Bayesian methods in bioinformatics and computational systems biology

D. J. Wilkinson*

January 27, 2007

Abstract

Bayesian methods are valuable, *inter alia*, whenever there is a need to extract information from data that is uncertain or subject to any kind of error or noise (including measurement error and experimental error, as well as noise or random variation intrinsic to the process of interest). Bayesian methods offer a number of advantages over more conventional statistical techniques that make them particularly appropriate for complex data. It is therefore no surprise that Bayesian methods are becoming more widely used in the fields of genetics, genomics, bioinformatics and computational systems biology, where making sense of complex noisy data is the norm. This review provides an introduction to the growing literature in this area, with particular emphasis on recent developments in Bayesian bioinformatics relevant to computational systems biology.

Keywords: Bayesian inference; computational systems biology; networks; graphical models; quantitative, predictive biology.

Darren Wilkinson is a Senior Lecturer in Statistics within the School of Mathematics & Statistics at Newcastle University. He has a background in computational Bayesian statistics, and in recent years has become increasingly interested in applications to statistical bioinformatics and computational systems biology. At Newcastle he is a member of the Centre for Integrated Systems Biology of Ageing and Nutrition (CISBAN) and the Systems Biology Resource Centre (SBRC).

*School of Mathematics & Statistics, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK. Tel: +44-191-2227320; Email: d.j.wilkinson@ncl.ac.uk

1 Introduction

Bioinformatics and computational systems biology are undergoing a Bayesian revolution similar to that already seen in genetics [1]. The reason is the same — biology is complex, and data is noisy. Traditional statistical techniques struggle to cope with complex nonlinear models that are only partially observed. Due to the fact that the Bayesian statistical paradigm is fully probabilistic, there is no fundamental distinction between any of the unknowns in a statistical model — parameters, hidden variables and observations are all treated together in a consistent manner — and it is from this that the power of the methodology is derived [2]. Provided that you can write down a statistical model relating the quantities you are interested in to the data you can observe (possibly via many unobserved intermediary variables), then you can (in principle) carry out Bayesian inference to extract the information in the data to give fully probabilistic information on all unobserved model variables. The main limiting factor in applying Bayesian methods is computational. For non-trivial problems, analytic approaches to Bayesian inference are not possible, and their numerical solution is often challenging due to the need to solve high-dimensional integration problems (which in the discrete case translate to combinatorial summation problems). Advances in the speed of commodity computing hardware in recent decades has been paralleled by developments in computationally intensive algorithms for Bayesian inference. Arguably the most important advance has been the development of a range of techniques based on Markov chain Monte Carlo (MCMC). The ideas originate from statistical physics [3], but are now widely used for Bayesian inference [4, 5]. Although by no means a panacea, carefully crafted MCMC algorithms executed on fast computers are able to solve a phenomenal range of problems that would have been considered completely intractable only a few years ago.

In the simplest (continuous) setting, we are interested in making inferences about the parameter vector ϕ of a probability (density) model $p(y|\phi)$ giving rise to an observed data vector y. If we treat the parameters as uncertain, and allocate to them a "prior" probability density $\pi(\phi)$, then Bayes theorem gives the "posterior" density

$$\pi(\phi|y) = \frac{\pi(\phi)p(y|\phi)}{p(y)},$$

where $p(y)$ is the marginal density for y obtained by integrating over the prior. Since $\pi(\phi|y)$ is regarded as a function of ϕ for fixed (observed) y, we can re-write this as

$$\pi(\phi|y) \propto \pi(\phi)p(y|\phi),$$

so that the posterior is proportional to the prior times the likelihood. Practical complications arise due to the fact that typically the normalising constant $p(y)$ is not known, and either $p(y|\phi)$ will not be known explicitly or marginalisation over some components of ϕ will be required. Whilst analytically intractable, these integration problems are typically amenable to a Monte Carlo or MCMC solution. In the high-dimensional context, it is often necessary to decompose the full problem according to the underlying conditional independence structure of the model, and it is in this context that *graphical models* [6] (also known as *conditional independence graphs*) are particularly useful. In non-statistical communities, the term *Bayes(ian) network* is often used to describe a discrete graphical model. However, it is important to note that graphical models can be used to describe any probabilistic conditional independence structure, and that many of the techniques that are often used to "learn" Bayesian networks are not Bayesian.

The simplest example of a MCMC method is the Gibbs sampler [7, 8]. Here a Markov chain is constructed with equilibrium distribution $\pi(\phi|y)$. Each iteration of the sampler involves cycling through each component of the p-dimensional vector ϕ in order and sampling from $\pi(\phi_i|\phi_{-i}, y)$, $i = 1, \ldots, p$, where ϕ_{-i} denotes the vector of all components of ϕ except ϕ_i. Knowledge of the conditional independence graph for the model can simplify the computation of these so-called *full-conditional* distributions. In many cases the full-conditionals will be straightforward to sample directly, but in others, a Metropolis-Hastings method will be required [9, 10]. Here a proposed new value is simulated from a largely arbitrary *proposal distribution*, $q(\phi_i^\star|\phi_i)$ and accepted with a probability carefully chosen to preserve the *detailed balance* of the chain. Many practical details of the method are presented in [11, 12].

2 Bioinformatics

2.1 Biological sequence analysis

One of the first areas to benefit from the application of Bayesian approaches was biological sequence analysis. Here it had already been recognised that working with probabilistic models was extremely useful [13]. Whilst for some simple hidden Markov models (HMMs) it is possible to estimate parameters using conventional statistical techniques (such as maximum likelihood via the EM algorithm) [14, 15], there are many interesting problems where a conventional approach would be inconvenient or unsatisfactory in terms of the information provided by the analysis; see [16] for a good introduction to the use of Bayesian methods in this area. Good examples of this include simultaneous multiple sequence alignment [17, 18], motif discovery and transcription factor binding

site prediction [19, 20] and protein secondary structure prediction [21]. One of the key benefits of the Bayesian approach is that it allows proper propagation of uncertainty across different levels of modelling. So whilst a traditional approach to phylogeny estimation would use a pre-calculated multiple alignment, uncertainty in the alignment will not propagate through to uncertainty in the phylogeny. In fact the converse is also true: models for alignment depend implicitly on an assumed phylogeny, so uncertainty in phylogeny induces alignment uncertainty. Using a Bayesian approach, simultaneous estimation is possible [22]. Even in the relatively simple context of HMM-based *ab initio* DNA sequence segmentation, the Bayesian approach enables the convenient inclusion of prior information, and provides much richer information about the model parameters [23]. Further, since uncertainty about model structure is treated consistently with parameter uncertainty in the Bayesian context, variable dimension algorithms such as reversible jump MCMC (RJMCMC) [24] can be used to estimate the number of segments and the order of the base dependence along with all other aspects of the model [25]. Liu and Logvinenko [26] provide a detailed review of Bayesian methods in sequence analysis.

2.2 Microarray data analysis

The analysis of gene microarray data [27] is another area where Bayesian methods have proven to offer many advantages over more conventional approaches [28, 29]. Although amenable to simple statistical analyses such as ANOVA, microarray data analysis is often broken down into a collection of distinct steps that fail to correctly propagate uncertainty. For example, a typical analysis may begin with some kind of normalisation process that produces "corrected" expression levels. These normalised data will then be subject to a secondary statistical analysis (such as identification of differentially expressed genes) that ignores any uncertainty in the normalisation processes. Often then the differentially expressed genes will be used for a further analysis that ignores the uncertainty in the identification procedure. Using Bayesian techniques it is possible to develop integrated models for the analysis of unnormalised cDNA microarray data that correctly propagate uncertainty across the various levels of analysis [30, 31]. Detailed modelling combined with a carefully designed experiment can allow coherent estimation of absolute transcript concentrations from cDNA array data [32, 33]. It is also much more convenient to pool information across multiple experiments and studies using a Bayesian approach [34]. For Affymetrix GeneChip data, developing probabilistic models of the hybridisation process down at the probe level again allows extraction of information likely to be missed using simpler stepwise approaches [35, 36]. Bayesian methods also offer advantages when clustering of expression profiles is felt to be relevant [37, 38, 39]. In fact, the initial task of segmentation and raw intensity estimation can also benefit from a Bayesian approach [40]. Further modelling approaches and applications are discussed in [41, 42, 43, 44, 45, 46]. Some recent developments in the field are described in [29], which also covers some proteomic applications.

2.3 Protein informatics

There are many applications of Bayesian techniques to problems in protein informatics. Down at the structure level, Bayesian techniques for site matching and alignment have been shown to be particularly valuable [47, 48, 49]. A Bayesian method for predicting protein–protein interactions from genomic data is given in [50]. Mass spectrometry data are widely used for understanding the peptide/protein composition of a sample, but these data are subject to many sources of variation, making Bayesian approaches to data analysis highly desirable. Some methods for processing "raw" spectra are discussed in [51, 52] in the volume [29]. Bayesian methods can also be useful in the context of mass spectrometry clustering and classification [53, 54], as well as protein identification [55, 56].

3 Computational systems biology

3.1 Introduction

The analysis of micro-array data is also central to much research in computational systems biology, although here the emphasis is slightly different. A major concern of computational systems biology is the development of dynamic predictive models of biological (especially genetic and biochemical) processes [57]. The first stage in this process is the identification of interacting partners (used in a loose sense). One approach to identifying gene–gene interactions is to attempt to use observed correlations in gene microarray data to infer networks of interaction.

3.2 Network inference

A variety of different approaches to network inference are possible, and many widely used techniques are fundamentally Bayesian in nature. Again, it is worth emphasising the apparent confusion between discrete Bayesian networks and more general Bayesian methods. The term "Bayes net" is generally used in non-statistical communities to refer to discrete probabilis-

tic graphical models, irrespective of whether the techniques used to analyse them are Bayesian. Despite some suggestions to the contrary in the literature, there is no need to discretise continuous data in order to learn a Bayesian network — only to learn a *discrete* Bayes net. As mentioned above, graphical models can be estimated without using Bayesian methods, but there are advantages in doing so. This is particularly true when the number of observations is small compared to the number of variables, which is typically the case in the context of microarray data analysis.

An early, influential paper on Bayesian networks for expression data was [58]; also see [59] for a more recent perspective. An approach based on manipulation experiments for inferring directed networks is described in [60]. An efficient method for inferring undirected Gaussian graphical models is described in [61]. More recently, a detailed comparison of various methods for static network inference has been carried out in [62]. Such methods do not have to be based on micro-array data. Typically, using more quantitative data on a (small) system of interest will lead to more reliable conclusions. Single-cell flow cytometry data is potentially useful in this context, and a strategy to using this for inferring network structure is described in [63]. It should be pointed out, however, that most of these papers are not especially Bayesian in their approach. More Bayesian approaches to the problem of inferring sparse undirected (Gaussian) graphical models are described in [64] and [65], based on earlier work for graphical Gaussian model selection [66], and these are likely to provide more robust inferences in high dimensional settings, particularly since most methods are able to provide marginal posterior probabilities for the presence of individual network edges.

Time-course expression data provide some information about system dynamics, and therefore dynamic network models provide a useful starting point for top-down systems biology modelling. Dynamic Bayesian networks (DBNs) have been widely used in this context; see [67, 68] for details. For dynamic networks based on linear Gaussian models a fast "Bayesian-inspired" algorithm has recently been proposed [69]. As for static networks, fully Bayesian approaches to this problem are likely to offer significant advantages, and are currently the subject of ongoing research.

Using Bayesian inference for integrating multiple sources of data offers great potential, but currently remains largely unexplored; see [70, 71, 72] for initial attempts and perspectives.

3.3 Quantitative network models

As has already been stated, a key aim of systems biology is to develop quantitative, dynamic models of biological processes of interest. One approach to this problem is to extend the top-down network models so that they provide some quantitative information regarding dynamics [73]. However, this approach has some shortcomings due to the fact that the elements of the model do not link directly to physical parameters of interest. There is therefore great interest in a different approach, based on using data to parameterise bottom-up mechanistic models of biological processes. Obviously, non-Bayesian approaches to this problem are possible [74, 75, 76], but are limited in terms of the information they can provide. Even in the context of deterministic models of biochemical networks based on ordinary differential equations (ODEs), there is considerable utility in using a Bayesian approach in order to properly address issues of noise modelling and parameter uncertainty [77, 78]. It is also possible to improve parameter estimation using proper prior modelling of parameter uncertainty [79].

A nice application of Bayesian modelling in the context of quantitative modelling is the Characterizing Loss Of Cell Cycle Synchrony (CLOCCS) model [80] for loss of synchrony in yeast populations. A simple application of this model is in the alignment of data sets collected under different conditions. However, this model can also be combined with population level data (such as gene expression array data) in order to recover information about single-cell dynamics from the population averaged data. This detailed modelling of both the process of interest and its relationship with the experimental data is a powerful technique in this context, and similar strategies are likely to lead to many other examples of extracting better information from high-throughput data.

There is increasing evidence that stochasticity plays an important role in intra-cellular processes [81], and there is therefore a great deal of interest in developing stochastic kinetic models of biological processes [82, 83, 84, 85]. Further, experimental technology is improving rapidly, so that (semi-)quantitative high-resolution single-cell data of the type that is most informative for the building of stochastic models is now realistically attainable [86]. Typically data is generated via fluorescence microscopy, then processed to extract gene expression time series [87]. Although fully-Bayesian approaches to this image-analysis step are likely to be extremely useful, such an techniques do not yet seem to have been described in the literature. Stochastic kinetic models are particularly difficult to estimate using non-Bayesian methods. A valiant attempt is described in [88], but the applicability of the methods described is limited due to the extent to which non-Bayesian methods can cope with hidden data. In particular, the parsimony assumptions that are typically required have the effect of downward-biasing of parameter estimates. However, whilst a fully Bayesian approach to inference for dis-

crete stochastic models is possible [85, 89], it is computationally problematic for models of realistic size and complexity. Also see [90] for a related approach. It turns out to be possible to instead work with a continuous (approximate) formulation of stochastic kinetics, known as the "chemical Langevin equation" [91, 85]. This model seems to be quite adequate for inferential purposes, and is advantageous due to the fact that inference for this diffusion approximation is more computationally amenable than for the discrete formulation. A basic inferential algorithm for this model is described in [92]. A better algorithm for models of this type, based on ideas of sequential Monte Carlo [93], is developed in [94], and applied to a general and flexible class of stochastic kinetic models in [95]. Finally, an efficient non-sequential MCMC algorithm for stochastic kinetic models is described in [96]. A recent review of fitting models to data by Jaqaman & Danuser [97] includes references to both the Bayesian and non-Bayesian literature.

There is another area of statistical methodology that has obvious applications to systems biology modelling: Bayesian analysis of computer code outputs (BACCO) [98]. Here, a complex (but typically, deterministic) computer simulation model is treated as a "black-box" from a statistical perspective, and the relationships between model inputs, outputs and experimental data are studied in a non-parametric way, often utilising Gaussian processes [99]. Although these techniques do not yet seem to have been applied to systems biology modelling problems, they have been applied to challenging problems in other application areas [100, 101], so it seems inevitable that as systems biology models become larger and more complex, and BACCO techniques become more sophisticated (better suited to high-dimensional inputs and outputs, and intrinsic stochasticity in the computer models), that applications of BACCO methods to problems in computational systems biology will become commonplace.

4 Discussion

It is impossible in an article of this nature to give a fully comprehensive review of all Bayesian work in bioinformatics. Here the focus has been on work which clearly demonstrates the advantages of the Bayesian approach, and that which is most directly relevant to the new science of computational systems biology. Of course this latter area is still an emerging field, and it is not yet clear which (if any) of the methods and techniques described here will stand the test of time. The main drawback of fully Bayesian methods are the computational demands associated with their computer implementation. This has so far limited their application to certain challenging problems in the bioinformatics arena (such as whole-genome annotation). The Bayesian framework provides a coherent mathematical solution to the problem, but not always an efficient computational algorithm for practical implementation. Even in difficult scenarios, however, probabilistic statistical models (such as Hidden Markov Models) are becoming the accepted framework for analysis [13], and used in conjunction with point estimation methods (such as the EM algorithm) for parameter fitting. However, experience from closely related disciplines suggests that fully Bayesian approaches will turn out to provide the most satisfactory solutions to the complex statistical inference problems which lie at the heart of computational systems biology. Improvements in computing hardware, the widespread availability of parallel computer clusters, and the development of computational Bayesian algorithms that are able to exploit them [102], mean that there is likely to be an increasing tendency to push for fully Bayesian solutions to the challenging inferential problems in this area, in order to maximise the information that can be extracted from expensive experimental data.

References

[1] M. A. Beaumont and B. Rannala. The Bayesian revolution in genetics. *Nature Reviews Genetics*, 5(4):251–261, 2004.

[2] A. O'Hagan and J. J. Forster. *Bayesian Inference*, volume 2B of *Kendall's Advanced Theory of Statistics*. Arnold, London, 2004.

[3] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

[4] D. Gamerman. *Markov Chain Monte Carlo*. Texts in Statistical Science. Chapman and Hall, New York, 1997.

[5] S. P. Brooks. Markov chain Monte Carlo method and its application. *The Statistician*, 47(1):69–100, 1998.

[6] S. L. Lauritzen. *Graphical Models*. Oxford Science Publications, Oxford, 1996.

[7] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[8] G. Cassella and E. I. George. Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

[9] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[10] L. Tierney. Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics*, 21:1701–1762, 1994.

[11] C. J. Geyer. Practical Markov chain Monte Carlo. *Statistical Science*, 7:473–511, 1992.

[12] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC Press, second edition, 2003.

[13] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, 1998.

[14] M. J. Bishop and E. A. Thompson. Maximum likelihood alignment of DNA sequences. *Journal of Molecular Biology*, 190:159–165, 1986.

[15] G. A. Churchill. Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology*, (51):79–94, 1989.

[16] J. S. Liu and C. E. Lawrence. Bayesian inference on biopolymer models. *Bioinformatics*, 15(1):38–52, 1999.

[17] J. S. Liu, A. F. Neuwald, and C. E. Lawrence. Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *Journal of the American Statistical Association*, 90:1156–1170, 1995.

[18] J. Liu, A. Neuwald, and C. Lawrence. Markovian structures in biological sequence alignments. *Journal of the American Statistical Association*, 94:1–15, 1999.

[19] Q. Zhou and J. S. Liu. Modelling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, 20(6):909–916, 2004.

[20] L. Narlikar, R. Gordân, U. Ohler, and A. J. Hartemink. Informative priors based on transcription factor structural class improve de novo motif discovery. *Bioinformatics*, 22:e384–e392, July 2006. ISMB06.

[21] S. C. Schmidler, J. S. Liu, and D. L. Brutlag. Bayesian segmentation of protein secondary structure. *Journal of Computational Biology*, 7:233–248, 2000.

[22] G. Lunter, I. Miklós, A. Drummond, J. L. Jensen, and J. Hein. Bayesian coestimation of phylogeny and sequence alignment. *BMC Bioinformatics*, 6(83), 2005.

[23] R. J. Boys, D. A. Henderson, and D. J. Wilkinson. Detecting homogeneous segments in DNA sequences by using hidden Markov models. *Journal of the Royal Statistical Society*, C:49(2):269–285, 2000.

[24] P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.

[25] R. J. Boys and D. A. Henderson. A Bayesian approach to DNA sequence segmentation (with discussion). *Biometrics*, 60:573–588, 2004.

[26] J. S. Liu and T. Logvinenko. Bayesian methods in biological sequence analysis. In D. J. Balding, M. Bishop, and C. Cannings, editors, *Handbook of Statistical Genetics*, chapter 3. Wiley, New York, second edition, 2003.

[27] T. P. Speed, editor. *Statistical Analysis of Gene Expression Microarray Data*. Chapman & Hall/CRC, Boca Raton, Florida, 2003.

[28] E. Wit and J. McClure. *Statistics for Microarrays: design, analysis and inference*. Wiley, New York, 2004.

[29] M. Vanucci, K.-A. Do, and P. Müller, editors. *Bayesian Inference for Gene Expression and Proteomics*. Cambridge University Press, New York, 2006.

[30] D. Zhang, M. T. Wells, C. D. Smart, and W. Fry. Bayesian normalization and identification for differential gene expression data. *Journal of Computational Biology*, 12(4):391–406, 2005.

[31] M. Bhattacharjee, C. C. Pritchard, P. S. Nelson, and E. Arjas. Bayesian integrated functional analysis of microarray data. *Bioinformatics*, 20(17):2943–2953, 2004.

[32] A. Frigessi, M. A. van de Wiel, M. Holden, D. H. Svendsrud, I. K. Glad, and H. Lyng. Genome-wide estimation of transcript concentrations from spotted cDNA microarray data. *Nucleic Acids Research*, 33(17), 2005.

[33] M. A. van de Wiel, M. Holden, I. K. Glad, H. Lyng, and A. Frigessi. Bayesian process-based modeling of two-channel microarray experiments: estimating absolute mRNA concentrations. In M. Vanucci, K.-A. Do, and P. Müller, editors, *Bayesian Inference for Gene Expression and Proteomics*. Cambridge University Press, New York, 2006.

[34] E. M. Conlon, J. J. Song, and J. S. Liu. Bayesian models for pooling microarray studies with multiple souces of variation. *BMC Bioinformatics*, 7(247), 2006.

[35] A-M. K. Hein, S. Richardson, H. C. Causton, G. K. Ambler, and P. J. Green. BGX: a fully Bayesian integrated approach to the analysis of Affymetrix GeneChip data. *Biostatistics*, 6(3):349–373, 2005.

[36] A. Lewin, S. Richardson, C. Marshall, A. Glazier, and T. Aitman. Bayesian modelling of differential gene expression. *Biometrics*, 62(1):10–18, 2006.

[37] K. Yeung, C. Fraley, A. Murua, A. Raftery, and W. Ruzzo. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17:977–987, 2001.

[38] J. C. Wakefield, C. Zhou, and S. G. Self. Modelling gene expression data over time: curve clustering with informative prior distributions. In J.-M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, editors, *Bayesian Statistics 7*, pages 721–732, Oxford, 2003. Oxford University Press.

[39] N. A. Heard, C. C. Holmes, D. A. Stephens, D. J. Hand, and G. Dimopoulos. Bayesian coclustering of Anopheles gene expression time series: study of immune defense response to multiple experimental challenges. *Proceedings of the National Acadamy of Sciences*, 102:16939–16944, 2005.

[40] R. Gottardo, J. Besag, M. Stephens, and A. Murua. Probabilistic segmentation and intensity estimation for microarray images. *Biostatistics*, 7:85–99, 2006.

[41] M. West, C. Blanchette, H. Dresden, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. R. Marks, and J. R. Nevins. Predicting the clinical status of human breast cancer utilizing gene expression profiles. *Proceedings of the National Acadamy of Sciences*, 98:11462–11467, 2001.

[42] J. G. Ibrahim, M.-H. Chen, and R. J. Gray. Bayesian models for gene expression with DNA microarray data. *Journal of the American Statistical Association*, 97(457):88–99, 2002.

[43] R. Gottardo, J. A. Pannucci, C. R. Kuske, and T. Brettin. Statistical analysis of microarray data: A Bayesian approach. *Biostatistics*, 4:597–620, 2003.

[44] B. J. A. Mertens. On the application of logistic regression modelling in microarray studies. In J.-M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, editors, *Bayesian Statistics 7*, pages 607–618, Oxford, 2003. Oxford University Press.

[45] M. West. Bayesian factor regression models in the "large p, small n" paradigm. In J.-M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, editors, *Bayesian Statistics 7*, pages 733–742, Oxford, 2003. Oxford University Press.

[46] R. Gottardo, A. E. Raftery, K. Y. Yeung, and R. E. Bumgarner. Bayesian robust inference for differential gene expression in microarrays with multiple samples. *Biometrics*, 62:10–18, 2006.

[47] P. J. Green and K. V. Mardia. Bayesian alignment using hierarchical models, with applications in protein bioinformatics. *Biometrika*, 93:235–254, 2006.

[48] S. C. Schmidler. Fast Bayesian shape matching using geometric algorithms. In J.-M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, editors, *Bayesian Statistics 8*, Oxford, 2007. Oxford University Press. In press.

[49] K. V. Mardia, P. J. Green, V. B. Nyirongo, N. D. Gold, and D. R. Westhead. Bayesian refinement of protein functional site matching. *BMC Bioinformatics*, 2007. In press.

[50] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. J. Krogan, S. Chung, A. Emili, M. Snyder, J. F. Greenblatt, and M. Gerstein. Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302(5644):449–453, 2003.

[51] J. S. Morris, P. J. Brown, K. Baggerly, and K. Coombes. Analysis of mass spectrometry data using Bayesian wavelet-based functional mixed models. In M. Vanucci, K.-A. Do, and P. Müller, editors, *Bayesian Inference for Gene Expression and Proteomics*, chapter 14. Cambridge University Press, New York, 2006.

[52] M. Clyde, L. House, and R. Wolpert. Nonparametric models for proteomic peak identification and quantification. In M. Vanucci, K.-A. Do, and P. Müller, editors, *Bayesian Inference for Gene Expression and Proteomics*, chapter 15. Cambridge University Press, New York, 2006.

[53] H. Bensmail, J. Golek, M. M. Moody, J. O. Semmes, and A. Haoudi. A novel approach for clustering proteomics data using Bayesian fast Fourier transform. *Bioinformatics*, 21(10):2210–2224, 2005.

[54] A. Saksena, D. Lucarelli, and I.-J. Wang. Bayesian model selection for mining mass spectrometry data. *Neural Networks*, 18:843–849, 2005.

[55] W. Zhang and B. T. Chait. Profound: an expert system for protein identification using mass spectrometric peptide mapping information. *Annals of Chemistry*, 72(11):2482–2489, 2000.

[56] S. S. Chen, E. W. Deutsch, E. C. Yi, X.-J. Li, D. R. Goodlett, and R. Aebersold. Improving mass and liquid chromatography based identification of proteins using Bayesian scoring. *Journal of Proteome Research*, 4(6):2174–2184, 2005.

[57] H. Kitano. Computational systems biology. *Nature*, 420(6912):206–210, 2002.

[58] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using Bayesian networks to analyse expression data. *Journal of Computational Biology*, 7:601–620, 2000.

[59] N. Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659):799–805, 2004.

[60] I. Pournara and L. Wernisch. Reconstruction of gene networks using Bayesian learning and manipulation experiments. *Bioinformatics*, 20(17):2934–2942, 2004.

[61] J. Schäfer and K. Strimmer. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6):754–764, 2005.

[62] A. V. Werhli, M. Grzegorczyk, and D. Husmeier. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks. *Bioinformatics*, 22:2523–2531, 2006.

[63] K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

[64] A. Dobra, C. Hans, B. Jones, J. R. Nevins, G. Yao, and M. West. Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, 90(1):196–212, 2004.

[65] B. Jones, C. Carvalho, A. Dobra, C. Hans, C. Carter, and M. West. Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science*, 20(4):388–400, 2005.

[66] P. Giudici and P. J. Green. Decomposable graphical Gaussian model determination. *Biometrika*, 86(4):785–801, 1999.

[67] D. Husmeier. Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics*, 19(17):2271–2282, 2003.

[68] J. Yu, V. A. Smith, P. P. Wang, A. J. Hartemink, and E. D. Jarvis. Advances to Bayesian network inference for generating causal networks from observational data. *Bioinformatics*, 20(18):3594–3603, 2004.

[69] R. Opgen-Rhein and K. Strimmer. Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process. In submission, 2006.

[70] I. Lee, S. V. Date, A. T. Adai, and E. M. Marcotte. A probabilistic functional network of yeast genes. *Science*, 306(5701):1555–1558, 2004.

[71] A. Bernard and A. J. Hartemink. Informative structure priors: joint learning of dynamic regulatory networks from multiple types of data. In R. Altman, A. K. Dunker, L. Hunter, T. Jung, and T. Klein, editors, *Pacific Symposium on Biocomputing 2005*, pages 459–470, New Jersey, 2005. World Scientific.

[72] M. West, G. S. Ginsburg, A. T. Huang, and J. R. Nevins. Embracing the complexity of genomic data for personalised medicine. *Genome Research*, 16:559–566, 2006.

[73] I. Nachman, A. Regev, and N. Friedman. Inferring quantitative models of regulatory networks from expression data. *Bioinformatics*, 20, supp. 1:i248–i256, 2004.

[74] M. Ronen, R. Rosenberg, B. I. Shraiman, and U. Alon. Assigning numbers to the arrows: parameterising a gene regulation network by using accurate expression kinetics. *Proceedings of the National Academy of Sciences*, 99:10555–10560, 2002.

[75] Moles. C. G., P. Mendes, and J.R. Banga. Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome Research*, 13:2467–2474, 2003.

[76] K. G. Gadkar, R. Gunawan, and F. J. Doyle III. Iterative approach to model identification of biological networks. *BMC Bioinformatics*, 6:155, 2005.

[77] K. S. Brown and J. P. Sethna. Statistical mechanical approaches to models with many poorly known parameters. *Physical Review E*, 68:021904, 2003.

[78] M. Barenco, D. Tomescu, D. Brewer, R. Callard, J. Stark, and M. Hubank. Ranked prediction of p53 targets using hidden variable dynamic modeling. *Genome Biology*, 7:R25, 2006.

[79] W. Liebermeister and E. Klipp. Biochemical networks with uncertain parameters. *IEE Systems Biology*, 152(3):97–107, 2005.

[80] D. Orlando, C. Lin, A. Bernard, E. S. Iversen, A. J. Hartemink, and S. B. Hasse. A probabilistic model for cell cycle distributions in synchrony experiments. *Cell Cycle*, 2007. In press.

[81] O. G. Bahcall. Single cell resolution in regulation of gene expression. *Molecular Systems Biology*, 2005. doi:10.1038/msb4100020.

[82] H. H. McAdams and A. Arkin. Stochastic mechanisms in gene expression. *Proceedings of the National Acadamy of Science USA*, 94:814–819, 1997.

[83] A. Arkin, J. Ross, and H. H. McAdams. Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected *Escherichia coli* cells. *Genetics*, 149:1633–1648, 1998.

[84] H. H. McAdams and A. Arkin. It's a noisy business: genetic regulation at the nanomolecular scale. *Trends in Genetics*, 15:65–69, 1999.

[85] D. J. Wilkinson. *Stochastic Modelling for Systems Biology*. Chapman & Hall/CRC Press, Boca Raton, Florida, 2006.

[86] R. Pepperkok and J. Ellenberg. High-throughput fluorescence microscopy for systems biology. *Nature Reviews Molecular Cell Biology*, 7:690–696, 2006.

[87] H. Shen, G. Nelson, D. E. Nelson, S. Kennedy, D. G. Spiller, T. Griffiths, N. Paton, S. G. Oliver, M. R. H. White, and D. B. Kell. Automated tracking of gene expression profiles in individual cells and cell compartments. *Journal of the Royal Society Interface*, 2006. In press.

[88] S. Reinker, R. M. Altman, and J. Timmer. Parameter estimation in stochastic biochemical reactions. *IEE Systems Biology*, 153(4):168–178, 2006.

[89] R. J. Boys, D. J. Wilkinson, and T. B. L. Kirkwood. Bayesian inference for a discretely observed stochastic kinetic model. In submission, 2004.

[90] G. A. Rempala, K. S. Ramos, and T. Kalbfleisch. A stochastic model of gene transcription: an application to L1 retrotransposition events. *Journal of Theoretical Biology*, 242(1):101–116, 2006.

[91] D. T. Gillespie. The chemical Langevin equation. *Journal of Chemical Physics*, 113(1):297–306, 2000.

[92] A. Golightly and D. J. Wilkinson. Bayesian inference for stochastic kinetic models using a diffusion approximation. *Biometrics*, 61(3):781–788, 2005.

[93] A. Doucet, N. de Freitas, and N. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer, New York, 2001.

[94] A. Golightly and D. J. Wilkinson. Bayesian sequential inference for nonlinear multivariate diffusions. *Statistics and Computing*, 16:323–338, 2006.

[95] A. Golightly and D. J. Wilkinson. Bayesian sequential inference for stochastic kinetic biochemical network models. *Journal of Computational Biology*, 13(3):838–851, 2006.

[96] A. Golightly and D. J. Wilkinson. Bayesian inference for nonlinear multivariate diffusion models observed with error. In submission, 2006.

[97] K. Jaqaman and G. Danuser. Linking data to models: data regression. *Nature Reviews Molecular Cell Biology*, 7(11):813–819, 2006.

[98] A. O'Hagan. Bayesian analysis of computer code outputs: a tutorial. *Reliability Engineering and System Safety*, 91:1290–1300, 2006.

[99] M. C. Kennedy and A. O'Hagan. Bayesian calibration of computer models (with discussion). *Journal of the Royal Statistical Society, Series B*, 63:425–464, 2001.

[100] M. Goldstein and J. Rougier. Bayes linear calibrated prediction for complex systems. *Journal of the American Statistical Association*, 101(475):1132–1143, 2006.

[101] P. G. Challenor, R. K. S. Hankin, and R. Marsh. Towards the probability of rapid climate change. In H. J. Schellnhuber, W. Cramer, N. Nakicenovic, T. Wigley, and G. Yohe, editors, *Avoiding Dangerous Climate Change*, pages 53–63. Cambridge University Press, 2006.

[102] D. J. Wilkinson. Parallel Bayesian computation. In E. J. Kontoghiorghes, editor, *Handbook of Parallel Computing and Statistics*, pages 481–512. Marcel Dekker/CRC Press, New York, 2005.