Constraint Satisfaction Programming in Gene Regulatory Networks

Merna Bibars
Systems & Biomedical Engineering
Cairo University
Cairo, Egypt
merna.bibars@gmail.com

Peter Emad Salah
Systems & Biomedical Engineering
Cairo University
Cairo, Egypt
petsalah3@gmail.com

Abdelrahman Hisham Mostafa
Systems & Biomedical Engineering
Cairo University
Cairo, Egypt
abdelrahman.hishammostafa@gmail.com

I. Introduction

Biological processes are those processes that are vital for an organism to live, and that shape its capacities for interacting with its environment. Biological processes include:

1) Homeostasis.
2) Metabolism.
3) Growth.
4) Adaptation.
5) Response to stimuli.
6) Reproduction.

Genes' cooperation and interactions form a dynamic network that brings about our vivid biodiversity. If one gene is regulated, then it affects its targets, which further affect their next targets, causing a domino effect towards the entire cellular environment. Gene networks have become a key tool for the understanding and modelling of complex biological processes. The term gene network, also called, Gene Regulatory Network (GRN), is used to describe complicated functional pathways in a given cell or tissue, which represent living processes. GRNs are models used to describe and predict dependencies between molecular entities. These are composed of nodes, representing genes, proteins, metabolites or RNA; and edges, which represent molecular relations, e.g. protein–DNA, protein–protein interactions or other relationships of several kind.

GRN reconstruction steps are based on the knowledge database discovery (KDD) workflow. KDD goes from input data preprocessing to the validation of generated models, often performed by data base search and comparison with prior experimental data. The process starts with the input data. This usually consists of gene expression datasets. After the input dataset is selected, it may be processed by any computational method in order to improve the quality of the study. Then, the processed data is used as input for a computational inference algorithm, which provides the resulting network. Finally, the obtained model (network) is optimized and validated so true biological insights can be obtained from it, by a comparison with real biological knowledge.

GRNs can be used to reveal numerous biological functions, which also include cancer initiation and progression. Through a GRN, one maybe able to better understand how genes and proteins interact, which further leads to a better understanding of functions of oncogenes and related pathways, offering a better discovery towards drug target identification.

II. Problem Description

The inference of a GRN is often accomplished through the use of gene expression data. So far, there are numerous computational methods and models developed for restoring GRNs in a real cellular environment. However, each of them have their own assumptions and methods, drawing different blueprints that the GRN described. There is still much confusion about the basic meaning of GRN, ways of assessment, and possible biomedical application. Typically, the relationships between genes are directional in nature and they can change over time or in response to external stimulus. GRN inference methods like CoExpression Networks and Relevance Networks, built with a correlation based method, provides coexpression or cofunctionality from a network scale with no directionality. So if no prior knowledge is provided (e.g.,a gene is confirmed to be a transcription factor in a two-gene interaction system), we cannot identify the causations and results of genes in a GRN. A regression-based approach estimates both the underlying associations among genes in a network and the association intensities. However,the model assumes that all the predictors follow a multivariate normal distribution, therefore it does not infer the directionality of associations. A Bayesian Network (BN) approach point out the directionality of each edge thus revealing a causation relationship among all genes. But, for one dataset, there might be more than one optimal BN structure that has the equivalent overall probabilities. Moreover, the BN graph doesn't allow the existence of feedback loops, which have been widely proven to exist in biological networks. This can be resolved by dynamic Bayesian network (DBN).

Furthermore, The complexity of biological systems and the lack of adequate data have posed many challenges to the inference problem.

GRN inference data sets are quite heterogeneous in nature, containing information which is limited and difficult to analyze. This reverberates on performance of GRN inference methods, which tend to be biased toward the type of data and experiments. To alleviate these difficulties several alternatives have been proposed, such as integrating heterogeneous data into the inference model, or integrating a collection of

predictions across different inference methods in Community Networks (CNs). The latter has the advantage of promoting the benefits of individual methods while smoothing out their drawbacks.

III. Literature Review

A. KDD workflow different approaches

1) ***Biological data: basic input for GRN inference****:* The two main data sources for GRN reconstruction are genome and transcriptome. The term "Genome" refers to the collection of genes comprised in a biological system. On the other hand, functional genomics or transcriptomics refers to the analysis of gene expression patterns and tries to find relationships between them and their biological background.

A different way to obtain biological data is using experimental design principles for GRN inference. Depending on the used approach, quality and quantity of the generated data may vary. Experimental design includes usually systematic perturbations e.g. shift between different environmental conditions, interventions at the genetic, transcriptomic, proteomic or metabolomic level. As a result, differential expression patterns can be found under imposed conditions. Measurements on perturbation experiments can be performed in a static (steady state) or time-course situation, the latter involves the use of dynamic programming. Depending on the knowledge to be achieved, the experimental set-up will vary and so will the choice between a static or a dynamic GRN architecture:

- Generation of static data comes with the assumption of an equilibrium or steady-state situation of the biological system. Depending on the case, the steady-state choice may miss critical dynamic events for reliable GRN construction i.e. dynamic changes occurring with time.
- On the other hand, time-series experiments, where samples are taken in a series of time-points after perturbation, constitute the dynamic approach. The experimental set-up determines the number of time-point measurements, thus, the data amount.

Finally, data preprocessing prior to GRN inference is a key step for GRN reconstruction and quality of outcome. Methods for this aim will depend on the type of data and the experimental design. There are two main sources of variability in GRN reconstruction: systematic errors (bias) and stochastic effects (noise). Systematic effects can be nearly removed through data normalization, since some genes expression can be very variable in one cell/ tissue type. On the other hand, replicates performance provides with repeated measurements to reduce stochastic effects.

2) ***Computational approaches for GRN construction****:* Main GRN inference methods can be summarized in:

- Information theory.
- Boolean networks.
- Differential equations models.
- Bayesian models.
- Neural models.

First, information theory models, Information theory-based networks are the most common type of networks due to their computational simplicity. They are also called coexpression networks since they establish gene–gene relationships. The main measures to determine the dependencies between genes are the correlation coefficients like Pearson, Spearman or Kendall coefficients. However, different measures like Euclidean distances or mutual information, were also applied for the inference of GRN. These models are suitable to cover different aspects of cells, which are here understood as time-varying living systems which perform complicated processes inside and between them.

Second, boolean networks, they are easy to implement and allow capturing the actual dynamical behaviour of GRN. Boolean networks represent genes by variables and their expression level is discretized into Boolean binary values: '0' for low values (silenced or nearly-silenced genes) or '1' for high values (activated genes). Boolean functions reconstruct the network compose directed graphs. Although straightforward and simple, Boolean networks find their main limitation in the discretization step. Gene expression is rarely a matter of fully-activation or fully-silencing, since there are often uncountable different gene states in between. Thus, important details of system behaviour might be lost. Nevertheless, Boolean networks are easy to interpret and they offer a simple dynamic approach for GRN.

Third, Ordinary differential equation (ODE), ODE approaches use continuous instead of discrete variables. This leads to a more accurate model, and enables the dynamic modelling of gene regulation. Differential equations will then represent changes in gene expression as a function of other genes expression and taking into account environmental factors, allowing a quantitative modelling. There are multiple solutions to ODE systems if no constraints are assessed. This is why, specifications and constraints representing prior knowledge (simplification, approximations or educated guesses among others) are required for the identification of model structure and parameters. A disadvantage of many ODE models is that these consider only linear models or just specific types of nonlinear functions, while regulatory processes are often characterized by complex non-linear dynamics.

Fourth, bayesian networks, they are one of the most used GRN inference architectures. They make use of the Bayes theorem of probability, then combining probability and graph theory to qualitatively model the properties of GRNs.

Fifth, neural networks. Recurrent neural network is a successful method for GRN inference, since it enables modelling of non-linear and dynamic interaction among genes. Neural models allow continuous variables and their outcome looks similar to the neural connections observed in natural processes.

Although GRN reconstruction is usually tackled by means of one of the models above (or a combination of them), there are some other GRN inference approaches that clearly differ from these methods.

3) ***Network optimization****:* First, a dimensionality reduction has to be performed on the large gene expression data in order

to reduce computational cost. Modelling biological system requires assumption-making to focus only on those specific aspects which are important for aim of the study. Feature selection and feature mapping help reducing model complexity by excluding non-relevant features in GRN inference. Upon feature selection, non-responsive or not well measured genes are removed from the data. It seeks for the minimization of the number of estimated parameters in order to improve performance and generalization of GRNs, solely using the data to deduce dependencies. While, in feature mapping, redundant information is removed.

Another way for structure optimization is scoring functions. Explicit structure optimization methods compare different topologies of GRN models by means of a scoring function, which helps achieving network sparseness. Several scoring criteria have been developed for the different inference method. Gene interactions are added or removed in order to obtain a better scored topology.

Alternatively, heuristic methods apply educated guesses to lead the search to the most likely solution. Search techniques add or remove connections in the network. The three main search techniques are:

- Forward selection which starts from a simple model and most important interactions are added first up to a certain limit.
- Backward elimination which starts from a highly connected model and less significant interactions are removed.
- Stepwise selection which combines both previous approaches.

*4) **GRN validation and appraisal of inference methods**:* Once the final network is obtained, its biological significance has to be tested because GRN-predicted interactions are not necessarily biologically meaningful. However, lack of validation does not necessarily mean 'not-biologically-meaningful' interactions, since, the validation methodology plays a crucial role but many interactions may have not been described yet. There are two main issues regarding network validation: (i) whether the inferred network provides good predictions on the experimental data (scientific validation) and (ii) whether the applied inference algorithm within a certain network model framework yields networks that are accurate relative to some criterion of goodness (inference validation). The boundaries between both approaches actually blur in practice, since validation of an inference model requires then scientific validation of the inference, and results of the later may be used to improve the inference method.

*5) **Quantitative evaluation of inference performance**:* GRNs can be evaluated using scoring methodologies which allow the comparison between different networks.The inferred network is compared with a reference network (Gold-Standard, closest to reality to the general knowledge) obtaining a quality measure. A Gold Standard enables the estimation of several metrics which would jointly provide an evaluation of model's goodness. This is certainly a key point in GRN inference, since data may provide a massive amount of possible interaction and only a few of them are true.

According to Schrynemackers et al., when the true Gold Standard is known, the inferred networks structure is compared to the first one using several metrics:

- True positive rate, sensitivity or recall.
- True negative rate or specificity.
- False positive rate.
- False negative rate.
- Precision.
- Rate of positive predictions.
- F-score.

These metrics are combined in the analysis of the inference performance. For this aim, several curves are displayed:

- Receiver operating characteristic curves.
- Precision-Recall (PR) curves.

B. CSP in GRN Inference

Constraint Technologies have been successfully applied in the field of System Biology. For example, Answer Set Programming has been adopted to address problems in network inconsistencies detection and in metabolic network analysis. CP has been investigated to reason over discrete network models, where GRNs are modeled using multi-valued variables and transition rules. In particular, CP is exploited to represent GRNs' possible dynamics. Also, CSP allows separation between prediction methods and model, is declarative and constraint expressions allow incremental model refinement.

A naive approach upon the GRN inference process would be the one of enumerating all possible directed and acyclic graphs (DAGs) for a given number of nodes, which is deemed brute-force search. However, the amount of possible DAGs for a given number of nodes grows exponentially, making this search problematic. Therefore, heuristics or/and constraints need to be applied to make the process more efficient.

F. Fioretto and E. Pontelli (2013) proposed a novel methodology based on CP to integrate community predictions. CP is a declarative problem-solving paradigm, where logical rules are used to model problem properties and to guide the construction of solutions. CP offered a natural environment where heterogeneous information can be actively handled. The use of constraint expressions allowed the incremental refinements of a model. The CN approach adopted in their work was built by combining four GRN inference procedures (TIGRESS, INFLEATOR, GENIE3, MI-based method (CLR)) and creating an inference ensemble. They used the -now decommissioned- GPDREAM web platform (http://dream.broadinstitute.org) to develop the predictions from each of these methods.

They calculated the average confidence value (Borda count) and the discrepancy value within the set of predictions given by the 4 methods for each edge to decide on the domain for each variable. The constraints employed were: Sparsity, Redundant Edge, Transcriptor Factor, and a Coregulator constraint. They implemented two search strategy to explore the

solution space, DFS and a Monte Carlo (MC)-based proplabeling tree exploration. They set a trial limit for the MC-based solution and a solution number limit for both strategies.

They proposed three criteria/estimators to compute the final GRN prediction: the max frequency (rewards the edge confidence value appearing with the highest frequency in the solution set.), average (average edge consensus among all solution in order to capture recurring predictive trends.), and weighted average.

To measure prediction accuracy against the corresponding reference network they adopted the AUC score. The MC search outperformed. The CCNs achieved higher average prediction accuracy for small and medium size networks, while performance improvements decreased for bigger networks. The application of additional constraints overcame this effect. Their approach consistently outperformed the consensus networks constructed by averaging individual edges ranks (up to 15.02% for small networks and 4.13% for big networks). They showed that knowledge specific about target networks could provide further improvements in the AUC measure.

IV. Methodology

A. Data

Gold standards datasets that are based on real world data have been systematically built up in both machine learning and system biology fields. ALARM network (Beinlich et al., 1989) is a popular tool that is firstly applied for comparison of reconstructed network in machine learning. DREAM project (Marbach et al., 2009; Prill et al., 2010; Marbach et al., 2010) consists of a series of well-studied regulatory network of prokaryotes, eukaryotes, including Escherichia coli (Simmons et al., 2008) and in silico microarray datasets of knock-down, knock-out and their associated transcriptome data are collected and published systematically.

B. CSP Formulation

From the literature review and problem description, we decided to use an inference ensemble method that combines the newest methods in GRN inference, then formulate that as a CSP problem.

A GRN can be described by a weighted directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is the set of regulatory elements of the network and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V} \times [0, 1]$ is the set of regulatory interactions. The presence of an edge $< s, t, w > \in \mathcal{E}$ indicates that an interaction between the regulatory elements s and t is present with confidence value w. The number $|\mathcal{V}|$ of regulatory elements of the GRN is referred to as its size.

In the problem of GRN inference, we are given the set of vertices $\mathcal{V}$ and a set of experiments describing the behavior of the regulatory elements. The goal is to accurately detect the set of regulatory interactions $\mathcal{E}$.

Constraint Programming (CP) is a declarative programming methodology commonly used to address combinatorial search problems. It focuses on capturing properties of the problem in the form of constraints, which are satisfied exclusively by solutions of the problem. A Constraint Satisfaction Problem (CSP) is formalized as a triple $< \mathcal{X}, \mathcal{D}, \mathcal{C} >$ where $\mathcal{X} =< x_1, x_2, ..., x_{n^2-n} >$ is an n-tuple of variables, $\mathcal{D} =< D_1, D_2, ..., D_n >$ is a corresponding n-tuple of domains, (and each D_i is a set of possible values for the variable x_i), and $\mathcal{C} =< C_1, C_2, ..., C_k >$ is a k-tuple of constraints. A constraint C_j over a set of variables $\mathcal{S}_j \subseteq \mathcal{X}$ is a subset of the Cartesian product of the domains of the variables in $\mathcal{S}_j$.

Given a set of n genes, a GRNi is a CSP $< \mathcal{X}, \mathcal{D}, \mathcal{C} >$, where

$$\mathcal{X} =< x_1, x_2, ..., x_{n^2-n} > \quad (1)$$

, each x_k is a regulatory relation, excluding self regulations. And

$$\mathcal{D} =< D_1, D_2, ..., D_{n^2-n} > \quad (2)$$

, with each $D_k = \{0, 1, ..., 100\}$, is the set of possible confidence values associated with such relation. And $\mathcal{C}$ is a list of constraints expressing properties of the GRNs. $x_{s \to t}$ is the variable associated with the regulatory relation "s regulates t" and $D_{s \to t}$ its domain. A solution to the above CSP defines a GRN prediction $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with $\mathcal{V} = \{0, 1, ...100\}$ and $\mathcal{E} = \{< s, w, t > |d(x_{<s,t>})| > 0\}$ where $w = d(x_{<s,t>})/100$.

The suggested constraints are: Sparsity, Redundant Edge, Transcriptor Factor, and a Coregulator constraints.

C. Performance Metrics and Evaluation

To prove the correctness and fidelity of the inferred GRN, a network validation is essential after the inference. For an undirected network, the inferred network model generated from validation data needs to be compared with the associated gold standard to see the consistency of the presence or absence of edges in gold standard. For a directed network, not only consistency of edge presence and absence, but also the consistency of edge directionality needs to be evaluated. True positive (TP), true negative (TN), false positive (FP), and false negative (FN) edge counts should be calculated. And precision (specificity) and recall (sensitivity), defined as $P = TP/(TP+FP)$ and $R = TP/(TP+FN)$ respectively, can be calculated and used as performance assessments.

A receiver-operating characteristic (ROC) curve (Balakrishnan, 1992) denotes the performance consistency of inference method under different conditions. In a ROC graph, y and x axis indicate true positive rate (TPR) and false positive rate (FPR), which can be further calculated with the respective formula: $TPR = TP/(TP+FN)$ and $FPR = FP/(FP+TN)$

References

[1] Fioretto, Ferdinando & Pontelli, Enrico. (2013). Constraint Programming in Community-Based Gene Regulatory Network Inference. 8130.10.1007/978-3-642-40708-6_11.

[2] Liu, Enze & Li, Lang & Cheng, Lijun. (2018). Gene Regulatory Network Review. 10.1016/B978-0-12-809633-8.20218-5.

[3] Delgado FM, Gómez-Vela F. Computational methods for Gene Regulatory Networks reconstruction and analysis: A review. Artif Intell Med. 2019 Apr;95:133-145. doi: 10.1016/j.artmed.2018.10.006. Epub 2018 Nov 9. PMID: 30420244.