

MANIA: A Gene Network Reverse Algorithm for Compounds Mode-of-Action and Genes Interactions Inference

Darong Lai^{1,2}, Hongtao Lu², Mario Lauria³, Diego di Bernardo³,
and Christine Nardini^{1,*}

¹ CAS-MPG PICB, Yue Yuan Road 320, Shanghai, PRC
darong.lai@gmail.com, christine@picb.ac.cn

² Dept. of computer science and engineering, Shanghai Jiao Tong University, PRC
lu-ht@cs.sjtu.edu.cn

³ TIGEM, 111 Via Pietro Castellino, Naples, Italy
{lauria,dibernardo}@tigem.it
www.picb.ac.cn/ClinicalGenomicNTW

Abstract. Understanding the complexity of the cellular machinery represents a grand challenge in molecular biology. To contribute to the deconvolution of this complexity, a novel inference algorithm based on linear ordinary differential equations is proposed, based on high-throughput gene expression data. The algorithm can infer (i) gene-gene interactions from steady state expression profiles *AND* (ii) mode-of-action of the components that can trigger changes in the system. Results demonstrate that the proposed algorithm can identify *both* information with high performances, thus overcoming the limitation of current algorithms that can infer reliably only one.

Keywords: gene network, gene expression, reverse engineering, Ordinary Differential Equations (ODE), compound mode-of-action.

1 Introduction

Thanks to the fast moving and recent advancements in technology, our society is assisting to an unprecedented high-throughput production of information coming from a variety of areas of human activity. This comprises, but is not limited to, economic, social and biological data. In particular, we focus our attention on biomolecular data. To deconvolute the structure underlying such data, cross fertilization from diverse areas of research, and notably the introduction of exact sciences in the realm of biology, has been a fundamental requirement to mine the complex interaction that explains the data we observe. However, the task is far from completed, and although economical, sociological and molecular systems own peculiar characteristics, advances in the deconvolution of the complexity in any of these areas bears the potential to significantly contribute to explain

* Corresponding author.

the complexity of the global system we live in. In the area of molecular biology several high-throughput platforms are quickly becoming available [1], however, gene expression data represents at the moment the most abundant source of molecular high-throughput information. This work focuses on the identification of networks of interaction among genes. Networks of interactions identify general relationships among the nodes of the network (genes), thus, a link in the network may not represent a physical interaction (carried on by intermediate molecules such as proteins). However, these algorithms can be extremely powerful in the initial characterization of unknown systems, taking advantage of low-cost, high-throughput screens and generating relevant *in silico* hypotheses that can be further and efficiently tested in *wet lab*. Moreover, these algorithms, thanks to their ability to reconstruct networks on genome-wide data, offer a systemic perspective of the interactions. Depending on the model adopted, these methods can infer a causality in the relationship (directed networks) or rather a simple 'connection' among items (undirected networks). Many methods [2] have been proposed to reverse-engineer gene expression data, that can either take advantage of the evolution in time of the state of the system (time series, i.e. [3]), or of different equilibrium states reached by the system (steady state, i.e. [4]). Our approach focuses on the latter, more abundant, steady state data. To achieve different equilibrium states of the system, the system is perturbed in different ways (i.e. knock-out, knock-down, alterations in the growing medium) and the resulting expression data is collected once the system has reached the novel equilibrium. Algorithms that handle these data typically output a representation of the gene network in the form of a graph or an adjacency matrix (here called A , [4,5,6]). These networks represent the relationships occurring among genes, and offer a first impression of the complex pathways that are being activated in the system under study. Alternatively these algorithms offer an estimation, for example in the form of a ranked list, of the genes that were directly affected by the perturbation in the experiments [7] (here called P). When the perturbation is obtained adding a compound in the environment of the cell, the genes identified by P represent the direct targets of the perturbing agents that have been used to alter the equilibrium. This identifies an information extremely valuable in areas such as *chemogenomics*, where the identification of a small molecule's direct target (also called *transcriptional perturbations*) can provide fundamental information on its use as a drug. Because of this, P is also known to represent the *mode-of-action* of the perturbing compound. So far, *a priori* knowledge of the direct targets of perturbation was required for a proper identification of A [4], or alternatively, the identification of an estimate of P was not able to produce a reliable representation of A [7], due to the high sensitivity of the algorithms to errors in P . With our novel approach we aim at the identification of both the gene network (A) and the single direct target matrix (P), overcoming the current limitation, while preserving and improving both performances. Our approach can handle efficiently experiments resulting in single transcriptional perturbations. Single transcriptional perturbations are useful to be quantified when the entity of a single gene knock-down is unknown and when the action

of perturbagens is supposed to target predominantly an individual (unknown) transcript or protein, rather than several elements of a pathway. In the following we present related methods (Section 2), details of our algorithm (Section 3), validation results (Section 4) and their interpretation (Section 5).

2 Related Work

Number of approaches are being designed and tested to uncover the complexity of molecular interactions. In the following we briefly describe currently used tools based on Bayesian theory (Banjo, [6]), information science approach (ARACNe, [5]) and ordinary differential equations (ODE, NIR and MNI [4,7]), that have proven to be useful in the identification of gene networks or compounds mode-of-action. All the above methods can handle steady-state data. Banjo [6] generates a network space and screens then the best network structures attributing the most appropriate conditional density function, by optimization of an objective function (Bayesian Dirichlet equivalence, or Bayesian information criterion). Banjo can reconstruct signed directed network indicating regulation among genes, but it cannot infer networks involving cycles (or loops). ARACNe (Algorithm for Reconstruction of Accurate Cellular Networks, [5]) is regarded as an information-theoretic approach to gene network inference. It computes mutual information (MI, [8]) for all pairs of genes profiles to estimate the independence between genes and uses strategies (Data Processing Inequality, DPI) to successfully filter out the number of false-positive interactions. ARACNe cannot reconstruct directed networks. Our approach is strongly rooted in two ODE-based methods previously developed and validated. Namely, we used as a starting point NIR [4] able to infer the network of genes interactions (A), provided the perturbations (P) are known, and MNI [7], able to rank the most likely direct target genes of perturbations (estimate of P). Briefly, these algorithms aim at the identification of the function that describes the variation of gene expression matrix \underline{x} over time $\underline{x}' = f(\underline{x}, \underline{p})$, with x representing the steady state expressions of the N genes involved in the network across M experiments, f is a non linear function that models how the expression values \underline{x} and M experiments provoking external influences \underline{p} modify the genes' activity. Assuming steady state and small perturbations, these equations can be linearized around the equilibrium state, and become, for a scalar element of the expression matrix $x'_{il} = \sum_j a_{ij}x_{jl} + p_{il} = \underline{a}_i^T \cdot \underline{x}_l + \underline{p}_{il}$ with $i, j = 1..N$ indicating genes and $l = 1..M$ experiments. In matricial form and at equilibrium this becomes $AX = -P$, with A being the N by N network matrix (a_{ij} represents the action of gene j on gene i), X the expression data (x_{il} represents the expression of gene i in experiment l) and P the matrix of transcriptional perturbations (p_{il} represents the transcriptional perturbation of gene i in experiment l), explaining the origin of our notations. These approaches assume that only a limited number of connections among genes are possible, to reflect the structure of the molecular pathways. Based on this sparsity assumption, NIR uses multiple linear regression to infer the connections among genes. Conversely, MNI is trained on the expression data in X to evaluate A and P

through an iterative process based on the minimization of an objective function (Sum of Square Errors, SSE).

3 Method

Our approach aims at identifying A and P based solely from the expression data X , thus overcoming the necessity to have *a priori* information on the direct target of the perturbation (P), which is very often an important unknown of the problem. To achieve this goal, we sought to chain the two algorithms in order to use the prediction of MNI to feed NIR and infer the network. To do so, our algorithm uses iteratively $M - 1$ experiments in X to predict the M -th column (experiment) of P , as a ranked list of most likely targets. Due to the intrinsic noise of the data and the limited deterministic predictive power of MNI, the reliable identification of A, P is not trivial, especially when predicting complex data, as it can be shown in Section 4, in the varying performances of MNI+NIR, which represent the trivial chaining of the 2 algorithms (output of MNI used directly as P for NIR, see Figure 1(a)). For this reason, other strategies had to be integrated, schematically shown in Figure 1(b). Based on previous acronyms (and on the obsessive search for the network identification) we call this new approach Mode of Action & Network Identification Approach (MANIA).

In this approach, an estimate of P is produced by MNI, called P_{MNI} , this matrix contains the top ranking perturbations (we tested 1 and 10 top best, parameter $topP$), while all other values of P_{MNI} are set to zero. When choosing the single top perturbation option ($topP = 1$), the algorithm should perform at its best, provided MNI reliably identifies the correct transcriptional perturbation as the most likely (i.e. the best prediction *is* indeed the gene target). In this case, in fact, no noise is added; however, we also tested the algorithm preserving the top 10 best predictions ($topP = 10$), to offer backup solutions in case MNI is not able to find the correct perturbation as first choice. The core step of the algorithm consists of the strategy used to *clean* P_{MNI} from the incorrect

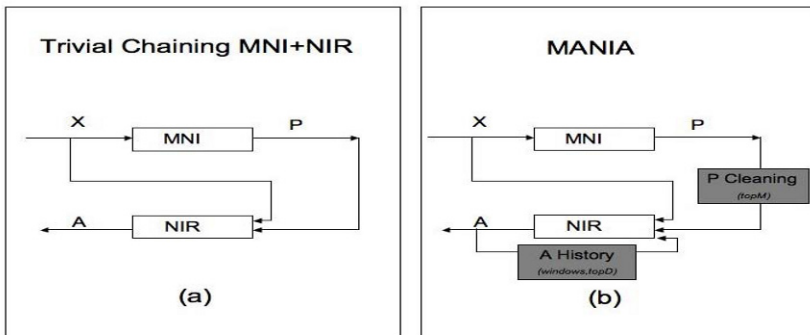


Fig. 1. Schematic view of the trivial chaining of MNI and NIR and of the strategies implemented in MANIA, discussed in Section 3

predictions, so that only the appropriate perturbation is used in NIR to predict A . This strategy consists of two steps. The first is the iterative computation of all the solutions for a given row of A , using all the predictions offered by P_{MNI} . The solutions are then ranked and only the $topM$ best are preserved (along with the corresponding perturbations) while computing the following rows. However, this step alone is not sufficient, since, often, the solution that minimizes the objective function (SSE) produces a local minimum of the objective function. Choosing this solution can result in the identification of a unique minimum for P and thus for all the rows of A . To overcome this issue, information about the previous rows computed in A are used. Thus, another parameter (*windows*) has been introduced to indicate the number of rows used as previous knowledge to calculate and minimize SSE. In particular, SSE is computed on all the $topM$ solutions as $X = -A_{tmp}^{-1}P$, where A_{tmp} is the identity matrix (self-relation is always assumed true) with the corresponding *windows* rows replaced by the solutions already computed. By construction A is always invertible. For each row of A and P only the best $topM$ solutions are preserved before computing the following rows of A, P . In our simulations, we set *windows* = 5 and *topM* = 200. In our experience, these values represent a good compromise between computation time and accuracy. Pseudocode in Algorithm 1 gives more details about the process.

4 Experimental Results

To validate our approach, we used two known benchmark datasets (here called Dataset 1 [2] and Dataset 2 [9]), and compared our performances to state-of-the-art algorithms briefly summarized in Table 1. These algorithms were used with their default parameters values.

Dataset 1. This dataset consists of 20 instances of expression matrices X with 100 genes and 100 experiments, obtained from 20 instances of network matrices A with sparsity 10 (indicating a maximum of 10 possible interactions for each gene), and single perturbation for P (identity matrix). Gaussian noise (10%) is added to expression data to better mimic real data. Performances are computed using positive predictive value (PPv, also called *accuracy*) defined as $TP/(TP + FP)$ and Sensitivity $TP/(TP + FN)$, where TP, FP and FN stand for True Positive, False Positive and False Negative, respectively. Results were averaged,

Table 1. Network inference algorithms used for performances comparison

Software	Download link	Model
BANJO	www.cs.duke.edu/~amink/software/banjo	Bayesian method
ARACNe	www.amdec-bioinfo.cu-genome.org/html	Information-theory method
MNI/NIR	http://dibernardo.tigem.it/wiki/index.php	ODE based method

Input: gene expression profiles matrix X , user defined parameters
Output: Adjacency matrix A of gene network; mode-of-action matrix P
 N : number of genes;
 M : number of experiments;
 $topP$: max number of perturbations proposed by MNI preserved in P_{MNI} per experiment;
 $topM$: max number of solutions preserved per each row of A ;
 $windows$: max number of previously computed row preserved;
for $i \leftarrow 1$ **to** M **do**
 Compute $P_{MNI}(:, i)$ with MNI from $X(:, [1..i - 1, i + 1..M])$;
end
Sort P_{MNI} columnwise in descending order $\rightarrow pIdx$ sorted index matrix of P_{MNI} ;
 $arrayA_{1:topM} \leftarrow \text{NULL}$;
 $arrayP_{1:topM} \leftarrow \text{NULL}$;
for $i \leftarrow 1$ **to** N **do**
 for $j \leftarrow 1$ **to** $topP$ **do**
 Create perturbation matrix $Pmat$ in two steps:
 (a). $Pmat \leftarrow O(\text{zero-matrix})$;
 (b). $Pmat(pIdx(j, m), m) \leftarrow P_{MNI}(pIdx(j, m), m)$, where $1 \leq m \leq M$;
 Get j -th perturbation vector $P_i(j, :) = Pmat(j, :)$;
 for $k \leftarrow 1$ **to** Total number of all combinations of non-zero perturbations in $P_i(j, :)$, say P_{ijk} **do**
 compute i -th row of A , i.e. A_{ijk} , with NIR and perturbation P_{ijk} ;
 compute $A_{tmp,ijk}$ as identity matrix with $windows$ rows of $arrayA_h(1 \leq h \leq topM)$ and A_{ijk} as i -th row;
 compute $P_{tmp,ijk}$ as zeros matrix with previous $windows$ rows of $arrayP_h(1 \leq h \leq topM)$ and P_{ijk} as i -th row;
 compute $SSE_{ijk} = (X - A_{tmp,ijk}^{-1} \cdot P_{tmp,ijk})^2$;
 end
 end
 Rank SSE and select $topM$ A_{ijk} solutions;
 for $h \leftarrow topM$ **do**
 $arrayA_h = [arrayA_h(1 : i - 1, :); A_{tmp,h}(i, :)]$;
 $arrayP_h = [arrayP_h(1 : i - 1, :); P_{tmp,h}(i, :)]$;
 end
 end
end
 $A = arrayA_1$;
 $P = arrayP_1$;

Algorithm 1. Pseudocode for MANIA. Matricial notations follow Matlab syntax: $M(:, i)$ indicates column i in matrix M , $M(i, :)$ indicates row i in matrix M .

and proved to be stable with $st.dev < 0.08$ in all cases (standard deviations of PPv/sensitivity for undirected networks are 0.07/0.07, and 0.06/0.06 for directed networks).

Dataset 2. This dataset comes from the Dream2 Competition (Heterozygous InSilico 1, Challenge 4) organized by the DREAM (Dialogue for Reverse Engineering Assessments and Methods, [9]) consortium, whose objective is to catalyze the interaction among researchers and improve progresses in the area of cellular network inference. Data was generated using simulations of biological interactions. Namely, the rate of synthesis of the mRNA of each gene is considered to be affected by the level of mRNA of other genes. For these reasons, this represents a valuable and challenging benchmark to test reverse engineering approaches. This dataset contains steady state levels of 50 genes of an hypothetical wild-type organism and 50 heterozygous knock-down strains. All ODE algorithms were tested assuming the number of connections associated with each gene (connectivity of the network, and sparsity of the matrix) is 10, including self-connection. Data were preprocessed with log-transformation of the expression ratio for each gene (knock-down vs wild-type strains). Ratios corresponding to null levels of expression in wild-type were treated as unknown values, and were set to zero as it was done in [4]. Standard deviation of each entry of the data matrix X was computed against the 25-nearest neighbors of the gene of interest, with the approach illustrated in [4]. Finally before computing A , the absolute value of P_{MNI} was normalized column-wise for numeric stability consideration, however this step does not affect the results. Besides evaluating PPv and Sensitivity (ROC curves), the adjacency matrix A was also scored following the procedure adopted in the DREAM 2 Challenge, after scoring the connections of A (normalizing the absolute values). Results were graded using the area under the curve (AUC) for ROC (false positive vs true positive rate) and precision-versus-recall curve (Prec vs Rec)) for the whole set of predictions. For the first

Table 2. Matrix A performance results of Dataset 1. *MNI+NIR* represents the trivial chaining of MNI and NIR, with no strategy to identify the best performances and keeping the single first best and 10 first best predictions of MNI (called respectively MNI+NIR1 and MNI+NIR10). The same values were used for *MANIA*. *Random* refers to the expected performances of an algorithm that selects pairs of genes randomly and then infers an edge between them.

Algorithm	Directed		Undirected	
	PPv	Sensitivity	PPv	Sensitivity
MNI+NIR1	0.84	0.75	0.86	0.76
MNI+NIR10	0.18	0.15	0.27	0.23
MANIA1	0.89	0.81	0.95	0.81
MANIA10	0.75	0.68	0.79	0.70
NIR	0.96	0.86	0.97	0.87
ARACNe	-	-	0.56	0.28
BANJO	0.42	-	0.71	0.00
Random	0.10	-	0.19	-

k predictions (ranked by score, and for predictions with the same score, taken in the order they were put in the prediction files), Precision was defined as the fraction of correct predictions to k , and Recall was the proportion of correct predictions out of all the possible true connections.

Figure 2 is the graphical version of the results of Table 3.

Table 4 and Figure 3 give the results produced by the algorithms for the performances on the directed network. Although MANIA always shows good

Table 3. Performance Results on Dataset 2 on A undirected network

Algorithm	Precision at n^{th} Correct Prediction				AUC	
	1st	2nd	5th	20th	Prec vs Rec Curve	ROC Curve
MNI+NIR1	1.0000	1.0000	0.8333	0.4651	0.2859	0.6965
MNI+NIR10	0.0909	0.0609	0.1219	0.1639	0.1158	0.6230
MANIA1	0.5000	0.4000	0.5556	0.5714	0.3513	0.7957
MANIA10	1.0000	1.0000	0.6250	0.5405	0.3014	0.7191
NIR	1.0000	1.0000	1.0000	1.0000	0.5968	0.8202
ARACNe	1.0000	1.0000	0.5000	0.3279	0.2143	0.6658
BANJO	1.0000	0.3333	0.3125	0.4167	0.1900	0.5925

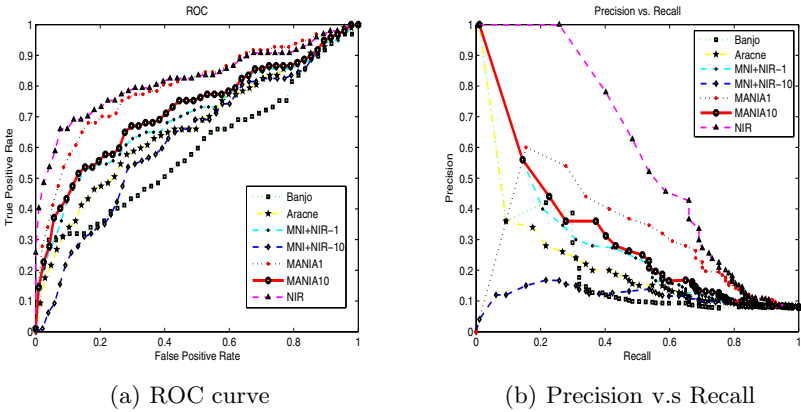


Fig. 2. Performances for A on Dataset 2. AUC Curves for undirected network.

Table 4. Performance Results on Dataset 2 on A directed network

Algorithm	Precision at n^{th} Correct Prediction				AUC	
	1st	2nd	5th	20th	Prec. vs Rec Curve	ROC Curve
MNI+NIR1	0.5000	0.6667	0.5556	0.3279	0.1921	0.6999
MNI+NIR10	0.1429	0.2222	0.1724	0.1835	0.1107	0.6864
MANIA1	0.5000	0.4000	0.6250	0.3846	0.2258	0.7877
MANIA10	1.0000	0.6667	0.5556	0.3448	0.2066	0.7232
NIR	1.0000	1.0000	1.0000	1.0000	0.5781	0.8314
BANJO	0.5000	0.6667	0.2174	0.0559	0.0724	0.5441

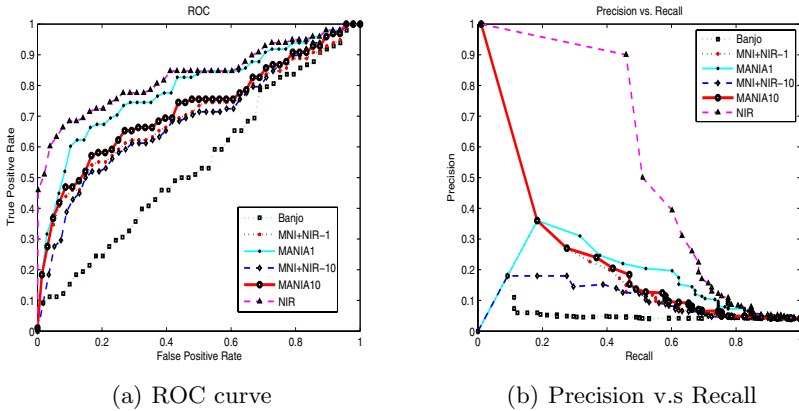


Fig. 3. Performances for A on Dataset 2. AUC Curves for directed network.

Table 5. Performance Results on Dataset 1 and Dataset 2 for the ODE-based algorithms that can predict P

	Dataset 1		Dataset 2	
	PPv	Sensitivity	PPv	Sensitivity
MNI+NIR1	0.870	0.870	0.540	0.540
MNI+NIR10	0.094	0.920	0.066	0.660
MANIA1	0.950	0.860	0.750	0.540
MANIA10	0.798	0.830	0.526	0.600

performances, in this test, interestingly, it also clearly outperforms other ODE-based algorithms. Possible reasons for this are discussed in Section 5.

5 Discussion

We have tested our approach against 4 different algorithms and across 2 datasets interpreted as directed and undirected networks. In general, MANIA can perform better than the state-of-the-art non-ODE approaches listed in Table 1, which were used by setting parameters to their default values, and comparably well or superiorly to ODE approaches as NIR or MNI+NIR. Our objective was to infer A with performances as close as possible to NIR, which we considered as our gold standard [2]. Before discussing further the performances, it is worth noting that comparison between MNI+NIR and MANIA were done with the purpose to assess the validity of the enhancements proposed, compared with our simpler idea of directly chaining the two approaches (MNI and NIR). We figured that, given the reasonably good performances of MNI on the identification of one perturbation, MNI+NIR would be advantaged when used with the parameter $topP = 1$, which offers to NIR the best possible P . In order to perform a fair comparison, MANIA was also tested with this value of the parameter, however, we expected

MANIA to perform better when it can take advantage of more proposed solutions. Our final goal was to assess if, despite the expected variable performances of the two algorithms with different parameter setting, MANIA could be able to identify solutions with global better performances. This is indeed true for the final output of A , and performances remain superior or comparable for the identification of P , indicating that the modification introduced in the multiple regression step defined in MANIA contribute to improve the final results. Overall MANIA has proved to be comparable or to outperform the simplified approach MNI+NIR, even when using identical parameter $topP$, thus, it guarantees more stable performances, and offers results comparable to NIR, with no need for *a priori* information on P . With respect to the identification of P , MANIA shows stable and robust results comparable or outperforming MNI+NIR. These results are confirmed when coming to the identification of A . In particular, when tested against simple models for simulations (Dataset 1) the performances of MANIA and MNI+NIR are inferior to NIR, however they do not degrade much and both algorithms have the fundamental advantage to infer A from expression data only. In the validation on Dataset 2 a more complex and realistic model, these trends are confirmed, with even less variance in the performances, and highlighting the superiority of MANIA. This supports the introduction of the algorithmic variations peculiar to MANIA. Since NIR was the best algorithm tested on this dataset in the Challenge DREAM 2, the possibility to preserve or degrade little the performances under more difficult conditions, represents an important achievement. In general, compared to the best performing algorithm NIR, MANIA has comparable performances in accuracy and the great advantage of not requiring *a priori* knowledge on the targets of the perturbations. Compared to NIR+MNI it has higher ability to remove the noise in matrix P , a characteristic that becomes more and more crucial when the network represents more complex interactions. Although at this level of analysis this is only speculation, it is quite reasonable to assume that real network are indeed complex ones. At the other hand, the lack of *a priori* information about P makes MANIA need to search for an optimal P in a space spanned by P_{MNI} , which increases its computation effort when compared to NIR or MNI+NIR. From the Algorithm in section 3, the main time cost is in the line of the computation of SSE involving inversion which takes $O(N^3)$, thus MANIA scales as $O(N^4)$. Therefore MANIA can infer networks with up to about thousands of vertices in reasonable time. However, MANIA is easily parallelized (row-wise computations of A), and can thus handle larger networks in reasonable time.

6 Conclusion

In this paper, a new reverse-engineering algorithm (MANIA) has been proposed, which effectively couples two assessed approaches MNI and NIR and overcomes their limitations, while preserving their performances. In our simulation experiments, we have shown that MANIA can identify the network of interactions among genes from steady state experiments provided single perturbations are

causing the expression variations. This covers several applications, like single gene knock-down or systematic small molecules testing [10], when assuming the perturbation affects a single target. These are two widely used experimental approaches with applications in chemo- and pharmaco-genomics and model organism research. Although MANIA performed encouragingly, it is worth noting that there is only one single gene perturbed in each experiment in the system under study. Our current work consists in the identification of a proper heuristic for extending this application to multiple perturbation targets and apply the validation to a larger variety of cases.

Acknowledgments. Darong Lai and Hongtao Lu are supported by the Specialized Research Fund for the Doctoral Program of Higher Education (SRFDP, No. 20050248048) and the Program for New Century Excellent Talents in University (NCET-05-0397).

References

1. Guiducci, C., Nardini, C.: High parallelism, portability and broad accessibility: Technologies for genomics. *ACM J* 4(1) Article 3 (2008)
2. Bansal, M., Belcastro, V., Ambesi-Impiombato, A., di Bernardo, D.: How to infer gene networks from expression profiles. *Mol. Syst. Biol.* 3 (2007)
3. Bansal, M., Gatta, G.D., di Bernardo, D.: Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics* 22(7), 815–822 (2006)
4. Gardner, T.S., di Bernardo, D., Lorenz, D., Collins, J.J.: Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* 301(5629), 102–105 (2003)
5. Margolin, A.A., Nemenman, I., Basso, K., Klein, U., Wiggins, C., Stolovitzky, G., Favera, R.D., Califano, A.: Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context (2004)
6. Yu, J., Smith, V.A., Wang, P.P., Hartemink, A.J., Jarvis, E.D.: Advances to bayesian network inference for generating causal networks from observational biological data. *Bioinformatics* 20(18), 3594–3603 (2004)
7. di Bernardo, D., Thompson, M.J., Gardner, T.S., Chobot, S.E., Eastwood, E.L., Wojtovich, A.P., Elliott, S.J., Schaus, S.E., Collins, J.J.: Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nat. Biotechnol.* 23, 377–383 (2005)
8. Cover, T.M., Thomas, J.A.: *Elements of Information Theory*. John Wiley and Sons, Chichester (2001)
9. <http://wiki.c2b2.columbia.edu/dream>
10. Lamb, J., Crawford, E.D., Peck, D., Modell, J.W., Wrobel, I.C.M.J., Lerner, J., Brunet, J.-P., Subramanian, A., Ross, K.N., Reich, M., Hieronymus, H., Wei, G., Armstrong, S.A., Haggarty, S.J., Clemons, P.A., Wei, R., Carr, S.A., Lander, E.S., Golub, T.R.: The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science* 313(5795), 1929–1935 (2006)