A Review of Bayesian Modelling Methods for Gene Regulatory Network Inference

Hamda B. Ajmal [1,2] **Michael G. Madden** [1,2]

[1]Data Science Institute , Insight SFI Research Centre for Data Analytics,
National University of Ireland Galway
[2]School of Computer Science, National University of Ireland Galway

With rapid advancements in genome sequencing technologies and a significant reduction in the associated costs, a vast amount of genomic and transcriptomic datasets have been created, covering thousands of genome sequences of a wide range of organisms [Barbosa et al., 2018]. Gene Regulatory Networks (GRNs) allow scientists to understand how various molecular components interact to collectively define the phenotype and physiology of the organism under study [Xing et al., 2018].

A GRN is defined as a network learned from gene expression data using statistical approaches [Emmert-Streib et al., 2014]. A GRN is represented as a graph with nodes and directed edges. The nodes represent the biological entities, e.g. genes, proteins, mRNAs, protein/protein complexes or cellular processes. Edges between nodes represent interactions or regulatory relationships between the nodes [Angelin-Bonnet et al., 2019]. These regulatory relationships correspond to molecular reactions between the biological entities through which the products of one gene can cause change in another. Inferring GRNs from gene expression data is one of the most challenging tasks of systems biology [Ivanov et al., 2016]. The prediction accuracy of models is negatively affected by factors including: the high dimensionality and sparsity of the gene expression data; multicollinearity; risks of over-fitting from multiple testing [Altman and Krzywinski, 2018]; and high complexity of the problem space. Conventional machine learning (ML) methods, that may work reasonably well on simple datasets, may become computationally infeasible to run on high-dimensional data. Inference of GRNs is considered to be an ill-posed problem [Ivanov et al., 2016]. Quality assessment and validation of the inferred GRNs is an additional challenge [Ivanov et al., 2016], due to scarcity of available information about the true structures of biological networks [Dojer et al., 2006].

In this talk, we describe GRNs and their importance in the field of biological and biomedical science. We review advancements in ML research for inferring GRNs from gene expression data. Computational methods of GRN inference are categorised into methods based on information theory models, Boolean networks, ordinary differential equation models, regression, trees, artificial neural networks (ANNs), deep learning and Bayesian networks (BNs). For each of these categories, we present a literature review of methods that have been proposed for GRN inference.

The main focus of this talk is on BNs and dynamic Bayesian networks (DBNs), which have been extensively used for GRN inference. A Bayesian network is probabilistic graphical model that consists of a directed acyclic graph (DAG) denoted by G and a set of conditional probability parameters denoted by θ. The DAG G is represented by a set of nodes X and a set of edges E that connect the nodes. Each node in a Bayesian network represents a distinct random variable and the directed edges signify the existence of conditional dependencies between linked variables. The strengths of the relationships between linked variables are expressed by forward conditional probabilities [Pearl, 2014]. Standard BNs do not provide a mechanism to represent temporal dependencies.

Dynamic Bayesian networks (DBNs) are an extension to BNs that model a dynamic process that evolves over time, in which a past state influences the current state [Enright and Madden, 2015]. In a DBN, time is discretized into a series of time-slices [Koller and Friedman, 2009]. Each time-slice represents the state of the variables at a certain point in time. A DBN contains an initial Bayesian network, denoted by G_0 and a transitional Bayesian network, denoted by $G_{\rightarrow}$ [Koller and Friedman, 2009]. G_0 is a standard Bayesian network. It consists of intra-time slice edges that represent non-temporal (instantaneous) relationships between the variables. $G_{\rightarrow}$ defines temporal relationships among variables across two adjacent time-slices. In $G_{\rightarrow}$, edges can exist between a node in the current time-slice to a node in the next time-slice. Such edges are called inter-time-slice edges. $G_{\rightarrow}$ may also contain edges between nodes in the same time-slice. Such edges are called intra-time slice edges. The intra-time slice edges of $G_{\rightarrow}$ correspond to the intra-time slice edges of G_0.

BNs have been extensively used to model gene data.

Accepted for the 16th Bayesian Modelling Applications Workshop (@UAI2022) (BMAW 2022).

Bayesian models are attractive for their ability to describe complex stochastic processes. All biological systems, including gene regulatory processes, evolve under stochastic processes [Donnet and Samson, 2013, Ajmal et al., 2017], therefore BNs are well-suited to represent them. An important advantage of using BNs to model GRNs is that they have to ability to work on locally interactive components with a relatively smaller number of variables [Friedman et al., 2000]. BNs can also combine prior knowledge to strengthen the causal relationships and avoid over-fitting. A GRN can be modelled as a BN. Each node in the Bayesian network represents the expression of a gene. The edges between the nodes represent the existence of a regulatory relationship between them. The conditional probability distribution of each node quantifies the strength of the influence of its regulators (parent nodes) on its value [Liu et al., 2016].

Boolean networks, although simple and easy to implement, can not cope up with noise and data uncertainty [Delgado and Gómez-Vela, 2019]. They require crude data discretization which often results in loss of information. ODE models use continuous variables to model the gene expression data [Delgado and Gómez-Vela, 2019]. ODE models represent how change in the expression of a gene is driven by the expression level of its regulator genes. They can also take into account the environmental factors to allow qualitative modelling [Delgado and Gómez-Vela, 2019]. ODE models assume that the observed dynamics of a system are exclusively driven by internal, deterministic mechanisms and there is no uncertainty in the process [Donnet and Samson, 2013]. This assumption does not hold true for biological systems [Donnet and Samson, 2013]. Most ODE modelling methods to elucidate GRNs consider only linear models or just specific types of non-linear functions [Voit, 2000], whereas gene regulatory processes are often driven by complex, non-linear dynamics [Delgado and Gómez-Vela, 2019]. ODE models can not scale up to a very large number of genes as it is hard to estimate the model parameters as the model dimension increases.

Information theory models are attractive because of their lower computational cost but they do not capture higher order conditional dependencies. Huynh-Thu and Sanguinetti [2019] have discussed advantages and disadvantages of regression-based models. As they note, regression-based methods are very popular for GRN inference. They can capture higher-order dependencies. They can also predict expression of a gene from expressions of a subset of genes. They are generally computationally intensive. Regression-based methods do not perform well on datasets with multiple strongly correlated covariates [Huynh-Thu and Sanguinetti, 2019], which is a common case in gene expression datasets due to their high dimensionality. Tree-based methods can be used model the multivariate effect of several genes on a target gene. The number and nature of the parameters is flexible. Compared to ODEs and BNs, they are computationally less expensive, hence they are more favourable to be applied on high-dimensional datasets [Huynh-Thu and Geurts, 2019]. Compared to other unsupervised methods, like ANNs, they have fewer parameters to be set [Huynh-Thu and Geurts, 2019].

Information theory models, ODE-based, regression-based and tree-based methods take a top-down approach for network construction. They start with a fully connected network and then apply some threshold to filter a set of edges [Huynh-Thu and Sanguinetti, 2019]. Choosing this threshold value is a non-trivial, challenging task [Huynh-Thu and Geurts, 2019]. In contrast, BNs construct joint probabilistic model out of local conditional independence terms [Huynh-Thu and Sanguinetti, 2019].

As noted by Koumakis [2020], the use of deep learning (DL) models is in the stage of infancy in bioinformatics research. Gene expression datasets usually have a relatively limited number of samples, but DL methods are notoriously data-hungry [Koumakis, 2020]. DL models are black-box. They also require a large number of hyper-parameters that need to be tuned. Finding the best configuration of these hyper-parameters is a non-trivial task. The results of DL methods tend to be sensitive to the choice hyper-parameters [Min et al., 2016].

BNs are more favourable for GRN modelling over ANN-based models due to their easy interpretation [Koumakis, 2020]. Finding a suitable BN structure is a computationally intensive task. One obvious limitation of BNs, is that one has to make fixed assumptions about the nature of interactions between the variables. Most BN based implementations to infer GRNs from gene expression data assume a linear model [Huynh-Thu and Sanguinetti, 2019].

Several BN structure learning methods have been proposed in the literature. We have created a catalogue of the most note-worthy methods. These are categorised into score-based, constraint-based and hybrid methods. Score-based methods apply general optimization techniques to BN structure learning [Koller and Friedman, 2009]. These methods have two components, a scoring objective and a search method. For each candidate DAG, a decomposable network score is calculated that reflects its goodness of fit to the data [Koller and Friedman, 2009]. A search method is applied to navigate through the space of possible network structures to find a structure that maximises the score [Koller and Friedman, 2009]. Search methods include exact algorithms or heuristic search methods. Some examples of score-based BN learning methods applied for GRN inference are Friedman et al. [1999], Ajmal and Madden [2021] and Yu et al. [2004], etc.

Constraint-based methods tend to learn the BN structure that best captures the independencies in the domain [Koller and Friedman, 2009]. These measure the conditional independence constraints with statistical tests and build a DAG

consistent with the corresponding d-separation statements.

Hybrid methods combine constraint-based and score-based methods. The first step is to apply conditional independence tests and rank all the possible edges in the network. Then, all the edges that rank below a certain threshold are excluded from the search space. This is followed by a score-based search to find a graph with oriented edges within the restricted search space [Tsamardinos et al., 2006].

To deal with scalability issues, many researchers have also proposed methods based on local search. These are also reviewed in our research. We also review methods for modelling time delays in gene regulation and for modelling time-varying GRN structures based on non-stationary DBN learning. To the best of our knowledge, this is a first study in the recent years that catalogues BN learning methods for GRN inference. In our talk, we also discuss the relative advantages and disadvantages of other methods and compare them to BN based methods.

References

Hamda B Ajmal and Michael G Madden. Dynamic bayesian network learning to infer sparse models from time series gene expression data. *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 2021.

Hamda B. Ajmal, Michael G. Madden, and Catherine G. Enright. Dealing with stochasticity in biological ODE models. In *Proc. WCB, co-located with 34th ICML*, Sydney, Australia, 2017.

Naomi Altman and Martin Krzywinski. The curse(s) of dimensionality. *Nature Methods*, 15:399–400, 2018.

Olivia Angelin-Bonnet, Patrick J. Biggs, and Matthieu Vignes. Gene regulatory networks: A primer in biological processes and statistical modelling. In Guido Sanguinetti and Vân Anh Huynh-Thu, editors, *Gene Regulatory Networks: Methods and Protocols*, pages 347–383. Springer, New York, 2019.

Sara Barbosa, Bastian Niebel, Sebastian Wolf, Klaus Mauch, and Ralf Takors. A guide to gene regulatory network inference for obtaining predictive solutions: Underlying assumptions and fundamental biological and data constraints. *Biosystems*, 174: 37–48, December 2018.

Fernando M Delgado and Francisco Gómez-Vela. Computational methods for gene regulatory networks reconstruction and analysis: A review. *Artif. Intell. Med.*, 95:133–145, 2019.

Norbert Dojer, Anna Gambin, Andrzej Mizera, Bartek Wilczyński, and Jerzy Tiuryn. Applying dynamic Bayesian networks to perturbed gene expression data. *BMC Bioinf.*, 7(249), 2006.

Sophie Donnet and Adeline Samson. A review on estimation of stochastic differential equations for pharmacokinetic/pharmacodynamic models. *Adv. Drug Delivery Rev.*, 65 (7):929–939, 2013.

Frank Emmert-Streib, Matthias Dehmer, and Benjamin Haibe-Kains. Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. *Front. Cell Dev. Biol.*, 2:38, 2014.

Catherine G Enright and Michael G Madden. Modelling and monitoring the individual patient in real time. In P.F. Lucas A. Hommersom, editor, *Foundations of Biomedical Knowledge Representation*, pages 107–136. Springer, 2015.

Nir Friedman, Iftach Nachman, and Dana Peér. Learning Bayesian network structure from massive datasets: the "Sparse Candidate" algorithm. In *Proc. Fifteenth Conf. Uncertainty Artif. Intell.*, pages 206–215, Stockholm, Sweden, 1999. Morgan Kaufmann Publishers Inc.

Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using Bayesian networks to analyze expression data. In *Proc. Fourth Annu. Int. Conf. Comput. Mol. Biol.*, RECOMB '00, pages 127–135, New York, NY, USA, 2000. Association for Computing Machinery.

VA Huynh-Thu and G Sanguinetti. *Gene regulatory network inference: An introductory survey*, volume 1883, pages 1–23. USA, 2019.

Vân Anh Huynh-Thu and Pierre Geurts. *Unsupervised Gene Network Inference with Decision Trees and Random Forests*, pages 195–215. Springer, 2019.

Ivan V. Ivanov, Xiaoning Qian, and Ranadip Pal. *Emerging Research in the Analysis and Modeling of Gene Regulatory Networks*. IGI Global, USA, 1st edition, 2016.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT press, 2009.

Lefteris Koumakis. Deep learning models in genomics; are we there yet? *Comput. Struct. Biotechnol. J.*, 18:1466–1473, 2020.

Fei Liu, Shao-Wu Zhang, Wei-Feng Guo, Ze-Gang Wei, and Luonan Chen. Inference of gene regulatory network based on local Bayesian networks. *PLOS Computat. Biol.*, 12(8), 2016.

Seonwoo Min, Byunghan Lee, and Sungroh Yoon. Deep learning in bioinformatics. *Briefings Bioinf.*, 18(5):851–869, 07 2016. ISSN 1467-5463.

Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 2014.

Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Mach. Learn.*, 65(1):31–78, 2006.

Eberhard O Voit. *Computational analysis of biochemical systems: a practical guide for biochemists and molecular biologists*. Cambridge University Press, 2000.

Linlin Xing, Maozu Guo, Xiaoyan Liu, Chunyu Wang, and Lei Zhang. Gene regulatory networks reconstruction using the flooding-pruning hill-climbing algorithm. *Genes*, 9(7):342, 2018.

Jing Yu, V Anne Smith, Paul P Wang, Alexander J Hartemink, and Erich D Jarvis. Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics*, 20(18):3594–3603, 2004.