Testing, Diagnosing, Repairing, and Predicting from Regulatory Networks and Datasets

by Torsten Schaub and Anne Siegel

We use expressive andhighly efficient tools from the area of Knowledge Representation for dealing with contradictions occurring when confronting observations in large-scale (omic) datasets with information carried by regulatory networks.

The availability of high-throughput methods in molecular biology has led to a tremendous increase of measurable data along with resulting knowledge repositories, gathered on the web usually within biological networks. However, both measurements and biological networks are prone to considerable incompleteness, heterogeneity, and mutual inconsistency, making it difficult to draw biologically meaningful conclusions in an automated way.

Further probabilistic and heuristic methods exploit disjunctive causal rules to derive regulatory networks from high-throughput -static- experimental data. For instance, disjunctive causal rules on influence graphs were originally introduced in random dynamical frameworks to study global properties of large-scale networks, using a probabilistic approach. These were demonstrated mainly on the transcriptional network of yeast. However, these methods are mostly data driven, and they lack the ability to perform corrections in a fast and global way. In contrast, efficient model-driven approaches based on model checkers - such as multi-valuated logical formalisms - are available to confront networks and measured data. These however, make use of time-series observations and can only be applied to small-scale parametered systems, since they need to consider the full dynamics of the system.

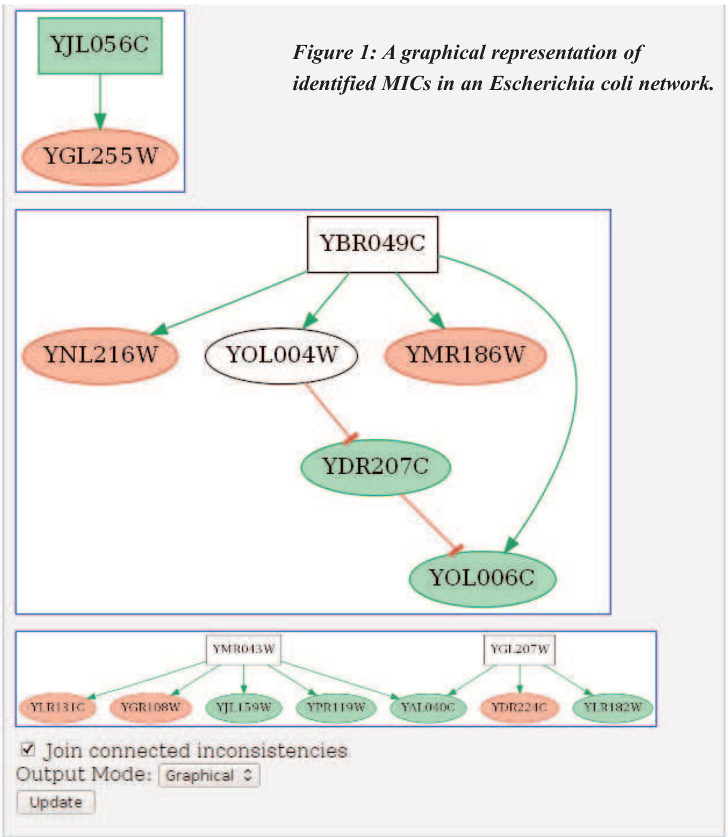
We have proposed an intermediate approach to perform diagnosis on large-scale static datasets. We use a Sign Consistency Model (SCM), imposing a collection of constraints on experimental measurements together with information on cellular regulations between network components.

The main advantage of SCM lies in its global approach to confronting networks and data, since the model allows the propagation of static information along the network and localization of contradictions between distant nodes. In contrast to available probabilistic methods, this model is particularly well-suited for dealing with qualitative knowledge (for instance, reactions lacking kinetic details) as well as incomplete and noisy data. Indeed, SCM is based on influence (or signed interaction) graphs, a common representation for a wide range of dynamical systems, lacking or abstracted from detailed quantitative descriptions.

By combining SCM with efficient Boolean constraints solvers, we address the problem of detecting, explaining, and repairing contradictions (called inconsistencies) in large-scale biological networks and datasets by introducing a declarative and highly efficient approach based on Answer Set Programming [1]. Moreover, our approach enables the prediction of unobserved variations and has shown an accuracy of over 90% on the entire network of E.Coli along with published experimental data. Notably, such genome-wide predictions can be computed in a few seconds.

From the application perspective, the distinguishing novel features of our approach are as follows: (i) it is fully automated, (ii) it is highly efficient, (iii) it deals with large-scale systems in a global way, (iv) it detects existing inconsistencies between networks and datasets, (v) it diagnoses inconsistencies by isolating their source, (vi) it offers a flexible concept of repair to overcome inconsistencies in biological networks, and finally (vii) it enables prediction of unobserved variations (even in the presence of inconsistency).

The efficiency of our approach stems from advanced Boolean Constraint Technology, allowing us to deal with



☑ Join connected inconsistencies
Output Mode: Graphical
Update

Figure 1: A graphical representation of identified MICs in an Escherichia coli network.

problems consisting of millions of variables. Although the basic tools [1] are implemented in C++ we haveimproved their accessibility by providing a Python library as well as a corresponding Web service [2].

Our project is a joint effort between the Knowledge Representation and Reasoning group [3] at the University of Potsdam and the SYMBIOSE Team [4] at IRISA and INRIA in Rennes. Our techniques have been developed in strong collaboration with the Max-Planck-Institute for Molecular Plant Physiology in Potsdam with the GoFORSYS Project [5] as well as Institut Cochin, Paris [6]. The members of the group include Martin Gebser, Carito Guziolowski, Jacques Nicolas, Max Ostrowski, Torsten Schaub, Anne Siegel, Sven Thiele, and Philippe Veber.

Links:

[1] http://en.wikipedia.org/wiki/Answer_set_programming
[2] http://potassco.sourceforge.net
[3] http://www.cs.uni-potsdam.de/bioasp/sign_consistency.html
[4] http://www.cs.uni-potsdam.de/wv
[5] http://www.irisa.fr/symbiose
[6] http://www.goforsys.org
[7] http://www.cochin.inserm.fr

Please contact:

Torsten Schaub
Universität Potsdam, Germany
E-mail: torsten@cs.uni-potsdam.de

Anne Siegel
CNRS/IRISA, Rennes, France
E-mail: Anne.Siegel@irisa.fr

MCMC Network: Graphical Interface for Bayesian Analysis of Metabolic Networks

by Eszter Friedman, István Miklós and Jotun Hein

The Data Mining and Web Search Group at the SZTAKI in collaboration with the Genome Analysis and Bioinformatics Group at the Department of Statistics, University of Oxford, developed a Bayesian Markov chain Monte Carlo tool for analysing the evolution of metabolic networks.

"Nothing in biology makes sense except in the light of evolution". The famous quote by Theodosius Dobzhansky (1900-1975) has been the central thesis of comparative bioinformatics. In this field, the biological function, structure or rules are inferred by comparing entities (DNA sequences, protein sequences, metabolic networks, etc.) from different species. The observed differences between the entities can be used for predicting the underlying function, structure or rule that would be too expensive and laborious to infer directly in lab. These comparative methods have been very successful in silico approaches, for example, in protein structure prediction. The idea can be used for inferring metabolic networks, too.

Metabolic networks are under continuous evolution. Most organisms share a common set of reactions as a part of their metabolic networks that relate to essential processes. A large proportion of reactions present in different organisms, however, are specific to the needs of individual organisms or tissues. The regions of metabolic networks corresponding to these non-essential reactions are under continuous evolution. By comparing metabolic networks from different species, we can find out which parts of the metabolic network are essential (ie those that are common in all networks) and which are non-essential (ie those that are missing in at least one of the networks). Sometimes there is more than one possible metabolic network that can synthesise or degrade a specific chemical. These alternative solutions can be transformed into each other, and the ensemble of all possible reactions form a complicated network (see Figure 1).

The central question is: what are the possible evolutionary pathways through which one metabolic network might evolve into another? This question is especially important to understand in the fight against drug-resistant bacteria. Drugs that are designed to protect us against illness-causing bacteria block an enzyme that catalyzes one of the reactions of the metabolic network of the bacteria. The bacteria, however, can avoid the effects of the drug by developing an alternative metabolic pathway. If we understand how the alternative pathways evolve we may be able to design a combination of drugs from which the bacteria cannot escape through the development of alternative pathways.

Analysis of past events is always coupled with some uncertainty about the nature and order of events that unfolded. It is therefore crucial to infer the evolution of metabolic networks using statistical methods that properly handle the uncertainty that inevitably occurs during analysis. Bayesian methods collect the a priori knowledge into an ensemble of distributions of random variables, set up a random model describing the changes, and calculate the posterior probabilities of what could happen. The relationship between the prior and posterior probabilities is described by the Bayes theorem as shown in Figure 2. Since the integral in the denominator is typically hard to calculate, and the Bayes theorem is often written in the form shown in Figure 3.

The Bayesian theorem in this form can be used in Monte Carlo methods to sample from the posterior distribution. The Markov chain Monte Carlo (MCMC) method sets up a Markov chain that converges to the desired distribution. After convergence, samples from the Markov chain follow the prescribed distribution.

MCMC Network implements the above-described Bayesian MCMC framework for inferring metabolic networks. We model the evolution of networks with a time-continuous Markov