Inferability and inference of gene regulatory networks

Author(s):
Ud-Dean, S.M. Minhaz

Publication Date:
2016

Permanent Link:

https://doi.org/10.3929/ethz-a-010686553 →

DISS. ETH NO. 23400

INFERABILITY AND INFERENCE OF GENE REGULATORY NETWORKS

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

Presented by

S M MINHAZ UD-DEAN

Master of Science in Appied Physics
Delft University of Technology
&
Master of Science in Photonics
Friedrich-Schiller University Jena

Born on 24.09.1985

Citizen of BANGLADESH

Accepted on the recommendation of

Prof. Dr. Rudiyanto Gunawan (examiner)
Prof. Dr. Konrad Hungerbuehler (co-examiner)

2016

Abstract

Understanding how different genes regulate each other is a prerequisite for systems-level modelling of molecular biology. Despite the development of a large number of inference algorithms, the inference of gene regulatory networks (GRN) from gene expression data remains an important unsolved problem of systems biology. The difficulty can be attributed to the fact that the accuracy of inferred GRNs depends not only on the ability of the algorithm to extract causal relationships from data, but also on the availability of relevant information in the data. Often it is not possible to accurately infer a GRN irrespective of the applied method because the available data do not contain sufficient information relevant to the causal structure of the network. This lack of relevant causal information causes the GRN inference problem to become underdetermined, and consequently there could be many equivalent (indistinguishable) solutions, implying that the GRN is not uniquely inferable. While the underdetermined nature of the GRN inference and its significance have been recognized in the community, the attention given to the inferability of GRNs still pales in comparison to the attention given to the development of inference algorithms.

The inferability of a GRN is determined by the causal information in a dataset, which in turn depends on the experiments performed to obtain the dataset. In consequence, during GRN inference it is necessary to take the performed experiments into consideration and analyze the inferability of the network from those experiments. Additionally, one should optimize the experiments in order to obtain new dataset that would improve the inferability of the network. Once again, the design of optimal experiments for GRN inference has received little attention compared to the development of inference methods. Motivated by these gaps, we sought to address the underdetermined issue of GRN inference in this thesis. Specifically, we have developed new framework and ensemble inference algorithms, called Transitive Reduction and

Closure Ensemble (TRaCE) for inferability analysis of gene regulatory networks from steady-state data of gene knock-out (KO) experiments. In addition, we have created REDuction of UnCertain Edges (REDUCE), an algorithm for designing optimal gene KO experiments.

Using data from a set of gene KO experiments, TRaCE and its extension TRaCE+ generate the upper bound (most complex network) and the lower bound (least complex network) of an ensemble of networks consistent with the data. The regulatory interactions that belong to the upper bound but not to the lower bound are the uncertain interactions that could not be verified by prior data. The number of uncertain interactions represents the inferability of the network, with fewer uncertain edges indicating better inferability. TRaCE analyzes the inferability of the GRN only accounting for the existence of regulatory interactions, while TRaCE+ improves upon TRaCE by considering both the existence and mode of the regulation (activation/repression). In consequence, usually TRaCE+ can extract more information from the same data.

In the case studies we applied TRaCE to analyze the inferability of random GRNs and the GRNs of *Escherichia coli* and yeast from single- and double-gene KO experiments. The results showed that, with the exception of networks with very few interactions, GRNs are typically not inferable. Using realistic simulation data, we compared the performance of TRaCE with the top performing methods of DREAM 4, a community-wide network inference challenge. The results demonstrated that TRaCE performs better than the top performing state-of-the-art gene regulatory network inference algorithms.

Our design of experiments algorithm, REDUCE employs uncertain gene interactions that could not be verified by available data during ensemble inference by TRaCE or TRaCE+. REDUCE generates the optimal gene KO experiment by maximizing the number of uncertain interactions that can be verified by the resulting data. For this purpose, we introduced the

concept of edge separatoid which gave a list of nodes (genes) that upon their removal would allow the verification of a specific gene regulation. We also developed an iterative inference strategy using TRaCE (or TRaCE+) and REDUCE. The iterative strategy involves estimating the bounds of a GRN by TRaCE (or TRaCE+), design of experiments by REDUCE, performing the designed experiment and using the data to update the ensemble bounds. Importantly, this strategy can infer the true network from error-free data, thereby resolving the issue of GRN inferability. Even using noisy data, the iterative strategy can converge to a single network. The case studies including the inference of *E. coli* GRN and DREAM 4 100-gene GRNs, demonstrated the efficacy of the iterative GRN inference. In comparison to systematic KOs, REDUCE could provide much higher information return per gene KO experiment and consequently more accurate GRN estimates. In the case studies involving the inference of the DREAM 4 100-gene GRNs, the iterative strategy required fewer iterations and KO experiments when using TRaCE+ compared to using TRaCE. The iterative inference of GRNs using TRaCE(+) and REDUCE represents an enabling tool for tackling the underdetermined GRN inference. Along with advances in gene deletion and automation technology, the iterative procedure brings an efficient and fully automated GRN inference closer to reality.

Zusammenfassung

Das Verständnis der regulatorischen Beziehungen zwischen verschiedenen Genen ist eine Voraussetzung zur Modellierung der Molekularbiologie auf der Systemebene. Trotz der Entwicklung einer Vielzahl von Inferenzalgorithmen, bleibt die Identifizierung der Genregulationsnetzweke (GRN) aus Genexpressionsdaten ein wichtiges ungelöstes Problem der Systembiologie. Die Schwierigkeit ist darauf zurückzuführen, dass die Genauigkeit der identifizierten GRN nicht nur von der Fähigkeit der Algorithmen aus den Daten kausale Beziehungen zu fördern abhängt, sondern auch durch die Verfügbarkeit relevanter Informationen in den Daten beinflusst wird. Unabhängig der angewandten Methoden ist es häufig unmöglich ein GRN mit Genauigkeit zu identifizieren, weil die verfügbaren Daten nicht ausreichend Informationen über die kausale Struktur des Netzwerks enthalten. Dieser Mangel an relevanten Kausalinformationen resultiert in einem unterbestimmten GRN Identifizierungsproblem. In Folge kann es mehrere äquivalente (ununterscheidbare) Lösungen geben, das heisst das GRN ist nicht eindeutig identifizierbar. Obwohl die Unterbestimmtheit der GRN Identifizierung und ihre Signifikanz von der wissenschaftlichen Gesellschaft anerkannt ist, verblasst die Aufmerksamkeit, welche der Identifizierbarkeit von GRN zuteil wird, im Vergleich zu der Aufmerksamkeit auf die Entwicklung von Inferenzalgorithmen.

Die Identifizierbarkeit eines GRNs wird durch die kausale Information in einem Datensatz bestimmt, welche wiederum von den Experimenten, mit denen der Datensatz erzeugt wurde, abhängt. Daher ist es notwendig die durchgeführten Experimente zu berücksichtigen, um die Identifizierbarkeit des Netzwerkes mit Bezug auf die Experimente zu analysieren. Weiter sollte man die Experimente optimieren, welche den neuen Datensatz erzeugen, der wiederum die Identifizierbarkeit des Netzwerks verbessern würde. Wiederum ist die Aufmerksamkeit, welche dem Design von Experimenten zur Identifizierung des GRNs zuteil wird, im Vergleich zur Entwicklung von Inferenzmethoden sehr klein. Motiviert durch diese Kenntnislücke, haben

wir in dieser Dissertation versucht die Unterbestimmtheit der GRN Identifizierung unter die Lupe zu nehmen. Genauer gesagt, haben wir einen neuen theoretischen Rahmen und neue Algorithmen zur Ensemble Inferenz entwickelt, welche unter dem Namen Transitive Reduction and Closure Ensemble (TRaCE) zur Analyse der Identifizierbarkeit der Genregulationsnetzweke aus Genexpressionsdaten im Gleichgewichtszustand von Gen-Knockout (KO) Experimenten dienen. Zusätzlich haben wir einen Algorithmus, REDuction of UnCertain Edges (REDUCE), zum Design optimaler KO Experimenten entwickelt.

Aus einem Datensatz einer Serie von KO Experimenten erzeugen TRaCE und seine Erweiterung TraCE+ die obere (das komplexeste Netzwerk) und die untere Grenze (das einfachste Netzwerk) eines Ensembles der Netzwerke, welche konsistent mit dem Datensatz sind. Die regulatorischen Beziehungen, welche zur oberen aber nicht zu der unteren Grenze gehören, sind die unbestimmten Beziehungen, welche aus dem Datensatz nicht verifizierbar sind. Die Zahl der unbestimmten Beziehungen stellt die Identifizierbarkeit des Netzwerks dar, wobei wenige unbestimmte Beziehungen für eine bessere Identifizierbarkeit stehen. Während die Analyse der Identifizierbarkeit des GRNs durch TRaCE nur die Existenz der regulatorischen Beziehungen betrachtet, stellt TRaCE+ eine Verbesserung von TRaCE dar, indem es nicht nur die Existenz, sondern auch den Typ (Aktivierung/Hemmung) der Regulation betrachtet. Folglich kann TRaCE+ aus demselben Datensatz mehr Information gewinnen als TRaCE.

In den Fallstudien haben wir TRaCE angewendet um die Identifizierbarkeit von zufälligen GRN und die GRN von *Escherichia coli* and *Saccharomyces cerevisiae* aus Einzeln- und Doppel-Gen KO Experimenten zu analysieren. Die Ergebnisse zeigen, dass GRN, ausser Netzwerke mit sehr wenigen Beziehungen, typischerweise nicht identifizierbar sind. Wir haben realistische Simultaionsdaten angewendet um die Leistung von TRaCE mit den Leistungen der besten Algorithmen von DREAM 4, eines gemeinschaftsweiten Wettbewerbs, zu vergleichen.

Die Ergebnisse zeigen, dass die Leistung von TRaCE besser als die Leistungen der stärksten bekannten GRN Inferenzalgorithmen ist.

Unser Algorithmus zum Design von Experimenten, REDUCE, verwendet die unbestimmten Beziehungen, welche aus den verfügbaren Daten durch TRaCE oder TRaCE+ nicht identifiziert werden konnten. REDUCE erzeugt das optimale Gen-KO Experiment durch Maximierung der Anzahl an unbestimmten Beziehungen, welche aus dem entstehenden Datensatz verifiziert werden kann. Zu diesem Zweck haben wir das Konzept des Kanten-Separatoids eingeführt, welcher eine Liste von Knoten (Genen) ergibt, durch deren KO eine spezifische Genregulation verifiziert werden kann. Wir haben ebenfalls eine iterative Inferenzstrategie, welche TRaCE (oder TRaCE+) und REDUCE verwendet, entwickelt. Diese iterative Strategie besteht aus der Schätzung der Grenzen eines GRNs durch TRaCE (oder TRaCE+), dem Design von Experimenten mittels REDUCE, der Durchführung der Experimente und der Anwendung der Daten zur Aktualisierung der Ensemblegrenzen. Es ist wesentlich, dass diese Strategie aus fehlerfreien Daten das echte Netzwerk identifizieren kann und somit das Problem der GRN Identifizierbarkeit löst. Sogar aus verrauschten Daten kann diese Strategie ein eindeutiges Netzwerk identifizieren. Die Fallstudien, einschließlich der Identifizierung des *E. coli* GRN und DREAM 4 100-Gen GRN, zeigen die Wirksamkeit dieser iterativen Strategie zur GRN Inferenz auf. Im Vergleich zu systematischen KOs, kann REDUCE deutlich mehr Information pro KO-Experiment generieren und damit genauere GRN erzeugen. In den Fallenstudien zur Inferenz von DREAM 4 100-gene GRN brauchte die iterative Strategie weniger Iterationen und KO-Experimente wenn TraCE+ statt TRaCE angewendet wurde. Die iterative Inferenz von GRN mit TRaCE(+) und REDUCE stellt ein Verfahren dar, welches es ermöglicht die Unterbestimmtheit der GRN-Inferenz zu bewältigen. Zusammen mit den Fortschritten in der Gen-Deletion und der Automatisierungstechnik, bringt das iterative Verfahren eine effiziente und vollständig automatisierte GRN Identifizierung der Realität näher.