Bayesian Approaches to Modeling Gene Regulatory Networks: A Literature Review

3 authors:

Chanchal K. Roy
University of Saskatchewan
193 PUBLICATIONS **4,314** CITATIONS
SEE PROFILE

Banani Roy
University of Saskatchewan
38 PUBLICATIONS **135** CITATIONS
SEE PROFILE

Mohammad Uddin
Universiti Tunku Abdul Rahman
22 PUBLICATIONS **137** CITATIONS
SEE PROFILE

Bayesian Approaches to Modeling Gene Regulatory Networks: A Literature Review

Chanchal Kumar Roy[1], Banani Roy[1] and Mohammad Gias Uddin[2]
[1]School of Computing, Queen's University, Kingston, Canada
[2]Department of Electrical and Computer Engineering, Queen's University, Kingston, Canada

Introduction

Responding to competition for space and nutrients in microResearch in molecular biology has traditionally focused on a single gene, a single protein or a single reaction at a time. The recent development of high-throughput methods and gene-expression arrays in particular, is leading to an unprecedented growth in available gene expression data. These data provide measurements of the expression levels of thousands of genes simultaneously, and is becoming an important tool in furthering the understanding of organisms at a cellular level. This understanding of organisms at a cellular level can help to predict genetic predisposition to disease and can serve as a set of diagnostic markers. This understanding can also assist in finding better treatments options for existing diseases (e.g., pharmacogenomics).

However, there is a need for methods that can handle this data reflecting the effective functional state of a large biological system, and that can analyze such large systems at some abstract level, without performing the exact biochemical reactions. At the very least, such methods could help in guiding the traditional pharmacological and biochemical approaches towards those genes of interest among the thousands of newly discovered genes. Ideally, a sufficiently predictive and explanatory model at an abstract level could obviate the need for an exact understanding of the system at the biochemical level having the ultimate goal to understand the exact systems in full detail.

Gene regulatory networks (GRNs) represent the dependencies of the different actors in a cell. GRNs determine the level of gene expression for each gene in the genome by observing whether a gene will be transcribed into RNA. A GRN consists of one or more input signaling pathways, several target genes, and the RNA and proteins produced from those target genes. In addition, such networks often include dynamic feedback loops that provide for further regulation of network activities. Knowledge gained from GRN will be important for understanding cellular processes, designing new strategies to combat disease and so on.

Several mathematical models exist that can describe GRNs in a precise and unambiguous manner. One of the most popular models is based on Bayesian Network. This is a graphical model that represent dependencies and conditional independencies among random variables. Due to its probabilistic nature, it succinctly captures the stochastic aspects of gene expression and noisy measurements.[1]

Here we review systems that are relying on Bayesian Networks for modeling genetic regulatory networks. We compare the different approaches, study their limitations, and suggest potential future research directions.

Evolution of Bayesian Approaches for Modeling GRNs

The application of Bayesian networks to modeling gene regulatory networks is an active research field. This section surveys some of the work in the area to date.

Bayesian Networks[1] were first suggested as a model for gene networks by Friedman *et al.*[2] They used it for learning/predicting relationships among genes from continuous gene expression data. For learning the GRN they have discretized gene expression data into 3 levels, i.e. -1(Unexpressed), 0(Normal), 1(Over-expressed), based on fixed thresholds. The work provided some promising results, but the authors concluded that in order to infer networks from expression data alone much larger datasets were required.[2] This approach was extended to allow for variable discretization by constructing sub-networks of interacting genes using hill climbing.[3] A significant improvement by Hartemink *et al.* was the incorporation of hidden nodes/variables into the network, which can capture the influence of currently unobserved factors (viewed as latent variables, e.g. protein levels) and make predictions. Moreover, these predictions can be verified later as data become available.[4]

Following these initial lines of research, Imoto et al. extended Friedman's continuous Bayesian network to handle non-linear relationships between genes, introducing the assumption that the parent genes do not depend linearly on the objective genes.[5] They also handle the possibility of cyclic regulations by adopting the dynamic Bayesian network model over time series data.[6]

During the past two years, most research in this area focused primarily on improving the performance of Bayesian approach for GRN modeling, by incorporating prior knowledge such as known dependencies between components of the system.[7, 8] In addition, Zou *et al.* introduced a new approach to increase the accuracy of learning GRNs and to improve computation complexity by limiting potential regulators to those genes with either earlier or simultaneous expression changes with respect to their target genes.[9]

There are several other works that extensively speak about the impact of expression data types (e.g., discrete vs. continuous data) on the results obtained[10,11], the effectiveness of using synthetic data with incremental process[12,13], the role of prior knowledge[14,15] to improve the Bayesian models, improving the performance of the model[20,21], and the feasibility of studying the hidden variables[22,23,24,25] of the model.

Observations, Challenges and Future Research Directions

We believe that several issues are left either unanswered or ambiguous for modeling GRN using Bayesian networks. Here we list some of the challenges the researchers are currently facing.

Discrete vs. Continues Expression Data: The first ambiguous issue is whether to use the discrete or the continuous expression data. In general, the discretization of expression data causes some information loss. The direct use of continuous data could overcome the problem of using discretization. The studies above do not provide information about difference in results when discretized data is used as opposed to the original continuous data. Ott *et al.* claimed that the data being discrete or continuous has no impact on the results obtained.[10] However, they do not show side-by-side results using discrete and continuous data to support this claim. On the other hand, Friedman claims that using continuous data will eventually lead to model the network with the noise in data (some examples of noise could be variation among patients and tumor locations or experimental noise).[11] It would be very useful if a case study on the same expression dataset is performed, explicitly comparing the two types of data: discretized and continuous.

Noise: Noise is inherent in microarray experiments. There are difference sources of noise in the experiment. Some noises exist due to the variation among patients and tumor locations, variation in the cellular composition of tumors, heterogeneity of the genetic material within tumor caused by genomic instability, whereas some noises stem from differences in sample preparation, and in variability in the experimental settings such as nonspecific cross hybridization, differences in the efficiency of labeling reactions and production differences between microarrays. The noise derived from experimental techniques is reproducible and its boundaries can be modeled, whereas the noise that stems from biological sources cannot be corrected but can be accounted for with statistics using replicates of the treatments or conditions.[3,6,11] Although Bayesian Network can handle noise due to its probabilistic nature, further investigation is required with the experiment of using discrete and continuous expression data as mentioned above.

Sample Size: The second major issue is concerning the sample size. None of the papers we have found suggests an estimate of a sufficient sample size for a good model. This is an important issue since obtaining data from microarray experiments is expensive and without a proper sample size it is not possible to obtain a good model that best fits our requirements.. Therefore, it is better to know the sample size in advance in order to have sufficient data for modeling the gene network accurately. In general, this can be done by simulating an artificial model to get a set of time series data. After that, the obtained artificial dataset is compared to the experimental dataset to determine the differences. These differences are then incorporated into a second iteration of the above process to modify the system in such a way that it better fits the experiment data. Therefore, for predicting a certain sample threshold, one can think of using synthetic data with incremental process to model the network.[11,12,13]

Prior Knowledge: There is relatively little work addressing the issue of prior knowledge. Some examples of studies in this direction include Hartemink *et al.* and Segal *et al.* that used binding site information as priors to improve their Bayesian models.[14,15] Imoto et al. used biological knowledge like protein-protein interactions, protein-DNA interactions, binding site information in the microarray data for improving their Bayesian model.[16] Therefore, it is advisable to gather existing biological knowledge as much as possible and then associate that knowledge in the modeling process.[17,18,19]

Performance: Performance is a major issue as it indicates how good the model is, in terms of computational complexity under various settings such as the size of the dataset, the connectivity of the network and so forth. It is not yet clear how better the obtained network is with respect to the computational complexity. Although, most of the papers have attempted to provide some contribution to this issue, none of them are successful to provide explicit results in support of their arguments. A good approach of improving the performance has been introduced by Ott *et al.*[8] Segal *et al.* have also used a motif finding approach to identify transcriptional modules which is a promising approach (at least a first step) to improve the performance of the inferred network.[20] Recently, Segal *et al.* have provided an extensive study on learning Module Network, a Bayesian network in which variables in the same module share parents and parameters. They have shown that the use of such networks greatly improves the performance of the learning procedure.[21] Using the approach of Segal *et al.*[20,21] one can greatly improve the speed of the learning process.

Number of Regulators of a Target Gene: In order to reduce the search space, most of the work described above tried to limit the number of potential regulators of each target gene. However, the methods were based on heuristics or assumptions and in most of the cases, the proposed algorithms hardly could have achieved dominating results. Moreover, most of the works claimed that their proposed algorithm usually finds pretty good solutions, but there is no proof that the solutions could not get arbitrarily bad. As mentioned before, Zou *et al.* have introduced a promising method of limiting the potential regulators to those genes whose expression change either preceded or occurred simultaneously to that of their target genes.[9] The problem of coping with feedback loop (where a target gene can in turn regulate its regulator) caused by mRNA half-lives can be efficiently handled with a technique using nonparametric regression for nonlinear modeling of GRNS by Kim *et al.*[6] Therefore, if it is possible to combine the methods of Zou *et al.*[9] and Kim *et al.*[6], the number of potential regulators for a target gene could be reduced with more accuracy and in the same time reducing the search space, thus improving the computational complexity.

States and Roles of the Regulators: Given a set of regulators for a target gene, it is important to understand at which stages of the gene expression the regulators are active or inactive. Predicting/learning the state of the regulators (i.e., active or inactive), will be very helpful. The next step is to find out whether an active regulator is going to act as an activator or an inhibitor. We know a regulator could play a role as an activator or an inhibitor which has two complementary impacts on the target gene. An approach to this issue of identifying activators or inhibitors and their roles in the modeling has been made by Noto *et al.*[22] However, their proposed approach was evaluated with some certain baseline(s) and there is no other sufficient evidence that their approach is a good one in all cases. They model the states of the regulators as hidden nodes in their network. Once the regulator states and roles are modeled as hidden nodes, one could use an efficient learning algorithm of the hidden nodes. Another recent work to this same direction is studied by Beal *et al.*[23] However, if we could have an efficient algorithm for learning hidden variables of Bayesian networks, most of problems pointed by Noto *et al.* and by Beal *et al.* could be avoided and the performance of the learning algorithm could be improved with better accuracy. Noto et al. have used the expectation maximization (EM) algorithm for learning hidden variables. This algorithm, however, can easily get trapped in suboptimal local maxima. Learning the model structure is even more challenging The structural EM algorithm can adapt the structure in the presence of hidden variables, but usually performs poorly without prior knowledge about the cardinality and location of the hidden variables. For overcoming such problems, recently, Elidan and Friedman have performed a comprehensive study for learning hidden variables[24] by using information bottleneck framework of Tishby *et al.*[25] For learning hidden variables, Elidan and Friedman's approach would be the promising option for researchers.

Conclusion

Genetic Network Modeling using microarray data is an active research area and the literature reports numerous algorithms/approaches as well as potential applications that can take advantage of microarray data in drug discovery and clinical diagnosis. This short review provides an introduction to the different Bayesian approaches currently used for modeling GRN. We have indicated some of the limitations, and suggested possible remedies as well as future research directions.

References

1. Heckerman D. A tutorial on learning with Bayesian networks. In Learning in Graphical Models, M. Jordan, ed. 1999; MIT Press.
2. Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. J. Comp. Biol. 2000; 7:601–620.
3. Pe'er D, Regev A, Elidan G, Friedman N. Inferring subnetworks from perturbed expression profiles. Bioinformatics. 2001; 17: S215–S224.
4. Hartemink AJ, Gifford DK, Jaakkola TS, Young RA. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. Pac. Symp. Biocomput, World Scientific Publishing Co. 2001; 6: 422-433.
5. Imoto S, Kim S, Goto T, Miyano S, Aburatani S, Tashiro K, Kuhara S. Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. J. Bioinformat. Computat. Biol. 2003; 1: 231–252.
6. Kim S, Imoto S, Miyano S. Dynamic Bayesian network and nonparametric regression for nonlinear modelling of gene networks from time series gene expression data. BioSystems. 2004; 75: 57–65.
7. Le PP, Bahl A, Unga LH. Using prior knowledge to improve genetic network reconstruction from microarray data. Silico Biology. 2004; 4: 0027.
8. Ott S, Hansen A, Kim S, Miyano S. Superiority of network motifs over optimal networks and an application to the revelation of gene network evolution. Bioinformatic. 2004; Vol. 21 (2), 227-238.
9. Zou M, Conzen S. A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. Bioinformatics. 2005; 21(1):71-79.
10. Ott S, Imoto S, Miyano S. Finding optimal models for small gene networks. Pac Symp Biocomput. 2004; 557-67.
11. Rogers S, Girolami M. A Bayesian regression approach to the inference of regulatory networks from gene expression data. Bioinformatics. 2005; 21(14):3131-3137.
12. Rice JJ, Tu Y, Stolovitzky G. Reconstructing biological networks using conditional correlation analysis. Bioinformatics, Advanced Access. 2004.
13. Husmeier D. Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. Bioinformatics. 2003; 19:2271–2282.
14. Hartemink J, Gifford DK, Jaakkola TS, Young RA. Combining location and expression data for principled discovery of genetic regulatory network models. Proceedings of Pacific Symposium on Biocomputing. 2002; 7:437 – 449.
15. Segal E, Barash Y, Simon I, Friedman N, Koller D. From promoter sequence to expression: a probabilistic framework, Bioinformatics, Proc. 6th Annual International Conference on Research in Computational Molecular Biology (RECOMB). 2002; 6:263 – 272.
16. Imoto S, Higuchi T, Goto T, Tashiro K, Kuhara S, Miyano S. Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. J. Bioinformat. Computat. Biol. 2004; 2(1):77-98
17. Shatkay H, Edwards S, Wilbur WJ, Boguski M. Genes, themes and microarrays. Proc. of the Int. Conf. on Intelligent Systems for Molecular Biology. 2000; 317-328.
18. Spieth C, Streichert F, Speer N, Zell A. Inferring regulatory systems with noisy pathway information. Proceedings of the German Conference on Bioinformatics. 2005; 193-203
19. Kyoto Encyclopedia of Genes and Genomes(KEGG): http://www.genome.ad.jp/kegg/
20. Segal E, Yelensky R, Koller D. Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. BIOINFORMATICS. 2003; 19(1):273–282.
21. Segal E, Pe'er D, Regev A, Koller D, Friedman N. Learning module networks. Journal of Machine Learning Research. 2005; 6:557–588.
22. Noto K, Craven K. Learning regulatory network models that represent regulator states and roles. RECOMB. 2004; 3318:52-64.
23. Beal M.A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. Bioinformatics. 2005; 21(3):349–356.
24. Elidan G, Friedman N. Learning Hidden Variable Networks: The information bottleneck approach. Journal of Machine Learning Research. 2005; 6: 81–127.
25. Tishby N, Pereira F, Bialek W. The information bottleneck method. In Proc. 37th Allerton Conference on Communication, Control and Computation. 1999; 368–377.