Chapter 20
Integrating Various Data Sources for Improved Quality in Reverse Engineering of Gene Regulatory Networks

Mika Gustafsson
Linköping University, Sweden

Michael Hörnquist
Linköping University, Sweden

ABSTRACT

In this chapter we outline a methodology to reverse engineer GRNs from various data sources within an **ODE** *framework. The methodology is generally applicable and is suitable to handle the broad error distribution present in microarrays. The main effort of this chapter is the exploration of a fully data driven approach to the* **integration** *problem in a "***soft evidence***" based way.* **Integration** *is here seen as the process of incorporation of uncertain a priori knowledge and is therefore only relied upon if it lowers the prediction error. An efficient implementation is carried out by a* **linear programming** *formulation. This LP problem is solved repeatedly with small modifications, from which we can benefit by restarting the primal* **simplex method** *from* **nearby solutions***, which enables a computational efficient execution. We perform a case study for data from the* **yeast cell cycle***, where all verified genes are putative regulators and the a priori knowledge consists of several types of binding data, text-mining and annotation knowledge.*

INTRODUCTION

Biological systems are intrinsically complex, still robust and at the same time able to quickly adapt to new situations. To understand, describe and model a wide range of biological systems −involving genes, proteins, metabolites and ecological food webs− networks have served as the unifying language (Barabasi *et al.* 2004). This description has often revealed a complex network topology. In the case of

DOI: 10.4018/978-1-60566-685-3.ch020

Gene Regulatory Networks (GRNs), some features are the existence of key genes regulating multiple processes ("hubs"), feed-back motifs and modularity enhancing the system robustness (Milo *et al.* 2002; Barabasi *et al.* 2004). Furthermore, the dynamical systems seem to be tuned to enable a stable system by keeping hubs repressed, but still flexible by utilizing, e.g., incoherent feed-back loops (Gustafsson *et al.* In press b, Ma'ayan *et al.* 2008). In addition to the architectural complications, we know that gene regulation is a non-linear process including combinatorial control, saturation and stochasticity. These pieces give raise to an extremely challenging modelling problem, which becomes even more complicated by the size of the genome.

Further, the experimental advancements in the last decades have resulted in a vast amount of large-scale data sets available through public databases. To infer a large-scale GRN it is of uttermost importance to take as much as possible of these data into account. Particularly informative for understanding genome-wide gene regulation is the interaction map between Transcription Factors (TFs) and their DNA binding regions. This information may give direct structural properties of the regulatory possibilities, e.g., the presence of a binding element upstream of gene of A for a TF which gene B codes for induces an enhanced possibility for regulation of gene A by gene B.

Other types of structural information may come from sequence based predictions, e.g., prediction of putative regulations from the TF binding sites (TFBS) and from common biological knowledge. The latter can be incorporated in a variety of ways, which may come from annotation knowledge or more "unclean" knowledge as text-mining. Annotation knowledge may be the collection of detailed knowledge from previous experiments, while text-mining may be a possibility to include the plethora of published biological papers in databases. On a more detailed causal level there is also a large number of time-series expression data sets for mRNA levels (see, e.g., Omnibus at Entrez (PubMed 2007) for collections at a unified format). However, although all these experiments are present on a large-scale, they are all typically several orders of magnitudes smaller than the number of presumptive regulators. Hence, all data at hand should be taken in consideration to overcome the indefiniteness of the reverse engineering problem. The greatest challenge in GRN inference to tackle is that the number of genes vastly exceeds the number of experiments, making it a tough statistical question. We should therefore strive to avoid introducing more entities in the model. Consequently, we project gene regulation onto the space of genes only, despite the fact that gene regulation is carried out from the interactions of mRNA molecules, proteins and metabolites (Brazhnik *et al.* 2002; Ptashne *et al.* 2002). Indeed, the obtained GRN is then an effective network of gene-to-gene interactions, where these interactions cannot be interpreted as biochemical reactions.

Reverse engineering of genome-scale GRNs is a grand challenge for system biologists, with a high potential for drug discovery. The challenge consists in taking many small pieces of information ranging from widely different experiment types and **prior knowledge** properly into account. However, most of the genome-wide experiments are associated with great uncertainties, thus connected with many false positive and negative regulations. Nevertheless, algorithms have gradually become more refined, from the first cluster analyses of gene expression data (Eisen *et al.* 1998), to more recent dynamic network inferences (Segal 2003; Luscombe *et al.* 2004; Bonneau *et al.* 2006; Gustafsson *et al.* 2005; Wang *et al.* 2006) taking more data into account (Luscombe *et al.* 2004; Bonneau *et al.* 2006). The next step to get more accurate descriptions of the GRNs is to carefully take different data sources into account, such as TF-bindings, protein-interactions, sequence information, literature knowledge and of course expression data. The introduction of several data types in the reverse engineering process enforces a method to weight the data types appropriately, i.e., to prioritize, filter and in some cases discard the data based on