Chapter 8
Probabilistic Graphical Modeling in Systems Biology: A Framework for Integrative Approaches

Christine Sinoquet

Abstract Systems biology may be defined as a discipline aiming at integrating various sources of heterogeneous data, with the objective to describe and predict the function of biological systems. The purpose is to cross many (possibly weak) evidences from several data types that describe different biological features of genes or proteins. Probabilistic graphical models offer an appealing framework for this objective. Through the thorough review of five selected examples, this chapter highlights how probabilistic graphical models can contribute to build the bridge between biology and computational modeling. In this methodological framework, the five cases illustrate three features of these models, which we discuss: flexibility, scalability and ability to combine heterogeneous sources of data. The applications covered address genetic association studies, identification of protein–protein interactions, identification of the target genes of transcription factors, inference of causal phenotype networks and protein function prediction.

Keywords Systems biology · Integrative approach · Integration of omics data · Heterogeneous sources of data · Computational modeling · Machine learning · Probabilistic framework · Probabilistic graphical model · Bayesian network · Markov random field

List of Acronyms

BN	Bayesian network
ChIP-chip	Chromatin immunoprecipitation on chip
ChIP-seq	Chromatin immunoprecipitation followed by sequencing
CPN	Causal phenotype network
DDI	Domain-domain interaction
DNA	Deoxyribonucleic acid

C. Sinoquet (✉)
LINA, UMR CNRS 6241, Université de Nantes, 2 rue de la Houssinière,
BP 92208 44322 Nantes Cedex, France
e-mail: christine.sinoquet@univ-nantes.fr

A. Prokop and B. Csukás (eds.), *Systems Biology*,
DOI: 10.1007/978-94-007-6803-1_8,

GA	Genetic architecture
GO	Gene ontology
GOS	GO sub-ontology
GWAS	Genome wide association study
MCMC	Monte Carlo Markov chain
MRF	Markov random field
MRF-MJM	MRF mixture joint model
PGM	Probabilistic graphical model
PPI	Protein–protein interaction
QTL	Quantitative trait loci
RNA	Ribonucleic acid
RNAi	RNA interference
ROC curve	Receiver operating characteristic curve
SMM	Standard mixture model
TF	Transcription factor

8.1 Introduction

In the machine learning domain, probabilistic graphical models provide a unified framework to both represent dependences between variables and model uncertain knowledge about the quantitative dependences between these variables. In the post-genomic era, the provision of voluminous and complex heterogeneous data by high-throughput omics technologies has brought increased attention to these models. Notably, their flexibility, scalability and ability to combine heterogeneous sources of data are expected to enhance the gain in biological and biomedical discoveries. Data integration is intended to make useful connections that could lead to novel biological knowledge.

Besides, if there is one area where transdisciplinarity is the daily lot, designing new computational methods based on advanced models devoted to applications in systems biology is this area. A constructive cooperation with a domain specialist requires ability to hold productive dialogue, which therefore demands a good understanding of the models by the non expert. Bridging the gap between biology, statistics and computer science is a condition to achieve progress in systems biology. Albeit dedicated to specific applications, the five models presented in this chapter remain general enough to help foster reflections about addressing other applications in systems biology, in an integrative framework.

Methods based on probabilistic graphical models (PGMs) may be complex and thus might be disconcerting to scientists non familiar with them, which is likely to hamper the dissemination of such methods. Thus there was a challenge in attempting to demystify the concepts and mechanisms behind such models in the

perspective of using them for systems biology. To this aim, this chapter was conceived as a thorough review of five illustrative approaches of the use of probabilistic graphical modeling as an integrative framework in systems biology. We first provide an intuitive presentation of the concept of conditional independence, which is the fundamental principle PGMs all rely on; then we introduce the Bayesian networks and the Markov random fields, which are the two classes of PGMs addressed in this chapter. Subsequently we present the five illustrations selected to cover major application fields in systems biology: (1) enhancement of genome-wide association studies with knowledge on biological pathways, (2) identification of protein–protein interactions, (3) identification of the regulatory target genes of transcription factors, (4) inference of causal relationships among phenotypes *via* integration of QTL genotypes, (5) prediction of protein function through ontology-enriched networks connecting multiple related species. A brief insight about the performance of each method is provided on the fly. We conclude this chapter highlighting the pros and cons of this modeling framework, when used for integration purpose in systems biology and we indicate some directions for future work.

The order of presentation for the contributions is not incidental: it puts forward an increasing gradient in the heterogeneity of the data sources integrated in the probabilistic framework. For example, approaches (2) and (3) both integrate information coming from gene ontologies but such information is used similarly to that coming from the other data sources. In contrast, accounting for this ontological knowledge thoroughly impacts the probabilistic inference scheme in the last approach. At the opposite extremity of the data integration spectrum, it is worth mentioning that PGMs provide the ability to integrate *meta-knowledge* about a *single* data source, at the genome-wide scale. An enlightening example is the modeling of genetic data, where the so-called linkage disequilibrium encompasses short-range, long-range and chromosome-wide dependences within these data [24–25]. Such meta-knowledge integration in a genetic association study aims at enhancing power and accuracy in identifying the causal factors of a disease [20, 35, 37]. In this book chapter, we focus on the integration of *multiple* data sources.

8.2 Preliminaries

In the present section, the concepts indispensable for further understanding are introduced in an intuitive manner. Besides, we highlight why probabilistic graphical models are appealing to model biological data in an integrative framework. Within the scope of this section, we suppose that the data available are as follows: p data samples are each described by n variables $X = \{X_1, \ldots, X_n\}$. In a general probabilistic framework, computing the joint probability distribution for large data is generally not tractable as, by virtue of the so-called product-rule, the only formalization applicable is

$$\mathbb{P}(X) = \mathbb{P}(X_1)\ \mathbb{P}(X_2 \mid X_1)\ \mathbb{P}(X_3 \mid X_1,\ X_2)\ \ldots\ \mathbb{P}(X_n \mid X_1,\ X_2,\ \ldots,\ X_{n-1}). \quad (8.1)$$

If we denote by x_i a value in the domain of the possible values for the random variable X_i, it should be noted that computing the probability distribution $\mathbb{P}(X)$ means that for any possible instantiation $x = (x_1, x_2,\ \ldots,\ x_n)$ of X, we know how to calculate $\mathbb{P}(x)$. It should also be kept in mind that from now on, symbols in lower cases will denote values taken by the variables. Third, the expression "probability distribution" will be reserved to discrete random variables whereas the expression "density probability" will be used for continuous random variables. In the above formula, $X_1, \ldots, X_r$, to be understood as the *event* $(X_1 = x_1,\ \ldots,\ X_r = x_r)$, denotes the joint observation of values $x_1,\ \ldots,\ x_r$. The product rule in Eq. 8.1 involves *conditional probabilities*.

The conditional probability[1] of event D_1 given event D_2, $\mathbb{P}(D_1 \mid D_2)$, is the probability of D_1 with the additional information that D_2 has already occurred. It is defined as:

$$\mathbb{P}(D_1 \mid D_2) = \frac{\mathbb{P}(D_1, D_2)}{\mathbb{P}(D_2)},\ \text{with}\ \mathbb{P}(D_2) \neq 0.$$

For instance, if D_1 and D_2 are two diseases, such that D_2 is observed with probability 0.05, and D_1 and D_2 are simultaneously observed with probability 0.001, then the onset probability for D_1, when D_2 is present, is 0.02.

Probabilistic graphical models are appealing models because they rely on *conditional independence*, to offer the immense advantage of a factorized formulation of probability distributions. Let us first introduce the concept of conditional independence. In the above example, suppose we calculate that the *prior probability* $\mathbb{P}(D_1)$ is equal to the *posterior probability* $\mathbb{P}(D_1 \mid D_2)$. Intuitively, this means that knowing whether D_2 occurs ($\mathbb{P}(D_1 \mid D_2)$) does not refine our knowledge about whether D_1 occurs. The diseases D_1 and D_2 are therefore *independent*: $D_1 \perp\!\!\!\perp D_2$. Conditional independence is a little bit more complex:

Definition 1 (*Conditional independence*) Given three variables A, B and C, conditional independence between A and B given the state of C ($A \perp\!\!\!\perp B \mid C$) is defined as: $\mathbb{P}(A \mid B,\ C) = \mathbb{P}(A \mid C)$ (with $\mathbb{P}(C) > 0$). The concept of conditional independence given a unique variable is easily extended to conditional independence given a set of variables.

Intuitively, A and B are conditionally independent given C ($A \perp\!\!\!\perp B \mid C$) if and only if, given any value of C, the probability distribution of A remains the same for all values of B: $\mathbb{P}(A \mid B = b_1,\ C = c) = \mathbb{P}(A \mid B = b_2,\ C = c) = \mathbb{P}(A \mid C = c)$. Suppose now that a third variable E measures the effects of the disease D_1, and that these effects cause the disease D_2 (symbolized through $D_1 \rightarrow E \rightarrow D_2$); indoubtedly, D_1 and D_2 are dependent; however, D_1 and D_2 are conditionally independent

[1] Depending on the context, the *conditional probability* of D_1 given D_2, $\mathbb{P}(D_1 \mid D_2)$, is also called the *posterior probability* of D_1 conditional on D_2.

given E (see Table 8.1). Intuitively, this means that the status of D_1 can be inferred from the status of E. Therefore, when dependences exist within data, conditional independence shields a given variable from the remaining variables, given some set of variables. Biological data are often described by a network, if not several networks, in the integrative framework. The conditional independence property is known as the *Markov property*. The Markov property is the corner stone for simplifying probability distributions, thus directly achieving tractability or making easier approximations to further obtain tractability. Intrinsically, all five models illustrated in this chapter rely on the Markov property, to infer knowledge from one or several biological networks.

Probabilistic graphical models (PGMs) provide a powerful framework for representing and reasoning with uncertainty and dependences. The qualitative part of a PGM is a graph $\mathcal{G}$ that encodes dependences (and independences) between the variables, represented by nodes in the graph. Uncertain knowledge about the qualitative dependences between the variables is formalized with the aid of probability distributions. Besides differences in their graphs, we now briefly show the variants of Markov property for the two kinds of probabilistic graphical model (PGMs) addressed in this chapter. One of the most popular kinds of PGMs is the Bayesian network (BN).

Table 8.1 Conditional independence of two variables D_1 and D_2 given a third-variable E

		D_2E	$D_2\bar{E}$	$\bar{D}_2E$	$\bar{D}_2\bar{E}$
D_1	0	120	40	40	20
	1	180	160	60	80

(a) Counts

		D_2E	$D_2\bar{E}$	$\bar{D}_2E$	$\bar{D}_2\bar{E}$
D_1	0	0.171	0.057	0.057	0.029
	1	0.257	0.229	0.086	0.114

(b) Joint distribution $\mathbb{P}(D1, D2, E)$

		D_2	
		0	1
D_1	1	0.086	0.229
	0	0.200	0.485

(c) Marginal distribution $\mathbb{P}(D1, D2)$

Conditional distributions

		E = 0 D_2	
		0	1
D_1	1	0.2	0.2
	0	0.8	0.8

(d) $\mathbb{P}(D1 \mid D2 = i, \mathbf{E = 0})$

		E = 1 D_2	
		0	1
D_1	1	0.4	0.4
	0	0.6	0.6

(e) $\mathbb{P}(D1 \mid D2 = i, \mathbf{E = 1})$

c Marginal probabilities are obtained through summing ("marginalizing") probabilities over the domain of E; $\mathbb{P}(D_1 \mid D_2) = \frac{\mathbb{P}(D_1, D_2)}{\mathbb{P}(D_2)} = \frac{0.485}{0.679} = 0.714 \neq \mathbb{P}(D_1) = 0.685$, thus D_1 and D_2 are dependent variables. **d** and **e** D_1 and D_2 are conditionally independent given E ($D_1 \perp\!\!\!\perp D_2 \mid E$) since the columns are identical within each table.

Definition 2 (*Bayesian network*) In a BN, the qualitative component is a directed acyclic graph (acyclic because no directed path $X_{i_1} \rightarrow X_{i_2} \rightarrow \cdots \rightarrow X_{i_r}$, where $i_1 = i_r$, is allowed). Conditional distributions are defined for each variable X_i: $\theta_i = [\mathbb{P}(X_i / Pa_{X_i})]$ where Pa_{X_i} denotes node i's parents. The local Markov property states that each variable is conditionally independent of its non-descendants given a known state of its parent variables: $X_i \perp\!\!\!\perp X \setminus desc(X_i) \mid Pa_{X_i}$, where the notation $X \setminus Y$ stands for the set $\{X_i \in X \text{ and } X_i \notin Y\}$ and $desc(X_i)$ is the set of descendants of X_i. The local Markov property entails that the joint distribution writes as a product of local distributions conditional on the parent variables:

$$\mathbb{P}(X) = \prod_{i \in \{1,\ldots,n\}} \theta_i.$$

Figure 8.1a shows a Bayesian network. Another widely used model is the Markov random field.

Definition 3 (*Markov random field*) In Markov random fields (MRFs), the qualitative component $\mathcal{G}$ is an undirected graph which may have cycles (that is (undirected) cycles $X_{i_1} - X_{i_2} - \cdots - X_{i_r}$, where $i_1 = i_r$, are allowed). The joint distribution is factorized over cliques "covering" the set X. A clique is defined by any set of pairwise connected nodes, such as $\{X_1, X_2\}$ or $\{X_1, X_2, X_4\}$ in Fig. 8.1b. A set of random variables X is an MRF if there exist so-called function potentials such that the joint distribution writes:

$$\mathbb{P}(X = x) = \frac{1}{Z} b(x)$$
$$\mathbb{P}(X = x) = \frac{1}{Z} \prod_{C \in cliques(\mathcal{G})} \varphi_C(x_C).$$

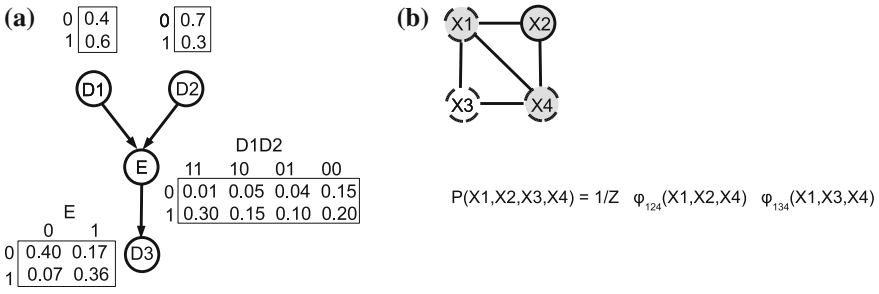

Fig. 8.1 Probabilistic graphical models. **a** Bayesian network. **b** Markov random field. **a** The prior probability distributions $\mathbb{P}(D_1)$ and $\mathbb{P}(D_2)$, and the conditional distributions $\mathbb{P}(E \mid D_1, D_2)$ and $\mathbb{P}(D_3 \mid E)$ are shown. The node E has two parents (D_1 and D_2). The node D_1 has one child (E) and two descendants (E and D_3). **b** The factorization of the joint distribution $\mathbb{P}(X_1, X_2, X_3, X_4)$ involves the potentials relative to the two cliques (X_1, X_2, X_4) and (X_1, X_3, X_4). The node X_1 has two neighbors: X_2 and X_3

There, x_C denotes some possible instantiation for the variables encompassed by clique C. Function φ_C is called a clique potential. Z is the normalizing function used to ensure that $\mathbb{P}$ be a probability distribution ($Z = \sum_x b(x)$ guarantees that $\sum_x \mathbb{P}(X = x) = 1$). In the case of the MRF, the local Markov property states that a variable is conditionally independent of all other variables given its set of neighbours N_i: $\mathbb{P}(X_i \mid X_{-i}) = \mathbb{P}(X_i \mid N_i)$, where X_{-i} designates the set X deprived of variable X_i.

Figure 8.1b shows a Markov random field. In particular, this chapter will refer to pairwise MRFs, which consider cliques of size 2 and whose joint distribution writes:

$$\mathbb{P}(X = x) = \frac{1}{Z} \prod_{i=1}^{n} \varphi_i(x_i) \prod_{(i,j)\in\mathscr{G}} \varphi_{i,j}(x_i,\ x_j). \tag{8.2}$$

Finally, we recall some additional notions to non specialists. Given a model M and the observed data D, according to Bayes theorem,[2] the relation between *posterior distribution*, *prior distribution* and *likelihood* writes: $\mathbb{P}(M \mid D) \propto \mathbb{P}(D \mid M)\ \mathbb{P}(M)$. The proportionality is explained by the fact that the probability to observe the data, $\mathbb{P}(D)$, is a constant. Model learning consists in evaluating how a candidate M fits the data D. Maximizing the likelihood $\mathbb{P}(D \mid M)$ is a standard procedure to achieve this purpose. Due to additional knowledge (D), the prior distribution $\mathbb{P}(M)$ is refined into the posterior distribution $\mathbb{P}(M \mid D)$. The reader is also reminded that $\mathscr{U}(a, b)$ designates the uniform probability distribution over interval $[a, b]$ and that $\mathscr{N}(\mu, \sigma^2)$ represents the normal (or Gaussian) probability distribution with mean μ and variance σ^2. The multivariate normal distribution is a generalization of the latter distribution to higher dimensions; then the normal distribution is summarized by a mean vector and a covariance matrix. To denote that a random variable A follows a given distribution, say $\mathscr{N}(\mu, \sigma^2)$, we will write: $A \sim \mathscr{N}(\mu, \sigma^2)$.

8.3 Enhancement of Genome-Wide Association Studies with Knowledge on Biological Pathways

To decipher the genetic causes of diseases, genome-wide association studies (GWASs) compare the genomes of affected people to those of unaffected. The aim is to identify associations between genetic variants and the disease. GWASs pose a formidable challenge since most of the time the effects from individual genetic variants are weak and the sample size is not large enough to guarantee sufficient power. To overcome this issue, various strategies have been proposed. Multilocus

[2] $\mathbb{P}(M \mid D)\ \mathbb{P}(D) = \mathbb{P}(D \mid M)\ \mathbb{P}(M)$.

association tests benefit from linkage disequilibrium—that is dependences existing within genetic data—by considering sets of correlated markers instead of single markers. An alternative lead lies in integrating evidences from external data sources, in the single locus approach. Various approaches based on the integration of prior biological knowledge were designed to prioritize candidate disease genes (see [16] for a survey). In GWASs, evidence from the gene level is recognized as the most promising. In particular, incorporating prior biological knowledge about pathways has a role to play [31, 40]: as genes interact with each other in biological pathways, they are likely to jointly affect disease susceptibility. However, so far, no GWAS approach had taken into account knowledge about *regulatory relationships* between genes of a given pathway. Not surprisingly, in this domain, the pioneering approach of Chen and collaborators takes full advantage of probabilistic graphical modeling [5].

In the following, we denote $S = \{S_1, \ldots, S_n\}$ the set of gene labels to be predicted based on the observed association data and the knowledge on the pathway topology. $S_i = 1$ states that gene i is associated with the disease; otherwise, the label is $S_i = 0$. Typically, the association data are p-values $P_1, \ldots, P_n$ resulting from n single-locus association tests. Usually, given some significant threshold P^*, $P_i < P^*$ (respectively $P_i \geq P^*$) indicates that S_i should be set to 1 (respectively 0). The probabilistic framework adopted by Chen and collaborators aims at improving the reliability in predicting the labels: the ultimate goal is thus to estimate the posterior distribution of S conditional on the data P, that is $\mathbb{P}(S \mid P)$. By virtue of the Bayes theorem, $\mathbb{P}(S \mid P) \propto \mathbb{P}(S)\ \mathbb{P}(P \mid S)$. The key to the prediction improvement by Chen et al. lies in the integration of knowledge on the pathway topology in the model: such knowledge is incorporated in the prior distribution $\mathbb{P}(S)$.

8.3.1 Exploiting Knowledge from the Gene Pathway

In the following, N_i denotes the set of the n_i neighbors of gene i in the pathway of concern; $\mathscr{G}$ denotes the pathway topology. To capture the idea that two neighbor genes i and j tend to share a common association status ($S_i = S_j$), Chen et al. adjust a nearest neighbor Gibbs measure [15] as follows:

$$\begin{aligned}\mathbb{P}(S = s \mid \theta_0) = &\frac{1}{Z} exp \\ &[h + \sum_i I_1(S_i) + \tau_0 \sum_{(i,j)\in\mathscr{G}} (w_i + w_j)\ I_0(S_i)\ I_0(S_j) \\ &+ \tau_1 \sum_{(i,j)\in\mathscr{G}} (w_i + w_j)\ I_1(S_i)\ I_1(S_j)].\end{aligned} \tag{8.3}$$

The symbol $s = (s_1, \ldots, s_n)$ denotes one label assignment (amongst the 2^n possible assignments), for instance $(0, 1, 1, \ldots, 1, 0)$. $\theta = (h, \tau_0, \tau_1)$ denotes hyperparameters fixed by the user. I_0 and I_1 are indicator functions, meaning that $I_1(S_i) = 1$

if $S_i = 1$ and $I_1(S_i) = 0$ otherwise, and, symmetrically $I_0(S_i) = 1$ if $S_i = 0$ and $I_0(S_i) = 0$ otherwise. Finally, the model in Eq. 8.3 also reflects the fact that genes showing many interactions in a pathway are likely to play a prominent role in a biological process; thus they are likely to exert a large influence. Consequently, weights w_is are incorporated in the model, that depend on the neighborhood sizes: $w_i = \sqrt{n_i}$ is an increasing function of the number n_i of neighbors of gene i.

Equation 8.3 formalizes the joint probability for S so that genes connected with each other tend to have the same labels, that is the same association status. The third term concerns all edges connecting neighbors sharing the common label 0. The fourth term concerns neighbors that share the label 1. Besides, τ_0 and τ_1 assign weights to such edges, depending on the shared labels. Positive parameters τ_0 and τ_1 will favor assignments s of S in which neighbor genes share the same label.

A property of nearest neighbor Gibbs measures is that they always define a Markov random field. In this case, the conditional independence assumption entails: $\mathbb{P}(S_i \mid S_{-i},\ \theta_0) = \mathbb{P}(S_i \mid S_{N_i},\ \theta_0)$, where we recall that $S_{-i} = (S_1,\ \ldots,\ S_{i-1},\ S_{i+1},\ \ldots,\ S_n)$. Besides, using Eq. 8.3, Chen et al. show that the conditional distribution $\mathbb{P}(S \mid S_{N_i},\ \theta_0)$ has a logistic regression form. A standard linear regression model is not convenient to represent a binary (0/1) variable B as $B = a_0 + a_1\ A_1 + a_2\ A_2 + \cdots + a_k\ A_k$, since the predictors A_i are unconstrained. Instead, one deals with $p = \mathbb{P}(B = 1)\ \in [0, 1]$ and a logit transformation is therefore required to apply a linear regression model to $logit(p) = log\ \left(\frac{p}{1-p}\right) \in]-\infty,\ +\infty[$. In the case illustrated, the logistic form is:

$$logit(\mathbb{P}(S_i \mid S_{N_i}, \theta_0)) = h + \tau_1 \left(w_i\ J_i^1 + \sum_{k \in N_i} w_k\ I_1(S_k) \right) - \tau_0 \left(w_i\ J_i^0 + \sum_{k \in N_i} w_k\ I_0(S_k) \right), \tag{8.4}$$

where $J_i^0 = \sum_{k \in N_i} I_0(S_k)$ and, similarly, $J_i^1 = \sum_{k \in N_i} I_1(S_k)$.

In the configuration where τ_0 and τ_1 are both null, all genes are interpreted as independent; the so-called intercept h then determines the posterior probability $\mathbb{P}(S_i \mid h, \tau_0 = \tau_1 = 0) = \frac{exp(h)}{1+exp(h)}$.

To recapitulate, the prior acknowledging for the pathway topology is the conditional distribution $\mathbb{P}(S \mid S_{N_i},\ \theta_0)$. This prior has the logistic regression form:

$$\begin{aligned}
&logit(\mathbb{P}(S_i \mid S_{N_i}, \theta_0)) = \beta_{i0} + \beta_{i1}\ S_1 + \cdots + \beta_{in}\ S_n \\
&\text{with} \\
&\beta_{i0} = h \\
&\beta_{ij} = 0\ \textit{if}\ i = j\ \textit{or}\ j \notin\ N_i \\
&\beta_{ij} = (w_i + w_j)\ (\tau_1\ I_1(S_j) - \tau_0\ I_0(S_j))\ \textit{otherwise}.
\end{aligned}$$

In the following, for concision, we will omit the references to S_{N_i} and θ_0 and the joint prior distribution will merely be denoted $\mathbb{P}(S)$ as in the end of the introductory paragraph of Sect. 8.3.

8.3.2 Posterior Distribution of Association Status

The posterior distribution integrates the knowledge about the pathway topology (from the prior) and the evidence from the observed association data (i.e. the p-values):

$$\mathbb{P}(S \mid P) \propto \mathbb{P}(S)\ \mathbb{P}(P \mid S). \tag{8.5}$$

A model remains to be defined for $\mathbb{P}(S \mid P)$. Chen and collaborators model instead $\mathbb{P}(S \mid Y)$, where any p-value P_i is converted into $Y_i = \Phi^{-1}(1 - P_i/2)$. Therein, Φ is the cumulative distribution function of $\mathcal{N}(0, 1)$. The justification for this conversion is simplification in further algebraic derivations. Then, the joint density of Y readily writes:

$$f(Y \mid S) = \prod_{i:S_i=0} f_0(Y_i) \prod_{i:S_i=1} f_1(Y_i),$$

where f_0 and f_1 respectively denote the distributions of Y_i under the null hypothesis and the hypothesis of association, that is: $f_0(Y_i) = \mathbb{P}(Y_i \mid S_i = 0)$ and $f_1(Y_i) = \mathbb{P}(Y_i \mid S_i = 1)$. Under the null hypothesis (no association, $S_i = 0$), any value in $[0, 1]$ is acceptable for the p-value (probability) P_i. P_i is therefore modeled to follow the uniform distribution $\mathcal{U}(0, 1)$. This setting entails that $f_0(Y_i)$ follows the Gaussian distribution $\mathcal{N}(0, 1)$. On the other hand, the unknown distribution of Y_i under the hypothesis of association is assumed to follow a Gaussian distribution: $f_1(Y_i) \sim \mathcal{N}(\mu_i, \sigma_i^2)$.

Under these settings, the algebraic derivation of the posterior distribution (see Eq. 8.5)

$$\mathbb{P}(S \mid Y) \propto \mathbb{P}(S)\mathbb{P}(Y \mid S)$$

shows that, similarly to the prior $\mathbb{P}(S)$, the posterior distribution $\mathbb{P}(S \mid Y)$ has a logistic regression form. The regression forms are identical in all points except for the intercept, which is now $h + log\ LR(Y_i)$ where $LR(Y_i) = \frac{f_1(Y_i)}{f_0(Y_i)}$ is the usual likelihood ratio. Importantly, the conditional independence assumption of the prior distribution is kept: $\mathbb{P}(S_i \mid S_{-i},\ \ldots) = \mathbb{P}(S_i \mid S_{N_i},\ \ldots)$.

Finally, the assignment of labels to the genes is performed by running an MCMC. The MCMC starts from some initial value $s^{(0)}$ assigned (at random) to S. Then step k sequentially updates the labels of the genes according to the following scheme:

$$logit(\mathbb{P}(s_i^{(k)} \mid Y, s_1^{(k)}, \ldots, s_{i-1}^{(k)}, s_{i+1}^{(k-1)}, \ldots, s_n^{(k-1)}) = \beta'_{i0} + \beta'_{i1}s_1 + \cdots + \beta'_{in}s_n.$$

An important point is that the conditional independence assumption in Eq. 8.4 holds for the posterior distribution, which is therefore also a Markov random field. The practical consequence is that the computation involved in the sampling of s_i only requires values s_j where j belongs to the neighborhood N_i: otherwise, the β'_{ij} coefficient is null if genes i and j are not neighbors in the pathway.

8.3.3 Performances

Incorporating prior biological knowledge to enhance GWASs is not new (see for instance Prioritizer [10], CANDID [13], CIPHER [42]). However, such approaches do not consider the functional relationships existing among genes. In contrast, the approach of Chen et al. takes advantage of knowledge on known associations to infer novel association knowledge on other genes related to the former through pathways.

The relevance of the model of Chen and collaborators was supported by a preliminary study. These authors considered 3,735 genes over 350 pathways. On the other hand, association results from a GWAS on Crohn's disease were available. In each of the pathways, the number of edges N_{++} with both extremities associated with the disease was observed. Over the 350 pathways, an overwhelming proportion of counts N_{++} showed exceptionally large values. This clearly confirms the hypothesis: in a given pathway, most often, genes that are associated with the disease are neighbors.

The approach of Chen et al. was then evaluated based on 289 pathways and GWAS data relative to Crohn's disease. Thirty-two genes associated with the disease were known (target genes). It was shown that ranking the genes according to their posterior probabilities is more faithful to the reality than ranking them based on their p-values. Finally, as expected, it was verified that compared to other genes in the pathway, the genes with an improved rank are more densely connected to target genes; besides, such genes are also more densily connected with each other.

8.4 Identification of Protein–Protein Interactions

Protein–protein interactions (PPIs) provide invaluable clues to help elucidate biological processes or cellular functions. Wetlab technologies such as co-affinity purification followed by mass spectrometry [12] may only provide PPI data with both low coverage and accuracy. *In silico* prediction of PPI networks falls into three categories: high-throughput data-based, sequence-based and ortholog-based methods. In the first category, for instance, correlation between mRNA expressions may suggest the existence of a PPI [7]. Sequence-based methods examine for example protein/domain structures [27], gene neighborhoods [21] and gene fusion events [9].[3] In ortholog-based methods, annotation transfer between genomes is the key to detect conserved PPIs—or interologs—*via* gene orthologs [46].

To face the ever-growing accumulation of high-dimensional data, combined with the apparition of new types of data, Xia and collaborators designed a flexible model, able to integrate up to 27 data sets of various data types. This model is

[3] Gene fusion is likely to detect a PPI since two proteins interacting in the genome of one species are more likely to be fused into one single protein in the genome of another species.

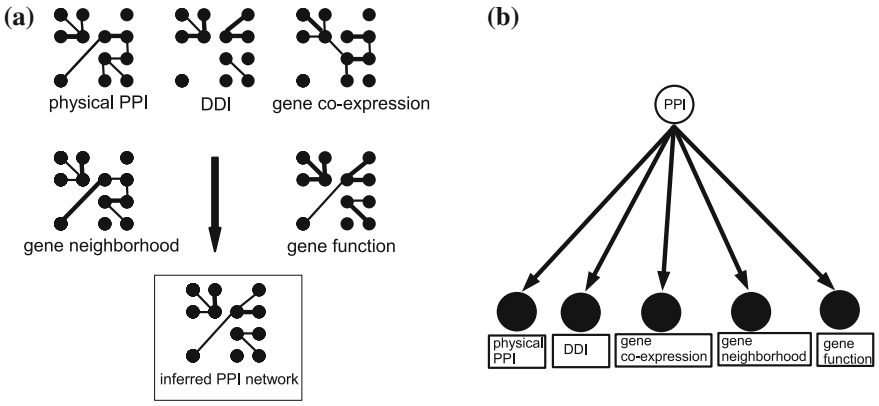

Fig. 8.2 Integration scheme in the approach of Xia et al. [43]. **a** The various data type evidences for the prediction of protein–protein interactions (PPI). **b** The naive Bayes classifier combining these data types. The variable class is binary: *PPI*/no *PPI*. *DDI:* domain-domain interaction.

based on a naive Bayes classifier [43]. A naive Bayes classifier is a Bayesian network whose elementary tree structure consists of a single parent (the class variable) and of children (here, the data types) that are independent of one another, given the class variable (see Fig. 8.2).

8.4.1 Coping with a Vast Spectrum of Heterogeneous Data Types

Before we explain thoroughly how the designed classifier, IntNetDB, integrates various types of data to find PPIs in the human genome, we emphasize the wide heterogeneity of the data types used by this system. In the most recent instance of IntNetDB, Xia and co-workers integrated 27 heterogeneous genomic, proteomic and functional data sets, encompassing 7 data types. For some model organisms (Yeast, Worm, Fruitfly...), annotations about **physical protein–protein interactions** are available. Mapping each such genome interactome to human genome through protein orthologs evidences interologs. **Domain-domain interactions** (DDIs) are known to mediate many protein–protein interactions. Structural domain information databases exist, that provide DDI scores. A DDI score is assigned to the pair of proteins that respectively harbour the two domains. **Gene co-expression** is often a reliable indicator for PPI.

In addition to data describing **gene fusion** and **gene neighborhood**, another data type also depicts gene contexts: **gene co-occurrence** is often indicative of a PPI; provided that the organism genome is fully sequenced, it is recognized that two interacting proteins are likely to be either both absent or both present in this genome [38]. On the other hand, proteins sharing the same biological function are

often involved in a PPI [33]. Thus, the **Gene Ontology** (GO) [36] provides supplementary evidence for PPI. Mapping human genes to orthologs in four model organisms (Yeast, Worm, Fruitfly and Mouse) was considered in [43].

Xia and co-workers also integrated two novel data types to help predict PPIs: **phenotypic distances** and **genetic interactions**. RNAi phenotype data have been used to predict PPIs for model organisms: under knock-out experiments, the respective phenotype profiles of interacting proteins tend to be similar. To transfer these phenotype data to human, Xia and collaborators mapped to their human orthologs the genes in the model organisms. Then similarity between the mapped phenotypes was assessed for the pair of genes in human. On the other hand, synthetic genetic analysis is a technology that was used in *Saccharomyces cerevisiae* to provide a global map of genetic interactions. Genetic interactions are recognized as high reliable indicators of PPIs. Xia and co-workers mapped the genetical interaction network of the Yeast model to human interologs.

8.4.2 Heterogeneous Data Integration by Naive Bayes Classifier

In the integrative model, each of the T data types used for the integration contributes an evidence e_i $(1 \leq i \leq T)$ for some given pair of proteins. To assess PPI for this pair of proteins, the likelihood ratio is

$$LR(e_1, \ldots, e_T) = \frac{\mathbb{P}(e_1, \ldots, e_T \mid PPI)}{\mathbb{P}(e_1, \ldots, e_T \mid \neg PPI)},$$

where $\mathbb{P}(e_1, \ldots, e_T \mid H)$ represents the probability that the evidence $(e_1, \ldots, e_T)$ has been observed under hypothesis H. The two alternative hypotheses we are interested in are PPI, the existence of a protein–protein interaction, and $\neg PPI$, the absence of such an interaction. Thus, if the numerator is significantly higher than the denominator, PPI will be assessed. Symmetrically, a low likelihood ratio will support the $\neg PPI$ hypothesis.

Under the assumption that the data sources are independent, the likelihood ratio writes as a product:

$$LR(e_1, \ldots, e_T) = \prod_{i=1}^{T} LR(e_i) = \prod_{i=1}^{T} \frac{\mathbb{P}(e_i \mid PPI)}{\mathbb{P}(e_i \mid \neg PPI)}.$$

The likelihood ratio for data type i providing evidence e_i is calculated from a set of assessed PPIs (positive set) and assessed counter-examples (negative set). The Human Protein Reference Database (HPRD) was used as the positive set; it references 19,438 experimentally verified PPIs for 5,983 proteins [32] (at the time of the integration by Xia et al.). The negative set was generated by Rhodes and coworkers [33]: it spans all pairwise combinations between two sets of proteins

located in two different subcellular compartments [plasma membrane (1,397 proteins) and nucleus (2,224 proteins) respectively]. Evaluating both positive and negative sets for each data type provides reference evidences, which allows to compute the desired likelihoods. Discretization into intervals is used for the purpose. Suppose we have to assess $\mathbb{P}(e_i \mid PPI)$ for some pair of proteins, where e_i is the evidence observed for this pair. The positive PPI reference set does not necessarily exhibit a protein pair showing the *exact* evidence e_i. Therefore, the value $\mathbb{P}(e_i \mid PPI)$ is replaced with $\mathbb{P}(I_{positive\ set}(e_i) \mid PPI)$, where $I_{positive\ set}(e_i)$ is an interval around e_i. This interval was obtained from the discretization into intervals of the evidences observed for the protein pairs of the positive PPI reference set. $\mathbb{P}(e_i \mid \neg PPI)$ is calculated similarly.

Care is required when several data sets contributing to the same data type are integrated. In this case, to avoid the bias due to dependence, the maximal likelihood (over the data sets) is retained for the data type.

8.4.3 Performances

The literature on alternative methods is vast. Machine learning methods addressing PPI prediction encompass Bayesian classifiers, decision trees, random forests, logistic regression and support vector machines. The reader is referred to [44] (for instance) for a recent overview of existing computational methods.

Two variants of the IntNetDB method were run. The two executions differed by the HPRD version (more than 10,000 newly annotated PPIs), the integration of three novel data types (phenotypic, genetic, gene context) in addition to PPI, GO, gene expression, DDI, and the incorporation of fourteen extra data sets. The comparison showed a drastical gain in coverage, for a similar ratio of true positives to false positives: the reinforced integration increased prediction coverage by fivefold (38,379 PPIs for 5,791 proteins versus 180,010 PPIs for 9,901 proteins). Besides, not only is the depicted probabilistic approach a simple yet efficient system to standardize the contributions of heterogeneous data types *via* likelihoods, it is also a flexible method: the combined likelihood easily supports the integration of any novel type of data.

8.5 Identification of the Regulatory Target Genes of Transcription Factors

A transcription factor (TF) is a protein that controls the expression of its target gene by binding to some specific DNA site located in the regulatory region of the gene. ChIP-chip and ChIP-seq techniques (Chromatin Immuno-Precipitation respectively followed by microarray gene expression measurements and by

massively parallel DNA sequencing) provide the genome-wide list of the physical *binding* sites, for a given TF. Exploiting *sequence* similarity to a consensus obtained for already known binding sites is also likely to pinpoint putative binding sites for the TF of interest. Another source of evidence lies in the variation in gene *expression* induced by knock-out or mutation of the gene coding for the TF. However, none of the above data types alone can achieve accurate and complete identification. First, high-throughput data are prone to present high noise level. Besides, ChIP-chip and ChIP-seq technologies only inform about physical DNA-TF interactions. Third, putative binding sites predicted based on sequence similarity with a canonical motif might actually not be bound by the TF of interest. Finally, variations in gene expression are equally observed for genes either directly or indirectly controled by a given TF. In the following, B, S and E will respectively stand for binding, sequence and expression data.

8.5.1 Integrating Multiple Genomic Data Sources with Multiple Gene Networks

To cross evidences from multiple types of genomic data, two categories of methods have been investigated. In regression approaches, where a data type is regressed against another, a large number of observations is required. This is a severe limitation in the case of gene expression microarray data. In mixture model[4] methods, the probabilistic framework allows inference based on the posterior probability of being a target conditional on the multiple data evidences. In the mixture model developed in [39], integration includes only two data types—(S, E) or (S, B) -. This model was further adapted in [30], to jointly handle the three data types B, S and E. So far, the mixture models used assumed conditional independence: conditional on a gene being a target or not, the different data types are independent. Nevertheless, for the pair (B, S), such an hypothesis is not consistent with experimental results: the higher the similarity with the canonical site (S), the higher the binding strength (B).

This section describes the model of Wei and Pan [41]. Therein, the multiple sources of genomic data are modeled through a multivariate normal mixture model, and integration of multiple gene networks with these genomic data types relies on a Markov random field (MRF). Besides relaxing the constraint on conditional independence of genomic data types, another major contribution of Wei and Pan lies in incorporating biological prior knowledge stating that neighboring genes tend to be co-regulated by a TF. Thus, not only does Wei and Pan's approach integrate several genomic data types; it allows to automatically incorporate knowledge from multiple gene networks (see Fig. 8.3). More and more gene

[4] A mixture model is a probabilistic model that represents a population of k groups, with random proportions $\pi_1, \ldots, \pi_k$.

networks are made available, such as protein–protein interaction networks . On the other hand, novel networks may be inferred, such as co-expression gene networks and networks derived from gene ontologies.

Without loss of generality, the presentation here restrains to the three genomic data types B, S and E. It is important to note that regions which associate with TFs according to ChIP-chip and ChIP-seq assays are not determined with single-nucleotide resolution. Wei and Pan computed the binding data (B) from ChIP-chip assay data: given two replicates in presence of the antibody appropriate for the TF of interest and given two control replicates, four log2 intensity ratios (LIR) were measured for the four combinations Immuno-Precipitation/control. The binding score B_i of a given gene i was computed as the average of the four LIR peaks on the coding region. If there were probes in the intergenic region, B_i was then calculated as the maximum of the average over the coding region and the average over the intergenic region. The sequence data (S) used by Wei and Pan was obtained as follows: first, a consensus sequence was produced from 10 known binding sites of the TF of interest; then the genome was scanned with respect to this consensus. Fixing a very low threshold allowed the detection of at least one match per gene. For gene i, S_i was calculated as the maximum of all its matching scores.

8.5.2 The Unified MRF-Based Mixture Joint Model

For a specified TF, and given a set of n genes, the aim is to estimate whether gene i is a target for a factor transcription of interest: $T_i = 1$ denotes a target, otherwise $T_i = 0$. The gene i is described by (B_i, S_i, E_i), summarizing observations for B, S and E data. In this approach, conditional normal distributions are described for the observed genomic data (B, S, E):

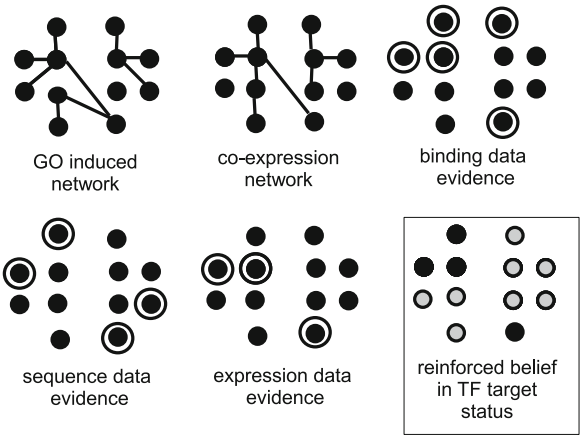

Fig. 8.3 Integration scheme for two networks and three types of data evidence in the approach of Wei and Pan [41]. A *circled node* indicates that strong evidence is observed for the corresponding node (gene). *Bottom right section* reinforced belief in transcription factor (*TF*) target status is indicated in *black*.

$$\mathbb{P}((B_i, S_i, E_i) \mid T_i = j) = \phi((B_i, S_i, E_i); \mu_j, \Sigma_j) = \phi_j(B_i, S_i, E_i),$$

where $j = 0, 1$ and ϕ is a trivariate normal density function of mean μ_j and covariance matrix Σ_j. A mixture model is then depicted as:

$$\mathbb{P}((B_i, S_i, E_i) \mid T_i) = (1 - \pi_1)\ \phi_0(B_i, S_i, E_i) + \pi_1\ \phi_1(B_i, S_i, E_i), \tag{8.6}$$

where $\pi_1 = \mathbb{P}(T_i = 1)$ is the prior probability of gene i being a target (and, symmetrically, $(1 - \pi_1)$ is the prior probability of gene i not being a target). The model in Eq. 8.6 has to be understood as the "superimposition" of two normal densities, one under the assumption that gene i is a target (ϕ_1), and one under the assumption that gene i is not a target (ϕ_0).

The knowledge from the N_{net} gene networks is incorporated through a Markov random field that rules the states $T_1, \ldots T_n$ of the n genes according to their N_{net} neighborhoods. Wei and Pan formalized an MRF-based mixture joint model (MRF-MJM), which writes as the following logistic regression model:

$$logit\left(\mathbb{P}(T_i = 1 \mid T_{\bigcup_{k=1}^{N_{net}} neigh(i,k)}, \theta)\right) = \gamma + \sum_{k=1}^{N_{net}} \beta_k\ (n_1(i,k) - n_0(i,k))/m(i,k), \tag{8.7}$$

where $neigh(i, k)$ designates the neighborhood of gene i in network k, parameter θ stands for $(\gamma, \beta_1, \ldots, \beta_{N_{net}})$, $n_j(i, k)$ is the number of genes in $neigh(i, k)$ that have state T_j $(j = 0, 1)$ and $m(i, k) = n_0(i, k) + n_1(i, k)$. The contribution of each network k is weighted by the non negative regression coefficient β_k, which therefore measures how informative network k is. In Eq. 8.7, conditioning by $T_{\bigcup_{k=1}^{N_{net}} neigh(i,k)}$ indicates that the TF target status of gene i depends on the statuses of all its neighbor genes, considered over all the N_{net} networks.

In this case, estimating the likelihood is intractable. In this framework, a tractable approximation to the joint distribution, the pseudolikelihood [1], is used instead. Tractability is ensured by the conditional independence assumption which leads to the following factorization:

$$\begin{aligned}\mathbb{P}(T) \simeq L_{pseudo}(T, \theta) &= \prod_{i=1}^{n} \mathbb{P}(T_i \mid T_{\bigcup_{k=1}^{N_{net}} neigh(i,k)}, \theta) \\ &= \prod_{i=1}^{n} \frac{exp\left(\gamma + \sum_{k=1}^{N_{net}} \beta_k\ (n_1(i,k) - n_0(i,k))/m(i,k)\right)}{1 + exp\left(\gamma + \sum_{k=1}^{N_{net}} \beta_k\ (n_1(i,k) - n_0(i,k))/m(i,k)\right)}.\end{aligned} \tag{8.8}$$

Besides the factorization, the transition from Eq. 8.7 to 8.8 uses the conversion $y = logit(x) = log\ \left(\frac{x}{1-x}\right) \Rightarrow x = \frac{e^y}{e^y+1}$.

The Bayes theorem states that $\mathbb{P}(T \mid (B,S,E)) \propto \mathbb{P}((B,S,E) \mid T)\ \mathbb{P}(T)$. The two ingredients on the right hand side are available from Eqs. 8.6 and 8.8, respectively. An MCMC is used to estimate the posterior probability of genes being targets of a specified TF.

8.5.3 Performances

A tremendous variety of alternative computational approaches are available. Some pointers to general surveys are provided in [29, p. 584]. In particular, Elnitski et al. wrote a summary on the synergism between *in silico*, *in vitro* and *in vivo* identification of TF binding sites [8]. On the other hand, the influential role of data integration is stressed in the surveys provided in [17, 28].

The MRF-MJM approach was evaluated with the LexA transcription factor of *Escherichia coli*. It was first noticed that allowing conditional dependence by assuming a general conditional variance structure in the MRF-MJM model does not increase the predictive power over assuming conditional independence. However, as binding data and sequence data are highly correlated for target genes, this result appears to go against intuition. It might be explained by moderate predictive power of sequence data and a simpler model in the assumption of conditional independence. All subsequent analyses were then run incorrectly assuming conditional independence.

Wei and Pan tested six different integration schemes. Six instances of the MRF-MJM approach, including simplified ones, were run: $(E; N_{CoE})$, $(E; N_{GO})$, $(E; N_{CoE} + N_{GO})$, where N_{CoE} and N_{GO} are gene networks respectively derived from gene co-expression and a gene ontology (GO), and the three previous instances with the full set of genomic data (B,S,E) instead of E. Besides, instances of the standard mixture model (SMM), which considers a single genomic data type, were also run for comparison: $SMM(B)$, $SMM(S)$, $SMM(E)$. The genes were ranked according to their posterior probabilities. The variation in the ranking across these instances was studied for the genes supported by experimental evidence or annotated with "strong evidence" in the RegulonDB database [11].

It was confirmed using ROC curves that mixed integration of both networks and various genomic data types greatly improves over considering a single genomic data type alone. Besides, in a mixed scheme, the improvement is less drastic when increasing the number of genomic data types or when increasing the number of networks. The GO-derived network constantly showed a β coefficient lower than the co-expression network's: it is explained by a higher connectivity of the GO network, which entails that a target and a non target genes are more likely to be neighbors in this network.

8.6 Inference of Causal Relationships Among Phenotypes via Integration of QTL Genotypes

A quantitative phenotype (or trait) is defined as any physical, physiological or biochemical quantitative feature that may be observed for organisms. Quantitative trait loci (QTL) mapping aims at identifying the genomic regions, or QTLs, where genotype variation is correlated with phenotype variation. Deciphering the causal relationships among *expression* traits involved in the same biological pathways—and therefore correlated—is a current research topic. To this aim, the identification of the eQTLs (expression QTLs) causal to each phenotype is of prime importance. In the following, we will denote by genetic architecture (GA) of a given phenotype the locations and effects of its (directly) causal QTLs. Conversely, GA inference has to benefit from the information borne by the network that links the phenotypes. Though, standard QTL mapping merely addresses one single trait at a time, not considering a possible causal network structure among traits. Thus, QTLs that exert a direct effect on the trait under study cannot be distinguished from QTLs with an indirect effect (see Fig. 8.4a). To reconstruct a causal phenotype network (CPN), several approaches in the literature include QTLs in a probabilistic framework. However, the common feature of these approaches lies in that GA inference and CPN reconstruction are conducted separately [3, 34, 47]. In general, the GA is first inferred, to further help the determination of the CPN. In the QTLnet approach, Chaibub Neto and co-authors pioneered the principle of joint inference of CPN and GA [4].

8.6.1 Joint Inference of Causal Phenotype Network and Genetic Architecture

Chaibub Neto et al. showed that performing the mapping analysis of a phenotype conditional on its parents in the CPN is the way to avoid detecting QTLs with indirect effects on this phenotype as directly causal QTLs. Namely, whereas standard mapping analysis would test the dependence between phenotype φ_1 and QTL candidate Q_1 ($\varphi_1 \perp\!\!\!\perp Q_1$), *conditional mapping* assesses or invalidates the dependence relation $\varphi_1 \perp\!\!\!\perp Q_1 | \mathbf{Pa}(\varphi_1)$ where $Pa(\varphi_1)$ is the set of parents of φ_1 in the CPN. As the CPN is itself unknown, the QTLNet approach jointly infers the CPN and the GA: the procedure iterates a process where updating the CPN alternates with updating the GA. Thus, GA inference will benefit from information on the CPN. The core idea is to learn a Bayesian network whose structure coincides with the candidate CPN, using the current information available about causal QTLs. It has to be noted that the central dogma of biology constrains unidirectionality for causality, from QTL to phenotype: arcs $\varphi \rightarrow Q$ are not allowed.

Adding information about causal QTLs is crucial to distinguish between candidate phenotype networks, when learning a phenotype network. The network

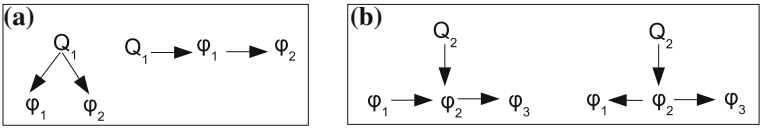

Fig. 8.4 Disambiguisation of causal relationships. **a** *Left* direct effect of QTL Q_1 on phenotype φ_2; *right* indirect effet of Q_1 on φ_2. In both cases, $\varphi_2 \not\perp\!\!\!\perp Q_1$. **b** The two models have the same joint probability $\mathbb{P}(\varphi_1,\ \varphi_2,\ \varphi_3)$ but have different conditional probabilities $\mathbb{P}(\varphi_1,\ \varphi_2,\ \varphi_3 \mid Q_2)$ given the QTL data

which best fits the data is that which maximizes some given criterion. In a probabilistic framework without QTL data integration, this criterion would rely on the joint probability $\mathbb{P}(\varphi_1,\ \ldots,\ \varphi_n)$. With QTL integration, the probability to be taken into account is $\mathbb{P}(\varphi_1,\ \ldots,\ \varphi_n \mid GA)$ where GA stands for the r QTLs available: $Q_1,\ \ldots,\ Q_r$. Let us consider a toy-example where the networks (1) $\varphi_1 \rightarrow \varphi_2 \rightarrow \varphi_3$ and (2) $\varphi_1 \leftarrow \varphi_2 \rightarrow \varphi_3$ cannot be distinguished without QTL data integration since they have the same joint probability $\mathbb{P}(\varphi_1,\ \varphi_2,\ \varphi_3)$.[5] We now incorporate QTL knowledge as Q_2 affecting φ_2 but neither φ_1 nor φ_3 directly and obtain two mixed models (see Fig. 8.4b). Then, the conditional probabilities of the two networks are: $\mathbb{P}_{(1)}(\varphi_1,\ \varphi_2,\ \varphi_3 \mid Q_2) = \mathbb{P}(Q_2)\ \mathbb{P}(\varphi_1)\ \ \mathbb{P}(\varphi_2 \mid Q_2,\ \varphi_1)\ \mathbb{P}(\varphi_3 \mid \varphi_2)$ and $\mathbb{P}_{(2)}(\varphi_1,\ \varphi_2,\ \varphi_3 \mid Q_2) = \mathbb{P}(Q_2)\ \mathbb{P}(\varphi_2 \mid Q_2)\ \mathbb{P}(\varphi_1 \mid \varphi_2)\ \mathbb{P}(\varphi_3 \mid \varphi_2)$. In the general case, the previous conditional probabilities are not equal.

8.6.2 The Mixed Model

To model continuous phenotypes that are involved in a causal phenotype network while also being under the dependence of discrete QTLs, a conditional Gaussian regression model is used: conditional on the genotypes and, possibly, covariates, the phenotypes follow a multivariate normal distribution.

Given n individuals, t phenotypes, let $\varphi = (\varphi_1, \ldots, \varphi_n)^T$ represent all phenotype values, with $\varphi_i = (\varphi_{1i}, \ldots, \varphi_{ti})^T$ representing the t phenotype values for individual i. Let $\epsilon_i = (\epsilon_{1i}, \ldots, \epsilon_{ti})^T$ be independent normal error terms. The regression model for the phenotype p of individual i writes:

$$\varphi_{pi} = \mu^*_{pi} + \sum_{v \in Pa(\varphi_p)} \beta_{pv}\ \varphi_{vi} + \epsilon_{pi}, \epsilon_{pi} \sim \mathcal{N}(0, \sigma^2_p). \tag{8.9}$$

The genetic contribution describes the effects of QTLs and possibly covariates: $\mu^*_{pi} = \mu_p + X_{pi}\ \theta_p$, where μ_p is the overall mean for phenotype p, X_{pi} represents the

[5] $\mathbb{P}_{(1)}(\varphi_1,\ \varphi_2,\ \varphi_3) = \mathbb{P}(\varphi_1)\ \mathbb{P}(\varphi_2 \mid \varphi_1)\ \mathbb{P}(\varphi_3 \mid \varphi_2)$ and $\mathbb{P}_{(2)}(\varphi_1,\ \varphi_2,\ \varphi_3) = \mathbb{P}(\varphi_2)\ \mathbb{P}(\varphi_1 \mid \varphi_2)\ \mathbb{P}(\varphi_3 \mid \varphi_2)$. Equality is assessed from the Bayes theorem.

row vector of genetic effect predictors derived from the QTL genotypes along with any covariates, and θ_p is a column vector of all genetic effects defining the genetic architecture of phenotype p augmented with any covariates. In the phenotypic contribution (second term of Eq. 8.9), $Pa(\varphi_p)$ designates the set of parents of phenotype p in the phenotype network and β_{pv} models the effect of parent phenotype v on phenotype p.

8.6.3 Causal Phenotype Network Reconstruction

Since the graph space grows super-exponentially with the number of phenotypes, reconstructing a CPN requires a heuristic. An MCMC is implemented, that combines sampling over CPN structures and QTL mapping. However, conceptually, a mixed structure $G = G_\varphi \cup GA$, is considered, which is the CPN G_φ augmented with the genetic architecture GA connecting QTLs to phenotypes (see Fig. 8.5). The posterior probability of a candidate G_φ is estimated as explained below.

From Eq. 8.9, we know that $\mathbb{P}(\varphi_{pi} \mid G_\varphi, GA, \gamma)$ is $\mathcal{N}(\mu^*_{pi} + \sum_{v \in Pa(\varphi_p)} \beta_{pv}\varphi_{vi}, \sigma^2_p)$,[6] where γ stands for the set of parameters of the mixed model (i.e. the coefficients β). Under the assumption of independence between the n individuals, the likelihood of the candidate CPN factorizes as:

$$\mathbb{P}(\varphi \mid G_\varphi, GA, \gamma) = \prod_{i=1}^{n} \prod_{p=1}^{t} \mathbb{P}(\varphi_{pi} \mid G_\varphi, GA, \gamma).$$

In this case, it is straightforward to compute the marginal likelihood by integrating the previous expression with respect to γ:

$$\mathbb{P}(\varphi \mid G_\varphi, GA) = \int_\gamma \mathbb{P}(\varphi \mid G_\varphi, GA, \gamma)\ \mathbb{P}(\gamma \mid G_\varphi, GA)\ d\gamma.$$

Finally, the posterior probability of structure G_φ conditional on the data may be computed from:

$$\mathbb{P}(G_\varphi \mid \varphi, GA) \propto \mathbb{P}(\varphi \mid G_\varphi, GA)\ \mathbb{P}(G_\varphi),$$

where $P(G_\varphi)$ is a prior on the CPNs.

Thus, integrating knowledge about QTLs actually modifies the likelihood landscape for the search space of G_φ structures.

To navigate in this search space, three moves are implemented in the MCMC scheme of Chaibub Neto et al.: addition of a directed edge, removal or direction

[6] If $X = y + E$, with $E \sim \mathcal{N}(0, \sigma^2)$, then $X \sim \mathcal{N}(y, \sigma^2)$.

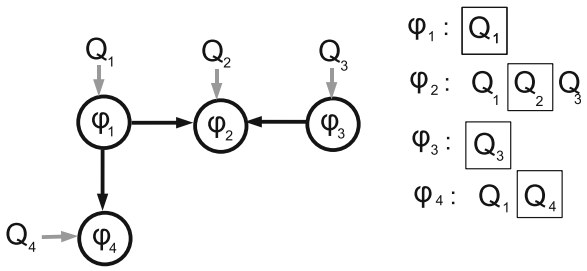

Fig. 8.5 The mixed model in the approach of Chaibub Neto et al. [4]. **a** The genetic architecture defines the QTL mapping (*arrows* in *light grey*); the causal network defines the dependences between the phenotypes (transcripts) (*arrows* in *dark*). **b** Comparison of the genetic architectures inferred without and with conditioning (in the latter case, the QTLs are framed) (see Sect. 8.6.1)

reversal. Subsequent to a move, (conditional) QTL mapping is replayed for those phenotypes whose set of parent nodes was modified by the move. Finally, a posterior probability for the causal relationship $\varphi_i \rightarrow \varphi_j$ $(1 \leq i, j \leq t)$ involving each pair of phenotypes is assessed through *Bayesian model averaging*: for each directed edge $\varphi_i \rightarrow \varphi_j$, the posterior probability is estimated as the frequence of occurrence observed over all the models sampled by the MCMC process.

8.6.4 Performances

First, 1,000 tests were performed based on simulated data generated under two conditions: respectively weak and strong dependences between the phenotypes and their eQTLs. The genetic architectures produced were compared with those obtained through standard QTL mapping. Conditional mapping (see first paragraph in Sect. 8.6.1) revealed the true architecture in both conditions. To estimate the quality of the phenotype network inference, the authors measured the frequency that the posterior probability of the true network was the highest, second highest, *etc*. Under the strong dependence condition, the true network is identified as the best one in 84 % of the cases. The results are more subdued under the weak dependence condition.

The QTLnet method was then used on real data (132 mice of a F2 intercross, 3,421 transcripts, 1,065 markers), to derive the causal phenotype network relative to 14 highly correlated transcripts. A consensus network was constructed through Bayesian model averaging. Interestingly, this consensus network suggests a key role of one of the transcripts in the regulation of the other transcripts in the phenotype network.

8.7 Prediction of Protein Function Through GO-Enriched Networks of Multiple Related Species

In Sect. 8.5, a gene ontology (GO) was used to derive a functional coupling gene network, to enhance the identification of transcription factor targets. Therein, a preprocessing step merely derived a gene network, based on some similarity measure in the ontology. In the present section, we outline an approach which benefits from GO knowledge on the fly. As the most developed biological ontology is the Gene Ontology [36], it is not surprising that this approach addresses the prediction of protein function.

Improving the coverage and accuracy for functional annotation of proteins is an active field in post-genomics research. On the one hand, only labor intensive small-scale experiments are able to provide direct evidence about the functions of proteins such as energy and RNA metabolism, signal transduction, translation initiation, enzymatic catalysis and immune response. In contrast, though numerous high-throughput technologies allow large-scale experimental investigations, the various types of molecular data but only yield indirect clues about protein function. To reach the objectives of coverage and accuracy, much is expected from computational methods.

Established prediction methods use sequence or structure similarity to transfer functional annotation from protein to protein [22]. However, it is well known that sequence similarity does not obligatorily entail functional identity. More reliable evidence is derived from indirect information provided by the biological context of the protein. Such contextual information includes physical protein–protein interactions (PPI), genetic interactions and co-expression of the genes coding for the proteins. These contextual data are commonly represented as networks. Thus, a wide category of methods predicts the function of a protein from the known functions of its neighbors in the network [2, 14, 45]. Besides, incorporation of heterogeneous data has been proven useful to increase the power of automated predictive systems [26].

Probabilistic graphical models offer an appealing framework to propagate functional annotations through neighborhoods; this explains that approaches based on these models are not new to protein function inference (e.g. [6, 19, 26]). However, severe limitations hamper these approaches in the (frequent) case of proteins that are isolated in the network or whose neighborhood is poorly annotated. Refined GO-based strategies have been proposed to overcome these issues. Amongst them, the probabilistic approach of Mitrofanova and collaborators combines random Markov models and Bayesian networks into a single model [23].

In classical approaches, probabilistic inference relies on partial knowledge of functional annotations to discover the missing functions by passing on and handling uncertain information over a large network. For instance, this network may be derived from knowledge on physical interactions (PPIs). One of the original concepts of Mitrofanova and co-workers' model lies in connecting the networks of two (or more) related species into a single computational model. The rationale

behind this approach exploits the fact that in most cases, proteins of different related species that share high similarity—orthologs—exerted the same established function before the speciation event. The second original concept of the approach described is the direct integration of an ontology or rather, of a sub-ontology (GOS), into the graphical model. This integration allows the simultaneous prediction for the multiple functional categories—or terms—described by the GOS. In the combined model, each protein is represented by its own GOS. As a consequence, during function inference, not only is the information passed between protein neighbors within a species, information also percolates within the GOS. Moreover, due to inter-species connections between orthologs, such information is diffused in an enlarged network.

8.7.1 The GO-Enriched Intra-Species Model

For the sake of a progressive exposition, we first present a model deprived of inter-species relationships. In the model, each protein is represented by a Bayesian network whose structure is a replicate of the GO sub-ontology (GOS) of interest (see Fig. 8.6a). Each protein has its own annotation (positive, negative, unknown) for each of the GOS terms. A positive annotation means that the protein has the function represented by the GOS term. The final objective of the probabilistic inference is to assign an annotation (positive/negative) to each term (GOS node) labeled unknown in the combined model. The GOS is a directed acyclic graph where the relationship between child c and parent p may be "IS A" or "IS PART OF". The GOS information is naturally modeled as a Bayesian network (BN). The so-called *true-path rule* for gene ontologies requires that if a protein i is positively annotated at a child node t (denoted by $x_i^t = +$), then it must also be at all the ancestor nodes of this child. Consequently, positive annotations may be expanded up within a GOS whereas negative annotations are expanded down if all the parent terms of a child term are annotated negative. It follows that conditional probabilities $\mathbb{P}(x_i^t = + \mid pa_{it})$ and $\mathbb{P}(x_i^t = - \mid pa_{it})$ need be estimated only if one parent at least is annotated positive within a possible assignment pa_{it} of the parents (for instance, $pa_{it} = (+, +, -)$ in the case when node t has three parents in the GOS).

On the other hand, a pairwise Markov random field (MRF) is used to encode connections between the proteins, based on some similarity measure between the proteins. Such measures may be derived from PPIs or orthology (i.e. sequence similarity). In the model resulting from GOS and MRF combination, a potential function, ψ^{intra}, is defined; this potential expresses the probability of joint annotation of two proteins i and j at a GOS term t, conditional on their being similar. In the case of a PPI-based measure, similar proteins are defined as interacting proteins: then, the probabilities $\psi^{intra}(x_i^t,\ x_j^t) = \mathbb{P}(x_i^t, x_j^t \mid interaction)$, with $x_i^t, x_j^t \in \{+, -\}$ are estimated from a training set. In the case of a sequence similarity-based measure, a potential is derived from a pairwise normalized BLAST score s_B:

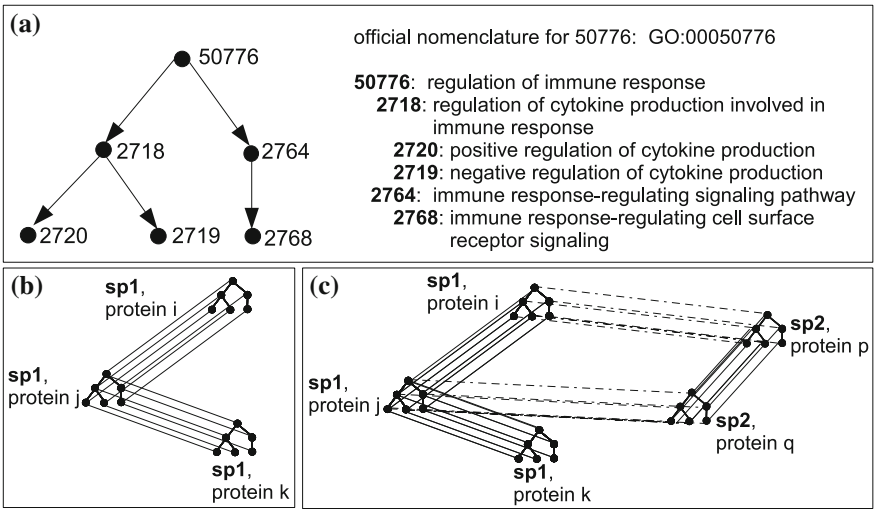

Fig. 8.6 The combined model of Mitrofanova et al. [23] for protein function prediction. **a** The Gene Ontology (GO) substructure. **b** The GO-enriched intra-species model. **c** The combined model obtained through inter-species homology (Two species are considered: sp1 and sp2)

$\psi^{intra}(+,+) = \psi^{intra}(-,-) = s_B(i,j); \quad \psi^{intra}(+,-) = \psi^{intra}(-,+) = 1 - s_B(i,j).$ If both similarity measures are available for a given pair of proteins, thus defining two potentials, the resulting potential is defined as the product of the two former.

The knowledge about the annotation information of protein i, at GOS term t is modeled through function ϕ: $\phi(+) = 1; \phi(-) = 0$ for a positive annotation; $\phi(-) = 1; \phi(+) = 0$ for a negative annotation; equiprobability for an unkown annotation ($\phi(?) = 0.5$).

The MRF and the GO-based BNs are combined into a single hybrid model [18]—(see Fig. 8.6b). Based on the material above defined, the joint distribution of the functional term annotations (X_i^t) over the set of proteins $\mathscr{P}$ is defined as a pairwise MRF distribution (see Eq. 8.2), whose statement is simplified as follows for the sake of conciseness:

$$\mathbb{P}(\{x_i^t\}_{t\in\mathscr{S}, i\in\mathscr{P}}) = \frac{1}{Z} \prod_{t\in\mathscr{S}} \prod_{i\in\mathscr{P}} \phi(x_i^t) \prod_{i,j\in edges(MRF(\mathscr{P}))} \psi^{intra}(x_i^t, x_j^t) \prod_{i\in\mathscr{P}} (x_i^t \mid pa_{it}). \tag{8.10}$$

$\mathscr{S}$ is the sub-ontology of interest and Z is the so-called normalizing constant (see Definition 2, Sect. 8.2). In the above joint distribution, it is easy to identify the contribution of the Markov random field defined by the similarity relation between proteins, and the contribution of the Bayesian networks. The flow of information about annotation is propagated through the hybrid model using a message-passing mechanism tailored for such hybrid models.

8.7.2 The GO-Enriched Inter-Species Model

The extension to the inter-species case is straightforward when two related species are considered. This time, when sequence similarity is ascertained for protein i in the first species and protein j in the second species, a corresponding potential function ψ^{inter} is defined. Similarly to the scheme described by Eq. 8.10, the combined model is merely augmented with undirected edges connecting the identical GOS term nodes of i and j proteins (see Fig. 8.6c). The joint distribution is readily extended as follows: a second intra-species factor is added, $\prod_{i\in\mathscr{P}'} \phi(x_i^t) \prod_{i,j\in edges(MRF(\mathscr{P}'))} \psi^{intra}(x_i^t, x_j^t)$ where $\mathscr{P}'$ is the set of proteins for the second species; an inter-species factor is also added, that accounts for the (valid) edges connecting some protein in $\mathscr{P}$ to some other similar protein in $\mathscr{P}'$: $\prod_{i,j\in edges(MRF(\mathscr{P}\cup\mathscr{P}'))} \psi^{inter}(x_i^t, x_j^t)$.

8.7.3 Performances

Mitrofanova and collaborators performed tests on Yeast and Fly. Respectively 6,008 and 12,199 proteins were considered for Yeast and Fly species. Various tests were performed based on (1) executions (S) of the approach without integration of the gene ontology, that is single-term prediction, and (2) runs (GO) of the approach with the integration of the gene ontology. Besides, annotation transfer by similarity was considered either within a single species (1), or within two species (2). We will denote S1, S2, GO1 and GO2 these four kinds of tests. Figure 8.7 recapitulates the experimental protocole.

The comparison S1 versus GO1 is meant to evaluate the impact of using a gene ontology. In this baseline test, predictions were then compared for a single term. The improvement in the prediction is outstanding in all cases: a gain of 26 % (from

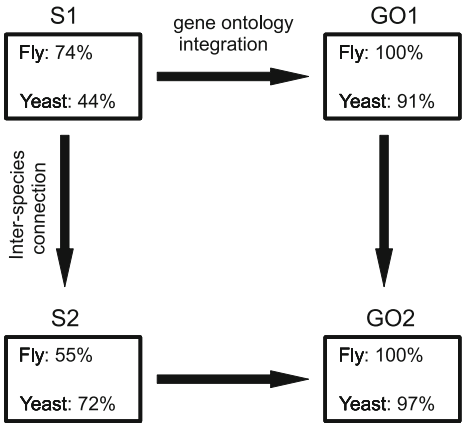

Fig. 8.7 Evaluation of the impact of gene ontology integration and inter-species connection on accuracy, in Mitrofanova et al.'s method. S1 denotes a run with intra-species connection only, whereas S2 indicates intra- and inter-species connection. GO_1 indicates the integration of gene ontology knowledge to the basic scheme S_1 (and symmetrically for GO_2 and S_2)

74 to 100 %) is observed for the Fly species; an increase of 47 % (from 44 to 91 %) is observed for the Yeast species. Thus, in the case of the Fly species, the integration of GO knowledge suffices to produce the accuracy of 100 %.

The comparison S1 versus S2 aims at evaluating the influence of annotation transfer between genomes, through inter-species connection. Mitrofanova and collaborators performed 5-cross validation for the Fly, Yeast and combined Yeast-Fly networks. The results are contrasted: an under-performance is obtained in the case of the Fly species, for which the accuracy decreases by 19 % (from 74 to 55 %); a gain of 28 % (from 44 to 72 %) is observed for the Yeast species. Thus, inter-species connection alone may be counter-productive (Fly). If a gain is observed through inter-species connection, it is more subdued than the gain obtained through integration of a gene ontology (Yeast).

The aim of comparing S2 against GO2 is to measure the impact of the integration of GO knowledge in presence of inter-species connection. This time, in the case of the Fly species, inter-species connection does not interfere to lower the performance, which confirms the prominent role of the GO integration (55 % to 100 %). A gain of 25 % (from 72 to 97 %) is observed for the Yeast species (to be compared to the increase from 44 to 91 % without inter-species connection).

A significant gain of 8 % (from 91 to 97 %) is thus observed for the Yeast species in the GO1 versus GO2 test.

The main conclusion is that the GO integration exerts the most influential role. Inter-species connection may perform worse than merely considering a single genome. However, it is always beneficial to integrate both GO knowledge and inter-species connection. Yeast species shows more substantial improvements compared to Fly species: this may be explained by the higher quality of Fly data and hence better neighborhoods for the Fly proteins. Annotation transfer is enhanced through two independent principles: simultaneous consideration of multiple but related functional GO categories, higher connectivity due to orthology or PPI knowledge. Expanded protein coverage is another observed advantage.

In the spirit of the comparison S2 versus GO2, Mitrofanova and collaborators also compared their full approach (GO2: GO integration and inter-species connection) to the method of Naria et al. [26] which can be seen as a variant of S2. The method of Naria et al. relies on a probabilistic Bayesian framework that integrates networks (e.g. PPI and/or expression networks) with categorical features (i.e. presence of protein domains, knockout phenotype (e.g. "starvation sensitivity") and cellular location categories). The case of lack of information about categorical features is taken into account in [26], which thus allows the comparison. Besides, for comparability, both PPI and sequence similarity were used to build the networks input to the two methods. The method of Mitrofanova et al. improves over that of Naria et al.: for the Fly species, the accuracies observed are respectively 100 and 45 %; for the Yeast species, the accuracies are 97 and 50 %. Again, GO integration is shown to play a more prominent role than inter-species connection. This improved performance can be attributed to the increased connectivity endowered in the GO structure. However, it has to be noted that the S2 executions of Mitrofanova and co-workers' method already outperformed the (S2)

runs of Naria et al.'s approach: 55 versus 45 % for the Fly species, and 72 versus 50 % for the Yeast species. It is difficult to speculate on the reasons why annotation information percolates more efficiently in the probabilistic model of [23] (without GO integration) than in that of [26]. Unfortunately, no common types of results are available (such as accuracy, false positive rate, or number of true positives) that could allow the comparison of the methods both at full integration level (GO2 for Mitrofanova et al.'s method, and integration of categorical features for Naria et al.'s approach.).

Finally, with a Gene Ontology subtree of size 8, the running times observed for each five-cross validation round on Yeast, Fly and Yeast-Fly models were 35, 59 and 28 mn on average on a standard personal computer. The third low execution time is explained by faster convergence in the combined network, probably due to denser sources of evidence.

8.8 Discussion and Future Directions

In this chapter, we have presented different approaches based on probabilistic graphical models, to illustrate the use of this class of models as an integrative framework for systems biology. In particular, various forms of Markov random fields were described, that were used to model the propensity to share a common state for neighbor nodes in a single network or in multiple networks. For instance, in the illustration devoted to genetic association study (GAS), the MRF models a single network—a biological pathway—and the state accounts for association with the disease.

One of the simplest Bayesian networks that can be imagined, the naive Bayes classifier, also represents one of the most flexible tools to integrate multiple data types. In this line, the method presented here to detect protein–protein interactions (PPIs) assigns equal weights to the genomic data types. However, a limitation lies in that the positive and negative sets of examples and counter-examples requested by this simple method do not necessarily benefit from equal covers across the data types.

Enhancement through mixed integration of genomic data types and gene networks is shown for the identification of the target genes of transcription factors (TF). It was emphasized that the key to improvement is much more mixing data sources than multiplying either the number of genomic data types or the number of networks. In the probabilistic model, a single Markov random field integrates and weights the contributions of the gene networks. Neighbor genes therein are expected to share a common state (target or non target). In the global model, genomic data types are integrated through a prior distribution. In the GAS application, the prior distribution accounted for the integration of pathway knowledge.

The gene networks mentioned above provide *qualitative* knowledge to rely upon. This time, for causal phenotype network (CPN) reconstruction, a conditional Gaussian regression model was used to integrate *quantitative* characteristics (continuous phenotypes) and *qualitative* assumptions (latent relationships between

the phenotypes). In contrast with the preceding approaches, prior knowledge—consisting in the genetic architecture (GA)—is not fixed from the start but is instead refined throughout the CPN inference procedure: feedback from the most recent incumbent CPN offers opportunity to update the GA and *vice versa*.

The second (PPIs) and third (TF target genes) models presented both rely on shared functional annotation. Raw data is used in the second model whereas the third one may incorporate a gene network induced from a gene ontology. In contrast, accounting for ontological knowledge thoroughly impacts the statistical inference scheme in the last approach presented, that addresses protein function prediction. This approach combines ontology replication with intra- and -inter-species homology knowledge. Again, as for the GAS illustration above, a Markov random field (MRF) is built from a known structure, here a network connecting similar proteins. Similarity is assessed from PPI knowledge as well as intra- and inter-species homology. Unlike the GAS approach, neighbors in the network tend to share a common hierarchy of function annotations instead of a single variable. The originality of the mixed model arises from the expansion of the protein nodes of the MRF into Bayesian networks (BNs), each replicating the gene ontology substructure. The completion by links between identical term nodes in similar protein meta-nodes provides a highly connected network. Thus boosted information propagation is expected.

Among the five integrative methods reviewed, the one addressing PPI prediction and the one predicting TF gene targets are perhaps the most exemplary in that they take advantage of various genomic data and/or networks. In the case of the TF gene target application, integrating genomic data and networks outstandingly improves the results but then, increasing the number of genomic data types or networks does not provide much improvement. On the other hand, the illustration on the prediction of protein functions reveals the prominent role of gene ontology (GO) knowledge. GO integration exerts the most influential role. However, in this context, it is always beneficial to integrate both GO knowledge and inter-species connection.

The previous paragraph raises in particular the question on the possible dependence of the various data sources and on how this dependence is ignored or modeled. In the illustration of the PPI detection, the naives Bayes classifier requires independence of the data types conditional on the state variable ($PPI/\neg PPI$). Robustness to deviation from this rule was not evaluated in this framework. However, in the case of another model and for another application (identification of TF target genes), the conclusion was that the simplifying assumption of conditional independence does not decrease performance. The PPI detection illustrates here a case where multiple data sets may be examined within a common data type. Retaining the empirical maximum likelihood computed over all data sets of the same data type avoids the dependence bias for this type. Again, an open question remains the significance of a high likelihood obtained for some data type if there are cover biases between data types, in terms of positive and negative sets.

Further progress in the field will mainly depend on improving implementations and allowing actual flexibility. For instance, MCMC implementations rely on hyperparameters whose tuning can hardly be delegated to the end-user. Besides, it

is worth examining how to incorporate additional biological knowledge in priors, as in the case of causal phenotype network inference. The reported advantages of probabilistic graphical networks in promoting highly integrative approaches combining various heterogeneous data sources may be sometimes offset by the computational burden. From the theoretical viewpoint, for all models presented here, generalization to multiple data types is straightforward. Mitrofanova et al.'s method readily generalizes to more than two species but scalability might be an issue. The method designed to predict protein functions was shown tractable for gene ontology substructures of size below 20, which might appear insufficient to some end-users and therefore requires further work. The next-generation sequencing era is also that of grid and cloud computing. For example, three of the models presented here use an MCMC scheme. MCMCs are amenable to distributed implementations. As more data and more data types will become available, adding a novel data type should be automatically handled by the models' implementations. Therefore, the dissemination in the biological community of integrated PGM-based approaches also implies that service-oriented integration accompanies theoretical developments.

Acknowledgments The author wishes to thank the anonymous reviewer for constructive comments on her manuscript, and feedback most helpful to produce the final version.

References

1. Besag J (1986) On the statistical analysis of dirty pictures. J Roy Statist Soc Ser B 48:259–302
2. Carroll S, Pavlovic V (2006) Protein classification using probabilistic chain graphs and the Gene Ontology structure. Bioinformatics 22(15):1871–1878
3. Chaibub Neto E, Ferrara CT, Attie AD, Yandell BS (2008) Inferring causal phenotype networks from segregating populations. Genetics 179(2):1089–1100. doi:10.1534/genetics.107.085167
4. Chaibub Neto E, Keller MP, Attie AD, Yandell BS (2010) Causal graphical models in systems genetics: a unified framework for joint inference of causal network and genetic architecture for correlated phenotypes. Ann Appl Stat 4(1):320–339
5. Chen M, Cho J, Zhao H (2011) Incorporating biological pathways via a Markov random field model in genome-wide association studies. PLOS Genet 7(4):e1001353. doi:10.1371/journal.pgen.1001353
6. Deng M, Chen T, Sun F (2003) An integrated probabilistic model for functional prediction of proteins. In: Proceedings of the seventh annual international conference on research in computational molecular biology (RECOMb), pp 95–103
7. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 95(25):14863–14868
8. Elnitski L, Jin VX, Farnham PJ, Jones SJ (2006) Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. Genome Res 16(12):1455–1464
9. Enright AJ, Iliopoulos I, Kyripides NC, Ouzounis CA (1999) Protein interaction maps for complete genomes based on gene fusion events. Nature 402:86–90

10. Franke L, van Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, Wijmenga C (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. Am J Hum Gen 78(6):1011–1025
11. Gama-Castro S, Jiménez-Jacinto V, Peralta-Gil M et al (2008) RegulonDB (version 6.0): Gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. Nucleic Acids Res 36:D120–D124. doi:10.1093/nar/gkm994
12. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM et al (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415(6868):141–147
13. Hutz JE, Kraja AT, McLeod HL, Province MA (2008) CANDID: a flexible method for prioritizing candidate genes for complex human traits. Genet Epidemiol 32(8):779–790
14. Karaoz U, Murali T, Letovsky S, Zheng Y, Ding C, Cantor CR, Kasif S (2004) Whole-genome annotation by using evidence integration in functional-linkage networks. Proc Natl Acad Sci USA 101:2888–2893
15. Kindermann R, Snell JL (1980) Markov random fields and their applications. American Mathematical Society
16. Köhler S, Bauer S, Horn D, Robinson PN (2008) Walking the interactome for prioritization of candidate disease genes. Am J Hum Genet 82:949–958
17. Ladunga I (2010) An overview of the computational analyses and discovery of transcription factor binding sites. Methods Mol Biol 674:1–22
18. Lauritzen SL (1996) Graphical models. Oxford University Press, New York
19. Letovsky S, Kasif S (2003) Predicting protein function from protein/protein interaction data: a probabilistic approach. Bioinformatics 19:i197–i204
20. Li H, Wei Z, Maris J (2010) A hidden Markov random field model for genome-wide association studies. Biostatistics 11:139–150
21. Marcotte EM (2000) Computational genetics: finding protein function by nonhomology methods. Curr Opin Struct Biol 10(3):359–365
22. Mering CV, Jensen LJ, Snel B et al (2005) String: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res 33:433–437
23. Mitrofanova A, Pavlovic V, Mishra B (2011) Prediction of protein functions with Gene Ontology and interspecies protein homology data. EEE/ACM Trans Comput Biol Bioinf 8(3):775–784
24. Mourad R, Sinoquet C, Leray P (2011) A hierarchical Bayesian network approach for linkage disequilibrium modeling and data-dimensionality reduction prior to genome-wide association studies. BMC Bioinform 12:16+
25. Mourad R, Sinoquet C, Dina C, Leray P (2011) Visualization of pairwise and multilocus linkage disequilibrium structure using latent forests. PLOS ONE 6(12):e27320
26. Nariai N, Kolaczyk ED, Kasif S (2007) Probabilistic protein function prediction from heterogeneous genome-wide data. PLOS ONE 2(3):e337
27. Ng SK, Zhang Z, Tan SH, Lin K (2003) InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes. Nucleic Acids Res 31(1):251–254
28. Nguyen TT, Androulakis IP (2009) Recent advances in the computational discovery of transcription factor binding sites. Algorithms 2(1):582–605. doi:10.3390/a2010582
29. Oshchepkov DY, Levitsky VG (2011) In silico prediction of transcriptional factor-binding sites. In: Series. Methods in molecular biology, vol 760, pp 251–267. doi:10.1007/978-1-61779-176-5_16
30. Pan W, Wei P, Khodursky A (2008) A parametric joint model of DNA-protein binding, gene expression and DNA sequence data to detect target genes of a transcription factor. Pacific Symp Biocomput 13:465–476
31. Peng G, Luo L, Siu H, Zhu Y et al (2010) Gene and pathway-based second-wave analysis of genome-wide association studies. Eur J Hum Genet 18:111–117

32. Peri S, Navarro JD, Amanchy R et al (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res 13(10):2363–2371
33. Rhodes DR, Tomlins SA, Varambally S et al (2005) Probabilistic model of the human protein-protein interaction network. Nature Biotechnol 23:951–959. doi:10.1038/nbt1103
34. Schadt EE, Lamb J, Yang X et al (2005) An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet 37(7):710–717
35. Sinoquet C, Mourad R, Leray P (2012) Forests of latent tree models for the detection of genetic associations. In: International conference on bioinformatics models, methods and algorithms (Bioinformatics), 5–14
36. The Gene Ontology Consortium, Ashburner M, Ball CA, Blake JA et al (2000) Gene Ontology: tool for the unification of biology. Nat Genet 25(1):25–29. doi:10.1038/75556
37. Verzilli CJ, Stallard N, Whittaker JC (2006) Bayesian graphical models for genome-wide association studies. Am J Hum Genet 79:100–112
38. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P (2002) Comparative assessment of large-scale data sets of protein-protein interactions. Nature 417(6887):399–403
39. Wang W, Cherry JM, Nochomovitz Y, Jolly E, Botstein D, Li H (2005) Inference of combinatorial regulation in Yeast transcriptional networks: a case study of sporulation. Proc Natl Acad Sci USA 102:1998–2003
40. Wang K, Li M, Bucan M (2007) Pathway-based approaches for analysis of genomewide associations studies. Am J Hum Genet 81:1278–1283
41. Wei P, Pan W (2012) Bayesian joint modeling of multiple gene networks and diverse genomic data to identify target genes of a transcription factor. Ann Appl Stat 6(1):334–355
42. Wu X, Jiang R, Zhang MQ, Li S (2008) Network-based global inference of human disease genes. Mol Syst Biol 4:189
43. Xia K, Dong D, Han J-DJ (2006) IntNetDB v1.0: an integrated protein-protein interaction network database generated by a probabilistic model. BMC Bioinform 7:508. doi:10.1186/1471-2105-7-508
44. Xia JF, Wang S-L, Lei Y-K (2010) Computational methods for the prediction of protein-protein interactions. Protein Pept Lett 17(9):1069–1078
45. Yosef N, Sharan R, Stafford Noble W (2008) Improved network-based identification of protein orthologs. Bioinformatics 24(16):i200–i206
46. Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, Han JD, Bertin N, Chung S, Vidal M, Gerstein M (2004) Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. Genome Res 14(6):1107–1118
47. Zhu J, Zhang B, Smith EN, Drees B, Brem RB, Kruglyak L, Bumgarner RE, Schadt EE (2008) Integrating large-scale functional genomic data to dissect the complexity of Yeast regulatory networks. Nat Genet 40(7):854–861. doi:10.1038/ng.167