

INFORMATION EXTRACTION FOR MODELING GENE EXPRESSIONS

E.P. van Someren^a, L.F.A. Wessels^{a,b} and M.J.T. Reinders^a

^aInformation and Communication Theory Group, ^bControl Engineering,
Faculty of Information Technology and Systems,
Delft University of Technology
Mekelweg 4, P.O.Box 5031, 2600 GA Delft, The Netherlands
E.P.vanSomeren@its.tudelft.nl

Recent bio-technological developments have made it possible to measure the activity levels of thousands of genes over time, giving biologists insight into the complex network of interactions between genes. Typically, such high-dimensional data-sets consists of many signals (genes) and relatively few time-points, making the inference of the exact network parameters impossible. Consequently, computational techniques are necessary that can cope with this dimensionality problem and extract as much useful information as possible from the data such that a basic sketch of the underlying interactions can be formed. In this paper, the regulatory interactions between genes are modeled as a linear genetic network, which enables the exact description of all solutions that match the given data-set. Biologically sound constraints, such as limited connectivity and redundancy can be incorporated to produce, based on this exact description, a network that still fits the data, but is as simple as possible in its structure.

INTRODUCTION

Traditionally, the research in molecular biology employed a reductionist approach, i.e. studying a single gene, protein or reaction at a time. With enzymatic or genetic reaction systems, this approach is problematic due to the high number of interactions and/or objects that need to be studied. Recent technological developments resulted in the availability of high-throughput gene expression assays that produce large-scale data about the activity levels of individual genes over time. Therefore, the need arises for data analysis methods that process this data

in a global fashion, and produce interpretable results at some intermediate level. In this paper we propose a modeling approach which represents genetic interactions as a connected network of nodes, generally known as a genetic network. The ultimate goal is the complete "reverse engineering" of the underlying regulatory interactions between genes.

The major problem concerning these measured data-sets is that they generally consist of hundreds to thousands of signals, measured on no more than twenty time-steps. From an information theoretic point of view this dimensionality problem will render any network model inferred from this data virtually meaningless. Due to the nature of the application and the costs involved in making a measurement this problem will not be solved by increasing the number of measurements. The answer to this problem lies in the extraction of useful information from the data by incorporating sensible constraints on the modeling process based on existing knowledge. For example, it is estimated that, on average, each gene interacts with four to eight other genes [1] and that biological networks usually contain a certain amount of redundancy [3]. This implies that the number of connections between nodes in the genetic network are limited and that the same connections are shared among nodes.

Currently, several different types of models are studied, like Boolean networks [8], Bayesian networks [6, 7], (Quasi)-Linear networks [4], Neural networks [10] and Differential Equations [2]. In this paper, a linear model is employed to model the regulating interactions between genes, because 1) it uses relatively few network parameters, 2) the parameters are interpretable, 3) it allows for an exact description of all parameters that fit the given data-set and 4) this exact description allows constraints to be applied without introducing an error in fitting the data. The aim of our approach is to yield a network model that visualizes both the *information* as well as the *uncertainty*.

THE LINEAR MODEL

Gene expression measurements can be represented in a so called gene expression matrix,

$$\mathbf{X} = [x_{i,t} \mid i \in 1, \dots, N \quad t \in 1, \dots, T] \quad (1)$$

, where each row, denoted by \mathbf{x}_i , represents the gene-profile of gene i taken over T time-points. The t -th column of \mathbf{X} is denoted by $\mathbf{x}(t) = [x_1(t), \dots, x_N(t)]^T$ and

determines the state of the system at time t . The linear model serves as a representation of the regulatory interaction between genes and follows the assumption that the activity level¹ of a gene at a certain point in time can be determined by the weighted sum of the activity levels of all genes at the previous time-point², i.e.

$$\mathbf{x}(t+1)^T = \mathbf{x}(t)^T \cdot \mathbf{R} \quad \forall t = 1, \dots, T-1 \quad (2)$$

The first goal is to find all weight-matrices \mathbf{R} that are consistent with our data, i.e. satisfy Eq. (2). In general, the weight-matrix will be under-constrained which means that there exist multiple solutions which can be written as a combination of a particular solution \mathbf{P} , a basis of homogeneous solutions \mathbf{H} and a set of free variables \mathbf{F} , i.e.

$$\mathbf{R} = \mathbf{P} + \mathbf{H} \cdot \mathbf{F} \quad (3)$$

The particular solution \mathbf{P} is *one* of the solutions that satisfies Eq. (2), while the homogeneous solution $\mathbf{H} \cdot \mathbf{F}$ reflects the remaining ambiguity in the data (that part of the weight-matrix that does not alter signal levels). The set of free variables \mathbf{F} reflects the degrees of freedom as each element can be substituted with any particular value without changing the estimation of the given data. For a given data-set the particular and homogeneous solution can be found by Gaussian elimination.

GENERALIZATION

For actual biologically measured gene-expression matrices, the amount of ambiguity, i.e. the number of columns in \mathbf{H} , denoted by M , is large as the number of measurements T is significantly less than the number of genes N . Although the ambiguity is exactly known it is hard to represent it in an interpretable way. To extract an interpretable result from the set of possible solutions, it is necessary to find the simplest model which will be a rough sketch of the underlying network. In [9] we exploited the fact that genes are redundant and consequently employed hierarchical clustering to group together genes with similar profiles and represented them with a prototype. Next, the relationships between these prototypical genes were learned, resulting in a model that sketches the basic in-

¹Throughout the paper we denote the activity level of a gene by the logarithm of the ratio between normal and sample mRNA level.

²In other words, relationships between genes are assumed to be stationary.

teractions among *groups of genes*. In this paper, we follow a different approach: i.e. we aim at minimizing the number of control actions that are exerted on each gene *individually*. As a result both the total number and the specific set of controlling genes will differ among each individual gene. This way, the number of time-points that are required to correctly identify the weights of a sparsely connected network remains on the order of the maximal connectivity. We will show that this approach elegantly alleviates the dimensionality problem and even more so than the clustering approach that we proposed in [9]. This approach minimizes the number of genes that control another gene, which is supported by biological evidence in [1].

OPTIMIZATION

The principle idea is that we manipulate the homogeneous solution, (Eq. 3), such that the number of zero elements along each column of the weight-matrix is maximized separately. In order to force a specific weight to be zero, i.e. $r_{i,j} = 0$, the following equation must be satisfied:

$$\mathbf{h}_i \cdot \mathbf{f}_j = -p_{i,j} \quad (4)$$

, where \mathbf{h}_i is the i -th row of \mathbf{H} and \mathbf{f}_j the j -th column of \mathbf{F} and $p_{i,j}$ is the (i, j) -th element of the particular solution \mathbf{P} . Thus, each weight-element that is forced to be zero will place certain restrictions on the free variables. In order to simultaneously force multiple elements in one weight-column to zero, say K elements in column j , a set of K simultaneous equations must be satisfied. Assume that such a particular set of elements is denoted by c_k ³ consisting of their row-indices m_k and column-index $n_k = j$, then:

$$\mathbf{H}_{m_k} \cdot \mathbf{f}_j = -\mathbf{p}_{c_k} \quad (5)$$

, where \mathbf{p}_{c_k} consists of the K elements in \mathbf{P} (with indices c_k) and \mathbf{H}_{m_k} consists of the K rows in \mathbf{H} (with row-indices m_k) that correspond to the K weights that are forced to be zero. If there exists a solution to Eq. (5) then this solution, denoted by \mathbf{f}_{c_k} , can, in turn, be described by a particular and a homogeneous solution (the

³Note that given K zeros and N genes, that k ranges from 1 to $\binom{N}{K}$.

latter might be empty), i.e. :

$$\mathbf{f}_{c_k} = \mathbf{f}_{c_k}^p + \mathbf{f}_{c_k}^h \cdot \mathbf{f}_{c_k}^f \quad (6)$$

Substitution in Eq. (3) results in the following change of the j -th column of the weight-matrix:

$$\mathbf{r}_j = (\mathbf{p}_j + \mathbf{H} \cdot \mathbf{f}_{c_k}^p) + (\mathbf{H} \cdot \mathbf{f}_{c_k}^h) \cdot \mathbf{f}^f \quad (7)$$

The first term in brackets denotes the new particular solution with elements c_k forced to zero, whereas the second term in brackets denotes the new basis of homogeneous solutions, reflecting the new degrees of freedom under the restriction that the desired elements c_k remain zero. Note that the number of columns in the new basis of homogeneous solutions as well as the remaining free variables in \mathbf{f}^f will always be equal to or less than the number of free variables in \mathbf{f}_j .

In essence, this approach reduces the ambiguity after introducing *information* about the value of some of the weight-values in \mathbf{R} . Our goal now is to find that set of variables \mathbf{F} that introduces the maximum number of zero weights out of the N weights in each column of weight-matrix \mathbf{R} so that \mathbf{R} still satisfies Eq. (2). Assume that there exists a solution that forces K elements out of the N in a column to zero. To find out exactly which set of K elements can be forced to zero we will need to test $\binom{N}{K}$ solutions of Eq. (5). When the number of zero's are not known beforehand, we need to test all combinations for all possible $K \leq N$: i.e. $\sum_{K=0}^N \binom{N}{K}$ tests. Fortunately, we don't need to follow this brute force strategy, because the particular solution \mathbf{P} already contains at least as many zero weights in each column as there are columns in the basis of homogeneous solutions \mathbf{H} , namely M and because we aim only to find more zero weights this gives us a lower bound of the K that we need to test. In fact, we only have to solve Eq. (5) for all $\binom{N}{K=M}$ possible combinations c_k . If a solution exists, Eq. (6) will give the corresponding set of free parameters \mathbf{f}_{c_k} . It now may turn out that different combinations will be caused by the same set of free parameters, i.e. $\mathbf{f}_{c_{k1}} = \mathbf{f}_{c_{k2}}$. In such a case the intersection of both solutions will cause more than M zeros, in fact as many as the number of elements in the union of the corresponding combinations c_{k1} and c_{k2} . Hence, choosing this intersecting solution for \mathbf{f}_j will yield a more optimal solution in terms of maximizing the zero weights. In general, we will choose for each column in \mathbf{F} , the free parameter set corresponding to that intersection point that occurs most often in the solutions, hence gives rise to the most zeros.

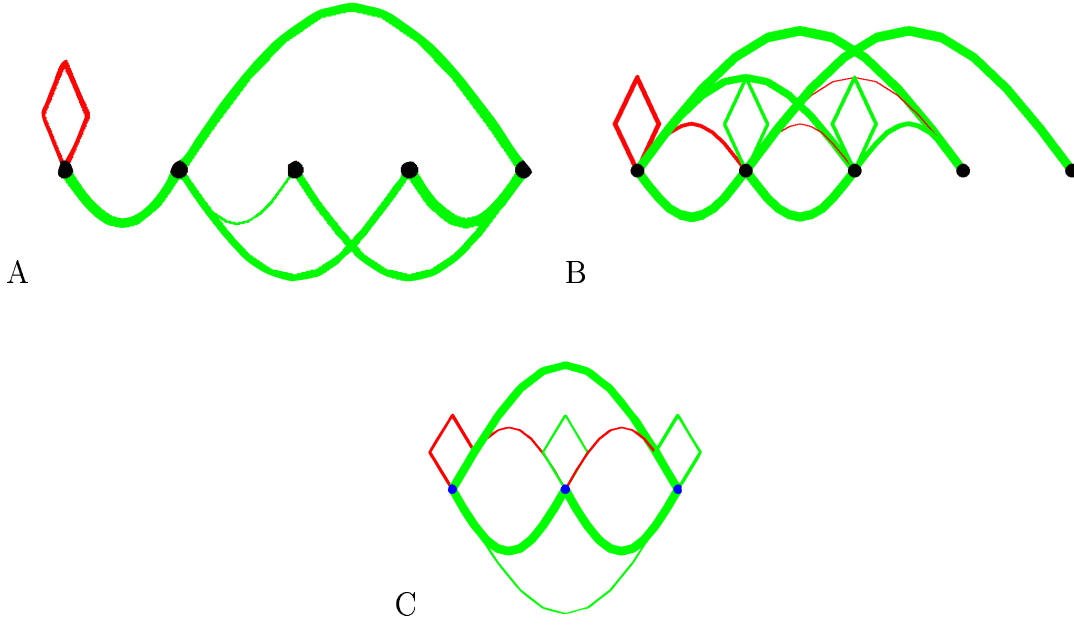


Figure 1: A) The original model. B) The particular solution. C) Clustering approach. Each dot represents a single gene and each arc between two dots represents the control from one gene to the other. The color indicates a positive (light) or negative (dark) control action. Arcs drawn above the genes indicate a direction from left to right, whereas arcs below go from right to left.

EXAMPLE

Our approach is illustrated on a sparse network that contains a feedback loop with redundancy, i.e. some genes share the same inputs and outputs. The corresponding weight-matrix is illustrated as a graph in Fig. 1.A. An initial state is initialized randomly and from this initial state consecutive states are computed by Eq. (2) using the given weight-matrix. In this way, a data-set is generated that consists of five signals with four time-points. By employing Gaussian elimination on this data-set we get a representation of the set of all weight-matrices that perfectly match this data, as follows:

$$\mathbf{R} = \begin{bmatrix} -0.5 & -0.41 & 0.8 & 1 & 0 \\ 1 & 0.3 & -0.12 & -0.14 & 1 \\ 0 & 0.9 & 0.39 & 0.49 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0.55 & -1 \\ -0.4 & 0.14 \\ -0.88 & -0.49 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \cdot [\mathbf{f}_1 \ \mathbf{f}_2 \ \mathbf{f}_3 \ \mathbf{f}_4 \ \mathbf{f}_5] \quad (8)$$

The particular solution is graphically depicted in the Fig. 1.B, illustrating that it is not the desired representation of the original model (i.e. the smallest number of connections between genes). The clustering approach, on the other hand, captures the basic loop but introduces additional relations, because it neglects the differences between individual genes in each group. Although, the current particular solution contains at least $M = 2$ zero weights for each column, our approach can find more! After calculating $\binom{5}{2} = 10$ possible combinations the following choice for \mathbf{F} results in the sparsest solution for \mathbf{R} .

$$\mathbf{F} = \begin{bmatrix} 0 & .75 & 0 & 0 & 0 \\ 0 & 0 & .8 & 1 & 0 \end{bmatrix} \quad (9)$$

Substitution in Eq. (8), results in the following sparse solution, which exactly equals the original model:

$$\mathbf{R} = \begin{bmatrix} -0.5 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & .25 & 0 & 0 & 0 \\ 0 & .75 & 0 & 0 & 0 \\ 0 & 0 & .8 & 1 & 0 \end{bmatrix} \quad (10)$$

In fact, it exactly identified the original model, simply because the original model only contains genes that are controlled by at most two other genes. Note that the original model was retrieved notwithstanding the fact that the number of observed time-points was smaller than the number of genes.

DISCUSSION

In this paper we presented a method to find the sparsest genetic network given insufficient time-response data. This method is motivated by the principle that simpler (sparser networks) explanations should be preferred if the data does not specify a unique solution. In addition, this is also in accordance with biological evidence indicating that genetic networks tend to be sparsely connected. A theoretical example was presented, and experiments were conducted on real data-sets, but the computational complexity proved to be still a major obstacle. This problem can be partially solved by preceding the application of the proposed method with a clustering step. This has the additional advantage that the number of

prototypes need not be reduced up to the point where the homogeneous solution becomes empty (previously this was required to obtain a unique solution). The clustering process can now be terminated when a biologically sound grouping is obtained, and the remaining degrees of freedom can be employed to optimize the sparseness of the genetic network.

REFERENCES

- [1] A. Arnone and B. Davidson. The hardwiring of development: Organization and function of genomic regulatory systems. *Development*, 1997.
- [2] T. Chen, H.L. He, and G.M. Church. Modeling gene expression with differential equations. *Pacific Symposium on Biocomputing '99*, 4:29–40, 1999.
- [3] P. D'Haeseleer, S. Liang, and R. Somogyi. Genetic network inference: From co-expression clustering to reverse engineering. *Submitted to Bioinformatics*, 2000.
- [4] P. D'Haeseleer, X. Wen, S. Fuhrman, and R. Somogyi. Linear modeling of mrna expression levels during cns development and injury. *Pacific Symposium on Biocomputing '99*, 4:41–52, 1999.
- [5] D. Sharp E. Mjolsness and J. Reinitz. A connectionist model of development. *Journal of Theoretical Biology*, 152, 1991.
- [6] N. Friedman, M. Goldszmidt, and A. Wyner. Data analysis with bayesian networks: A bootstrap approach. *Proc. Fifteenth Conf. on Uncertainty in Artificial Intelligence (UAI)*, 1999.
- [7] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyze expression data. *Submitted*, 1999.
- [8] S. Liang, S. Fuhrman, and R. Somogyi. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pacific Symposium on Biocomputing '98*, 3:18–29, 1998.
- [9] E.P. van Someren, L.F.A. Wessels, and M.J.T. Reinders. Linear modeling of genetic networks from experimental data. *Accepted for ISMB2000*, 2000.
- [10] D.C. Weaver, C.T. Workman, and G.D. Stormo. Modeling regulatory networks with weight matrices. *Pacific Symposium on Biocomputing '99*, 4:112–123, 1999.