BMC Research Notes

BioMed Central

Technical Note

Model selection in the reconstruction of regulatory networks from time-series data

Eugene Novikov* and Emmanuel Barillot

Address: Service Bioinformatique, Institut Curie, 26 Rue d'Ulm, 75248 Paris Cedex 05, France

Email: Eugene Novikov* - Eugene.Novikov@curie.fr; Emmanuel Barillot - Emmanuel.Barillot@curie.fr

* Corresponding author

Published: 5 May 2009

BMC Research Notes 2009, **2**:68 doi:10.1186/1756-0500-2-68

Received: 24 January 2009

Accepted: 5 May 2009

This article is available from: http://www.biomedcentral.com/1756-0500/2/68

Abstract

Background: A widely used approach to reconstruct regulatory networks from time-series data is based on the first-order, linear ordinary differential equations. This approach is justified if it is applied to system relaxations after weak perturbations. However, weak perturbations may not be informative enough to reveal network structures. Other approaches are based on specific models of gene regulation and therefore are of limited applicability.

Findings: We have developed a generalized approach for the reconstruction of regulatory networks from time-series data. This approach uses elements of control theory and the state-space formalism to approximate interactions between two observable nodes (*e.g.* measured genes). This leads to a reconstruction model formulated in terms of integral equations with flexible kernel functions. We propose a library of kernel functions that can be used for the first insights into network structures.

Conclusion: We have found that the appropriate kernel function significantly increases the accuracy of network reconstruction. The best kernel can be selected using prior information on a few nodes' interactions. We have shown that it may be already possible to select models ensuring reasonable performance even with as small as two known interactions. The developed approaches have been tested with simulated and experimental data.

Findings

Two sources of experimental data are generally used in the reconstruction of regulatory networks: steady-state and time-series experiments. Steady-state data [1,2] are generated by measuring the expression levels of every gene (or protein concentrations) when a system relaxes into a steady state after a perturbation. There are many publications [3-5] reporting different methods for the network reconstruction from the steady-state data. Time-series data represent the expression levels measured at a number of time points following global or local perturbations of a system [6,7]. If these perturbations do not bring the system far from a steady state, the relaxation into the steady state is approximated by a set of the first-order, linear ordinary differential equations (LODE) [6,8,9]. Time-series experiments do not require as many perturbations as steady-state experiments, thus avoiding perturbations that may be not easy to design [10,11]. Moreover, analysis of time-series data allows us to investigate the dynamics of regulatory interactions, which is not possible from the steady-state data.

However, it has been shown [4,5] that the network reconstruction is more difficult from the time-series data than

from the steady-state data. The authors have envisaged two possibilities to improve the reconstruction. One is to collect more time series from additional perturbations. The other one is to perform time-series experiments where an investigated system demonstrates richer dynamics. The latter case is advantageous because it may generate more informative data without performing extra experiments. This can be done either by applying stronger perturbations or by monitoring system dynamics controlled by internal factors (*e.g.* cell-cycle processes). In both cases, the LODE models can hardly be justified as it is difficult to ensure that a system does not strongly deviate from a steady state. More sophisticated system dynamics needs more detailed formalizations on gene/molecular interactions. Many attempts to improve the basic LODE model can be found in recent publications [12-14]. In most cases, the authors suggest to model the combined regulatory effect of a number of regulatory factors by a particular non-linear function. Additionally, the second-order differential equations are sometimes invoked to reproduce gene expression profiles [14,15].

In this paper, we are looking for a generic approach to approximate interactions between the observable nodes in a network. The generic approach allows us to systematically apply specific models and, eventually, to define the most appropriate model using available experimental data and, possibly, prior knowledge on the nodes' interactions. The developed approaches were tested with simulated and experimental data.

Mathematical framework

We apply elements of control theory [16] to develop a generalized model of the network dynamics. A regulatory network (Fig. 1) is represented as a bipartite graph with two types of nodes: observable nodes reproducing measurable characteristics (*e.g.* gene expression levels), and non-observable, or control, nodes controlling the interactions between the observable nodes. Each control node i can be modelled as:

$$Y_I(\cdot) = F_i\{W_i, Y_o(\cdot)\} \tag{1}$$

where F_i is a functional reproducing behaviour, $Y_I(\cdot)$, of a set of observable nodes I based on signals, $Y_O(\cdot)$, from a, possibly different, set of observable nodes O, and W_i is a vector of "internal" parameters of control node i. Note that some non-trivial behaviour can be assigned to the observable nodes as well. It may account for instrumental distortions, specifics of image processing, normalization, etc.

The goal of the network reconstruction is to identify parameters W_i encoding for the interactions between the observable nodes. For that, functional F_i in (1) has to be further developed. It is frequently assumed that the cooperative regulatory contribution from different observable nodes is a sum of the contributions from each node, so that equation (1) can be written as:

$$\gamma_i(t) = \sum_{j=1}^{n} F_{ij}\{W_{ij}, \gamma_j(t)\} + b_i(t, t_0) \tag{2}$$

where n is the number of observable nodes, $\gamma_i(t)$ is the measured response of observable node i, F_{ij} is a functional characterized by a set of parameters W_{ij} converting measured profile, $\gamma_j(t)$, at node j to measured profile, $\gamma_i(t)$, at node i, and $b_i(t, t_0)$ is the output of non-regulated observable node i. We consider pair-wise controls F_{ij} as linear, continuous, time-invariant, finite-dimensional, single input-single output control systems that can be modelled using the state-space formalism:

$$\begin{aligned} \dot{X}_{ij}(t) &= A_{ij}X_{ij}(t) + B_{ij}\gamma_j(t) \\ \gamma_i(t) &= C_{ij}X_{ij}(t) \end{aligned} \tag{3}$$

where $X_{ij}(t)$ is the state vector and A_{ij} is the state matrix, $\gamma_j(t)$ is the input value and B_{ij} is the input vector, $\gamma_j(t)$ is the output value and C_{ij} is the output vector. We also assume that F_{ij} are in a steady state prior to the input perturbation $\gamma_j(t)$ starting at time $t = t_0$, that is $X_{ij}(t_0) = 0$. Integrating (3) and combining n regulatory inputs as in (2) yields

$$\gamma_i(t) = \sum_{j=1}^{n} \int_{t_0}^{t} w_{ij}(t - x)\gamma_j(x)dx + b_i(t, t_0) \tag{4}$$

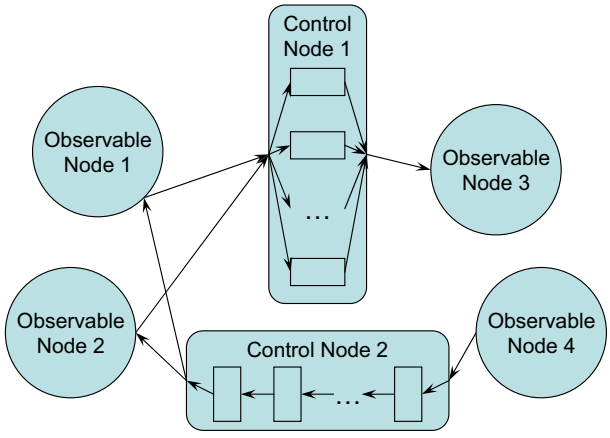

Figure 1
Regulatory network with four observable and two control nodes.

with $w_{ij}(t) = C_{ij}exp(tA_{ij})B_{ij}$ representing the influence of node j on the regulation of node i. Although every link (control node) is unique and should be modelled in a specific way, little prior knowledge on molecular interactions does not allow us to postulate specific models for every link. Therefore, we are looking for universal models that can approximate any control node.

The LODE regulatory model is widely used in the network reconstruction [6,8,9]. It can be obtained from (4), if we set $w_{ij}(t) = const = w_{ij}$ and $b_i(t, t_0) = const \times t = b_i t$:

$$\frac{d\gamma_i(t)}{dt} = \sum_{j=1}^{n} w_{ij}\gamma_j(t) + b_i \qquad (5)$$

This model approximates system relaxation into a steady state after a small perturbation. However, it is difficult to confirm that perturbations are small enough to justify model (5).

Equation (4) allows us to create a number of less restrictive models that can cover broader spectrum of dynamical behaviours. These models can integrate prior knowledge or can be further refined in experimental data analysis. In this report, we use the following representations for $w_{ij}(t)$:

$$w_{ij}(t) = \sum_{l=1}^{L} u_{l,ij} t^{l-1} \qquad (6)$$

$$w_{ij}(t) = \sum_{l=1}^{L} u_{l,ij} \exp\{-t/\tau_l\} \qquad (7)$$

$$w_{ij}(t) = \sum_{l=1}^{L} u_{l,ij} \{1 + t/\tau_l\}^{-1} \qquad (8)$$

where L is the number of terms, $u_{l,ij}$ are the coefficients encoding for the regulation of node i by node j and τ_l are the characteristics times that can be either set as prior values or estimated from experimental data. The background functions $b_i(t, t_0)$ can also be developed, but we will keep them constant as, with little data, more complicated models for $b_i(t, t_0)$ can fit the data without identifying any link.

We have devised a library of eight models (Table 1) to be tested and compared. Rationale for using the selected kernel functions is given in [Additional file 1].

Discussion on the parameter identifiability for the developed models can be found in [Additional file 2].

Network reconstruction is done by fitting the developed models to experimental data. Among different fitting strategies [17], the forward selection (FS) algorithm has shown reasonable performance, in particular for sparse networks, and therefore, it has been adopted in this paper. We refer to [18] for the details on the implementation of the FS algorithm. A more robust modification of the FS algorithm has also been tested as described in [Additional file 3].

Table 1: Kernel functions

Equation	$w_{ij}(t)$	Model
(6)	$u_{1,ij}$	P1
	$u_{1,ij} + u_{2,ijt}$	P2
(7)	$u_{1,ij}\exp\{-t/(0.1T)\}$	E1
	$u_{1,ij}\exp\{-t/(0.9T)\}$	E2
	$u_{1,ij}\exp\{-t/(0.1T)\} + u_{2,ij}\exp\{-t/(0.9T)\}$	E3
(8)	$u_{1,ij}(1 + t/(0.1T))^{-1}$	I1
	$u_{1,ij}(1 + t/(0.9T))^{-1}$	I2
	$u_{1,ij}(1 + t/(0.1T))^{-1} + u_{2,ij}(1 + t/(0.9T))^{-1}$	I3

We can use prior knowledge on the nodes' interactions to select the best network reconstruction model from the pre-defined library (Table 1). We look for the kernel function $w_{ij}(t)$ that reconstructs the prior links with the highest accuracy. The description of the adaptive model selection (AMS) algorithm can be found in [Additional file 4].

Testing

We compared the performances of the eight kernel functions from Table 1 as well as the LODE regulatory model (5) using simulated and experimental data. Three artificial systems were used for testing: the oscillating network in *E. coli*, called repressilator [19], the mitogen-activated protein kinase (MAPK) cascade [20] and the glycolysis pathway in yeast [21]. We also used the yeast (*Saccharomyces cerevisiae*) cell cycle microarray time-series data [22] to demonstrate applicability of the developed approach to real experimental data. The positive predictive value (PPV) and sensitivity (Se) were applied to estimate the performance. Further details on the artificial and real systems used for testing and description of the testing procedure can be found in [Additional file 5].

The dependencies of PPV on the total number of links are presented in Fig. 2. The Se values at 50 generated links are collected in Table 2. Among the three artificial systems, the choice of a model was the most critical for the *E. coli* repressilator. In this case, the best reconstruction was

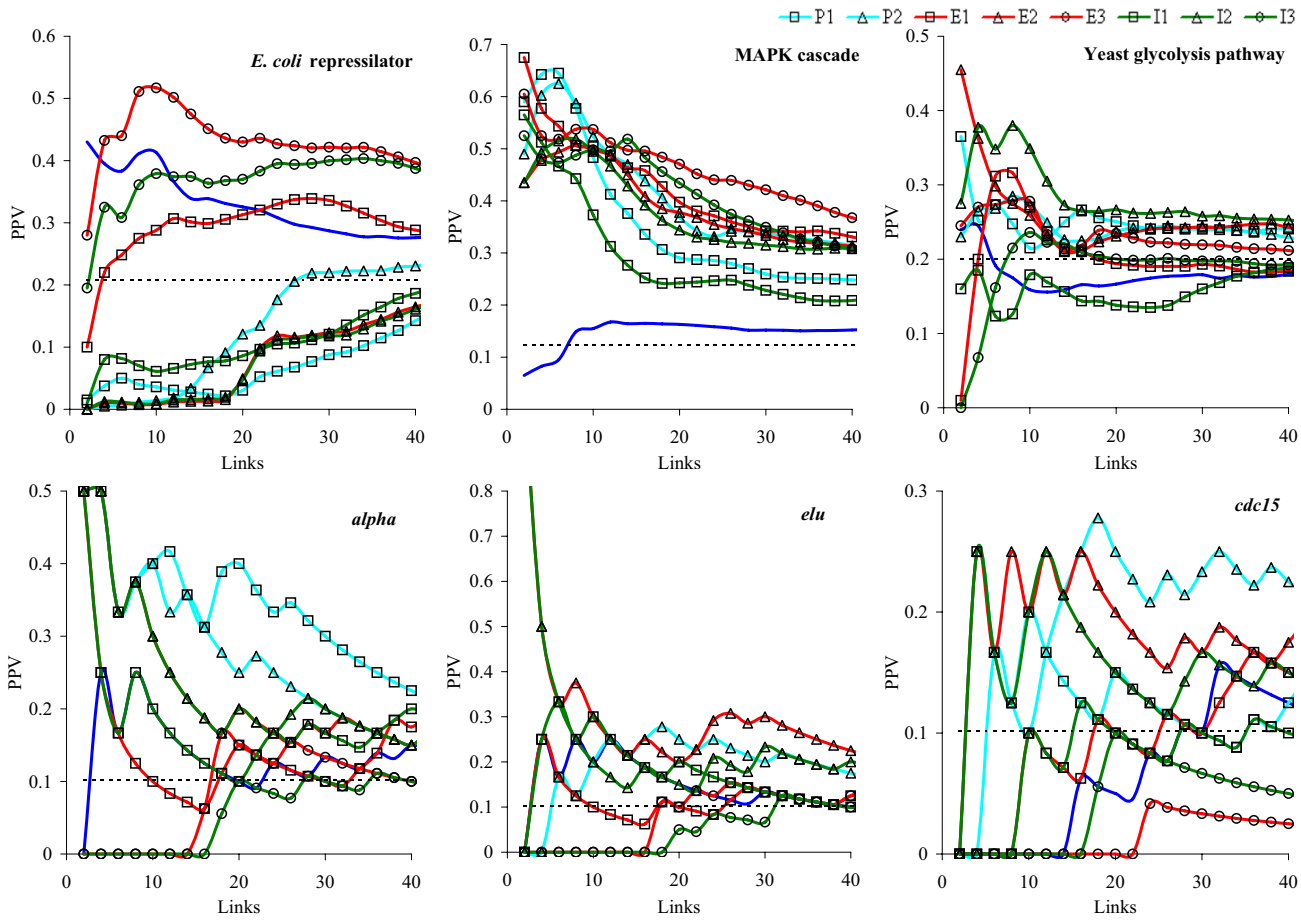

Figure 2
The average dependencies of PPV on the total number of links for the three artificial systems and for the three yeast cell cycle microarray time-series datasets. Blue line corresponds to the LODE model and dashed black line corresponds to random prediction. Confidence intervals for the obtained estimates are too narrow to be recognizable in the graphs and therefore not shown.

achieved with the bi-exponential E3 model. The LODE model performed better than random reconstruction but still worse than E3. All tested kernels were significantly better than random link assignment for the MAPK cascade. All kernels also outperformed the LODE model in this case. However, there is still a notable (and statistically significant) difference between the kernels. The yeast glycolysis network (Fig. 2c) was the most difficult to reconstruct because many times series were similar and hardly distinguishable by the reconstruction algorithm. Nevertheless, several models (P1, P2, E2, E3, and I2) demonstrated the performance different from random. The LODE model could not outperform the random prediction in this case.

For the yeast cell cycle time-series data, the polynomial models (P1 and P2) were the most powerful. For the *alpha* dataset and for the *elu* dataset, P1 had the highest performance whereas P2 was the most accurate for *cdc15*. Note that, in each case, the best performing models (P1 and P2) also outperformed the LODE model. Comparing different experiments, we see that *cdc15* led to less accurate predictions. This indicates that this experiment requires more elaborated reconstruction models or more representative datasets.

From Fig. 2 and Table 2, we can conclude that the "optimal" models were different for the artificial and real systems. The obtained results suggest that no unique model exists to ensure reasonable performance for different systems and therefore the most appropriate models should be searched for each system.

We applied the AMS algorithm [Additional file 4] to the same three artificial systems and three experimental datasets. As at each run the prior links were different, the

Table 2: Se at 50 generated links for the three artificial systems (*E. COLI* repressilator (A), MAPK cascade (B) and yeast glycolysis pathway (C)) and three yeast cell cycle microarray time-series datasets

Models	*A*	B	C	*alpha*	*elu*	*cdc15*
LODE	0.46	0.12	0.16	0.23	0.19	0.27
P1	0.32	0.19	0.20	0.35	0.42	0.27
P2	0.41	0.23	0.18	0.35	0.31	0.35
E1	0.47	0.25	0.16	0.38	0.31	0.23
E2	0.32	0.24	0.20	0.31	0.35	0.31
E3	0.60	0.27	0.17	0.15	0.27	0.08
I1	0.35	0.18	0.18	0.31	0.23	0.15
I2	0.32	0.24	0.21	0.27	0.35	0.27
I3	0.59	0.23	0.16	0.19	0.19	0.12

For the artificial systems, the Se values were averaged over 100 runs of the simulation procedure. Model definitions (P1, P2, E1, E2, E3, I1, I2 and I3) are given in Table 1.

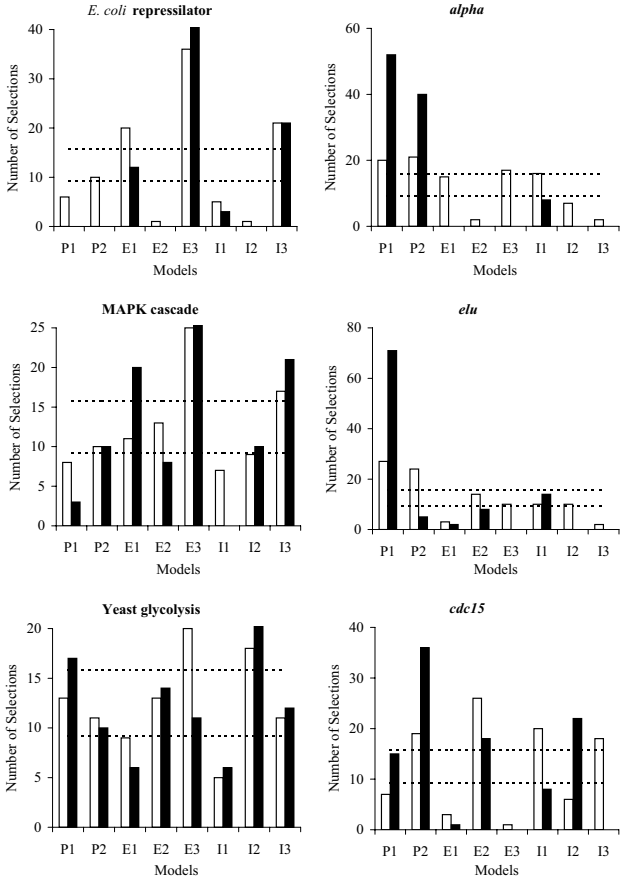

Figure 3
Adaptive model selection. Number of times each model from Table 1 has been selected in 100 runs of the simulation procedure by the AMS algorithm based on 2 (empty bars) and 10 (filled bars) prior links. Confidence intervals for the random model selection are indicated by dashed lines.

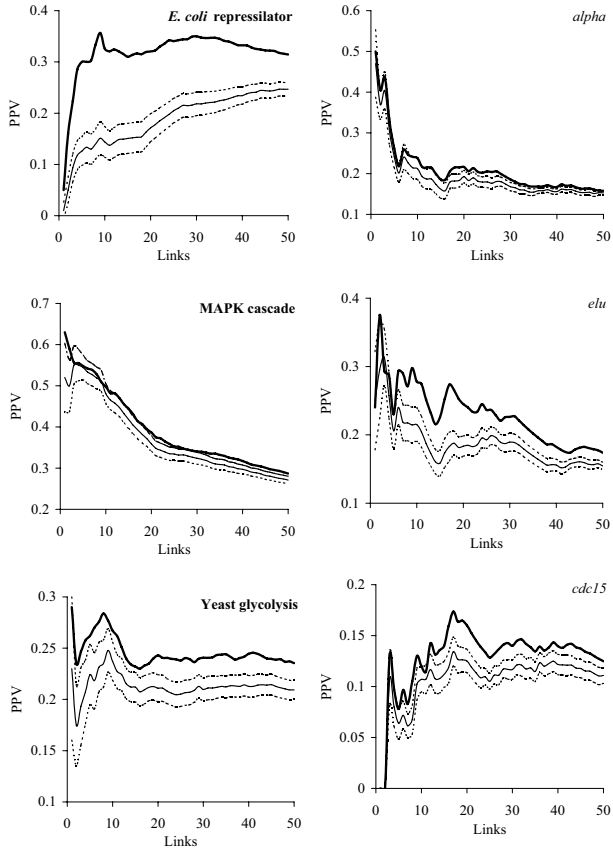

Figure 4
The dependencies of PPV on the total number of links for the AMS algorithm (with two prior links). Thick line – PPV by the AMS algorithm; thin line – PPV after random model selection. Confidence intervals for PPV after random model selection are shown as dashed lines.

selected model might also be different. Therefore, we counted number of times each model from Table 1 was selected in the 100 runs. The results for 2 and 10 prior links are shown in Fig. 3. We found that the higher performing models from Fig. 2 were selected more often than the lower performing ones. Moreover, reasonable model recognition could be already achieved with only two prior links. As expected, the increase in the number of prior links led to better model identification.

However, in some cases with two prior links, the AMS algorithm relatively often selected the models that were rather poor as judged by the results presented in Fig. 2. For example, for the artificial yeast glycolysis pathway or real *alpha* dataset, the bi-exponential E3 model was selected almost as often as other, better performing, models. This indicates that the E3 model was more adequate just for certain links and not for any link in the networks. Therefore, we can conclude that the network reconstruction

model should be link-specific, that is different models may be assigned to different links.

As the AMS algorithm may select poor performing models, the overall performance of the network reconstruction is lower than for the best performing model. However, even with as small as two prior links, AMS is already better than random model selection, as illustrated in Fig. 4. If the performance of different models is not very different (as for the MAPK cascade), the prediction of the AMS algorithm is close to random. If, however, a certain model demonstrates clear advantage (as, for example, for the *E. coli* repressilator), the AMS algorithm can identify this model leading to the performance substantially higher than by random selection.

The performance of the AMS algorithm using independent set of artificial data described in [5] is presented in [Additional file 6].

Conclusion

We have presented a generalized approach for the regulatory network reconstruction, that gives us an easy possibility to create and to test different inference models and, potentially, to identify appropriate models from experimental data. We have shown that even with as small as two prior links it is already possible to select models ensuring reasonable performance. Further discussion and perspectives for further research are given in [Additional file 7].

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

EN developed the model, performed software implementation and drafted the manuscript. EB conceived of the study and participated in coordination. All authors read and approved the final manuscript.

Additional material

Additional file 1

***Kernel functions**. Rationale for using the kernel functions from Table 1.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1756-0500-2-68-S1.pdf]

Additional file 2

***Identifiability note**. Discussion on the parameter identifiability for the developed models.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1756-0500-2-68-S2.pdf]

Additional file 3

***Modified forward selection (FS) algorithm**. Description and testing of the modified version of the FS algorithm.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1756-0500-2-68-S3.pdf]

Additional file 4

***Adaptive model selection (AMS)**. Description of the AMS algorithm to identify the kernel function that reconstructs the prior links with the highest accuracy.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1756-0500-2-68-S4.pdf]

Additional file 5

***Simulated and experimental data**. Details on the artificial and real systems used for testing and description of the testing procedure.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1756-0500-2-68-S5.pdf]

Additional file 6

***Independent artificial data**. Testing of the AMS algorithm using independent set of artificial data described in [5].*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1756-0500-2-68-S6.pdf]

Additional file 7

***Discussion**. Discussion and perspectives for further research.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1756-0500-2-68-S7.pdf]

Acknowledgements

This work was supported by the Institut Curie and the Ligue Nationale Contre le Cancer. E.N. and E.B. are members of the Equipe Biologie des Systèmes from the Service de Bioinformatique of Institut Curie, équipe labellisée par La Ligue Nationale Contre le Cancer.

References

1. Wagner A: **How to reconstruct a large genetic network from *N* gene perturbations in fewer than *N2* easy steps.** *Bioinformatics* 2001, **17:**1183-1197.
2. Ideker TE, Thorsson V, Karp RM: **Discovery of regulatory interactions through perturbation: inference and experimental design.** *Pac Symp Biocomput* 2000, **5:**302-313.
3. Werhli AV, Grzegorczyk M, Husmeier D: **Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks.** *Bioinformatics* 2006, **22:**2523-2531.
4. Soranzo N, Bianconi G, Altafini C: **Comparing association network algorithms for reverse engineering of large-scale gene regulatory networks: synthetic versus real data.** *Bioinformatics* 2007, **23:**1640-1647.
5. Bansal M, Belcastro V, Ambesi-Impiombato A, di Bernardo D: **How to infer gene networks from expression profiles.** *Molecular Systems Biology* 2007, **3:**78.

6. Bansal M, Gatta GD, di Bernardo D: **Inference of gene regulatory networks and compound mode of action from time course gene expression profiles.** *Bioinformatics* 2006, **22:**815-822.
7. MacCarthy T, Pomiankowski A, Seymour R: **Using large-scale perturbations in gene network reconstruction.** *BMC Bioinformatics* 2005, **6:**11.
8. D'haeseleer P, Liang S, Somogyi R: **Genetic network inference: from co-expression clustering to reverse engineering.** *Bioinformatics* 2000, **16:**707-726.
9. Kim J, Bates DG, Postlethwaite I, Heslop-Harrison P, Cho KH: **Least-squares methods for identifying biochemical regulatory networks from noisy measurements.** *BMC Bioinformatics* 2007, **8:**8.
10. Basso K, Margolin AA, Stolovitzky G, Klein U, Della-Favera R, Galifano A: **Reverse engineering of regulatory networks in human B cells.** *Nature Genetics* 2005, **37:**382-390.
11. Dojer N, Gambin A, Mizera A, WilczyMski B, Tiuryn J: **Applying dynamic Bayesian networks to perturbed gene expression data.** *BMC Bioinformatics* 2006, **7:**249.
12. Bonneau R, Reiss DJ, Shannon P, Facciotti M, Hood L, Baliga NS, Thorsson V: **The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets *de novo*.** *Genome Biology* 2006, **7:**R36.
13. Vu TT, Vohradsky J: **Nonlinear differential equation model for quantification of transcriptional regulation applied to microarray data of Saccharomyces cerevisiae.** *Nucleic Acids Research* 2007, **35:**279-287.
14. Chang WC, Li CW, Chen BS: **Quantitative inference of dynamic regulatory pathways via microarray data.** *BMC Bioinformatics* 2005, **6:**44.
15. Perrin BE, Ralaivola L, Mazurie A, Bottani S, Mallet J, d'Alché-Buc F: **Gene networks inference using dynamic Bayesian networks.** *Bioinformatics* 2003, **19:**ii138-ii148.
16. Sontag ED: **Mathematical Control Theory: Deterministic Finite Dimensional Systems.** 2nd edition. Springer, New York; 1998.
17. van Someren EP, Wessels LFA, Reinders MJT, Backer E: **Searching for limited connectivity in genetic network models.** *Proceedings of the Second International Conference on Systems Biology* 2001:222-230.
18. Novikov E, Barillot E: **Regulatory network reconstruction using an integral additive model with flexible kernel functions.** *BMC Systems Biology* 2008, **2:**8.
19. Elowitz MB, Leibler S: **A synthetic oscillatory network of transcriptional regulators.** *Nature* 2000, **403:**335-338.
20. Huang CHF, Ferrell JE Jr: **Ultrasensitivity in the mitogen-activated protein kinase cascade.** *Proc Natl Acad Sci USA* 1996, **93:**10078-10083.
21. Pritchard L, Kell DB: **Schemes of flux control in a model of Saccharomyces cerevisiae glycolysis.** *Eur J Biochem* 2002, **269:**3894-3904.
22. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: **Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization.** *Mol Biol Cell* 1998, **9:**3273-3297.