COMPARISON OF STATISTICAL AND OPTIMIZATION-BASED METHODS FOR DATA-DRIVEN NETWORK RECONSTRUCTION OF BIOCHEMICAL SYSTEMS

Behrang Asadi [a,1], Daniel M. Tartakovsky[a,2]
[a]Department of Mechanical and Aerospace Engineering
University of California, San Diego, La Jolla, CA, USA
E-mail: [basadi, dmt]@ucsd.edu
[1] Equal effort.
[2] Corresponding author; Department of Mechanical and Aerospace Engineering, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0411, Phone: (858) 534-1375, Fax: (858) 534-7599.

Mano Ram Maurya[b,1], Shankar Subramaniam [b,c,3]
[b]Department of Bioengineering
[c]Department of Chemistry & Biochemistry
University of California, San Diego, La Jolla, CA, USA
E-mail: [mmaurya, shankar]@ucsd.edu
[3] Corresponding author; Department of Bioengineering, University of California San Diego, 9500 Gilman Drive, La Jolla, CA, 92093-0412, Phone: (858) 822-0986, Fax: (858) 822-5722.

ABSTRACT
Data-driven reconstruction of biological networks is a crucial step towards making sense of large volumes of biological data. While several methods have been developed recently for reconstruction of the networks, no comprehensive study has been carried out to compare these characteristically different methods in terms of their performance with regard to important aspects such as incomplete data-sets and noisy data. In this paper we have applied and compared four methods, viz. least squares (LS), principal component regression (PCR), linear matrix inequalities (LMI), and Least Absolute Shrinkage and Selection Operator (LASSO), on a real data set and a synthetic data set with respect to important metrics. This comparison gives us an insight into when to choose an appropriate approach for reconstruction of networks based on *a priori* properties of experimental data.

KEY WORDS
Network reconstruction, least squares, principal component regression, linear matrix inequality, LASSO.

1. Introduction

Understanding the topology of networks from experimental data has recently received considerable attention in the recent decade particularly from system biologists and researchers in the field of bioinformatics and biotechnology [1]. Reconstruction of biological networks is a crucial topic in this field due to its role in interpretation of biological data.

Given the importance of network reconstruction, various methods have been introduced and are being developed for reconstruction of static, dynamic, and static-dynamic networks [2-4]. Optimization-based approaches such as method of leas-squares [5], dimensionality reduction methods such as principal components regressions (PCR) and partial least-squares (PLS) that are integrated with statistical significant tests [6-8], partial-correlation-related [9-11], Bayesian networks analysis [12-15], and hybrid methods such as "linear matrix inequalities" (LMI) [16] and "Least Absolute Shrinkage and Selection Operator" (LASSO) [17, 18] are examples of well-developed methods. Through appropriate formulation these approaches can be tailored for static or temporal (dynamic) data. A well-organized review of these methods is provided in [1].

Although several approaches are introduced in the literature to tackle challenges in biological network reconstruction, there has been no systematic effort to compare the performance of available methods on actual data-sets in terms of properties of the data pattern with respect to amount of missing data, or level of noise included in the dataset. In the present work, we provide an organized comparison of four popular methods for reconstruction of data driven biological networks with respect to their computational complexity and their robustness in identifying the true network with different levels of missing/unavailable and noisy data generated from real and simulated experiments.

The organization of this paper is as following: Section 2 briefly describes the methods implemented in this paper. Section 3 presents the results of implementation of the four methods for the reconstruction of the networks on actual data (phosphoprotein signaling and cytokine measurements in RAW 264.7 cells by the Alliance for Cellular Signaling (AfCS)) with partially known network and simulated data for which the true network (or model coefficients) is known. In section 4 we evaluate the performance of these methods for different levels of missing data and the level of noise. Section 5 provides a summary and conclusions.

2. Methods

In this section the four methods are succinctly described. These descriptions also shed light on the conceptual differences among the four methods in tackling the

problem of network reconstruction. The scope of the methods is linear input/output mapping of static or dynamic data.

2.1 Standard Least Squares

Standard least squares is a method for estimating the unknown coefficients (or parameters) of a linear model such that the sum of squares of deviations from observed response is minimized. This method is one of the oldest techniques in modern statistics[19].

Let $X_{m\times n}$ be an input data set (each column normalized to zero-mean and unit standard deviation) and $Y_{m\times p}$ (mean-centered) be the corresponding observed response (outputs). For simplicity, assume that p = 1 (else the procedure can be repeated on each output individually). Suppose that $\hat{B}$ is the candidate estimate for the parameter B in the linear (affine) system: $Y = XB$. Then the linear regression model of the system becomes

$$Y = X\hat{B} + \varepsilon \tag{1}$$

where ε is the residual vector. The objective is to minimize the Euclidean norm of residual vector in the following equation:

$$\hat{B} = \arg\min\{\varepsilon^2 = (Y - X\hat{B})^T (Y - X\hat{B})\} \tag{2}$$

The least squares solution to (2) is:

$$\hat{B} = (X^T X)^{-1} X^T Y \tag{3}$$

2.2 Principal Component Regression (PCR)

Principal component regression, which is based on principal component analysis, is required when $X^T X$ is (nearly) singular so that one or more of its eigenvalues are (close to) zero. Then, the principal components corresponding to only the first several eigenvalues (starting with the largest) are used.

The procedure of PCR is as follows:

1) Given the normalized input data $X_{m\times n}$ and mean-centered output data $Y_{m\times p}$, let $\Gamma_k = \{\gamma_j, j = 1,...k\}$ be the set of k largest eigenvalues and $V_k = \{v_j, j = 1,...k\}$ be the set of corresponding eigenvectors of the covariance matrix $C = XX^T/(m-1)$. Calculate the matrix of latent variables T_k:
$$T_k = X \times V_k \tag{4}$$

2) Create the PCR model based on k latent variables:
$$B_k = V_k \times \Gamma_k^{-1} \times T^T \times Y \;\&$$
$$RMSE_{PCR} = std(Y - Y_p);\; Y_p = X \times B_k \tag{5}$$

where RMSE is the square-root of mean-squared-error. The number of latent variables in PCR can be either based on cross-validation or on the basis of fraction of cumulative variance (say $0.8 < r < 0.95$) captured.

Partial least squares (PLS) is a method similar to PCR with the difference being that both *X* and *Y* matrices (instead of only *X*) are used to construct the set of linear combinations of significant inputs for regression. A detailed description of the method of PLS is presented in [20, 21]. With either of these methods, the coefficients *B* can be tested for their statistical significance by estimating the standard deviation of coefficients (σ_B) of the model and then comparing their ratio using a two-tailed t-test. For the PCR method, concisely, $\sigma_{B,k} \approx diag(V_k \times \Gamma_k^{-1} \times V_k{}^T)^{1/2} \times RMSE_{PCR}$ and $r_{j,k} = B_{j,k}/\sigma_{B,j,k}$ for jth input when *k* latent vectors are used. Average of $r_{j,k}$ over *k s/t*: $0.8 < r$ (fractional cumulative variance) < 0.95 is computed and if it is greater than $tinv(1-\alpha/2, v); v = m-k-1$ (inverse of cumulative t-distribution; $\alpha = 0.01$ for a significance level of 0.99), the input is considered significant [8].

2.3 Least Absolute Shrinkage and Selection Operator (LASSO)

In LASSO the problem of reconstruction is cast into an optimization problem of the form (2) with an additional nonlinear constraint. An abstract formulation of the LASSO is given as the following:

$$\hat{B} = \arg\min\{\varepsilon^2 = (Y - X\hat{B})^T (Y - X\hat{B})\} \text{ s/t } \sum_j \left|\hat{b}_j\right| \le t \tag{6}$$

where parameter *t* handles the amount of shrinkage in the estimation of parameters $\hat{B}$. The constraint imposed on optimization problem (6) shrinks the absolute value of some parameters and set the rest to zero, hence, extracts the descriptive features of the model. A quadratic programming approach (interior-point method) is used to solve the constrained optimization problem (6) [17].

2.4 Linear Matrix Inequalities (LMI)

The basic idea of this method is to convert a nonlinear optimization problem into a linear optimization problem [22]. This method has also been used to reconstruct and minimize dynamic networks [16, 23]. For a fair comparison of different methods applied in this paper, LMI method is also applied to static data. Problem (2) may be modified into:

$$\min_{B\in\mathbb{R}^{n\times p}}(\varepsilon) \text{ s/t } (Y - X\hat{B})(Y - X\hat{B})^T < \varepsilon I_{m\times m} \quad (7)$$

The constraint imposed on (7) is nonlinear with respect to $\hat{B}$. Congruence transformation converts problem (7) to into the following LMI representation:

$$\begin{pmatrix} -\varepsilon I_{m\times m} & Y - X\hat{B} \\ (Y - X\hat{B})^T & -I_{p\times p} \end{pmatrix} < 0 \quad (8)$$

Pre-existing knowledge (e.g., $b_{12} > 0$ or $b_{31} = 0$) can be added in the form of LMI constraints in the following format:

$$V_i^T B U_j + U_j^T B^T V_i = (><)0 \quad (9)$$

where $V_i = \begin{cases} v_r = 0, r \neq i \\ v_r = 1, r = i \end{cases}$ and $U_i = \begin{cases} u_r = 0, r \neq i \\ u_r = 1, r = i \end{cases}$ are respectively $n\times 1$ and $p\times 1$ column vectors to constrain $\hat{b}_{ij}^{\ th}$ element.

Suppose that the normalized matrix of parameters, $\bar{\hat{B}}$, is calculated as

$$\bar{\hat{b}}_{ij} = \left|\hat{b}_{ij}\right| / \left(\left\|\hat{b}_{i..}\right\|_2 \left\|\hat{b}_{:j}\right\|_2\right) \quad (10)$$

i.e., by dividing each element by the L_2-norm of its row and column. If a value of $\bar{\hat{b}}_{ij}$ becomes smaller than a threshold (say, r_{LMI}), then the corresponding parameter is nullified (insignificant). Further discussion can be found in [16].

2.5 Metrics for comparing the methods

Two data sets have been used to evaluate the performance of the four methods presented in Section 2. First set is experimental data measured on macrophage cells (Phosphoprotein (PP) vs Cytokine [24]) and the second set consists of synthetic data generated in Matlab. We build the model using 80% of the data-set (called training set) and use whole data-set to validate the model (called test set). Root-Mean-Squared-Error (RMSE) on the test set, and the number and the identity of the significant predictors (model parameters) selected are used as metrics to evaluate the performance of each method.

3. Results

The results of the implementation of the described methods are presented below.

3.1 Comparison on PP/Cytokine Data

The PP/Cytokine data set has 22 inputs and 6 outputs. RMSE of the resulting model via each method was calculated for all the outputs. Figure 1 shows a scatter-plot of the predicted outputs vs. experimental values for the LS and PCR methods. $\sigma = RMSE_{LS}$. For ease of visualization, σ and 2σ bands are also shown (dashed and dotted lines, respectively) in Figure 1. Further information regarding the performance of the models will be brought into the text separately. Table 1 lists the RMSE for the six outputs for the four methods. As expected, the LS method has the smallest RMSE, but other three methods (PCR, LMI and LASSO) are also comparable.

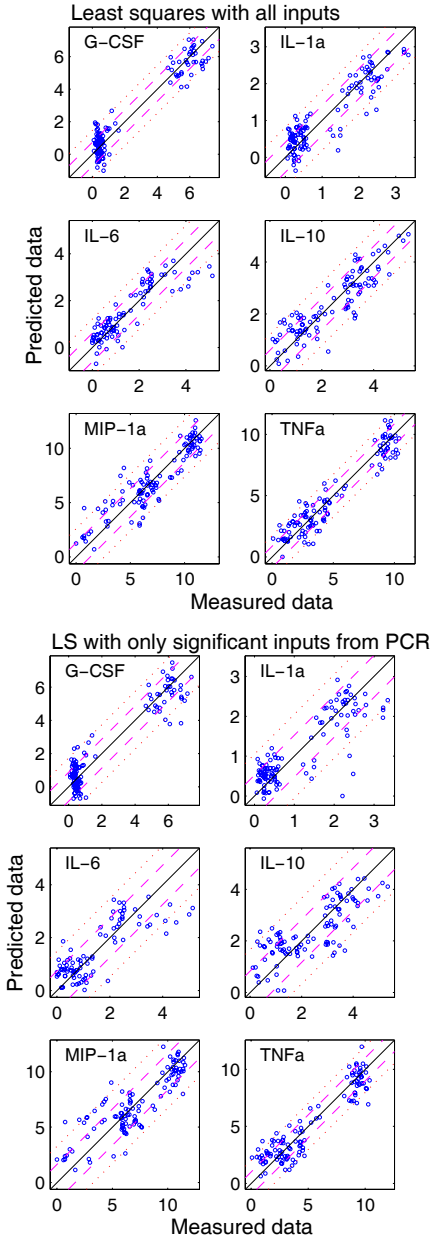

Figure 1. Actual response vs. predicted response by LS and PCR methods for PP/cytokine data. Dotted and dashed lines represent 1σ and 2σ bands, respectively.

Table 1. RMSE on training set for different methods (PP/cytokine data).

Output	1	2	3	4	5	6
LSQ	0.73	0.40	0.60	0.60	1.30	0.99
PCR	0.78	0.44	0.75	0.76	0.96	1.06
LASSO	0.76	0.41	0.61	0.61	1.31	1.01
LMI	0.74	0.41	0.61	0.61	1.31	0.99

Each method used a different strategy to identify the more informative/significant variables for building the model (Section 2). The number of significant variables depends on the selection criterion (or tuning parameter) used in the methods. For PCR, $0.8 < r < 0.95$ is used to capture 80-95% of the variance in the input data and the significance is based on the average ratio for t-test to get a stable estimate. In LASSO, the criterion is set to $t = 0.66$ for nullifying 33% of the smallest estimated parameters of the resulting model. For LMI, the threshold $r_{LMI} = 0.3$ is used. For each method the number of significant variables identified and used to build the model is listed in Table 2. PCR tends to retain lesser number of inputs. LASSO tends to retain more inputs (depends on the value of t). Overall, LASSO and LMI are comparable.

Table 2. Number of significant inputs for each output (PP/cytokine data).

Output	PCR	LASSO	LMI
G-CSF	12	10	14
IL-1a	12	14	15
IL-6	6	18	13
IL-10	7	15	15
MIP-1a	11	18	17
RANTES	9	12	12
TNFa	12	15	17

3.2 Comparison on synthetic noisy data

The four methods are applied on synthetic data with 22 inputs and 1 output. The true coefficients for the inputs (about 1/3rd) are made zero to test the methods if they identify them as insignificant. Here we also study the effect of increasing noise in the output data. Four outputs with 5, 10, 20 and 40% noise levels, respectively, are generated from the noise-free (true) output. Figure 2 shows the fit of predicted vs. supplied (noisy) output data for the LMI method. Increase in the noise is evident. Some of this noise has contaminated the predictions since the fraction of data points within the 2σ bands is about the same for all noise levels. In terms of RMSE (Table 3), LS performs better than PCR, LMI and LASSO. Since the inputs were independent, $RMSE_{LS}$ is comparable to the standard deviation of the noise added to the outputs. $RMSE_{LMI}$ is very close to $RMSE_{LS}$.

4. Discussion

In order to evaluate the accuracy of each method in estimating the model parameters, for the case of synthetic data set with known model coefficients, the computed/estimated parameter values have been compared with their true (known) values and their deviations expressed as mean(abs(b_{method}/b_{true} -1)) are listed in Table 4. b_{method} and b_{true} are the estimated (using the "method" method) and the true values of the parameters for a chosen output, respectively. The "mean" is computed over the coefficients for all the n inputs.

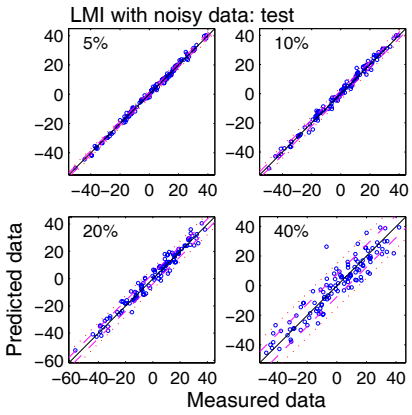

Figure 2. LMI on synthetic noisy data: Predicted versus supplied (noisy) data. Dotted and dashed lines represent 1σ and 2σ bands, respectively.

Table 3. RMSE on all data: methods vs. noise level.

Noise %	5	10	20	40
LS	0.9	2.0	3.7	7.3
PCR	3.3	3.6	5.0	8.6
LASSO	4.4	4.8	6.0	9.5
LMI	1.4	2.1	3.8	7.5

Table 4. Fractional error in estimating the parameters: methods vs. noise level (synthetic data).

Noise %	5	10	20	40
PCR	0.09	0.09	0.11	0.11
LASSO	0.47	0.47	0.46	0.42
LMI	0.21	0.18	0.23	0.72

Next, we explore the effect of the amount of training data used on the prediction accuracy (through RMSE), both for the real and the synthetic data.

4.1 Effect of missing data: real data-set

To test the effect of missing data, the output GCSF from the real data set is chosen. 0-60% data, in increments of 5%, was assumed to be missing. The remaining data was used for training and RMSE was computed on the test (missing) data. This was repeated 10 times by choosing

the selected fraction of data randomly, and average RMSE was computed. Figure 3 shows average RMSE for the real data. With increasing level of missing data, prediction accuracy deteriorates as expected. Table 5 lists the fractional standard deviation (std($RMSE_{0\text{-}60\%}$)/$RMSE_{0\%}$) and fractional maximum deviation (max($RMSE_{x\%}$ - $RMSE_{0\%}$)/$RMSE_{0\%}$) as compared to no missing data. PCR and LASSO are more robust than LS and LMI.

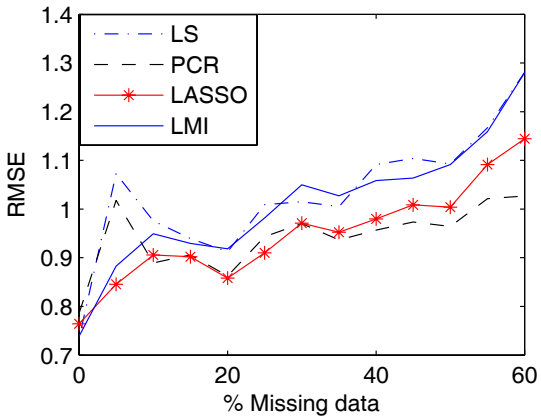

Figure. 3 RMSE versus percentage of missing data for different methods on PP/Cytokine data.

Table 5. Effect of missing data with PP/cytokine data (validation set)

Method	Fractional standard deviation	Fractional max deviation
LS	0.08	0.68
PCR	0.05	0.23
LASSO	0.04	0.42
LMI	0.05	0.56

4.2 Effect of missing data: synthetic data

Figure 4 shows the same comparison for the synthetic data with 20% noise. The qualitative nature of the behavior differs from that for the real data. It is surprising that the performance of LS is the best and that of LASSO is poorest as the amount of missing data increases. LS and PCR show stable performance.

Overall, PCR performs well on both data sets, leaving LASSO slightly behind for the real data and LMI for the synthetic data. Performance of PCR is robust whereas that of LS, LASSO and LMI appears to be dependent on some inherent characteristics of the data, which need to be explored further. Excellent performance of LS method on the synthetic data also deserves further investigation.

4.3 Computing time

The simulations were run on a Dual-Core Intel Pentium IV processor with 2.66GHz processing speed per core, 4MB of cache, and 3 GB of RAM. An estimate of CPU time used by each method is summarized in Table 6. A trade-off between robustness and computing time is apparent for the four methods. Based on the two case studies (real data and synthetic data), PCR is robust as well as fast as compared to LASSO and LMI. LASSO and LMI are slower than LS and PCR method by a factor of hundred for the sizes of the datasets chosen. Further studies on the effect of the size of the system are underway. This comparison can help the users choose the right method for a specific application.

Table. 6 Processing time vs. methods.

Method	Processing time(sec)
LS	0.1491
PCR	0.2484
LASSO	6.5657
LMI	37.595

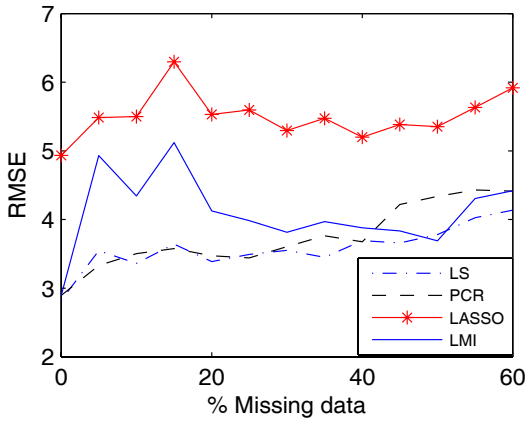

Figure. 4 RMSE versus percentage of missing data for different methods on synthetic date with 20% noise level.

5. Conclusion

Four methods for reconstruction of networks (LS, PCR, LASSO, and LMI) have been described and implemented on two different data-sets. First data-set comprised of a experimental data on phosphoproteins/cytokines and the second data-set was synthesized artificially. The least-squares method is naturally the best in terms of goodness of fit, but the other three methods are better in capturing most of the true inputs/predictors for the outputs. PCR performed better than other methods for the synthetic data with increasing levels of noise. The effect of missing data was also investigated in this work and our analysis demonstrates that the PCR technique is the most robust method for both the real data and synthetic data with medium level of noise. Not surprising is that fact the LS and PCR are the fastest and LMI is the slowest. There are several unanswered questions such as why LASSO performed poorly on synthetic data with increasingly more missing data. Another interesting question is the dependence of the performance on the type/characteristics of the noise. We are currently investigating these issues.

Acknowledgements

This research was supported by the National Heart, Lung and Blood Institute (NHLBI) grant 5 R33 HL087375-02 (SS), National Science Foundation (NSF) grant DBI-0641037 (SS) and the NSF collaborative grant DBI-0835541 (SS).

Author Contributions

Research design: SS, DMT, MRM. Implementation of algorithms and methods: BA, MRM. Wrote manuscript: BA, MRM. Revision/supervision: SS, DMT, MRM.

References

1. Maurya, M. and S. Subramaniam, *Computational Challenges in Systems Biology*, in *Systems Biomedicine: Concepts and Perspectives*, E. Liu and D. Lauffenburger, Editors. 2009, Academic Press: SAN DIEGO. p. 177-223.
2. Iwasaki, Y. and H.A. Simon, *Causality in Device Behavior.* Artificial Intelligence, 1986. **29**(1): p. 3-32.
3. Uckun, S., *Model-Based Reasoning in Biomedicine.* Critical Reviews in Biomedical Engineering, 1992. **19**(4): p. 261-292.
4. Maurya, M.R., R. Rengaswamy, and V. Venkatasubramanian, *A systematic framework for the development and analysis of signed digraphs for chemical processes. 1. Algorithms and analysis.* Industrial & Engineering Chemistry Research, 2003. **42**(20): p. 4789-4810.
5. Maurya, M.R., et al., *Mixed-integer nonlinear optimisation approach to coarse-graining biochemical networks.* IET – Systems Biology, 2009. **3**(1): p. 24-39.
6. Gupta, S., M.R. Maurya, and S. Subramaniam, *Identification of crosstalk between phosphoprotein signaling pathways in RAW 264.7 macrophage cells.* PLoS Comput Biol, 2010. **6**(1): p. e1000654.
7. Wu, Y., G.L. Johnson, and S.M. Gomez, *Data-driven modeling of cellular stimulation, signaling and output response in RAW 264.7 cells.* J Mol Signal, 2008. **3**: p. 11.
8. Pradervand, S., M.R. Maurya, and S. Subramaniam, *Identification of signaling components required for the prediction of cytokine release in RAW 264.7 macrophages.* Genome Biology, 2006. **7**(2): p. R11.
9. Camacho, D., et al., *Comparison of reverse-engineering methods using an in silico network.* Ann N Y Acad Sci, 2007. **1115**: p. 73-89.
10. de la Fuente, A., et al., *Discovery of meaningful associations in genomic data using partial correlation coefficients.* Bioinformatics, 2004. **20**(18): p. 3565-3574.
11. Aburatani, S., et al., *Gene systems network inferred from expression profiles in hepatocellular carcinogenesis by graphical Gaussian model.* EURASIP J Bioinform Syst Biol, 2007: p. 47214.
12. Janes, K.A., et al., *Systems model of signaling identifies a molecular basis set for cytokine-induced apoptosis.* Science, 2005. **310**(5754): p. 1646-1653.
13. Sachs, K., *Causal protein-signaling networks derived from multiparameter single-cell data.* Science, 2005. **308**(5721): p. 523-529.
14. Hartemink, A.J., et al., *Bayesian methods for elucidating genetic regulatory networks.* Ieee Intelligent Systems, 2002. **17**(2): p. 37-43.
15. Yu, J., et al., *Advances to Bayesian network inference for generating causal networks from observational biological data.* Bioinformatics, 2004. **20**(18): p. 3594-3603.
16. Cosentino, C., et al., *Linear matrix inequalities approach to reconstruction of biological networks.* IET Systems Biology, 2007. **1**(3): p. 164-173.
17. Tibshirani, R., *Regression shrinkage and selection via the Lasso.* Journal of the Royal Statistical Society Series B-Methodological, 1996. **58**(1): p. 267-288.
18. Bonneau, R., et al., *The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo.* Genome Biology, 2006. **7**(5): p. R36.
19. Geladi, P. and B.R. Kowalski, *Partial Least-Squares Regression - a Tutorial.* Analytica Chimica Acta, 1986. **185**: p. 1-17.
20. Helland, I.S., *Partial Least-Squares Regression and Statistical-Models.* Scandinavian Journal of Statistics, 1990. **17**(2): p. 97-114.
21. Gerlach, R.W., B.R. Kowalski, and H.O.A. Wold, *Partial Least-Squares Path Modeling with Latent-Variables.* Analytica Chimica Acta-Computer Techniques and Optimization, 1979. **3**(4): p. 417-421.
22. Vandenberghe, L., S. Boyd, and S.P. Wu, *Determinant maximization with linear matrix inequality constraints.* Siam Journal on Matrix Analysis and Applications, 1998. **19**(2): p. 499-533.
23. Julius, A., et al., *Genetic network identification using convex programming.* IET Systems Biology, 2009. **3**(3): p. 155-166.
24. *The Alliance for Cellular Signaling (AfCS)* Available from: http://www.signaling-gateway.org.