Front. Biol. 2010, 5(2): 98–104
DOI 10.1007/s11515-010-0023-8

REVIEW

Application of Bayesian networks on large-scale biological data

Yi LIU (✉), Jing-Dong J. HAN (✉)

Chinese Academy of Sciences Key Laboratory of Molecular Developmental Biology, Center for Molecular Systems Biology, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China

Abstract The investigation of the interplay between genes, proteins, metabolites and diseases plays a central role in molecular and cellular biology. Whole genome sequencing has made it possible to examine the behavior of all the genes in a genome by high-throughput experimental techniques and to pinpoint molecular interactions on a genome-wide scale, which form the backbone of systems biology. In particular, Bayesian network (BN) is a powerful tool for the ab-initial identification of causal and non-causal relationships between biological factors directly from experimental data. However, scalability is a crucial issue when we try to apply BNs to infer such interactions. In this paper, we not only introduce the Bayesian network formalism and its applications in systems biology, but also review recent technical developments for scaling up or speeding up the structural learning of BNs, which is important for the discovery of causal knowledge from large-scale biological datasets. Specifically, we highlight the basic idea, relative pros and cons of each technique and discuss possible ways to combine different algorithms towards making BN learning more accurate and much faster.

Keywords Bayesian networks (BN), large-scale biological data

1 Introduction

During the past decade, the fast emergence and development of high-throughput experimental techniques has been an important impetus and distinct hallmark of systems biology. Different from traditional molecular biology experimental protocols, these new techniques or machineries often generate gigabytes or even terabytes of data, typically covering parts of or even the whole genome rather than merely involving a single gene or pathway. For example, microarray experiments are able to measure the expression of thousands of genes simultaneously. As a result, by performing microarray experiments at several planned time points, the expression of all the genes in a genome can be comprehensively profiled over a developmental process. Such data contains valuable information as to which genes' up or down regulation is responsible for a particular developmental stage, and, how these genes are dynamically regulated by a set of key transcription factors. The distinct feature of microarray technique is that it has a rich-data output, i.e., paralleling a large number of traditional molecular biology experiments. A handful of other new, fast emerging high-throughput experimental techniques are also available to systems biology researchers: the RNA deep short-read sequencing technique (RNA-seq) for more accurate gene expression profiling, the yeast-two-hybrid (Y2H) and Co-immunoprecipitation followed by mass spectrometry (Co-IP MS) techniques for protein-protein interaction screening; the Chromatin Immunoprecipitation followed by microarray (ChIP-chip) or deep sequencing (ChIP-seq) methods for quantifying the distributions of transcription factors (TFs) and histone-modifications; the comparative genomic hybridization arrays (Array-CGH) and single nucleotide polymorphic allele (SNP) arrays for the identification of DNA copy number variations and single nucleotide polymorphisms, just to mention a few. These high-throughput techniques enable the genome-wide investigation of biological systems from genomic, proteomic, metabolic, epigenetic and other perspectives possible. However, it is not straightforward to extract or derive the detailed biological knowledge, which is typically embodied as various types of causal or non-causal interactions between a set of factors, from the huge volumes and potentially heterogeneous types of data. As a result, data mining and

Received January 4, 2010; accepted February 2, 2010

E-mail: yiliu@genetics.ac.cn, jdhan@genetics.ac.cn

knowledge discovery algorithms play an important role in this process, and conversely, the proper interpretation and analysis of these large-scale biological data sets also poses many challenging problems for algorithm development. This interesting crosstalk between systems biology and machine learning is gradually recognized by many researches in the two fields.

2 Probabilistic models and Bayesian networks

There are many data mining and knowledge discovery algorithms which are designed to extract a certain kind of regularity from data. In particular, the logic-based approaches have been widely used since the early development of artificial intelligence. However, these methods typically require some rather strong form of *a priori* knowledge and are generally vulnerable to the uncertainties in the data. Unfortunately, in systems biology problems, people usually do not have much prior knowledge about how the genes and transcription factors being investigated causally influence each other. To account for the intrinsic limitations of logic-based methods, probabilistic models are more frequently exploited to better represent the uncertainties in the data (Koller and Friedman, 2009). In general, there are two main types of probabilistic models: discriminative and generative. Discriminative models mainly target at making future predictions, e.g., dealing with classification or regression problems; while the principal focus of generative models is to explain the regularities in the data. In this paper, our aim is to uncover the relations between various factors in the huge volume of data generated by high throughput techniques in systems biology. Therefore, our discussion is mainly restricted to generative models.

Specifically, there are two major classes of probabilistic generative models, namely Markov networks (MN, a.k.a., Markov random fields) and Bayesian Networks (BN) (Koller and Friedman, 2009). Both of the two classes of models have graphical representations, where each node in a graph corresponds to a factor being investigated, and the edges in the graph signify the interactions between these factors. More precisely, it is the absence of edges in these graphs represents the *conditional independency* relations among sets of nodes, thereby encoding the structural relationships between these factors. Here, that A and B are conditionally independent of C means that given the third set of variables (C), the joint probability distribution of two sets of variables $P(A, B|C)$ can be factorized into the product of two probability distributions $P(A|C)P(B|C)$, one for each set of variables. Also, to simplify the presentation, in the rest of this paper, we denote all the factors being investigated or all nodes in a graph collectively as a "domain". Besides the independency oracles that can be directly read out from the graph, each graphical model has a set of parameters for specifying a number of local probability distributions, which multiplicatively defines an explicit, joint probability density function of the domain (which always satisfies these conditional independence relations by construction). In particular, the local probability models are the un-normalized functions for positive Markov networks; however, the local probability functions for Bayesian networks (which specify the distribution of each node conditioned on its parents) are always normalized. With this distinction, the joint distribution of a Bayesian network is always normalized, while an extra term (called the partition function) has to be introduced to a Markov network for normalizing the joint distribution. Partly because of this specific dependency structure, the edges of Bayesian networks are always directed, and there is no directed cycle in the graphs to prevent the recurrence of information flow. As a result, the structure of a Bayesian network is a *directed acyclic graph* (*DAG*), where edges are directed from parent nodes to their children. Conversely, the edges of a Markov network are undirected, and there is no restriction for the presence of loops.

Based on the causal interpretation of Bayesian networks, a directed edge can be seen as a causal influence from the start node to the end node. Therefore, BNs could be used to represent important causal interactions in systems biology, such as which transcription factors regulate the expression level of a downstream gene, and which SNPs are responsible for a heritable disease. Also note that there might be some recruiting order for some seemingly non-causal protein-protein interactions. Therefore, their relationship can be modeled as a causal interaction. However, if a Markov network is used for representing the dependency relations in the domain, there is no way to model all the important causal information. Due to this observation, Bayesian networks are widely used as a modeling and representation language in systems biology. An example of a Bayesian network structure is illustrated in Fig. 1, in which the interdependencies between nodes can be read out from the graph directly.

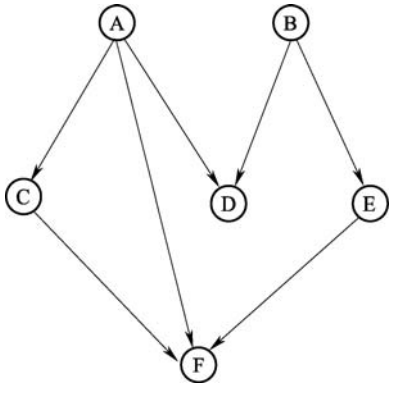

Fig. 1 The structure of a simple Bayesian network with six nodes. There is no directed cycle in the graph and it is easy to derive the dependency relation from the graph, e.g., F depends on {A, C, E}.

3 Learning of Bayesian network structures

As we have discussed above, to define a Bayesian network, one has to specify a directed acyclic graph as well as the local probability distribution of each node conditioned on its parents. When this is done, the BN fully determines the joint probability distribution over all the variables in the domain. As a result, we can answer various probabilistic queries (probabilistic inference) about the domain by conditioning and marginalizing some sets of variables (Koller and Friedman, 2009). To do this, we do not need to first write down the joint distribution explicitly and then perform the computations. In fact, the conditional independence properties of the distribution as implied by the graph can be exploited to make the computation much faster. We will not go through the details of such BN inference algorithms here, but only briefly mention a few well-known ones: (1) Belief propagation, which is a message passing algorithm for performing exact inference when the skeleton (not considering the direction of edges) of a BN is a "polytree", and for performing approximate inference when the skeleton of a BN contains loops (Pearl, 1988). (2) The junction tree algorithm, which can be used to perform exact inference for arbitrary BNs (Lauritzen and Spiegelhalter, 1988). However, the computational complexity can be exponentially high if the diameter of a junction tree is large. (3) The Gibbs sampling based inference algorithm, which can be employed to trade off between the inference accuracy and the computational complexity by sampling an appropriate number of particles (Geman and Geman, 1984).

When we have a well-defined BN at hand, these BN inference algorithms can be used to quantify the probabilities of various events of interest. However, this is not of uttermost importance in systems biology investigation since there are still a large number of unknown questions in both the qualitative and the structural aspects. The quantitative modeling of the system will make sense only if these unknown structural relationships and dependencies are well resolved. In fact, when the structure of a BN is known, the parameters of the local conditional probability distributions (CPDs) for each family (a node and its parents) can be computed by simply counting the number of data items in different configurations (Heckerman, 1999). As a result, the most important thing in applying BNs to systems biology is to infer the BN structure from data. With this information, we can not only answer many qualitative questions, but also lay the foundation for further quantitative reasoning.

Briefly, the algorithms for the structural learning of BNs can be divided into three major classes: constraints-based, scoring-based and hybrid learning algorithms. In the following sections, we are going to discuss various issues about these algorithms.

3.1 Constraints-based learning algorithms

Since a BN encodes a number of conditional independence relations (by the d-separation criterion), constraints-based learning algorithms try to recover the BN structure by performing a number of conditional independence tests on the training data. Theoretically, this reverse engineering process is largely based on the *faithfulness assumption* (Koller and Friedman, 2009), which states that *if and only if* the conditional independence relations implied by the true BN are satisfied by the training data.

The most prominent examples of constraints-based learning algorithms are the Inductive Causation (IC) algorithm (Pearl and Verma, 1991), PC (named by its inventors P. Spirtes and C. Glymour) algorithm and variants of the PC algorithm (Spirtes et al., 2001). Briefly, these algorithms iteratively remove an edge from an initially fully connected undirected graph if a conditioning set is found to make a pair of nodes conditionally independent. The output of this learning stage is the BN skeletion, i.e., the undirected graph formed by removing the arrows of the directed edges in a BN. In the second step, the results of these conditioning tests can be used again to determine the directions of these edges. The advantage of PC algorithm over IC algorithm is at the reduced number of statistical tests by organizing their ordering properly.

Although the PC algorithm is proved to be consistent in the asymptotical sense, they are not very appropriate for the structural learning of BNs from biological data. This is because real world biological data-sets are often very noisy, and the conditional independence tests do not always yield the correct answer based on limited training data. As a result, some redundant edges might appear in the DAG while some true edges might be missing. Furthermore, the directions of some edges could be wrong.

Two classes of approaches have been proposed to improve the performance of PC algorithm. The first one is to further reduce the number of statistical tests in learning the BN structure. In a recent work, a divide and conquer algorithm is proposed to learn BNs based on recursive vertex set decomposition (Xie and Geng, 2008).

The second class of algorithms improves the accuracy of the PC algorithm by first learning the skeleton of the BN, which is essentially an intermediate result of the PC/IC and the recursive decomposition algorithm. Then, the undirected skeletal graph can be used to constrain the search space of the scoring-based BN-structural learning algorithms (Tsamardinos et al., 2006). The technical details of these hybrid approaches will be presented shortly after introducing the scoring based methods.

3.2 Scoring Bayesian network structures

One of the major problems with statistical tests is that the results are typically "yes" or "no". However, real-world

data sets are often noisy and having a limited number of cases, in which many statistical tests have no definitive answer. This problem could be more severe in the structural learning of BNs since the power of statistical tests could be further reduced in the context of multiple tests. A better way to handle the ambiguity with limited data is to compare many plausible BNs and find the one that best fit the data. This idea motivates the development of scoring-based BN learning algorithms (Heckerman et al., 1995). Unfortunately, it can be shown that for any dataset, regardless of the direction of edges, fully connected BNs have the best fitness with any training data. This is an example of a general problem with many machine learning algorithms: overfitting. Briefly, this phenomenon occurs because the model (BN) not only captures the regularities in the underlying probability distribution, but also many nuisance factors in this particular dataset. The latter part deteriorates the generalization performance of the model (BN) on other datasets which are sampled from the same distribution. As a result, it is desirable to build a model with the best tradeoff in its fitness to training data and relative low complexity. Ideally, one can obtain the best BN by first enumerating all plausible DAGs, learning the parameters of each DAG and then testing their generalization performance using a cross-validation procedure. However, it can be shown that the number of DAGs is super-exponential to the cardinality of the variable set. Thus, this naïve brute-force approach is not technically feasible. To design a tractable approach using the same line of thinking, one has to solve two problems: (1) How to evaluate the goodness of a BN (i.e., its generalization performance) efficiently? and (2) How to search for the best DAG efficiently? We will address the two questions in the following paragraphs.

Now we discuss the solution to the first question. Briefly, the goodness of a BN-structure can be defined as its fitness to the training data minus the model complexity. The first term is defined as the log likelihood of the data given the maximum likelihood estimation (MSE) of the parameters in a BN (which parameterize the local conditional probability distributions (CPDs) for each node), while the complexity is signified by the number of parameters which fully specify the model (Heckerman, 1999). There are basically three ways to connect the two terms in order to define a sound scoring function, the *Akaike Information Criterion* (AIC) (Akaike, 1974), the *Bayesian Information Criterion* (BIC) (Schwarz, 1978) and the *minimum description length* (MDL) principle (Grünwald, 2007). It can be proven that if some particular data coding scheme is used by the MDL principle, it is equivalent to the BIC criterion.

Note that both of the BIC/MDL scoring functions are only consistent in the asymptotical sense, i.e., when there is infinite number of training data. Consequently, these scoring functions are not precise in practical scenarios where the number of training data is limited. In this case, using the *Bayesian's formula* to derive an exact scoring criterion is better than the approximate scoring functions above (Heckerman et al., 1995). In fact, the Bayesian scoring approach has close-form solutions for discrete (Heckerman et al., 1995) and linear-Gaussian BNs (Geiger and Heckerman, 1995) when some extra assumptions are made. Thus, the high computational complexity with this approach can be fully avoided.

3.3 Bayesian network learning algorithms based on scoring functions

When the scoring function is specified, the task of learning the BN structure is reduced to the problem of finding an optimal Directed Acyclic Graph which maximizes the function. However, the number of BNs is super-exponential to the number of variables in the domain. So naïve exhaustive search is only applicable to only a few variables. In practice, there are two ways to get around this problem. The first approach is to employ heuristic search techniques to find a close to optimal solution without traversing a huge number of DAGs (Heckerman et al., 1995). Here, TABU strategy can be used in the greedy ascend search (Cvijovic and Klinowski, 1995) to partially alleviate the local optima problem.

Some properties of the scoring function and local search can be used to speed up the BN structural learning process. Common BN scoring functions, such as the AIC, BIC/MDL, Bayesian or the BDe metric, can be decomposed into terms that only involve one node and its parents (family). As a result, it can be shown that for each insertion/deletion/edge reversal operation, the total score of the DAG can be computed conveniently by only updating the local score of one or two families. Moreover, it can be shown that a simply caching technique can be used to make further speeding up since most score changes keep invariant after a graph operation (Friedman, 1997).

The above heurist search methods have been very successful in the structural learning of BNs, but there is no warrantee as to the precision of the learning results. In recent years, it has been shown that exact BN structural learning is even possible for medium-sized datasets, such as 25 numbers of variables. This intriguing theoretical advance, again, arises from the fact that the BN scoring function is decomposable. Briefly, exact BN learning methods exploit this particular structural property of scoring functions to design efficient dynamic programming algorithms in order to avoid unnecessary and/or repeated computations. In a recent work, it has been shown that the exact maximal scoring network can be constructed in a relatively simple manner (Silander and Myllymäki, 2006).

There is, however, another class of exact BN structural learning algorithm which is able to generate more robust learning results (Koivisto and Sood, 2004; Koivisto, 2006). The basic idea is different from traditional BN

structural learning algorithms which are formulated as model selection problems, i.e., to search for the best scoring DAG. In this new class of algorithms, the structural learning task is formulated as a Bayesian model averaging problem, i.e., computing the probability of the presence of each edge in a BN by averaging over all possible BN structures. This is also achieved by using a (more complicated) dynamic programming algorithm (Koivisto and Sood, 2004; Koivisto, 2006).

To make BN structural learning scalable to large datasets, some individual algorithmic steps can be sped up further. First, to compute the AIC, BIC/MDL, Bayesian or the BDe metric, it is necessary to count the number of data instances for different configurations of each node's family. This is the major computational burden for learning small or medium-sized networks. To address this problem, a particular data structure, namely the "ADtree" has been proposed to speed up this process (Moore and Lee, 1998).

When the number of variables is large, the acyclicity checking for each graph operation will dominate the computational time of heuristic search based BN structural learning. This problem can be mitigated based on the observation that only a local change takes place after each graph operation, so that we need not check the validity of a BN from scratch at each time, but could leverage on previous results. An efficient algorithm for check the acyclicity of a graph has been proposed based on this observation (Giudici and Castelo, 2003).

3.4 Hybrid Bayesian network learning algorithms

Although we have introduced many techniques for speeding up the scoring-based methods for learning BN structures, there are still a lot of problems to be solved when we confront a large number of domain variables. First, heurist search techniques are prone to get trapped into poor local minima if a large number of local steps are needed to traverse between different DAG configurations. Second, the algorithm has to evaluate more graph operations before deciding the next move. The increased number of graph operation and local score evaluation per operation has severely limited the scalability of scoring based BN-structure search. To this end, hybrid learning algorithms (Tsamardinos et al., 2006) have been proposed to speed up this process by eliminating the search space using constraint-based learning techniques first.

Specifically, in the first stage of hybrid learning algorithm, we learn an undirected skeleton of a BN or a superset of the undirected edges. Then, in the second stage, we perform the scoring based BN structure learning, but restrict the scope of heuristic search to be within the undirected skeleton learnt in the first stage. In this way, the number of valid graph operation at each local search step is greatly reduced, which not only speed up the learning process, but also prevent the search from entering incorrect spaces.

In fact, as we have discussed above, the first step of the IC/PC/recursive decomposition algorithm is to learn the undirected edges (skeleton) of the BN, i.e., the parents/children set of each node. Therefore, it suffices to use the intermediate result of these algorithms in the first stage of hybrid learning. However, the constraints-based learning algorithms are not good at deciding the direction of edges and the identified BN skeleton may contain a few false positive edges. Nevertheless, the scoring-guided structure search algorithms are able to address these problems well at the second stage. As a result, if properly designed, the hybrid learning algorithms can achieve high accuracy with relative small time cost.

In this class of methods, the MMHC algorithm (Tsamardinos et al., 2006) proposes a new max-min search heuristic, which allows the relevant variables entering the parent/children set more quickly than the PC algorithm.

Similarly, the PCMB algorithm (Peña et al., 2007) also exploits the max-min heuristic and extends the PC sets learning to the Markov blankets (the minimum set of variables which are able to separate the dependency between a node and the rest of nodes in a BN) learning. This work also studies the case where the faithfulness assumption is not satisfied. In a recent work (Fu and Desmarais, 2008), the authors show that the breadth-first search strategy used in the PC algorithm is more effective than the max-min strategy on some datasets. Unfortunately, the Markov Blanket discovery algorithm described in their paper is not fully correct.

Note that the basic idea of hybrid learning approach allows the incorporation of general feature selection methods in the structural discovery of BNs. For example, the LARS/Lasso algorithm (Efron et al., 2004) can be used to narrow down the number of edges in the candidate set for real-valued data. As a direction for future work, it is interesting to establish the connection between current statistical tests based Markov blanket set discovery algorithms and related feature selection algorithms.

3.5 Deriving causal knowledge from BN structures

There are two types of interactions in systems biology, non-causal and causal. Examples of the first class includes undirected influences such as protein-protein interactions in a complex, while the second class covers nearly all regulatory relations, e.g., the activity of one molecule influences or decides the activity of the second one. Due to the importance of resolving the ambiguities of the directionalities of causal relations in forming biological hypothesis, BNs have been widely used for knowledge discovery on systems biology datasets. However, cares must be taken in making causal explanations as there are many subtle issues involved. In this paper, we will not address this issue in great depth, but just mention a few important points briefly.

Note that different BNs in an equivalence class encode

the same set of conditional independency semantics. As a result, we could not distinguish them in light of the training data without making any *a priori* assumption (Verma and Pearl, 1991). Indeed, it can be shown that these BNs have the same set of skeletons, i.e., they are fully identical regardless of the directions of edges. However, when our concern is about causality, there will be ambiguities in deciding the direction of some of the edges. In fact, if we define compelled edges to be the ones whose directions are fixed for all the BNs in an equivalence class, and *vice versa* for non-compelled edges, we can only extract causal information from compelled ones while deferring those non-compelled edges to an *interventional* study to resolve their causality ambiguity (Chickering, 1995).

Things could be better when we have some *a priori* knowledge about the domain. If such knowledge is consistent with that of the compelled edges and it resolves the directionalities of some of the non-compelled edges, we can extract more causal information from a BN structure. That is, the restriction on the directionality of some non-compelled edges can be propagated on the BN structure so that the directions of other non-compelled edges are also fixed. Such causal inferences can be performed automatically using a set of propagation rules (Meek, 1995).

3.6 Biological data suitable for Bayesian network analysis

Because BN analysis is essentially to infer conditional dependency and independency of various events, the more incidences obtained for the events, the more accurate the inference will be. Especially when background noise in the dataset is large, such as for nearly all biological datasets, the number of data points or incidences needed for inference will be even larger. This was, in fact, the major limitation of its application in the previously commonly used microarray gene expression analyses, where there are a large number of variables (genes) and often very limited number of measurements (incidences). However, with recent development of next-generation deep sequencing technology, the number of data points measured might not be a limiting factor. For example, in ChIP-seq experiments, the scenario is exactly the opposite, where the number of variables is small (related TF or chromatin modifications), but the number of measurements is large ($>15\,000$ genes or other features measured at a genome-wide level). Thus ChIP-seq experiments provide an ideal study case to infer causal relationships between TFs, chromatin modifications and gene expressions by BN (Yu et al., 2008; van Steensel et al., 2010). Using such a model, we have inferred that among 20 different histone methylations, only trimethylation on lysine 4 of histone H3 (H3K4me3) and trimethylation on lysine 27 residue of histone H3 (H3K27me3) are directly causal to Pol II binding to gene promoters, and to downstream gene expressions (Fig. 2a). We also found that H4K20me3, which is known to be a mark induced by DNA repair, at the promoter regions can promote the conversion of H3K27me3 mark to H3K9me3 (Fig. 2b). H3K27me3 is a reversible transcription repression mark most commonly found in normal stem cells, whereas H3K9me3 is a rather irreversible heterochromatin mark more enriched in cancer stem cells. Therefore, such causal relationship inferred implicates a role of DNA repair in the conversion of normal stem cell epigenetic signature to cancer stem cell signature (Yu et al., 2008).

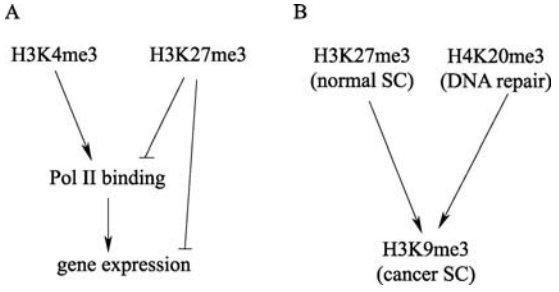

Fig. 2 Causal relationship inference by Bayesian network analysis using ChIP-seq data. A: Promoter H3K4me3 is inferred to stimulate Pol II binding, which in turn activates downstream gene expression, whereas H3K27me3 is repressive to both Pol II binding and gene expression. Arrows indicate activation and barbed lines indicate repression. B: DNA-repair induced H4K20me3 mark is inferred to help converting the promoter H3K27me3 mark (a normal stem cell enriched epigenetic signature) to H3K9me3 mark (a cancer stem cell enriched epigenetic signature).

Another advantage of deep-sequencing is that, similar to SAGE, the intensity measurements are digital and more comparable between different laboratories. Therefore, even for RNA-seq experiments measuring gene expression changes, given the data generated by many laboratories in the world, data might be accumulated fast enough for reliable BN analysis. In the meantime, the development of better variable selection algorithms and knowledge-based constrained BN inference algorithms will help to reduce the number of variables and the number of conditional dependencies to be learned and ultimately making the genome-wide reconstruction of causal relationships a practical possibility.

4 Conclusions

In this paper, we have introduced the important role of a special class of probabilistic model, namely Bayesian networks, in discovering causal knowledge from large-scale systems biology data sets. Specifically, we reviewed state-of-the-art techniques for learning BN structures from data, including constraints-based, scoring-guided heuristic search and hybrid learning algorithms, with an emphasis on the basic idea and practical issues. Moreover, we also discussed many useful techniques and tricks for scaling up

these algorithms to large datasets. Due to space limitation, not all important algorithms and applications of BNs are addressed in this review. Other than the common application of BN learning in gene expression data, it has recently been applied to infer the dependency or causal relationships of various chromatin binding factors and epigenetic modifications (Yu et al., 2008; van Steensel et al., 2010). However, most of the BN algorithms have not yet been implemented into ready to use packages. Therefore, an integrated coherent causality discovery pipeline would prove to be very useful for a number of important systems biology applications.

Acknowledgements We thank the support from the China National Science Foundation (Grant No. 30890033, 30588001 and 30620120433), Chinese Ministry of Science and Technology (No. 2006CB910700) to J.D.J.H.

References

Akaike H (1974). A new look at the statistical model identification. IEEE Trans Automat Control, 19(6): 716–723

Chickering D M (1995). A Transformational Characterization of Equivalent Bayesian Network Structures. Proc 11th Ann Conf Uncertainty Artif Intell, 87–98

Cvijovic D, Klinowski J (1995). Taboo search - an approach to the multiple minima problem. Science, 267: 664–666

Efron B, Hastie T, Johnstone I, Tibshirani R (2004). Least angle regression. Ann Statis, 32: 407–499

Friedman N (1997). Learning Belief Networks in the Presence of Missing Values and Hidden Variables. Proc 14th Intl Conf Mach Learn, 125–133

Fu S, Desmarais M (2008). Fast Markov Blanket Discovery Algorithm Via Local Learning within Single Pass. Canadian Conf AI, 96–107

Geiger D, Heckerman D (1995). Learning Gaussian Networks. Proc 10th Ann Conf Uncertainty Artif Intell, 235–243

Geman S, Geman D (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans Automat Control, 6(6): 721–741

Giudici P, Castelo R (2003). Improving Markov chain Monte Carlo Model search for data mining. Mach Learn, 50(1–2): 127–158

Grünwald P (2007). The Minimum Description Length principle. Cambridge, MA: MIT Press

Heckerman D (1999). A Tutorial on Learning with Bayesian Networks. In: Jordan M, ed. Learning in Graphical Models. Cambridge, MA: MIT Press

Heckerman D, Geiger D, Chickering D M (1995). Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. Machine Learning, 20(3): 197–243

Koivisto M (2006). Advances in Exact Bayesian Structure Discovery in Bayesian Networks. Proc 22nd Conf Uncertainty Artif Intell

Koivisto M, Sood K (2004). Exact Bayesian structure discovery in Bayesian networks. J Mach Learn Res, 5: 549–573

Koller D, Friedman N (2009). Probabilistic Graphical Models: Principles and Techniques. Cambridge, MA: MIT Press

Lauritzen S L, Spiegelhalter D J (1988). Local computations with probabilities on graphical structures and their application to expert systems. J Royal Statist Society. Series B (Methodological), 50(2): 157–224

Meek C (1995). Causal inference and causal explanation with background knowledge. Proc 11th Ann Conf Uncertainty Artif Intell: 403–410

Moore A W, Lee M S (1998). Cached sufficient statistics for efficient machine learning with large datasets. J Artif Intell Res (JAIR)8: 67–91

Peña J M, Nilsson R, Björkegren J, Tegnér J (2007). Towards scalable and data efficient learning of Markov boundaries. Intl J Approx Reasoning, 45(2): 211–232

Pearl J (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Fransisco, CA: Morgan Kaufmann Publishers

Pearl J, Verma T (1991). A Theory of Inferred Causation. Proc 2nd Intl Conf Princip Knowledge Representation and Reasoning (KR'91): 441–452

Schwarz G E (1978). Estimating the dimension of a model. Ann Statis, 6 (2): 461–464

Silander T, Myllymäki P (2006). A Simple Approach for Finding the Globally Optimal Bayesian Network Structure. Proc 22nd Conf Uncertainty Artif Intell

Spirtes P, Glymour C, Scheines R (2001). Causation, Prediction, and Search, 2nd ed. Cambridge, MA: MIT Press

Tsamardinos I, Brown L E, Aliferis C F (2006). The max-min hill-climbing Bayesian network structure learning algorithm. Mach Learn, 65(1): 31–78

van Steensel B, Braunschweig U, Filion G J, Chen M, van Bemmel J G, Ideker T (2010). Bayesian network analysis of targeting interactions in chromatin. Genome Res, 20: 190–200

Verma T, Pearl J (1991). Equivalence and synthesis of causal models. Proc Sixth Ann Conf Uncertainty Artif Intell, 255–270

Xie X, Geng Z (2008). A recursive method for structural learning of directed acyclic graphs. J Mach Learn Res, 9: 459–483

Yu H, Zhu S S, Zhou B, Xue H L, Han J D J (2008). Inferring causal relationships among different histone modifications and gene expression. Genome Res, 18(8): 1314–1324