C. (2000) A knowledge model for analysis and simulation of regulatory networks. *Bioinformatics*, **16**, 1120–1128.

[MEDLINE Abstract](#)

Sprinzak,E. and Margalit,H. (2001) Correlated sequence-signatures as markers of protein–protein interaction. *J. Mol. Biol.*, **311**, 681–692.

[MEDLINE Abstract](#)

Tong,A.H., Drees,B., Nardelli,G., Bader,G.D., Brannetti,B., Castagnoli,L., Evangelista,M., Ferracuti,S., Nelson,B., Paoluzi,S. *et al.* (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science*, **295**, 321–324.

[MEDLINE Abstract](#)

**Abstract**

**Introduction**

**Probabilistic Model**

**Network inference...**

**Three parts of the...**

**Results and Discussion**

**Acknowledgements**

**References**



**GO BACK**

**CLOSE FILE**