Robust Target Identification for Drug Discovery

Elling W. Jacobsen, Torbjörn E. M. Nordling

Dep of Automatic Control, KTH Royal Institute of Technology,
Stockholm 10044, Sweden
Dept. of Mech. Eng., National Cheng Kung University, Tainan 70101,
Taiwan (e-mail: jacobsen@kth.se, tn@kth.se)

Abstract:
A key step in the development of new pharmaceutical drugs is that of identifying direct targets of the bioactive compounds, and distinguishing these from all other gene products that respond indirectly to the drug targets. Currently, dominating approaches to this problem are based on often time consuming and costly experimental methods aimed at locating physical bindings of the corresponding small molecule to proteins or DNA sequences. In this paper we consider target identification based on time-series expression data of the corresponding gene regulatory network, using perturbation with the active compound only. As we show, the problem of identifying the direct targets can then be cast as a linear regression problem and, in principle, be accomplished with a number of samples equal to the number of involved genes and bioactive compounds. However, the regression matrix will typically be highly ill-conditioned and the target identification therefore prone even to small measurement uncertainties. In order to provide a label of confidence for the target identification, we consider conditions that can be used to quantify the robustness of the identification of individual drug targets with respect to uncertainty in the expression data. For this purpose, we cast the uncertain regression problem as a robust rank problem and employ SVD or the structured singular value to compute the robust rank. The proposed method is illustrated by application to a small scale gene regulatory network synthesised in yeast to serve as a benchmark problem in network inference.

Keywords: Systems biology, systems medicine, target identification, network inference, robust regression, gene regulatory networks, drug discovery

1. INTRODUCTION

When developing new pharmaceutical drugs, an important task is optimization of the therapeutic efficacy while keeping undesirable side effects at a minimum. For this purpose, it is important to determine the so called mode of action of the drug, i.e., the molecular targets that mediate the therapeutic effects as well as the side effects [Hopkins, 2008]. Drug discovery can follow various paths, but all can more or less be divided into two fundamental approaches [Schenone et al., 2013]. The first approach, based on biochemical assays, first determines a desirable protein target and this protein is then exposed to various small molecules to detect molecules that can bind to and modify the protein [Schenone et al., 2013]. In the second approach, involving phenotypic assays, small molecules are added to cells or animals and their impact on the phenotype of interest is measured. For compounds with a desirable impact on the phenotype, a subsequent time consuming and costly task is target identification, i.e., determination of the genes or gene products that are directly perturbed by the added active compound. Target identification is the process of identifying the direct molecular target - for example a protein or a nucleic acid - of a small molecule [Schenone et al., 2013]. From a systems perspective, adding a small molecule drug to a living cell is equivalent to directly perturbing one or more nodes in an intricate network of genes and gene products. In general, the result of this perturbation is a response in essentially all nodes of the network (all genes and their products) and the task of target identification is to identify the direct target nodes and distinguish these from the nodes that respond indirectly to the target nodes.

The traditional approach to target identification is to employ various experimental methods aimed at locating the physical binding sites of the active compound [Ursu and Waldmann, 2015]. Apart from being a costly and time consuming process, and usually the rate limiting step in drug development [Ursu and Waldmann, 2015], physical binding does not necessarily imply an observable effect on the protein or gene activity. Hence, it it preferable to identify targets based on actual response data from the cell. This has led several authors to propose model based methods in which response data are used to infer network models, including the direct targets of external perturbations, for the underlying gene regulatory networks. Previous work in this area, e.g., [Gardner et al., 2003, diBernardo, 2005, Woo, 2015], have assumed steady-state response data only, in which it is necessary to have availability of a model of the gene regulatory network (GRN) prior to performing target identification from responses.
to the active compound of interest. Obtaining the GRN model from steady-state response data requires a large number of experiments, in principle equal to the number of genes of interest, in which a different perturbation (e.g., RNAi) is used in every experiment and this makes also these methods costly and time consuming. Furthermore, the final step in which the targets are identified from the GRN in combination with response data from the active compound is in general highly sensitive to uncertainties in the obtained GRN model.

While previous work in this area have assumed the availability of a network model describing the interactions among all genes and gene products, we will in this paper consider the feasibility of determining the direct targets based on response data obtained from a single perturbation experiment with the active compound only. For this purpose, we employ a linear dynamic model structure for the gene regulatory network and formulate the combined network inference and target identification problem as a single linear regression problem. As we show, target identification is indeed feasible under some mild controllability conditions on the gene regulatory network. In principle, in the absence of uncertainty, both the GRN and the direct targets can be identified after having collected a number of samples equal to the number of genes and active compounds. However, the collected response matrices will in general be strongly ill-conditioned and the identified targets therefore sensitive to small uncertainties in the model structure and measurement data. To provide a label of confidence we introduce a model for the measurement uncertainty and formulate a robust rank condition for target identification at a certain level of confidence. As we demonstrate, correctly inferring the GRN model is not a prerequisite for robust target identification, i.e., target identification will in general require less samples than what is required for the overall GRN inference.

We start by formulating a general network model and the problem at hand in the context of this model, and show that in principle both the network interactions as well as the direct targets can be identified given a certain number of samples. Since target identification is a structural identification problem, rather than a parametric problem, we introduce a geometric perspective to the formulated regression problem to derive conditions for when an edge (link) should be included in a network model. These conditions involve the rank of certain matrices constructed from the collected response data and the rank will in general be prone to small uncertainties in the elements. We therefore introduce an uncertainty model and formulate a robust rank problem using the structured singular value or the singular value decomposition depending on the nature of the uncertainty model employed. The solution of the robust rank problem provides confidence scores for the inferred edges, including the edges from external perturbations corresponding to the direct targets. We finally illustrate the proposed robust target identification method on a small gene regulatory network synthesised in yeast to serve as a benchmark problem in network inference [Cantone, 2009], before concluding the paper.

2. A LINEAR MODEL FOR TARGET IDENTIFICATION

Within a cell, genes interact by regulating each others activity at the transcriptional or post-transcriptional level. These regulatory interactions imply that the genes, and their products, form a network in which the nodes represent the gene activities (RNA or protein concentrations) and the edges represent regulatory mechanisms. As an example, consider the IRMA network [Cantone, 2009] in Figure 1, in which 5 genes interact with each other and Galactose acts as a small molecule perturbation directly targeting the Gal80 gene activity. The aim in target identification is to identify which genes or proteins the small molecule directly binds to and thereby modifies the activity of. For this purpose we here consider formulating a linear model of the network and then formulate the target identification as a regression problem based on time-series response data obtained from the intact network in the cell.

![Fig. 1. The engineered IRMA gene regulatory network involving 5 genes.](image)

We consider a biological system in a particular physiological state. Then it is reasonable to approximate the gene regulatory network using a system of linear first order ODEs, i.e., a linear state space model

\[
\dot{x}(t) = Ax(t) + Bp(t) \\
y(t) = Cx(t)
\]

The state vector \(x(t) = [x_1(t), \ldots, x_n(t)]\) contains the value of each gene activity at time \(t\), \(p(t)\) is an external perturbation and \(y(t)\) is the measurement of \(x(t)\), all defined relative to their corresponding steady-state values. A non-zero element \(a_{ij}\) in the interaction matrix \(A\) corresponds to direct regulation of gene \(i\) by gene \(j\). Similarly, a non-zero element \(b_i\) in the \(B\)-matrix corresponds to indirect effect of the external perturbation (small molecules) on the activity of gene \(i\). We will assume that all genes are directly measured, implying that the \(C\)-matrix is the identity matrix. Note that we do not include any uncertainty in the model at this stage, but that we will introduce uncertainty descriptions in section 4.

While the model class we consider here is linear, also certain nonlinear models of relevance to biochemical systems may be accommodated within this framework. For
instance, the mass action like model \( \dot{x}_i = p_i \prod_j x_j^{n_j} - d_i x_i \) can be transformed into the linear form (2) by the transformation \( \tilde{x}_i = \log x_i, \forall i \) and \( \tilde{p}_i = \log p_i, \forall i \).

The aim in target identification is now to determine, based on response data, non-zero elements in the \( B \)-matrix of the network model (2). Assuming steady-state response data only, the problem will be underdetermined unless the interaction matrix \( A \) is known. In this case it is therefore necessary to first perform \( n \) steady-state experiments with different external perturbations in each experiment in order to determine the \( A \)-matrix. Note that for all these experiments, the corresponding \( B \) matrix of each perturbation must also be known. This makes target identification based on steady-state response data quite a demanding task. We therefore consider here what can be achieved using time-series data instead.

With time series data, the relation between subsequent samples and the perturbations is given by the discrete time version of (2)

\[
\Delta x_{k+1} = (A_d - I) x_k + B_d p_k \tag{3}
\]

\[
y_k = x_k \tag{4}
\]

where \( \Delta x_{k+1} = x_{k+1} - x_k \). Note that we here are concerned with identifying the structure of \( B \) (and possibly \( A \)) in (2), and that in general the structure of \( B_d \) and \( A_d - I \) will differ from their continuous time counterparts. If we consider zero-order hold discretisation of (2), then \( A_d = e^{AT}, B_d = \exp(\{(A_d - I)B \}T) \) where \( T \) is the sampling time. Here \( A_d \) and \( B_d \) will in general be full matrices, and hence any structure in \( A \) and \( B \) is lost in sampling. This may seem discouraging for identifying the structure of the network from sampled data. However, first note that \( (A_d - I)/T, B_d/T \) will asymptotically converge to \( A, B \) as the sampling time \( T \) goes to zero. This can easily be seen from the Taylor series representation of the matrix exponential

\[
e^{AT} = I + AT + \frac{1}{2} A^2 T^2 + \ldots
\]

and \( A_d - I \) converges to \( AT \) and \( B_d \) to \( BT \) as \( T \to 0 \) and we hence recover the structure of \( A \) and \( B \) in the discrete time model. Thus, with reasonably fast sampling, zero elements in \( A, B \) will correspond to relatively small elements in \( A_d, B_d \) and with the introduction of uncertainty and robustness analysis considered below, these can then not be distinguished from other sources of uncertainty. We therefore assume here that we are concerned with identifying the non-zero elements of \( B_d \), where non-zero later will be defined as robustly non-zero. Also note that, due to degradation, the diagonal elements of \( A \) and hence \( A_d \) will always be non-zero, why the structure of \( A_d \) and \( A_d - I \) will be the same. Below we therefore simply consider \( A_d \).

Consider now an experiment in which we add the compound of interest to the cell and measure the response in the genes, or gene products, of interest with a sample time \( T \). We collect \( m \) samples, and store these in the matrix \( Y \in \mathbb{R}^{n \times m} \) and the perturbations in \( P \in \mathbb{R}^m \). Also introduce the matrices

\[
M = \begin{bmatrix} y_{m-1} & y_{m-2} & \cdots & y_1 \\ p_{m-1} & p_{m-2} & \cdots & p_1 \end{bmatrix}
\]

Then we can write

\[
R = [A_d \ B_d] M \tag{5}
\]

Trivially, if the number of samples \( m = n + 1 \) and the matrix \( M \) is invertible, we can determine \( A_d, B_d \) from

\[
[A_d \ B_d] = R M^{-1}
\]

Invertibility of \( M \) can be guaranteed under some mild controllability condition on the system and perturbation [Schmidt et al., 2005]. Thus, in the absence of uncertainty in measurements and model it is possible to obtain the matrix \( B_d \) from a single experiment with addition of the active compound of interest only and collection of \( n + 1 \) samples where \( n \) is the number of genes in the network. The targets, corresponding to non-zero elements in \( B \) in (2), can then be obtained from an inverse zero order hold transformation (assuming perturbations are kept fixed between samples).

The system identification approach discussed above can be useful in cases where a predictive model is the main objective. However, in target identification, we are primarily concerned with identifying the structure of the \( B \)-matrix, and predictive capabilities and parameter estimates are less of a concern. With only small uncertainties in the model and/or measurements, the structure of the underlying network is lost in \( B_d \), and eventually \( B \), as these in general then will both be full matrices. A possible remedy could be to employ a statistic model and estimate the covariances of the estimated parameters. However, the covariances are given by a so called marginal estimator, which does not account for alternative models \([Ad, Bd]\) that can explain the data. They should therefore not be used to make statements about the uncertainty of individual edges, since it is a statement about the existence of an alternative model.

A popular approach to identify structures of sparse models is through regularisation of the corresponding regression problem, e.g., by penalising non-zero elements using an \( l_0 \) or \( l_1 \)-norm penalty, e.g., as in LASSO [Tibshirani, 1996]. However, as shown in Nordling and Jacobsen [2011], while this has proven a useful method for obtaining sparse models as such, it will in general not provide a good estimate of the true network structure but rather result in a large number of both false positives (identified edges not existing in the system) and false negatives (missed edges). See also Tjärnberg et al. [2015]. To overcome these problems, we here adopt the geometric view of the regression problem presented in Nordling and Jacobsen [2011] to determine conditions on when an edge, i.e., a non-zero parameter in \( A \) or \( B \), should be included in the network model. We then add uncertainty to the regression problem and extend the derived conditions to this case, resulting in a robust rank problem which we solve using results from robust control theory.

3. A GEOMETRIC PERSPECTIVE ON LINEAR REGRESSION

Consider rewriting the identification problem in (5) as a standard linear regression problem

\[
\Phi \Theta = \Xi \tag{6}
\]
Here $\Phi = M^T$, $\Xi = R^T$ and $\Theta = [A_d \ B_d]^T$. The regression problem can be solved independently for the columns of $\Theta$ and $\Xi$

$$\Phi \theta_j = \xi_j$$

(7)

Here the columns $\phi_{ij}$ of $\Phi$ and the vector $\xi_j$ are $m-1$-dimensional vectors, where $m$ is the number of samples. The vector $\xi_j$ is then a linear combination of the vectors $\phi_{ij}$, $i = 1, n+1$ and the parameters in $\theta_j$ act as corresponding weights.

Identification of the structure of $A_d, B_d$ corresponds to determining which parameters in $\theta_j$ that should be non-zero, i.e., which regressor vectors $\phi_{ij}$ that are needed to explain the regressand $\xi_j$. Any regressor $\phi_{ij}$ which spans a unique direction in $m-1$-dimensional space and which furthermore is present in the regressand $\xi_j$ must be included and hence the corresponding parameter $\theta_{ij}$ be non-zero. Likewise, if $\phi_{ij}$ spans a unique direction and this direction is absent in $\xi_j$, then the parameter $\theta_{ij}$ must be zero.

A simple test to determine if parameter $\theta_{ij}$ should be non-zero is thus to create the matrix [Nordling and Jacobsen, 2011, Nordling, 2013]

$$\Psi_{ij} = [\psi_1 \ ... \ \psi_{i-1} \ \psi_{i+1} \ ... \ \psi_{n+1} \ \xi_j]$$

(8)

If the matrix $\Psi_{ij}$ has full rank $n+1$, then the parameter $\theta_{ij}$ should be non-zero, i.e., the corresponding edge in the network exists. Note that the parameters of the $n+1$th column of $\Theta$ correspond to the $B_d$ matrix of the network model (4) which is of main concern in target identification.

4. A ROBUST RANK CRITERION FOR TARGET IDENTIFICATION

The geometric interpretation of the linear regression problem presented above provides some insight into the structure identification problem which will prove useful when we have uncertainty in the regressor $\Phi$ and regressand $\Xi$. Consider adding uncertainty to $\Psi_{ij}$ according to

$$\hat{\Psi}_{ij} = \Psi_{ij} + V_{ij}$$

(9)

where $\hat{\Psi}_{ij}$ is a set of matrices assumed to include the true matrix and $\Psi_{ij}$ is the corresponding measurement matrix. The matrix $V_{ij}$ represents the uncertainty and is as such unknown. To characterise the uncertainty we use a norm-bound in which we either bound the matrix norm $\|V_{ij}\|_2 \leq \nu$ or the norm of the individual columns $\|v_{ij}\| \leq \nu$. Bounding the size of the columns corresponds to independent uncertainty in the measurement of each individual gene (regressor), but we also include the matrix norm bound as this leads to a simpler test for robust rank as shown below.

Introducing uncertainty in (9) implies that we will have a set of network models rather than a single model as considered in (5). If a regressor $\phi_{ij}$ is need to explain $\xi_j$, for all models in the set, then we say that the corresponding parameter $\theta_{ij}$ is robustly non-zero and the corresponding edge corresponds to a robust target identification. The test for a robust edge, or target, is that the matrix $\hat{\Psi}_{ij}$ has full rank $n+1$ for all allowable $V_{ij}$. That is, we need to derive a test for robust rank of a matrix.

If we first consider the case in which the uncertainty is such that $\|V_{ij}\|_2 \leq \nu$, then a sufficient condition for robust rank of $\hat{\Psi}_{ij}$ is that the $n+1$th singular value of $\Psi_{ij}$ is larger than $\nu$, i.e.,

$$\sigma_{n+1}(\Psi_{ij}) > \nu$$

(10)

With $j = n + 1$, Equation (10) represents a sufficient condition for target identification of gene $i$ with respect to the added active compound in the case where the uncertainty matrix $V_{ij}$ is norm-bounded.

As stated above, it will often be more natural to consider independent measurement uncertainty for each gene, corresponding to norm-bounding the columns of the uncertainty matrix $V_{ij}$. For simplicity, we here assume that the measurements have been scaled so that the same norm-bound $\nu$ applies to all columns of $V_{ij}$. For this case we employ the structured singular value to compute the robust rank, and just state the result without proof here. For proofs we refer to [Nordling, 2013]. The matrix $\Psi_{ij}$ will have full rank $n+1$ for all models in the set generated by $\|v_{ij}\|_2 \leq \nu$ if the confidence score

$$\gamma_{ij} = \frac{1}{\mu_A(\Psi_{ij})} > 1$$

(11)

where $\mu_A$ is the structured singular value computed for a diagonal real matrix $A = diag(\delta_1, ..., \delta_{n+1})$ with $|\delta_i| < \nu$.

With $j = n + 1$, Equation (11) represents a sufficient condition for target identification of gene $i$ wrt to the added active compound in the case where the uncertainty of the columns of $V_{ij}$ are norm-bounded.

We next illustrate the usefulness of the proposed method through application to target identification of the IRMA network.

5. EXAMPLE: TARGET IDENTIFICATION IN IRMA NETWORK

![Fig. 2. The network structure of the IRMA network. We consider a small molecule binding directly to gene 2 (GAL4) and gene 5 (GAL80).](image)

The IRMA network is a 5 gene network synthesized in yeast for the purpose of benchmarking network inference methods [Cantone, 2009]. The network structure is illustrated in Figure 1 and Figure 2. Since the strength of the
network interactions are largely unknown we here consider a model with the same structure as the IRMA network but with unit strength for all edges. The linear model is then

$$\dot{y} = \begin{bmatrix} -1 & 0 & 1 & -1 & 0 \\ 1 & -1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & -1 & 1 & 0 & -1 \end{bmatrix} y(t) + \begin{bmatrix} 0 \\ 10 \\ 0 \\ 0 \\ 5 \end{bmatrix} p(t) \quad (12)$$

Thus, we consider perturbation with a small molecule that directly binds to gene 2 and gene 5, having twice the strength on gene 2. The aim is then to infer non-zero elements in $B$ from experimental data obtained after perturbations with the small molecule only. For this purpose we perform an experiment in which we sample with $T = 0.1h$ and add a new random perturbation in the small molecule every three samples. Without uncertainty, target identification is as stated above feasible with only 7 samples. However, due to uncertainty more samples will probably be required to make a robust target identification. We therefore collect a total of 60 samples and then compute the robustness measure presented above as a function of the number of samples to see how many samples are required for robust target identification. As uncertainty, we use norm bounded uncertainty on the measurement of each gene, that is, we bound the 2-norm of the columns of $V_{ij}$ as

$$\|v_{ij}\|_2 \leq 0.05\sqrt{m}$$

where $m$ is the number of samples considered. This corresponds approximately to allowing 0.05 additive uncertainty in each gene measurement in each sample.

As stated above, if we neglect uncertainty, a unique network model $A_B, B_T$ can be inferred with only $m = 7$ samples. However, computing the corresponding robustness measure (11) based on the structured singular value we obtain the following confidence scores for the $B_T$-matrix

$$\gamma_{16} = \begin{bmatrix} 0.0266 & 0.0288 & 0.0301 & 0.0114 & 0.0289 \end{bmatrix}$$

For robust target identification these measures should be above 1 and hence we see that we are far from robust target identification after 7 samples. Figure 3 shows the confidence scores $\gamma_{16}$ for target identification, i.e., robust inference of the $B_T$-matrix elements, as a function of the number of samples collected and as can be seen the $\gamma$-scores for some of the targets exceed 1 after about 50 samples. The two targets that eventually are robustly identified with $\gamma$ scores larger than 1 are correctly genes 2 and 5. Note that they are identified with approximately the same value of the robustness confidence score $\gamma$ despite having quite different strengths in the response to the added small molecule. For both targets we need 53 samples for robust target identification for which the $\gamma$-scores obtained are

$$\gamma_{16}^{m=53} = \begin{bmatrix} 0.0606 & 1.0336 & 0.3090 & 0.0751 & 1.0556 \end{bmatrix}$$

We note that the $\gamma$-value for gene 3, albeit being much less than 1, is significantly larger than those for genes 1 and 4, and this is probably explained by the fact that the corresponding element in the discretised $B_T$-matrix is distinctly different from zero. However, even if we increase the number of samples further, only genes 2 and 5 are identified as targets.

We also computed the corresponding robustness measure for the network interactions in the $A$ matrix, and interest-

6. SUMMARY AND CONCLUSIONS

We have in this paper considered the problem of target identification in gene regulatory networks based on a single experiment with perturbation by the active compound of interest only. By posing the target identification problem as a linear regression problem and computing the robust
rank of a matrix consisting of the relevant regressors and regressands, we showed that robust target identification indeed is feasible from time-series data from a single experiment. In general, significantly more samples than the number of nodes in the network will be required for robust target identification. The effectiveness of the method for identifying the targets was demonstrated on a small gene regulatory network synthesised in yeast, and the two direct targets were correctly identified.

REFERENCES


