Reconstructing Linear Gene Regulatory Networks

Jochen Supper, Christian Spieth, and Andreas Zell

Centre for Bioinformatics (ZBIT), University of Tübingen, Germany
jochen.supper@uni-tuebingen.de

Abstract. The ability to measure the transcriptional response after a stimulus has drawn much attention to the underlying gene regulatory networks. Here, we evaluate the application of methods to reconstruct gene regulatory networks by applying them to the SOS response of *E. coli*, the budding yeast cell cycle and *in silico* models. For each network we define an *a priori* validation network, where each interaction is justified by at least one publication. In addition to the existing methods, we propose a SVD based method (NSS). Overall, most reconstruction methods perform well on *in silico* data sets, both in terms of topological reconstruction and predictability. For biological data sets the application of reconstruction methods is suitable to predict the expression of genes, whereas the topological reconstruction is only satisfactory with steady-state measurements. Surprisingly, the performance measured on *in silico* data does not correspond with the performance measured on biological data.

1 Introduction

Measuring the transcriptional response of cellular processes under various conditions provides valuable insight into the global behavior. Such measurements can be analyzed by clustering, thereby providing undirected gene relations. In order to model gene regulatory networks (GRN), directed relations between genes have to be considered. Different approaches, describing how the expression of the genes is controlled, have been proposed. GRNs can be represented by interconnections between genes, each indicating that one gene influences the expression of another gene. The functionality associated with the gene-gene interconnections ranges from simple Boolean interactions to complex differential equations [1–3].

The properties of GRNs have been analyzed on a global scale [4] and recently also on the individual gene scale [5]. The global organization of GRNs seems to follow a power-law distribution, where most genes are controlled by a small number of genes and a few genes are highly connected. At this point it is important to stress that the global structure is a static property of the GRN, and cellular perturbations only activate a subset of this network. Therefore, expression profiles only cover a subset of all possible cellular conditions and thus provide

E. Marchiori, J.H. Moore, and J.C. Rajapakse (Eds.): EvoBIO 2007, LNCS 4447, pp. 270–279, 2007.

only partial information about the underlying regulatory program. On a single gene scale quantitative characterization of the promoter activity has been investigated [6]. The common picture from these investigations is, that genes are activated at a certain threshold, then the dose-response curve progresses linearly until it saturates at its maximal fold-change. These investigations allow general questions regarding the proper conceptual and mathematical representation as well as the sufficient amount of detail required for modeling.

The major problem for the reconstruction of GRNs is that it is very hard to validate the hypothesized networks. This makes it difficult to compare methods and yet to judge if certain approaches are helpful at all. In previous publications, GRN models have been validated by co-citation [7], cross-validation [8] or on artificial data [9]. Despite these efforts, no topological validation on biological models, in which each interconnection is *a priori* justified by published work, has been assessed for different reconstruction methods. In this work, we present two biological networks that have been investigated thoroughly, with plenty of known interactions, as well as an *in silico* GRN model. In addition to the topological validation, we predict the gene expression by 5-fold cross-validation.

Several methods have been developed to reverse engineer genetic networks. Since the proper modeling terms of GRNs are not yet known, we avoid specific assumptions by using standard methods and reconstruction algorithms. Therefore, we apply linear regression, greedy search, exhaustive search and methods developed by Weaver *et al.*[10] and Someren *et al.*[11]. The description of these linear models is straightforward, whereas non-linear interactions can be modeled by applying a preprocessing step. Thereby, the gene-gene dependence can be altered, such that the interaction terms are non-linear and multiplicative. To investigate the influence of the preprocessing, we incorporate and evaluate different types of preprocessing.

In addition to comparing different reconstruction algorithms, we propose our own reconstruction method for GRNs, the Null Space Solver (NSS). The core algorithm was previously described by Supper *et al.*[9]. To reconstruct the GRN, we employ a heuristic based on Singular Value Decomposition (SVD). Different groups have previously used SVD to reconstruct GRN [3]. Our approach searches the solution space in an efficient way by exploiting the properties of the SVD. The search steps are probabilistic and provide an ensemble of networks, which we rank according to different criteria. Our first criteria is to choose the solution with the minimal number of interactions. If several such solutions exist, we rank these according to the 1-Norm.

Besides choosing the proper reconstruction method it is critical to obtain applicable data sets. Transcriptional response data is characterized by a large number of measured genes along with a small sampling rate. Consequently, the number of features is high compared to the number of sampling points, rendering the reconstruction of a GRN ambiguous. By restricting ourselves to sub-networks we relax this problem so that the number of measurements is in the range of the number of genes. We also perform a reconstruction on an additional data set containing 800 genes for which this relaxation is not possible.