

An Indian-Australian research partnership

Project title: Modeling and inference of gene regulatory network

Project number: IMURA0150

Monash University supervisors: Dr. Madhu Chetty

Monash University contact: Dr. Madhu Chetty; Email: madhu.chetty@infotech.monash.edu.au

IITB supervisors: Prof. Pramod Wangikar

IITB contact: Prof Pramod Wangikar; Email: pramodw@che.iitb.ac.in

Research Academy theme/s

1. Advanced computational engineering, computational intelligence and simulation
2. Biotechnology and Bioinformatics

The research problem

The difficulties of genetic network inference include:

- (1) Inference of the regulatory networks based on microarray data means treating other entities, such as proteins and metabolites, as hidden variables. This will produce uncertain results.
- (2) It is required to consider the combinatorial nature of gene regulation (one gene might be regulated by multiple gene products).
- (3) The number of measurements (arrays) is very limited compared to the large number of objects (genes). This is true especially for complex models and large-scale networks. For example, there are 13,600 genes for fruit fly, 20,000 to 25,000 for Human and 45,000 for rice. But the number of samples is very limited for a specific experiment (usually only several or tens of samples).
- (4) The gene expression measurements are noisy, due to variations among different individuals, low quantities of some RNAs and measurement errors. Also, data from different experiments may not be directly comparable.
- (5) The gene interactions may happen within specific time periods and conditions instead of across the whole expression profiles. Catching these transient interactions is challenging.
- (6) The data may be under sampled. Some fast changing information may not be captured.
- (7) The exact mechanisms of regulatory interactions are usually unclear.
- (8) It is challenging to integrate prior knowledge or to resolve conflicts during network inference. Much more research is needed in this area.

We have identified several research problems in modelling and inference of gene regulatory network from microarray data.

Problem 1 *Development of model for large scale gene regulatory network inference*

Problem 2 *Improving the proposed method for prediction of causal structure of the GRN*

Problem 3 *Identifying the feedback loop in GRN, incorporating dynamicity and identifying transient gene interactions.*

Problem 4 *To infer more detailed gene regulatory relationships and hypotheses*

Problem 5 *Reducing Computation time and cost*

If the student has a strong background in System Theory and wishes to work with deterministic modelling approach, e.g. representing the system as a set of non-linear differential equations, this can also be considered. As we know, the non-linear differential equations are useful by "Control System Engineers" to model many complex large scale systems.

Project aims

- To develop a novel causal model for Gene regulatory network for identifying regulatory interactions.
- To refine the model further to identify spurious causality relationship in the model by the application of say a technique such as Minimum Message Length (MML) technique
- To investigate model enhancement for incorporating feedback, transient gene interaction and dynamicity.
- To improve computational speed of the model by application of distributed and grid computing techniques.

Expected outcomes

Genetic regulation plays a fundamental role in biological processes. Regulatory systems cannot simply be described as an assembly of genes and proteins and diagrams of their interconnections. Many analysis techniques, as for example cluster analysis, only provide 'correlations' between genes and do not provide insight into causal relations between the genes in a regulatory network. An important option for the analysis of regulatory control systems are simulation models in combination with optimization algorithms such as GA. In this project we develop a novel causal model for simulating regulatory networks that are capable of quantitatively reproducing spatial and temporal expression patterns in developmental processes. De la Fuente et al had proposed gene regulatory inference using partial correlation and d-separation theory. Their work mainly focuses on explaining how and to what extent partial correlation can lead to identifying gene interactions, which inspired us to propose this model. Major challenges faced in this work are large scale network inference, identifying direct and indirect causal relationships, identifying feedback loop gene interactions and transient gene interactions. Search and optimization algorithms (genetic algorithms) will be used in combination with the model to explore large spaces of regulatory networks and to infer actual gene regulatory networks from expression data.

The primary outcome of this study will be a generalized computational model for inferring gene regulatory network from microarray data. This work will be highly significant for at least three reasons. Firstly, it will represent a significant step towards a systems-level understanding of the eukaryotic such as yeast cell cycle. When we are able to infer system dynamics, protocols, and the design principles of the eukaryotes, we will be able to extend to larger datasets (which can represent mammalian targets). Secondly, it will provide novel approach/methods (causal modeling) that will underpin the modeling of systems-level biological processes and underlying mechanisms. Thirdly, the statistically supported gene network model established in this study will seed similar studies of other organisms and could expedite the discovery of new strategies for a wide variety of diseases.

Which of the above Theme does this project address?

The project is multidisciplinary and falls in a new area of 'System Biology' and addresses two main themes: 1) Advanced computational engineering, computational intelligence and simulation. ii) Biotechnology and Bioinformatics.

How will the project address the Goals of the above Themes?

The project is a cutting edge and multidisciplinary project.

It is said (David Knuth comment), "There are millions and millions of unsolved problems. Biology is so digital, and incredibly complicated, but incredibly useful. The trouble with biology is that, if you have to work as a biologist, your experiments take you three years and then, one night, the electricity goes off and all the things die! You start over. In computers we can create our own worlds ...

It is hard for me to say confidently that, after fifty more years of explosive growth of computer science, there will still be a lot of fascinating unsolved problems at peoples' fingertips, that it won't be pretty much working on refinements of well-explored things. Maybe all of the simple stuff and the really great stuff has been discovered. It may not be true, but I can't predict an unending growth. I can't be as confident about computer science as I can about biology.

Computational Biology easily has 500 years of exciting problems to work on..."