International Conference on Control, Engineering & Information Technology (CEIT'13)
Proceedings Engineering & Technology - Vol.1, pp.38-42, 2013

A Combined Constraint Approach for Inference of Sparse Large-Scale Biomolecular Networks

Mohammed Mohammed-Rasheed[1], Nidhal Bouaynaya[2] and Hassan Fathallah-Shaykh[3]
[1] Department of Applied Science, University of Arkansas at Little Rock
[2] Department of Systems Engineering, University of Arkansas at Little Rock
[3] Department of Neurology, University of Alabama at Birmingham

***Abstract*—In this paper, we address the problem of inferring sparse biomolecular networks from a limited number of noisy measurements. We model the gene expression dynamics using a system of ordinary differential equations corrupted by additive white noise. We derive the gene interactions that maximize the likelihood function while constraining the network to be sparse. We consider a convex combination of three l_1, l_2 regularization terms that take into account (i) the sparsity of the network, (ii) irrelevant predictors and (iii) irrelevant responses. The last two constraints are especially important when inferring large-scale networks. We further propose a procedure to recover weak interactions based on the Bayesian information criterion. It has been shown that weak interactions are important to preserve the structure of functional linkages among pathways. We conduct Monte Carlo simulations to study the effect of the model parameters and the number of measurements on the error rate. Our simulation results show that the proposed approach to estimate the strength of molecular interactions from noisy measurements outperforms the l_1-constrained maximum likelihood method.**

I. Introduction

Inference of genetic regulatory networks from gene and protein expression profiles is an important problem in genomic signal processing and systems biology as it lies at the bottleneck of genetic-based therapies [1], [2], [3], [4], [5], [6], [7]. Inference or reverse-engineering of the underlying genetic regulatory network can be harnessed into educated intervention of diseases including cancer. Different models and methods have been used in the literature to construct genetic networks from high-throughput data, including Gaussian graphical models [8], Boolean networks [9], Bayesian networks [10], dynamic Bayesian networks [11] and ordinary differential equations (ODEs) [7] [12].

Differential equation models of gene regulatory networks are generally preferable to graphical models for numerous reasons. Graphical models may be undirected (e.g., correlation-based networks), acyclic (e.g., Bayesian networks) and cannot distinguish between stimulative and repressive relations because they model probabilistic dependencies among variables. On the other hand, differential equations result in model networks that are directed, allow for feedback loops, and categorize stimulations and inhibitions. The ability to recognize if a gene interaction is stimulative or inhibitory is particularly crucial in understanding transcriptional regulatory interactions and designing appropriate drug targets. Indeed, from a pharmacological viewpoint, a drug target is either inhibited or activated by drug molecules, e.g., small organic molecules, antibodies, therapeutic proteins. Furthermore, ODE models can be used to predict the behavior of the network under different conditions, e.g., gene knockout, treatment with an external agent, etc. However, the ODE models considered previously in the literature are deterministic. Genomic data, on the other hand, is very noisy and thus a meaningful estimation method must take into account the noise in the data and must possess a certain degree of stability for a range of noise levels. In addition to the noise, the number of measurements or samples is smaller than the number of genes, which makes the system under-determined and thus not identifiable. One strategy to obtain a meaningful formulation of the problem is to assume a number of predictors for each gene (i.e., assume a number of zeros for the connectivity matrix) and carry out a least-squares estimation [13], [14]. In most practical cases, however, no a priori knowledge is available about the true number of predictors for each gene. This led some researchers to resort to combinatorial approaches, where they try all possible combinations of predictors and select the number of zeros that corresponds to the minimum least-squares error [1].

In this paper, we model the gene interactions using a linear ODE system [4] with an additive noise term. Following the work in [15], we formulate the network parameter estimation as a constrained multivariate regression problem. We reduce the number of necessary observations by constraining the estimation problem using a combined constraint function that includes l_1 and l_2 regularization terms. The purpose of the constraints is to impose sparsity in order to identify "master predictors. Bimolecular networks are known to be sparse, i.e., a molecule (gene or protein) usually interacts with only a small subset of the total number of molecules in the network [4]. In addition, we impose constraints to identify irrelevant predictors (zero columns in the connectivity matrix) and irrelevant responses or molecules that cannot be predicted by the considered set of molecules (zero rows in the connectivity matrix). These two constraints are especially important to include when estimating large-scale networks. We solve the convex constrained regression problem using standard convex optimization techniques. We further use the Bayesian Information Criterion (BIC) to uncover weak interactions and reduce the false negative rate. Specifically, we use the BIC to resolve the ambiguity about weak interactions, whether they reflect true interactions or artifacts of the imperfections in the

data and the model. This is in contrast to the thresholding approach adopted in the literature, where interactions below a fixed (arbitrary) threshold are simply disregarded [13], [4], [2], [14]. We assess the performance of the proposed algorithm by generating synthetic biomolecular networks with various sizes and noise levels. We use Monte-Carlo simulations to assess the performance with respect to the effect of the tuning parameters, the network sizes, the number of measurements and the sparseness level of the network. Our simulation results show that the proposed method outperforms the classical unconstrained maximum likelihood approach.

II. Gene Network Model

We consider a linear dynamical system in which a gene regulatory network of N genes can be described by a linear ordinary differential equation model of the form

$$\frac{dx_i}{dt}(t_k) = \sum_{j=1}^{n} a_{ij}x_j(t_k) + b_i u(t_k) + \varepsilon(t_k), \tag{1}$$

where $i = 1,...,N$, $t_k = 1,...,M$ and M denotes the number of measurements or time points with $M < N$. $\frac{dx_i}{dt}$ is the rate of change in the concentration of gene (mRNA) i. a_{ij} is the influence of gene j on gene i. $u(t_k)$ is the amount of perturbation at time t_k, b_i is the effect of the external perturbation on the gene i and ε is the error or noise term due to imperfections in the data and the model The linearity of the model can be justified when the system is operating in the vicinity of its steady-state. We introduce the variable y_i

$$y_i(t) = \frac{dx_i(t)}{dt} - b_i u(t). \tag{2}$$

Equation (1) can be written in matrix form as:

$$\boldsymbol{y}_k = A\boldsymbol{x}_k + \boldsymbol{\varepsilon}_k, \tag{3}$$

where $A \in \mathbb{R}^{N\times N}$ is the gene-gene interaction matrix. For notation simplicity, we write $\boldsymbol{x}_k$ to denote the $N \times 1$ vector $\boldsymbol{x}(t_k)$. Writing Eq. (3) for all times $k = 1,\cdots,M$, we obtain

$$Y = AX + E, \tag{4}$$

where $X \in \mathbb{R}^{N\times M} = [\boldsymbol{x}_1,\boldsymbol{x}_2,...,\boldsymbol{x}_M]$, $Y \in \mathbb{R}^{N\times M} = [\boldsymbol{y}_1,\boldsymbol{y}_2,...,\boldsymbol{y}_M]^t$, $E \in \mathbb{R}^{N\times M} = [\boldsymbol{\varepsilon}_1,\boldsymbol{\varepsilon}_2,...,\boldsymbol{\varepsilon}_M]$. That is, every column of Y,X and E represents a single experiment. The aim of the network inference problem is to estimate the matrix A given the under-determined system in (4).

We pursue a probabilistic approach and assume that the measurement noise $\boldsymbol{\varepsilon}_1,\cdots,\boldsymbol{\varepsilon}_M$ is i.i.d. $\mathcal{N}(0,\sigma^2 I)$ with I being the identity matrix and σ^2 represents the noise power. The negative log likelihood function of (4) can be shown to be given by [15]

$$\text{-}l(A) = \text{Tr}[\frac{1}{M\sigma^2}(Y-AX)(Y-AX)^t] + 2N\ln\sigma, \tag{5}$$

where $\text{Tr}(X)$ denotes the trace of matrix X. The maximum likelihood estimate of the connectivity matrix A is therefore given by

$$\hat{A} = \underset{A}{\text{argmin}}\ \text{Tr}[(Y-AX)(Y-AX)^t] \tag{6}$$

The optimization problem in (6) has infinitely many solutions because the system $Y = AX$ is under-determined. In order to obtain a meaningful estimator, we constraint the problem by exploiting the fact that bimolecular networks are sparse.

III. Constrained Maximum Likelihood Estimation

A. The constrained maximum likelihood

Spareness is one of the important characteristics of gene interaction networks [2]. It is known from optimization theory that the l_p norm leads to sparse solution for $p \leq 1$ [16]. In fact, the smaller p is, the sparser the solution. At the limit, $p = 0$ and the l_0-norm counts the number of non-zero elements. However, it is only for $p = 1$ that the norm is convex and hence computationally tractable. Moreover, an interesting result due to Donoho and Huo [17] shows that the solution of an l_1 norm optimization problem in some cases coincides with the l_0 norm. We, therefore, use the l_1-norm to impose sparsity on the elements of the matrix A. In addition, we also constrain the sum of the l_2-norms of the columns and the rows of A. These additional constraints introduce zeros for all entries in some columns and rows, respectively, of A, meaning that some predictors are irrelevant for all responses and some responses are irrelevant to the model [18]. Such constraints are especially important to introduce when inferring large-scale networks from high-throughput datasets, where irrelevant predictors and responses may arise.

The constrained negative likelihood function is, therefore, given by

$$\begin{aligned} f(A) &= Tr[\frac{1}{M\sigma^2}(Y-AX)(Y-AX)^t] - 2N\ln|\sigma^2| \\ &+ \alpha\sum_{i=1}^{N}\sum_{j=1}^{N}|a_{i,j}| + \beta\sum_{i=1}^{N}(\sum_{j=1}^{N}a_{i,j}^2)^{\frac{1}{2}} \\ &+ \gamma\sum_{j=1}^{N}(\sum_{i=1}^{N}a_{i,j}^2)^{\frac{1}{2}} \end{aligned} \tag{7}$$

where $\alpha \geq 0, \beta \geq 0, \gamma \geq 0$ and $\alpha+\beta+\gamma = 1$. Observe that the l_1-norm imposes sparsity while the l_2-norms are used to force an entire column or row to be zero. The constrained maximum likelihood of A is therefore given by

$$\hat{A} = \underset{A}{\text{argmin}}\, f(A). \tag{8}$$

The optimization problem in (8) is convex, and thus admits a global solution. Moreover, it can be solved efficiently using standard convex optimization techniques [19].

B. Recovering weak interactions

Due to the imperfections in the noise-corrupted data and the linearity of the model, and given the sparsity constraints, some entries in the estimated matrix A may be small. The standard approach in the literature in this case has been to set up an arbitrary threshold and set to zero all entries that fall below the threshold [13], [4], [2], [14]. However, we argue that arbitrary thresholding may lead to disregarding true interactions, thus increasing the false negative rate, and especially leading to the

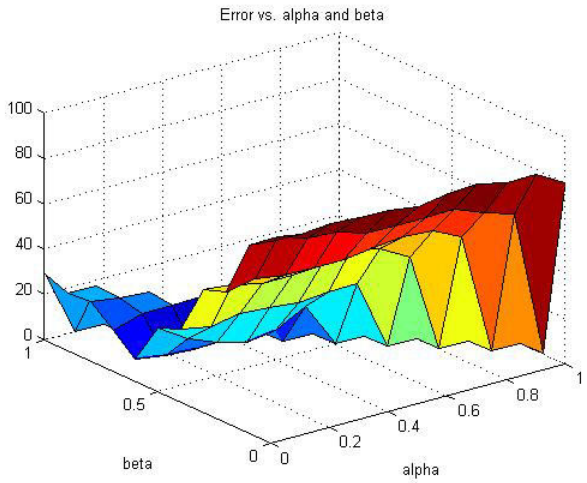

Fig. 1. Estimation error vs. the parameters α and β for a 10×10 matrix with two zero columns corresponding to two irrelevant predictors. The combined constraint strategy outperforms the l_1-norm alone: the minimum error is obtained at $\alpha = 0.6, \beta = 0.3, \gamma = 0.1$.

wrong network connectivity and hence the wrong biomolecular dynamics. Weak molecular interactions can be physiologically relevant. A recent study of genetic networks has established that weak interactions are important to preserve the structure of functional linkages among pathways [20]. Multiple gene knockouts involved in weak interactions could thus have a strong effect on genetic regulation and cellular functions [20]. Thus, the ability to distinguish true weak interactions from noise artifacts cannot be overstated.

We propose to use the Bayesian Information Criterion to achieve this goal [21]. The BIC method assesses the goodness of a model fit penalized for the number of estimated parameters. It is based, in part, on the likelihood function. The BIC function is given by [22]

$$BIC = -2\log p(x|k) + k\log(n), \tag{9}$$

where x is the observed data, k the number of free parameters to be estimated, $p(x|k)$ is the probability of the observed data given the number of parameters and n is the number of observations or the sample size. In our framework, given the constrained maximum likelihood estimate of the matrix A, we consider the entries or interactions below a specified threshold. To answer whether these entries are valid interactions or false positives, we compute the BIC of the data and compare it with the BIC corresponding to setting these entries to zero in a combinatorial way. The configuration corresponding to the smallest BIC is selected. In the Simulations Section we describe a greedy approach in order to find the smallest BIC matrix in an efficient manner for large-scale networks.

IV. Simulation Results

We generate synthetic gene networks using the linear model with additive noise in (1). We perform Monte Carlo simulations to assess the performance of the proposed method with respect to its parameters. All simulations are conducted in MATLAB. We first consider the sparsity tuning parameters α and β in (7) (observe that $\gamma = 1 - \alpha - \beta$). Before the BIC step, the errors are first computed according to the following thresholding procedure [2]

$$E = \sum_{i=1}^{n}\sum_{j=1}^{n} e_{ij}, \quad \text{with}$$
$$e_{ij} = \begin{cases} 1, & \text{if} \quad |A_{R,ij} - A_{T,ij}| > \delta, \\ 0, & \text{otherwise}, \end{cases} \tag{10}$$

where A_R and A_T denote, respectively, the estimated and true connectivity matrices, and δ is a fixed threshold. Figure 1 shows the error curve as a function of the parameters α and β. We observe that the minimum error is obtained for a value of α between 0 and 1, namely $\alpha = 0.6$ and $beta = 0.3$. In particular, the combined constraints outperforms the l_1 constraint alone (which corresponds to the case, $\alpha = 1, \beta = 0, \gamma = 0$). The matrix in Fig. 1 has two zero columns, which model the case where two genes are irrelevant predictors for the rest of the genes. In this case, the l_2-norm constraint on the columns improves the estimation by introducing zeros for all entries in some columns of A. Next, we investigate the effect of the number of measurements. As expected and shown in Fig. 2, the estimation accuracy increases when the number of measurements increases.

The interpretation of the estimation errors depends on the threshold in (10). The common procedure adopted in the literature is that values of the estimated matrix A that are smaller than the threshold are set to zero. This procedure is justified due to the presence of noise in the data. Small values of connectivity are considered mere noise and numerical errors. However, it has been recently argued that bimolecular networks have weak interactions that are essential to their functionalities [20]. Therefore, recovering such interactions may be critical to identifying the correct topology of the network and understanding its dynamics. The presence of noise makes this task quite difficult because of noise overfitting

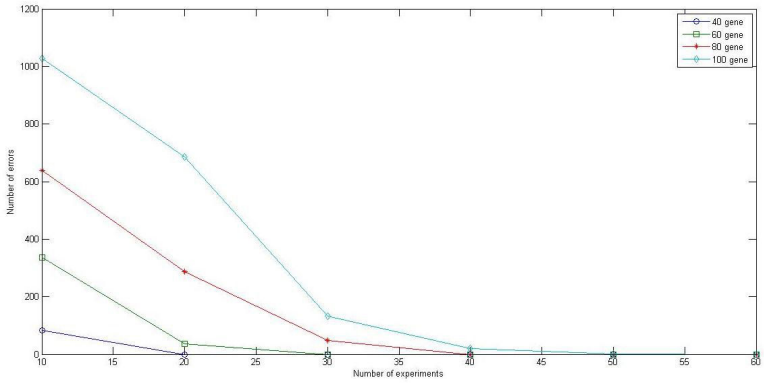

Fig. 2. The error rate as a function of the number of measurements for different network sizes.

problems, which will increase the false positive rate. As an example, consider the following "true" connectivity matrix $\begin{pmatrix} * & 0 & 0.26 \\ 0 & * & * \\ * & 0.34 & 0 \end{pmatrix}$, where the "*"'s denote entries above the threshold 0.5 and there are two small (but non-zero) entries, 0.26 and 0.34, below the threshold. Assume that the ML estimate of this matrix is given by $\begin{pmatrix} * & 0.17 & 0.21 \\ 0.1 & * & * \\ * & 0.30 & 0.09 \end{pmatrix}$ Following the thresholding practice in the literature, where all values below the threshold of 0.5 are set to zero, the optimal estimate is given by $\begin{pmatrix} * & 0 & 0 \\ 0 & * & * \\ * & 0 & 0 \end{pmatrix}$. That is the two weak interactions are mistakenly assumed to be zero; thus increasing the false negative rate. In order to solve this issue in the light of new developments showing the importance of weak interactions in biological networks [20], we use the BIC to decide whether the detected weak interactions are true interactions or artifacts of the noise level. We call an interaction weak if it falls below a specified threshold. We propose a combinatorial procedure, where we compute the BIC of the maximum likelihood estimate matrix and all matrices where each combinatorial combination of weak interactions are set to zero. For instance, assume that the ML estimate, $\hat{A}$, has two weak interactions. We compute the BIC of the four matrices, $\hat{A}$, $\hat{A}_1$, $\hat{A}_2$ and $\hat{A}_3$, where $\hat{A}_1$ is the ML estimate with one of the weak interactions set to zero, $\hat{A}_2$ sets the second weak interaction to zero, and $\hat{A}_3$ sets both weak interactions to zero. The matrix corresponding to the lowest BIC is considered to be the optimal estimate. In this procedure, all other parameters (above the threshold) are left intact. The BIC computation is combinatorial, and thus computationally infeasible for large number of genes N. We, therefore, consider a greedy search approach. We choose a configuration at random for the weak interactions. We attempt to reduce the BIC by changing each of the elements. This process is continued until we find a configuration, for which no further reduction in the BIC can be achieved. We repeat this procedure many times starting from different initial configurations, and choose the matrix that corresponds to the smallest BIC. Figure 3 shows the reduction in the error rate obtained by applying the BIC procedure. Figure 3 shows the reduction in the error rate after applying the BIC for a 10×10 network.

V. Conclusion

In this paper, we tackled the problem of inferring bi-molecular networks from an under-determined set of ordinary differential equations, modeling the molecular interactions. The challenges of this problem are of at least three types: First, the data is noisy and the linear model may be inaccurate or incomplete. Second, in large-scale inference problems, there may be irrelevant predictors, i.e., molecules or genes that do not contribute to the dynamics of the other molecules, or molecules that are not predicted by the considered set of genes. Thirdly, given the noise level in the data, we would like to recover weak "true" interactions by reducing the false negative rate. We solve the first issue by considering a maximum likelihood approach that takes into account the noise statistics. We solve the second issue by constraining the maximum likelihood estimate using a convex combination of three constraints, the l_1-norm to impose sparsity, the sum of the l_2-norms of the columns to uncover irrelevant predictors and the sum of the l_2-norms of the rows to identify the nodes that are not predicted by the considered set of molecules. The third issue of recovering weak interactions is considered by applying a combinatorial (in low-size networks) or a greedy approach (in large size networks) Bayesian information criterion to the constrained maximum likelihood estimate to reduce the false negative rate due to thresholding. Our simulation results on synthetically generated networks with varying sizes show the improved accuracy of the proposed approach compared to a standard l_1-constrained maximum likelihood approach followed by thresholding.

Acknowledgment

This project is supported by Award Number $R01GM096191$ from the National Institute Of General Medical Sciences

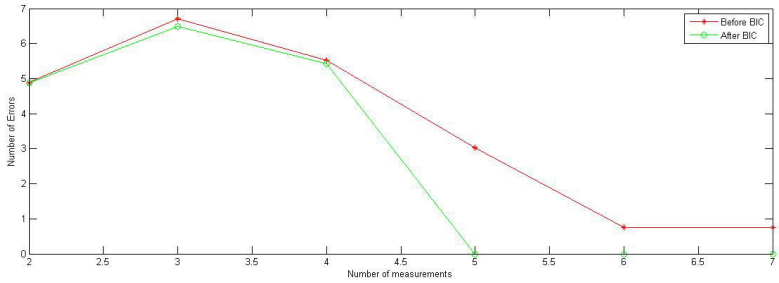

Fig. 3. Reduction of the error after applying the Bayesian Information Criterion

(NIH/NIGMS). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute Of General Medical Sciences or the National Institutes of Health.

REFERENCES

[1] T. Gardner, D. di Bernardo, D. Lorenz, and J. Collins, "Inferring genetic networks and identifying compound mode of action via expression profiling," *Science Signalling*, vol. 301, no. 5629, p. 102, 2003.

[2] M. Yeung, J. Tegnér, and J. Collins, "Reverse engineering gene networks using singular value decomposition and robust regression," *Proceedings of the National Academy of Sciences*, vol. 99, no. 9, pp. 6163–6168, 2002.

[3] J. Tegner, M. Yeung, J. Hasty, and J. Collins, "Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling," *Proceedings of the National Academy of Sciences*, vol. 100, no. 10, pp. 5944–5949, 2003.

[4] M. Bansal, G. Della Gatta, and D. Di Bernardo, "Inference of gene regulatory networks and compound mode of action from time course gene expression profiles," *Bioinformatics*, vol. 22, no. 7, pp. 815–822, 2006.

[5] H. Sabrina, K. Aneta, N. Zoran, and K. Jürgen, "Unraveling gene regulatory networks from time-resolved gene expression data–a measures comparison study," *BMC Bioinformatics*, vol. 12.

[6] X. Zhang, W. Cheng, J. Listgarten, C. Kadie, S. Huang, W. Wang, and D. Heckerman, "Learning transcriptional regulatory relationships using sparse graphical models," *PloS one*, vol. 7, no. 5, p. e35762, 2012.

[7] A. Meister, B. Choi, W. Wong, *et al.*, "Learning a nonlinear gene regulation model from perturbed steady-state data," *arXiv preprint arXiv:1207.3137*, 2012.

[8] Z. Liu, W. Zhang, K. Horimoto, and L. Chen, "A gaussian graphical model for identifying significantly responsive regulatory networks from time series gene expression data," in *Systems Biology (ISB), 2012 IEEE 6th International Conference on*, pp. 142–147, IEEE, 2012.

[9] S. A. Kauffman, *The Origins of Order: Self-Organization and Selection in Evolution*. New York: Oxford University Press, 1993.

[10] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian networks to analyze expression data," *Journal of Computational Biology*, vol. 7, no. 3, pp. 601–620, 2000.

[11] K. P. Murphy, *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, University of California, Berkeley, July 2002.

[12] S. Aluru, *Handbook of computational molecular biology*, vol. 9. Chapman & Hall/Crc, 2005.

[13] T. Chen, H. He, G. Church, *et al.*, "Modeling gene expression with differential equations," in *Pacific symposium on biocomputing*, vol. 4, p. 4, 1999.

[14] M. De Hoon, S. Imoto, K. Kobayashi, N. Ogasawara, and S. Miyano, "Inferring gene regulatory networks from time-ordered gene expression data of bacillus subtilis using differential equations," in *Biocomputing 2003: Proc. Pacific Symposium*, vol. 8, pp. 17–28, 2002.

[15] G. Rasool, N. Bouaynaya, H. Fathallah-Shaykh, and D. Schonfeld, "Inference of genetic regulatory networks using regularized likelihood with covariance estimation," in *IEEE Statistical Signal Processing Workshop*, pp. 560 – 563, August 2012.

[16] D. Donoho, "For most large underdetermined systems of linear equations the minimal 1-norm solution is also the sparsest solution," *Communications on pure and applied mathematics*, vol. 59, no. 6, pp. 797–829, 2006.

[17] D. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *Information Theory, IEEE Transactions on*, vol. 47, no. 7, pp. 2845–2862, 2001.

[18] J. Peng, J. Zhu, A. Bergamaschi, W. Han, D. Noh, J. Pollack, and P. Wang, "Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer," *The Annals of Applied Statistics*, vol. 4, no. 1, pp. 53–77, 2010.

[19] M. Grant, S. Boyd, and Y. Ye, "Cvx: Matlab software for disciplined convex programming," *Online accessiable: http://stanford. edu/~ boyd/cvx*, 2008.

[20] J. Poyatos, "The balance of weak and strong interactions in genetic networks," *PloS one*, vol. 6, no. 2, p. e14598, 2011.

[21] D. Posada and T. Buckley, "Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests," *Systematic biology*, vol. 53, no. 5, pp. 793–808, 2004.

[22] J. Chen and Z. Chen, "Extended bayesian information criteria for model selection with large model spaces," *Biometrika*, vol. 95, no. 3, pp. 759–771, 2008.