OPEN ACCESS Freely available online

Message From ISCB

Getting Started in Probabilistic Graphical Models

Edoardo M. Airoldi

Probabilistic graphical models (PGMs) have become a popular tool for computational analysis of biological data in a variety of domains. But, what exactly are they and how do they work? How can we use PGMs to discover patterns that are biologically relevant? And to what extent can PGMs help us formulate new hypotheses that are testable at the bench? This Message sketches out some answers and illustrates the main ideas behind the statistical approach to biological pattern discovery.

Introduction

Probabilistic graphical models offer a common conceptual architecture where biological and mathematical objects can be expressed with a common, intuitive formalism. This enables effective communication between scientists across the mathematical divide by fostering substantive debate in the context of a scientific problem, and ultimately facilitates the joint development of statistical and computational tools for quantitative data analysis. A number of success stories have appeared over the years [1–4]. Today, probabilistic graphical models promise to play a major role in the resolution of many intriguing conundrums in the biological sciences. The goal of this short article is to be a dense, informative introduction to *the language* of probabilistic graphical models, for beginners, with *pointers* to successful applications in selected areas of biology. The exposition introduces the essential concepts involved in PGMs in the context of the various stages of a typical collaboration between natural and computational scientists, and discusses the aspects to which each scientist should contribute to carry out the data analysis successfully using PGMs.

Let us start by considering a specific problem in transcriptional regulation. Given measurements about the abundance of gene transcripts in retinal cells across stages of development, we would like to discover which functional processes are relevant for development, and reveal which ones are most important at which stage. To develop a PGM to address this problem, we begin by identifying the biological objects that would appear in a cartoon model of how cellular development impacts transcription. In this illustrative example, we have genes and functional processes/contexts. It is reasonable to assume that each gene will participate in multiple functional processes, although typically in a small number of them, and that not all functional processes will be important at all stages of development. We then assess what aspects of the problem we can probe directly, with experimental techniques, and what aspects we cannot. In the example, while an abundance of gene transcripts can be obtained, for instance, via SAGE (serial analysis of gene expression), it is harder to measure functional processes. However, the latter could be operationally defined as sets of genes that share a similar temporal regulation pattern; this definition has the advantage of creating a connection between membership of genes to functional processes (i.e., an unobservable mapping) and similarity of the temporal expression profiles (i.e., observable quantities). The establishment of connections between those biological objects that we can probe and those that we cannot ends a first conceptual effort.

A cartoon model of how cellular development impacts transcription is now specified in terms of genes and their abundance, functional processes, and membership of genes to functional processes. Next we translate the biological players and the connections we established among them into mathematical quantities (i.e., random variables) and connections among them (i.e., statistical dependencies). This translation specifies the model structure. At this stage, we rely on biological intuitions to fine-tune the model, for instance, by deciding which sources of variability in the measurements carry information about the latent variables and which do not—if the temporal expression profiles of genes A and B are similar on a relative scale, but their absolute abundance is quite different, should we believe that they both participate in the same functional processes? Last, we assign numerical values to those quantities that are unknown in the final model specifications (i.e., we fit the model to the data) and we use them to develop biological intuitions in the context of the original problem. (Functional aspects of retinal development, in mouse, are fully addressed in [5].)

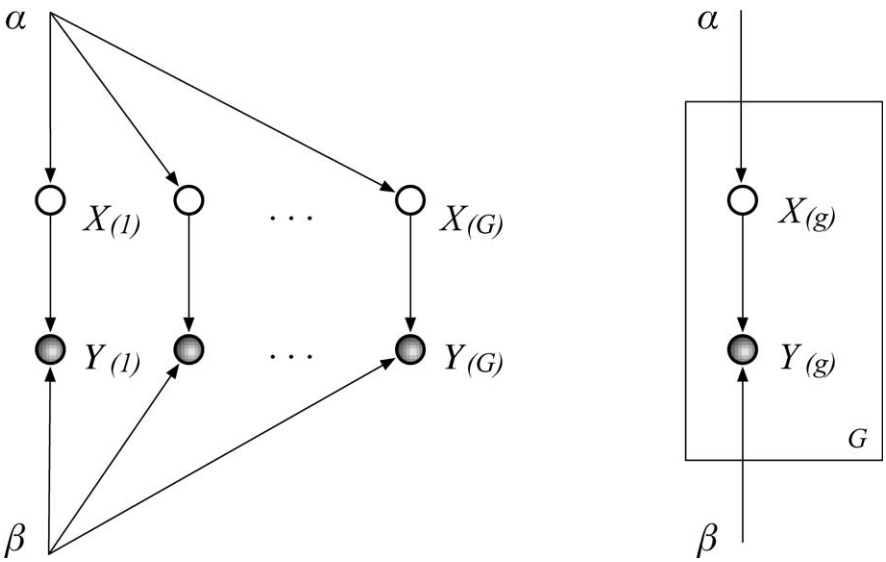
In the following, we briefly introduce the basic mathematical quantities that enable the translation of a cartoon model of biology into a PGM, and we

Editor: William Noble, University of Washington, United States of America

Citation: Airoldi EM (2007) Getting started in probabilistic graphical models. PLoS Comput Biol 3(12): e252. doi:10.1371/journal.pcbi.0030252

Abbreviations: EM, expectation–maximization; MCMC, Monte Carlo Markov chain; PGM, probabilistic graphical model

Edoardo M. Airoldi is with the Lewis-Sigler Institute for Integrative Genomics and the Computer Science Department, Princeton University, Princeton, New Jersey, United States of America. E-mail: eairoldi@princeton.edu



doi:10.1371/journal.pcbi.0030252.g001

Figure 1. Two Equivalent Representations of the Same Probabilistic Graphical Model
The left panel shows the full model, and the right panel shows the same model expressed in compact form. Nodes denote random variables; observed random variables are shaded while latent random variables are not; edges denote possible dependences. The box in the right panel is called a *plate*; it denotes independent and identically distributed replicates.

review strategies to assign numerical values to the unknown quantities underlying any PGM that are most likely given the observations. We conclude with an overview of selected applications, complete with pointers to published work.

The Basics

A probabilistic graphical model defines a family of probability distributions that can be represented in terms of a graph. Nodes in the graph correspond to random variables; its structure translates into statistical dependencies (among such variables) that drive the computation of joint, conditional, and marginal probabilities of interest [6]. In applications, most of the (node-specific) random variables are chosen to express the variability of an observed quantity, such as the expression of a specific gene measured under a certain condition. Some random variables, however, may specify unobserved quantities that are believed to influence the observable outcomes of a given experiment, such as which cellular processes were active at the time measurements were taken. The (directed or undirected) arcs of the graph specify the biological hypotheses about how observable and latent quantities influence one another. A set of constants underlying the distributions of the random variables completes the picture. These constants are referred to as *parameters* in the frequentist paradigm and as *hyper-parameters* in the Bayesian paradigm. (See [7], pp. 185–189, for a discussion of when the distinction matters in practice, with examples.)

Figure 1 shows an example of a probabilistic graphical model for gene expression. (We note that there is a considerable overlap between the class of probabilistic graphical models and the class of Bayesian networks. A number of scholars choose to refer to PGMs that *can be represented as directed acyclic graphs, with nodes corresponding to discrete-valued random variables, encoding observed measurements, and no latent variables* as Bayesian networks.) The observed expression of a gene, $Y(g)$, depends on the latent functional process it is involved in, $X(g)$. The underlying constants, (α,β), control the probability that any given functional process is active and the probability of observing expression of a certain magnitude, respectively. The left panel shows the full model, and the right panel shows the same model expressed in compact form.

The *likelihood function*, or the probability of the measurements given the underlying constants, is the main quantity of interest in PGMs. It summarizes how well the observations are explained by the specific PGM that is identified by a given value of the underlying constants. The likelihood can be computed using the structural hypotheses encoded by the graph, and the probability distributions specified for the nodes. Continuing the example, the likelihood corresponding to the model in Figure 1 is computed as follows:

$$Pr(Y|\,\alpha,\beta) = \int_x \Pr(Y, X|\,\alpha,\beta)dX \quad (1)$$

$$= \int_X \prod_{g=1}^{G}[Pr(Y(g)|X(g),\beta) \cdot \Pr(X(g)|\,\alpha)]dX \quad (2)$$

$$\equiv \ell(Y|\,\Theta), \quad (3)$$

for $\Theta \equiv (\alpha,\beta)$. The joint probability of measurements and latent variables given the underlying constants, that is, the integrand on the right-end side of Equation 1, is often referred to as the *complete likelihood function* in the literature—an important quantity in the statistical treatment of PGMs with latent variables.

Estimation and Inference

A family of PGMs is *fit to the data* to find likely values for its underlying constants and likely distributions for its latent variables. This process boils down to an optimization problem where the objective function is based on the likelihood. Considered jointly, the estimation and inference tasks identify a specific model in the family of PGMs that is defined by the assumptions on the graph and the random variables, which successfully summarizes the variability of the observations.

In the language of the statistical literature, we distinguish the task of *estimating* the underlying constants (i.e., the parameters in a frequentist statistical setting, or the hyper-parameters in a Bayesian statistical setting) of a probabilistic graphical model, from the task of *inferring* the distributions of the latent variables given the observations. Let us consider strategies to address the latter task first. The choice among the many strategies available is often informed by the

complexity of the model, and in particular by whether the integral on the right-end side of Equation 1 can be computed in closed form. Exact inference is available for models that belong to special families [6]. Focusing on the biology of the problem, however, often leads to a model structure and probabilistic specifications that cannot be subsumed under any special family. The likelihood is intractable in many such cases—that is, the integral in Equation 1 cannot be solved in closed form—and we resort to approximations. Below, we briefly survey the intuitions behind three popular strategies to perform approximate inference in PGMs: Monte Carlo Markov chains (sampling-based), and expectation–maximization (EM) and variational methods (optimization-based).

Monte Carlo Markov chain (MCMC) techniques such as the Gibbs or Metropolis-Hastings samplers can be used to explore the joint posterior distribution of the latent variables [8,9]. Although the likelihood is intractable, the complete likelihood $Pr\ (Y,X \mid \alpha,\beta)$ can be easily computed for the large majority of PGMs. The main concept behind MCMC schemes is to work with the complete likelihood, and to reduce the full joint posterior to lower-dimensional conditional distributions—on individual, or blocks of latent variables—that we can sample from. Samples from the joint posterior are then obtained by composing conditional samples. The Gibbs sampler, for instance, requires that one can sample from all univariate, full-conditional distributions:

$$Pr(X(g)|X(-g), Y, \alpha, \beta), \text{for } g = 1, ..., G, \tag{4}$$

where $X(-g)$ is the collection of random variables X without $X(g)$. The Metropolis-Hastings sampler requires that one can at least compute a quantity proportional to the desired posterior—samples are drawn from an arbitrary *proposal distribution* and are accepted or rejected using a formula that depends on the proposal. Other sampling-based algorithms such as particle filters can be used to perform inference in PGMs of sequential observations [10].

The two alternatives to sampling we survey here aim at approximating the integral on the right-end side of Equation 1. The main idea shared by both approaches is to find a lower bound for the likelihood, $\ell\ (Y\ |\Theta)$, making use of Jensen's inequality and of an arbitrary distribution on the latent variables $q(X)$:

$$\log \ell\ (\mathrm{Y}|\Theta) = \log \int_x \Pr(\mathrm{Y}, \mathrm{X}|\alpha, \beta)\mathrm{dX}$$

$$= \log \int_x \mathrm{q}(\mathrm{X}) \cdot\ \Pr\ (\mathrm{Y}, \mathrm{X}|\alpha, \beta)/\mathrm{q}(\mathrm{X})\mathrm{dX}$$

$$\text{(for any q)}$$

$$\geq \int_x \mathrm{q}(\mathrm{X}) \cdot\ \log \Pr(\mathrm{Y}, \mathrm{X}|\alpha, \beta)/\mathrm{q}(\mathrm{X})\mathrm{dX}$$

$$\text{(Jensen's inequality)}$$

$$= \mathrm{E_q}[\log \Pr(\mathrm{Y}, \mathrm{X}|\Theta)] - \log \mathrm{q}(\mathrm{X}) \equiv \mathrm{L}(\mathrm{q}, \Theta) \tag{5}$$

In EM, the lower bound $\mathrm{L}\ (q, \Theta)$ is iteratively maximized with respect to Θ, in the M step, and q in the E step [11]. In particular, at the t-th iteration of the E step the q distribution must satisfy the following equation:

$$q^{(\mathrm{t})} = Pr(X|Y, \Theta^{(t-1)}), \tag{6}$$

That is, we set the arbitrary distribution q equal to the posterior distribution of the latent variables given the data and the estimates of the parameters at the previous iteration. Unfortunately, it is not always possible to express the distribution $q^{(t)}$ in Equation 6 in analytic form. In such cases, a variational approximation to the EM [12] can be obtained by defining a parametric approximation to the posterior in Equation 6, denoted by $\tilde{q} \equiv q_\Delta\ (X)$, which involves an extra set of *variational parameters,* Δ, and leads to an approximate lower bound for the likelihood $L_\Delta\ (q, \Theta)$. At the t-th iteration of the E step, we then minimize the Kullback-Leibler divergence between $q^{(t)}$ and $q_\Delta^{(t)}$, with respect to Δ, using the data—this is equivalent to maximizing the approximate lower bound for the likelihood, $L_\Delta\ (q, \Theta)$ with respect to Δ. The optimal parametric approximation can be thought of as an approximate posterior distribution for the latent variables in the sense that it depends on the data Y, although indirectly, $q^{(t)} \approx q^{(t)}_{\Delta^*(Y)}(X) = Pr\ (X \mid Y)$.

Let us now return to the task of estimating the constants underlying a PGM; few established strategies exist. The estimates for the underlying constants may be chosen, for instance, to maximize the likelihood, or to match empirical and theoretical moments of the random variables that correspond to measurements ([7], pp. 120–124). Alternatively, when the likelihood is too difficult or expensive to compute, an approximation, $L_\Delta \approx \ell$, or a lower bound, $L \leq \ell$, for the likelihood can be used as a surrogate. These alternatives and others are sometimes referred to as empirical Bayes estimates in the context of nontrivial probabilistic graphical models ([13], Chapter 3).

Popular software packages that implement a language to specify and fit PGMs are available. For MCMC, see BUGS [14]; for variational inference, see VIBES [15].

Applications

With the technical machinery we just introduced, we are now ready to bring the biological intuition back into the picture. Let us continue with the transcriptional regulation example. In the PGM of Figure 1, the expression of gene g may be encoded by a real-valued random variable $Y(g)$. The mixed membership of gene g to nonobservable biological contexts may be encoded by the nonzero components of a latent random vector, $X(g)$. The number of latent biological contexts we ask the PGM to infer, denoted by K, is an important quantity in this model, which we discuss later—briefly, the value of K specifies the *dimensionality* of this PGM, that is, the number of components of the vector-valued latent variables, $X(g)$. The two constants (α,β) may be used to encode biological constraints. For instance, α may be used to introduce a notion of biological parsimony in the form of a probabilistic (soft) constraint on the number of biological contexts each gene may participate in, and β may be used to specify gene expression patterns in the form of differential expression levels across those experimental conditions for which microarray measurements were taken—alternative pattern specifications and parameterizations exist [5]. For any given number of latent biological contexts, K, the PGM is fit to the data. Estimation and

inference will assign numerical values to the unknown quantities $(\vec{X},\alpha,\beta)$. These quantities provide us with *model-based* and *observation-induced* summaries of the data. In the example, for instance, while β summarizes gene expression patterns that summarize the main trends of transcription in a collection of microarrays, the values assigned to the latent variables, $X(g)$, provide gene-specific information that can be used for making fine-grained predictions.

In the last stage of the analysis, we assess the biological relevance of the patterns we inferred from the data (such as the biological contexts, or gene coexpression patterns, in the example) to make sure the model is capturing the signal we set out to capture, and we use the inferred patterns to gain insights into the problem. Assessment of biological relevance can be qualitative or quantitative. Qualitative methods such as visual inspection are typically useful for focused scientific endeavors; for instance, whenever a biological problem targets a small set of genes or a specific cellular process or component, or a signaling pathway. Quantitative methods are necessary for genome-wide scientific endeavors, and typically rely on knowledge-based repositories and ontologies (such as gene ontology [16]) and bioinformatics tools to carry out the evaluation [17,18]. Arguably, in any given application, the more interpretable the patterns are, in terms of functional processes and other biological concepts of interest, the better the family of PGMs captures some aspects of biology that may be relevant for the understanding of the phenomenon under investigation, and that *are not directly measurable* with experimental techniques.

Moving a step forward, the goodness of model fit is often taken as a measure of *how well* the data support structural biological hypotheses encoded by the cartoon model of biology that was used to posit a given family of PGMs. Measures of goodness of model fit include the Bayesian information criterion, the held-out likelihood obtained using bootstrap or cross-validation techniques, measures of predictive power such as the predictive R^2 in linear regression, or other quantities, depending on the goals of the analysis. (These measures can also be used to select the dimensionality, K, of the PGM in the example.) The goodness of fit, along with the substantive value of the inferred patterns, should inform a critical review of the biological assumptions underlying the initial cartoon model, and possibly suggest new hypotheses—testable either with new statistical analyses, or with new experimental probes at the bench. In this sense, probabilistic graphical models contribute to an iterative process of scientific discovery, where statistical and biological thinking are intertwined as both cause and effect.

There is a rich history of applied research that leverages the probabilistic graphical models approach outlined above to problems in the biological sciences. It includes a model for inferring the ancestral population structure of individuals starting from a collection of multilocus genotype measurements [2] and a model for inferring HIV mutation patterns from longitudinal clonal sequence data [19]; the former model is closely related to the classic probabilistic graphical models to infer phylogenetic trees [1,20] and to recent extensions, in particular, that take into account the dependence among the bases at neighboring sites [21,22]. Models for sequence analysis are well-established in the community [4,23]; more recently, the connection between sequence information and gene expression has been investigated using probabilistic graphical models as well [24,25]. Other applications of this research include: a model for predicting the clinical status of breast cancer using gene expression profiles [26]; a model for facilitating content browsing of biomedical literature about the nematode *Caenorhabditis elegans* [27]; a model for inferring the location of chromosome aberrations from array-based comparative genomic hybridization measurements [28], and an extension that leverages array-based comparative genomic hybridization profiles from multiple individuals to recover shared aberration patterns [29]; a model for reconstructing features of the internal organization of the cell from the nested structure of observed perturbation effects, such as those measured via high-dimensional phenotype screens [30]; a model for inferring proteins' multiple functional roles from a large collection of manually curated protein interactions, as well as cross-talk patterns among proteins that participate in distinct functional processes [31]; and a model for inferring temporal patterns of coexpressed genes from time-course expression data measured via SAGE and microarray technologies [5].

Note that the graphical representation of a family of PGMs goes only so far in specifying the model; it's informative, but not exhaustive. Probabilistic assumptions and some features of the sampling scheme cannot be specified by the graph. Such subtle variants typically make a significant difference in applications.

Conclusions

Probabilistic graphical models offer a common conceptual architecture where biological and mathematical objects can be expressed with a common, intuitive formalism. This enables effective communication between scientists across the mathematical divide by fostering substantive debate in the context of a scientific problem, and ultimately facilitates the joint development of statistical and computational tools for quantitative data analysis. In other words, probabilistic graphical models provide a bridge between biology and statistical computations. These models recently earned a spot at the center stage of modern (computational) biology by furthering our ability to probe data for biological hypotheses, and will undoubtedly play an important role in resolving many intriguing conundrums in the biological sciences, in the future. ■

Acknowledgments

The author thanks Florian Markowetz, Chad Myers, David Hess, and Olga Troyanskaya at Princeton University, and Eric Xing at Carnegie Mellon University, for comments on an early draft of this manuscript.

Author contributions. EMA wrote the paper.

Funding. This research was partly supported by United States National Institute of General Medical Sciences Center of Excellence grant P50 GM071508, by National Science Foundation grants DBI-0546275 and IIS-0513552, and by National Institutes of Health grant R01 GM071966.

Competing Interests. The author has declared that there are no competing interests.

References

1. Felsenstein J (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. J Mol Evol 17: 368–376.
2. Pritchard J, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945–959.
3. Friedman N (2004) Inferring cellular networks using probabilistic graphical models. Science 303: 799–805.
4. Xing EP, Karp RM (2004) MotifPrototyper: A profile Bayesian model for motif family. Proc Natl Acad Sci U S A 101: 10523–10528.
5. Airoldi EM, Fienberg SE, Xing EP (2006) Mixed membership analysis of expression studies: Attribute data. Available: http://arxiv.org/abs/0711.2520/. Accessed 20 November 2007.
6. Jordan MI (2004) Graphical models. Statistical Science 19: 140–155.
7. Wasserman L (2004) All of statistics. New York: Springer-Verlag.
8. Gelman A, Carlin J, Stern H, Rubin D (1995) Bayesian data analysis. London: Chapman & Hall.
9. Robert C, Casella G (2005) Monte Carlo statistical methods. Springer texts in statistics. Corrected second edition. New York: Springer-Verlag.
10. Liu JS (2001) Monte Carlo strategies in scientific computing. New York: Springer-Verlag.
11. Dempster A, Laird N, Rubin D (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc [Series B] 39: 1–38.
12. Jordan M, Ghahramani Z, Jaakkola T, Saul L (1999) Introduction to variational methods for graphical models. Machine Learning 37: 183–233.
13. Carlin BP, Louis TA (2005) Bayes and empirical Bayes methods for data analysis. Second edition. London: Chapman & Hall.
14. Lunn DJ, Thomas A, Best NG, Spiegelhalter DJ (2000) WinBUGS: A Bayesian modeling framework: Concepts, structure and extensibility. Statistics and Computing 10: 321–333. Available: http://www.mrc-bsu.cam.ac.uk/bugs/. Accessed 8 November 2007.
15. Bishop C, Spiegelhalter D, Winn J (2003) VIBES: A variational inference engine for Bayesian networks. In: Becker S, Thrun S, Obermayer K, editors. Advances in neural information processing systems 15. Cambridge (Massachusetts): MIT Press. pp. 777–784. Available: http://vibes.sourceforge.net/. Accessed 8 November 2007.
16. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: Tool for the unification of biology. The gene ontology consortium. Nat Genet 25: 25–29.
17. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, et al. (2004) GO::TermFinder—Open source software for accessing Gene Ontology terms associated with a list of genes. Bioinformatics 20: 3710–3715.
18. Myers CL, Barret DA, Hibbs MA, Huttenhower C, Troyanskaya OG (2006) Finding function: An evaluation framework for functional genomics. BMC Genomics 7: 187.
19. Beerenwinkel N, Drton M (2007) A mutagenetic tree hidden Markov model for longitudinal clonal HIV sequence data. Biostatistics 8: 53–71.
20. Felsenstein J, Churchill GA (1996) A hidden Markov model approach to variation among sites in rate of evolution. Mol Biol Evol 13: 93–104.
21. McAuliffe JD, Pachter L, Jordan MI (2004) Multiple-sequence functional annotation and the generalized hidden Markov phylogeny. Bioinformatics 20: 1850–1860.
22. Siepel A, Haussler D (2004) Combining phylogenetic and hidden Markov models in biosequence analysis. J Comput Biol 11: 413–428.
23. Durbin R, Eddy S, Krogh A, Mitchison G (1998) Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge: Cambridge University Press.
24. Segal E, Yelensky R, Koller D (2003) Genome-wide discovery of transcriptional modules from dna sequence and gene expression. Bioinformatics 19 (Supplement 1): i273–282.
25. Beer MA, Tavazoie S (2004) Predicting gene expression from sequence. Cell 117: 185–198.
26. West M, Blanchette C, Dressman H, Huang E, Ishida S, et al. (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. Proc Natl Acad Sci U S A 98: 11462–11467.
27. Blei DM, Franks K, Jordan MI, Mian IS (2006) Statistical modeling of biomedical corpora: Mining the Caenorhabditis genetic center bibliography for genes related to life span. BMC Bioinformatics 7: 250.
28. Myers CL, Dunham MJ, Kung SY, Troyanskaya OG (2004) Accurate detection of aneuploidies in array CGH and gene expression microarray data. Bioinformatics 20: 3533–3543.
29. Shah SP, Lam WL, Ng RT, Murphy KP (2007) Modeling recurrent DNA copy number alterations in array CGH data. Bioinformatics 23: i450–i458.
30. Markowetz F, Kostka D, Troyanskaya OG, Spang R (2007) Nested effects models for high-dimensional phenotyping screens. Bioinformatics 23: i305–i312.
31. Airoldi EM, Blei DM, Fienberg SE, Xing EP (2006) Mixed membership analysis of high-throughput interaction studies: Relational data. Available: http://arxiv.org/abs/0706.0294/. Accessed 20 November 2007.