Bioinformatics, 31(13), 2015, 2151–2158
doi: 10.1093/bioinformatics/btv075
Advance Access Publication Date: 24 February 2015
Original Paper

OXFORD

Systems biology

GeNOSA: inferring and experimentally supporting quantitative gene regulatory networks in prokaryotes

Yi-Hsiung Chen[1], Chi-Dung Yang[2], Ching-Ping Tseng[2], Hsien-Da Huang[1,2] and Shinn-Ying Ho[1,2,*]

[1]Institute of Bioinformatics and Systems Biology and [2]Department of Biological Science and Technology, National Chiao Tung University, Hsinchu 300, Taiwan, Republic of China

*To whom correspondence should be addressed.

Associate Editor: Igor Jurisica

Received on March 19, 2014; revised on January 4, 2015; accepted on January 30, 2015

Abstract

Motivation: The establishment of quantitative gene regulatory networks (qGRNs) through existing network component analysis (NCA) approaches suffers from shortcomings such as usage limitations of problem constraints and the instability of inferred qGRNs. The proposed GeNOSA framework uses a global optimization algorithm (OptNCA) to cope with the stringent limitations of NCA approaches in large-scale qGRNs.

Results: OptNCA performs well against existing NCA-derived algorithms in terms of utilization of connectivity information and reconstruction accuracy of inferred GRNs using synthetic and real *Escherichia coli* datasets. For comparisons with other non-NCA-derived algorithms, OptNCA without using known qualitative regulations is also evaluated in terms of qualitative assessments using a synthetic *Saccharomyces cerevisiae* dataset of the DREAM3 challenges. We successfully demonstrate GeNOSA in several applications including deducing condition-dependent regulations, establishing high-consensus qGRNs and validating a sub-network experimentally for dose–response and time–course microarray data, and discovering and experimentally confirming a novel regulation of CRP on AscG.

Availability and implementation: All datasets and the GeNOSA framework are freely available from http://e045.life.nctu.edu.tw/GeNOSA.

Contact: syho@mail.nctu.edu.tw

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

In systems biology, inferring gene regulatory networks (GRNs) has focused on identifying effective mathematical models that accurately depict the variation of gene expressions through the use of mathematical equations (Marbach *et al.*, 2012). In these equations, the values of model parameters are estimated through the use of microarray technologies and biological domain knowledge of connectivity information from customized biological experiments (Cooke *et al.*, 2009; di Bernardo *et al.*, 2005) and reliable databases [e.g. RegulonDB (Gama-Castro *et al.*, 2011) and EcoCyc (Keseler *et al.*, 2013)] remains challengeable (Marbach *et al.*, 2012). The general procedure for reconstructing GRNs is illustrated in Figure 1. Various microarray experiments are conducted for some control factors, such as knockout, drug treatment or overexpression, to investigate biological functions using gene expression analysis. To accurately reconstruct GRNs from multiple microarray profiles of dose–response or time–course data, prior connectivity information from public databases of considered species (e.g. RegulonDB and EcoCyc for *Escherichia coli*) that decreases the number of regulations to be estimated is used to reduce the solution space for inferring

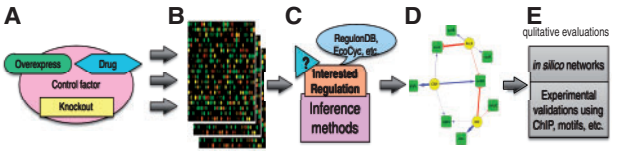

Fig. 1. Overview of a general procedure for reconstructing GRNs. (**A**) Microarray experiments are conducted for some control factors (e.g. knockout, drug treatment or overexpression). (**B**) mRNA expression profiles are obtained from dose–response or time–course microarray data. (**C**) Inference methods reconstruct GRNs with interested regulation from the mRNA expression profiles. (**D**) The mathematical model with parameter values is obtained for a GRN with qualitative regulation. (**E**) Qualitative evaluations using *in silico* networks or biological experiments on interactions of reconstructed GRNs

GRNs. Once the mathematic model of the GRN is established, one can deduce and analyze interested regulation through qualitative assessments regardless of regulatory roles as either activation or repression in GRNs. Qualitative GRNs have been inferred by numerous methods (Marbach *et al.*, 2012) but quantitative analysis and further experimental support is still lacking (Bourdon *et al.*, 2011; Dybas *et al.*, 2008). Quantitative GRNs have been widely investigated since Pan *et al.* (2007) proposed an approach for inferring GRNs with quantitative transcription rates and realistic descriptions of how a gene's regulators influence its expression level. Quantitative regulation information can help biologists reveal hidden knowledge in complicated and large-scale GRNs (Bar-Joseph *et al.*, 2012).

The network component analysis (NCA) approach is a well-known model-based decomposition method for inferring GRNs to deduce valuable information related to the transcription factor activity (TFA) and control strength (CS) of TF-gene connectivity networks (Liao *et al.*, 2003). TFA is defined as the concentration of its subpopulation capable of DNA-binding domains (Chang *et al.*, 2008; Kao *et al.*, 2004) and is difficult to measure experimentally owing to the post-transcriptional and post-translational modifications (Chang *et al.*, 2008; Yang *et al.*, 2005). Estimating the quantities of TFAs provides a basis for investigating perturbations caused by drug effects, genetic mutations, over-expression or complex environmental challenges. The linear model of NCA is described in Eq. (1) where the levels of gene expression are determined by CS and TFA, respectively, represented using the matrices $[A]$ and $[P]$ (Liao *et al.*, 2003),

$$[E] = [A][P] + [\Gamma] \tag{1}$$

The values in the matrix $[E]$ (size $M \times T$) are the log ratios of expression values of M genes and T time points measured using microarray technology under specific conditions. The connectivity matrix $[A]$ (size $M \times N$) is composed of the regulation between M genes and N TFs, and the matrix $[P]$ (size $N \times T$) represents activities of the N TFs on T time points. The matrix $[\Gamma]$ represents the measurement noise or other factors that influence gene expression but are not directly related to this linear model. The NCA-derived algorithms must obey the three identifiability criteria serving as constraints (Chang *et al.*, 2008; Liao *et al.*, 2003; Tran *et al.*, 2010): (i) the matrix $[A]$ must have full-column rank; (ii) any subset of the matrix $[A]$ still has full-column rank and (iii) the matrix $[P]$ must have full-row rank to uniquely decompose the matrix $[E]$ into matrices $[a]$ and $[p]$ by minimizing the matrix $[\Gamma]$ through the following objective function (Liao *et al.*, 2003).

$$\min_{[a],[p]} f = \|[E] - [a][p]\|^2, \; s.t. \; [a] \in Z_0 \tag{2}$$

The set Z_0 is the topologies induced by the network connectivity pattern. The inputs of the NCA-derived algorithms are the matrices $[E]$ and $[A]$ given from gene expression profiles and connectivity information, respectively. The outputs are the values in the matrices $[a]$ and $[p]$. To satisfy the three criteria, some valuable connectivity information (such as the known TF-gene regulations) and some TFs and genes in the GRN of interest would be ignored.

The existing NCA algorithm (Liao *et al.*, 2003) adopts a two-step iterative method to satisfy the three aforementioned criteria, but the obtained solution has two shortcomings: (i) the method adopts a local search methodology (Chang *et al.*, 2008; Tran *et al.*, 2005) and (ii) the solution is computationally instable depending on initial values in matrix $[A]$ (Chang *et al.*, 2008; Galbraith *et al.*, 2006; Tran *et al.*, 2005). The subsequently developed NCA algorithm incorporating the Tikhonov regularization method (Tikhonov and Arsenin, 1977) can improve the numerical sensitivity induced by the ill-conditioned matrices. However, the existing NCA-derived algorithms cannot estimate so many TFAs due to insufficient transcriptome data points in the third criterion. Hence, an enhanced version of the NCA algorithm revises the third criterion, thus relaxing the theoretical data point limitation (Galbraith *et al.*, 2006). The NCA algorithm with the revised third criterion and Tikhonov regularization is named NCAr in this work. However, the third criterion is difficult to meet because a single gene can be regulated by a large number of TFs in higher eukaryotes (Tran *et al.*, 2010). Therefore, the NCA-derived algorithms have to prune some TFs and genes in the original GRN of interest to make the GRN NCA-complaint and thus yield a unique solution. Therefore, the trimming schema was used to extend the applicability of the NCA model to the realm of mammalian regulatory network analysis (Tran *et al.*, 2010). In addition, FastNCA (Chang *et al.*, 2008), non-iterative NCA (NI-NCA; Jacklin *et al.*, 2012), and ROBNCA (Noor *et al.*, 2013) focused on improving the efficiency of matrix decomposition in the NCA model when applied to large-scale GRNs. ROBNCA was more accurate than FastNCA and NI-NCA, irrespective of varying noise, correlation and/or amount of outliers for synthetic datasets (Noor *et al.*, 2013). Numerous studies have successfully used the NCA approach for qualitative GRNs reconstruction, although the NCA algorithms are still subject to limitations in the matrix $[A]$ (Misra and Sriram, 2013; Shao *et al.*, 2012; Tran *et al.*, 2010, 2012; Ye *et al.*, 2009) and instability in the matrix $[P]$ with insufficient connectivity information (Chang *et al.*, 2008; Misra and Sriram, 2013; Noor *et al.*, 2013; Ye *et al.*, 2009). Theoretically, the inferred GRNs using existing NCA-derived algorithms need to be further refined by biologists and confirmed through biological knowledge and empirical evidence.

This work proposes a framework (named Gene Networks via Orthogonal Simulated Annealing, GeNOSA) for inferring quantitative GRNs from multiple microarray profiles of dose–response or time–course data using a global optimization algorithm (named OptNCA). OptNCA bases on orthogonal simulated annealing (OSA) and the best use of prior knowledge for the initial setting of the connectivity matrix $[A]$ without matrix reduction to confine the solution space of the decomposition problem of the NCA model (Fig. 2). We demonstrate the effectiveness of this framework in several respects. First, we adopt NCA-compliant datasets (Liao *et al.*, 2003) as benchmarks and show that OptNCA is superior to existing NCA-derived algorithms in terms of the true network *in silico*. Moreover, OptNCA without using known qualitative regulations is also evaluated in terms of qualitative assessments using one of the widely used datasets of DREAM3 challenges. Second, we show the good quality of the inferred GRNs using GeNOSA to deduce condition-dependent regulations of a CRP-regulated GRN. Third, we apply GeNOSA to establish quantitative GRNs from

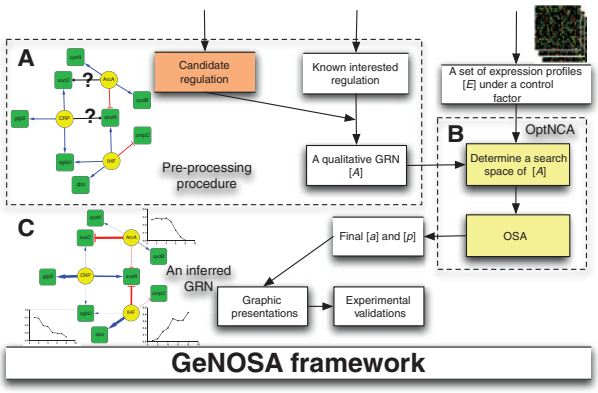

Fig. 2. System diagram of the GeNOSA framework: (**A**) the pre-processing procedure takes a qualitative GRN with initial regulation settings from biological domain knowledge of connectivity information; (**B**) the OptNCA algorithm decomposes the matrix [*E*] to obtain matrices [*a*] and [*p*] by minimizing LSE. The OptNCA algorithm obtains final matrices [*a*] and [*p*] as an inferred GRN from a number of independent runs using OSA, and (**C**) quantitative results of the inferred GRN are also shown in graphic presentation

dose–response or time–course microarray data, and validate quantitative regulations in a small sub-network *in vitro*. The wide applicability of GeNOSA suggests that this framework is not only of theoretical use in the reverse engineering of quantitative GRNs but also can be practically applied in real-world GRNs.

2 Methods

2.1 Proposed algorithm OptNCA

The performance of the OptNCA algorithm benefits mainly from the optimization algorithm OSA and available information of TF-gene regulation. According to the NCA model with knowing qualitative regulation, we encode the solution representation and design an objective function to make the best use of OSA. The initial setting of the matrix [*A*] uses known regulation between genes and TFs collected from the literature and existing databases for *E. coli*, such as RegulonDB (Gama-Castro *et al.*, 2011) and EcoCyc (Keseler *et al.*, 2013). The initial values of variables in the matrix [*A*] are one of the four values 1, −1, 0 and 99, respectively, representing up regulation, down regulation, no regulation and unknown regulation. The zero values in the matrix [*A*] remain unchanged and the corresponding variables are not encoded into the solution representation. The magnitude of the variables with initial values of 1 and −1 are tuned using OSA without changing their signs. If the recorded regulation of a variable in the matrix [*A*] is dual, the variable is assigned with an initial value of 99 and can be positive or negative real values in the feasible range optimized by OSA. The setting of variable ranges in matrices [*A*] and [*P*] depends on the given matrix [*E*] with prior knowledge or problem-specific constraints. Since the values in the matrix [*P*] for real profiles are difficult to measure directly by biological experiments, we assume the variables in the matrix [*P*] have real values in the range of [−1, 1] unless specified otherwise, and these values (representing the relative activities of TFs in log ratio for different phrases) are initialized with zero. The reconstruction of a GRN is formulated as an optimization problem where the least square error (LSE) between the experimental and estimated gene expressions is minimized as in the objective function Eq. (2) originally proposed by Liao *et al.* (2003). OptNCA aims to obtain accurate signs of variables with dual regulations and magnitudes of variables in matrices [*a*] and [*p*] using an efficient optimization approach while considering noisy gene expressions. The OptNCA algorithm using OSA aims to find a nearly optimal solution to the large-scale optimization problem. The values in matrices [*a*] and [*p*] are obtained from decoding a current solution of OSA. The OSA used in OptNCA is described below.

There are three essential parts to be specified before performing the procedure of OSA to solve this optimization problem using the NCA model: (i) representation of a solution, (ii) objective function and (iii) cooling schedule. Generally, the cooling schedule in OSA consists of four parameters: (i) the temperature T, (ii) the cooling rate CR, (iii) the search radius R and (iv) the S number of steps for each temperature. The proper values of parameters are problem-dependent using OptNCA. Let X be a candidate solution consisting of all variables in matrices [*a*] and [*p*].

1. Initialize a current solution X according to the initial settings of variables in the NCA model and four parameters T, CR, R and S of OSA. Let the count index $C = 0$.
2. Generate two temporary solutions X_1 and X_2 using X for perturbation.
3. Divide X, X_1 and X_2 into a number of groups of variables and apply the orthogonal experimental design (OED) with a three-level orthogonal array to obtain a candidate solution Q which is a potentially good combination of these variables' values (Ho *et al.*, 2006).
4. Accept Q as the new X with probability $P(Q)$:

$$P(Q) = \begin{cases} 1, & \text{if } f(Q) \leq f(X) \\ \exp\left(\dfrac{f(X) - f(Q)}{T}\right), & \text{if } f(Q) > f(X) \end{cases} \quad (3)$$

5. Increase the value of C by one. If $C < S$ go to step 2.
6. Multiply the values of T and R by CR and reset C to zero.
7. If a pre-specified stopping criterion is met (In this work, the variance of the fitness values is less than 10^{-4} over 50 iterations), decode X to obtain the solutions for matrices [*a*] and [*p*] and stop the algorithm. Otherwise, go to step 2.

The matrix decomposition problem incorporated with the NCA model for reconstructing large-scale GRNs is very intractable and is subject to computational instability due to multiple local optima (Chang *et al.*, 2008; Tran *et al.*, 2005). The three optimization metaheuristics, OSA, Intelligent evolutionary algorithm (IEA; Ho *et al.*, 2004) and Orthogonal particle swarm optimization (OPSO; Ho *et al.*, 2008) based on OED are useful for solving large parameter optimization problems. OSA is a single solution based metaheuristic, and IEA and OPSO are population-based metaheuristics (Boussaid *et al.*, 2013). When the connectivity information is used to substantially reduce a search space in inferring large GRNs, OSA outperforms IEA and OPSO in a limited computation time that benefits from a good initial solution. Considering the rapid increase of available connectivity information accompanied with high-throughput techniques and the proposed GeNOSA framework for inferring large-scale GRNs, we adopt OSA in OptNCA. OSA aims to escape local optima to find a globally optimal solution (Ho *et al.*, 2006).

2.2 Evaluation of inferred GRNs

Pursuing high quality of inferred GRNs requires an efficient computation method to obtain accurate values of matrices [*A*] and [*P*]. The LSE error between the true and estimated profiles is commonly used to evaluate the search ability of computation methods. However,

the quality of the mathematic model for GRNs cannot be assessed by only LSE. A suitable measurement is needed to evaluate the inferred models of GRN. The TFA for describing the regulation strength (RS) between TFs and its target genes is time-variant or dose-dependent. To provide a simple measurement for comparing two inferred models of GRN, we summarize the TFAs of all time points or doses. The value of RS_{mn} defined in Eq. (4) illustrates the RS of the nth TF to the mth gene over T experiments for each TF-gene regulation.

$$RS_{mn} = \sum_{t=1}^{T} A_{mn} P_{nt} \tag{4}$$

The variables A_{mn} and P_{nt} respectively denote each value in matrices $[a]$ and $[p]$. Furthermore, false prediction rate (FPR) is an evaluation function defined in Eq. (5) to evaluate models of GRNs

$$FPR = \frac{\sum_{m=1}^{M}\sum_{n=1}^{N} Sign(RS_{mn}, RS'_{mn})}{N_Z} \tag{5}$$

where RS_{mn} and RS'_{mn} respectively denote the true and estimated RS. The N_Z is the total number of true TF-gene regulations with non-zero RS. The *Sign* function returns one if the signs of true and estimated values are different; otherwise it returns zero. We examine the signs of RS s to assess the correctness of the inferred GRNs compared to the high-trust regulation between TFs and genes. The method of obtaining a low FPR value can help to identify a correct TF-gene topological structure to justify unknown or questionable regulations from the literature or published databases.

2.3 Evaluation of dual regulation with adaptive thresholds

Most of TF-gene interactions were condition-dependent meaning that the regulatory roles of these TFs change with environmental conditions (Brynildsen *et al.*, 2006; Harbison *et al.*, 2004; Tran *et al.*, 2010), thus increasing the difficulty of GRN reconstruction. GeNOSA provides three comprehensive steps to deduce behaviors of dual regulation for a specific condition. First, the inferred TFA for these dual-regulatory genes must find a consensus of activity values with a low standard deviation in 30 independent runs of OptNCA. Second, insignificantly expressed genes should be omitted by determining adaptive thresholds of the summation of gene expression levels (E_{RS}) and their standard deviation (E_{SD}) over experiments from microarray data. Third, two variables CS_{avg} and S_c were used thresholds to assess the correctness of the predicted regulation of GeNOSA.

$$CS_{avg,mn} = \frac{\sum_{r=1}^{S_{c,mn}} A_{r,mn}}{S_{c,mn}},\ s.t.\ A_{r,mn} \in AS_{mn}. \tag{6}$$

The variable $S_{c,mn}$ is the large one of the two respective numbers of activation and repression in the 30 CSs and the set AS_{mn} consists of a number $S_{c,mn}$ of CSs within 30 independent runs of OptNCA. In the microarray DR-cAMP, we used thresholds ($|E_{RS}| > 1.0$ and $E_{SD} > 0.1$ for at least 1-fold changes and low experimental variability over dosages of expression levels, respectively) to filter out insignificantly expressed genes from microarray data, and then we reserved highly consistent regulations with thresholds at $|CS_{avg}| > 1.0$ and $S_c = 100\%$ (Supplementary Dataset S2). Results show that highly reliable regulation depends on the thresholds determined in these three steps. Hence, dual regulations with high confidence could be confirmed as activators or repressors under cAMP dosages by these thresholds.

3 Results and discussion

Various datasets are designed to compare OptNCA with existing NCA-derived algorithms by using LSE and FPR. The NCA (Liao *et al.*, 2003), NCAr (Tran *et al.*, 2005), FastNCA (Chang *et al.*, 2008), NI-NCA (Jacklin *et al.*, 2012), ROBNCA (Noor *et al.*, 2013) and NARROMI (Zhang *et al.*, 2013) algorithms are freely available from their websites. The OptNCA parameters used in this work are $T = 1000$, $CR = 0.999$, $R = 0.05$, and $S = 1000$. Because the non-zero values in matrices $[A]$ and $[P]$ of these four synthetic datasets have real values within the range $[-10, 10]$, and NCA-derived algorithms estimate solutions without retaining the sign of the values, we set the variables with an initial value of 99 in the same range for OptNCA.

3.1 Efficiency and accuracy comparisons among NCA-derived algorithms and OptNCA

For the *in silico* experiments, we performed 30 independent runs using NCAr, FastNCA, NI-NCA, ROBNCA and OptNCA with four designed datasets (Section 1.1 of Supplementary Materials). The results shown in Table 1 reveal that OptNCA is superior to all the NCA-derived algorithms while considering the mean accuracies of qualitative and quantitative regulations in terms of FPR and LSE, respectively. Although FPRs of FastNCA and NI-NCA are lower than those of NCAr in Table 1, LSEs of these two algorithms are much higher than those of NCAr revealing that the values of quantitative regulations do not approximate expression profiles perfectly. It is better to consider both FPR and LSE simultaneously. In general, NCAr is better than FastNCA, NI-NCA and ROBNCA considering the three datasets (Kao_PNAS, Kao_Silico.R, and Kao_Silico.UR). Hence, the results of NCAr and OptNCA are compared and analyzed further in detail. The values of LSE and FPR obtained using OptNCA are substantially reduced for the real dataset Kao_PNAS. Gaussian noises with levels of 5 and 10% were added to the expression profile $[E]$ for Kao_PNAS. For matrices $[A]$ and $[P]$ we used the same parameter settings, which are required to meet all the identifiability criteria for the NCAr algorithm. The results show that OptNCA outperforms NCAr in terms of accuracy and standard deviation for various noise levels.

We also compare OptNCA with NCAr using three synthetic datasets based on the Kao_PNAS with known solutions. The values of LSE and FPR obtained by OptNCA on Kao_Silico.R and Kao_Silico.UR are 0.0003 ± 0.00 and 68.12 ± 7.39, and 0.43 ± 1.31 and $0.00 \pm 0.00\%$, respectively. Considering the results of NCAr (0.56 ± 0.10 and 13656.1 ± 0.00 for LSE, and 19.64 ± 0.77 and $8.57 \pm 0.00\%$ for FPR), OptNCA performs much better than NCAr in terms of LSE and FPR. The high LSE causes bias in quantitative analysis of inferred GRNs, and even FPR is low. The comparable improvements of FPR with minimized LSE had qualitative and quantitative impacts on the inferred GRNs. For the synthetic dataset Silico_30 of an NCA-noncompliant network, OptNCA can minimize LSE (0.0004 ± 0.00) for approximating expression profiles. Although OptNCA can obtain a nearly optimal solution to the NCA model decomposition problem defined in Eq. (1), its FPR is as high as $23.28 \pm 4.20\%$. The result indicates that best profile fitting cannot guarantee correctness of the inferred GRN model, especially when the degree of freedom in the matrix $[A]$ is

Table 1. Comparisons among NCAr, FastNCA, NI-NCA, ROBNCA and OptNCA using real and synthetic datasets

Datasets	NCAr		OptNCA		FastNCA		NI-NCA		ROBNCA	
	LSE	FPR (%)	LSE	FPR (%)	LSE	FPR (%)	LSE	FPR (%)	LSE	FPR (%)
Kao_PNAS	14.66 ± 0.03	40.93 ± 0.38	12.91 ± 0.09	30.98 ± 1.10	74.95	26.43	87.34	26.43	23.25	41.43
Kao_PNAS (5%)	16.44 ± 0.05	41.74 ± 0.74	15.60 ± 0.12	32.55 ± 1.33	69.01	35.71	75.31	26.43	24.72	46.43
Kao_PNAS (10%)	23.39 ± 0.83	42.43 ± 2.82	21.12 ± 0.16	33.90 ± 4.24	71.90	44.29	88.42	27.14	27.15	42.14
Kao_Silico.R	0.56 ± 0.10	19.64 ± 0.77	0.0003 ± 0.00	0.43 ± 1.31	21.89	23.57	39.33	27.86	2.86	37.86
Kao_Silico.UR	13656.1 ± 0.00	8.57 ± 0.00	68.12 ± 7.39	0.00 ± 0.00	135251	50.00	457746	38.57	25147	47.14
Silico_30	–	–	0.0004 ± 0.00	23.28 ± 4.20	–	–	–	–	–	–

The mean ± standard deviation (SD) is given for each measurement. The results obtained by OptNCA are better than those using NCAr, FastNCA, NI-NCA and ROBNCA in terms of LSE and FPR. FastNCA, NI-NCA and ROBNCA are deterministic algorithms resulting in SD = 0.

substantially increased. OptNCA can use connectivity information to decrease the degree of freedom in the matrix [*A*]. In conclusion, OptNCA is more stable and finds better solutions than NCAr in terms of LSE and FPR for these four datasets. Regarding the TFAs estimated by OptNCA and NCAr for Kao_PNAS, the similarity of the TFA profiles between these two algorithms is high except those of CRP and NarL. The Pearson's correlation coefficients (PCCs) for CRP and NarL are 0.622 and 0.027, respectively.

OptNCA and NCAr were compared using Kao_Silico.R and Kao_Silico.UR with known values in matrices [*E*], [*A*] and [*P*] for advanced accuracy analysis. We analyzed the values in the matrix [*P*] for these two datasets and found that the trends of TFAs are similar in terms of both signs and magnitudes, but there are only two exceptions from these two experiments (Supplementary Fig. S1). The averaged PCCs for matrices [*A*] and [*P*] on both Kao_Silico.R and Kao_Silico.UR using NCAr and OptNCA are as high as 0.914 and 0.978 (Supplementary Table S1), respectively. One exception is the TFA of NarL in that PCC = 0.634 for NCAr and 1.000 for OptNCA on Kao_Silico.R (Supplementary Fig. S1A). Similarly, the other exception on Kao_Silico.UR is RpoE and its TFA profile obviously exhibits an opposite trend (PCC = −0.999) using NCAr while OptNCA has PCC = 1.000 (Supplementary Fig. S1B and Dataset S1). In general, OptNCA and NCAr performed well in estimating TFAs on Kao_Silico.R (Supplementary Fig. S1A). However, NCAr obtained larger magnitudes of TFAs on Kao_Silico.UR compared to the true TFAs. OptNCA with PCC = 1.000 is substantially superior to NCAr with PCC = 0.857 (Supplementary Fig. S1B and Dataset S1).

3.2 Comparisons with non-NCA-derived algorithms

Although OptNCA aims to utilize known qualitative regulations to infer reliable quantitative GRNs (Fig. 2), OptNCA can also deduce qualitative regulations by assigning initial values of 99 to variables in the matrix [*A*]. In this work, the reference network Yeast1 from the fourth challenge of DREAM3 (Marbach *et al.*, 2010) was used for further evaluation and comparison. Although these simulated data sets likely exhibit properties that are different from real data, it is instructive for evaluation as it is widely used in the literature on network inference. To infer qualitative GRNs, the threshold value of RS for discriminating the regulation from non-regulation is set to 0.1, 0.05 and 0.01 for the network sizes of 10, 50 and 100, respectively.

The results of non-NCA-derived algorithms were obtained from NARROMI (Zhang *et al.*, 2013), as shown in Table 2. The following measures, true positive rate (TPR), false positive rate (FPoR), positive predictive value (PPV), accuracy (ACC), Matthews Coefficient Constant (MCC) and the area under receiver operating characteristic curve (AUC), were used (Zhang *et al.*, 2013). The effectiveness of NARROMI (Zhang *et al.*, 2013) was confirmed through cross-validation results on various datasets suggesting that NARROMI integrally outperformed previous non-NCA-derived methods, such as LP (Wang *et al.*, 2006), RO (Zhang *et al.*, 2013), ARACNE (Margolin *et al.*, 2006), GENIE3 (Huynh-Thu *et al.*, 2010) and LASSO (Geeven *et al.*, 2012). However, performance of these algorithms diverse according to measures and datasets used (Zhang *et al.*, 2013). OptNCA outperforms LASSO, LP and RO in the case of network size 10 in terms of TPR, FPoR, PPV, ACC and MCC and performs well only in TPR and FPoR for the network size of 50. The other non-NCA-derived algorithms outperform OptNCA in deducing qualitative regulations on the two datasets.

Table 2. Comparison of various methods on networks (Yeast1) with sizes 10 and 50 in the fourth challenge of DREAM3

Method	TPR	FPoR	PPV	ACC	MCC	AUC
Size 10						
LASSO	0.600	0.837	0.082	0.211	−0.191	0.703
LP	0.100	0.412	0.029	0.533	−0.202	0.738
RO	0.100	0.500	0.024	0.456	−0.252	0.798
ARACNE	0.900	0.112	0.500	0.888	0.618	0.930
GENIE3_FR_sqrt	0.700	0.112	0.437	0.867	0.483	0.919
GENIE3_FR_all	0.700	0.138	0.389	0.844	0.442	0.894
NARROMI	0.700	0.050	0.636	0.922	0.623	0.938
OptNCA	0.800	0.375	0.211	0.644	0.270	0.713
Size 50						
LASSO	0.351	0.129	0.081	0.855	0.113	0.711
LP	0.389	0.085	0.130	0.899	0.182	0.669
RO	0.494	0.131	0.109	0.857	0.181	0.727
ARACNE	0.597	0.082	0.192	0.908	0.303	0.832
GENIE3_FR_sqrt	0.481	0.078	0.167	0.908	0.245	0.843
GENIE3_FR_all	0.442	0.073	0.164	0.912	0.231	0.796
NARROMI	0.532	0.062	0.217	0.925	0.307	0.839
OptNCA	0.533	0.293	0.056	0.701	0.091	0.620

LP, RO, ARACNE, GENIE3, LASSO and NARROMI represent methods based on linear programming, recursive optimization using ODE, mutual information (MI), random forests, regression model and recursive optimization and MI, respectively. OptNCA based on OSA optimization and the NCA model.

OptNCA can infer regulatory roles (either activation or repression) of TFs. Therefore, we further utilized the qualitative regulations of Yeast1 from the gold standard of DREAM3 to infer quantitative GRNs. The error rates of predicted regulatory roles for the network sizes of 10, 50 and 100 are 0, 23.5 and 28.9%, respectively. OptNCA performs well in inferring small-scale quantitative GRNs. Notably, no result of predicted regulatory roles on the same datasets is available for the methods in Table 2.

3.3 Deduce behaviors of dual-regulatory genes of CRP

In the qualitative GRN CRP-ReDB72 (Section 2, Supplementary Materials), there were 13 dual-regulations where 42 regulations were regulated by CRP under some specific culture conditions. Adaptive thresholds for evaluating dual regulation were used to assess the deductive performance of OptNCA for the 42 CRP-regulated genes from these dual regulations in the cAMP dose–response microarray under aerobic growth conditions (DR-cAMP) to reveal the regulatory roles of CRP within seven relative concentrations (Section 1.2, Supplementary Materials). First, the estimated activity of CRP was stable over 30 independent runs of OptNCA, and the averaged standard deviation for seven concentrations was less than 0.04. This result suggested that the estimated activity of CRP was highly consistent under this condition. The averaged PCC between estimated values of CS and expression levels of genes in seven cAMP concentrations was 0.788, indicating that the deduction substantially agreed with the experimental results. Second, insignificantly expressed genes are rarely informative and we thus obtained 21 significantly expressed genes for further analysis. Third, we used a majority rule on the 30 estimated values of CS to determine the regulatory role of CRP on these 21 genes, and then calculated mean values considering the values of CS with consistent signs for the regulatory role, which represent the fold change contributed to gene expressions by the activity of CRP. Six deduction regulations were correctly confirmed among the nine regulations with $|CS| < 1$, and the other 12 deduction regulations with $|CS| \geq 1$ (greater than 1-fold change) were all correctly confirmed by analyzing gene expression in the microarray DR-cAMP. The 12 genes with substantially identified regulations are *araJ*, CytR, *dadA*, *dadX*, DgsA, MalI, *nagB*, *nagE*, *nupG*, *proP*, *tsx* and *udp* (Supplementary Table S2).

3.4 Application to infer and validate quantitative GRNs

GeNOSA was used to reconstruct two quantitative GRNs for a CRP-mediated network (CRP-ybiT) from the dose–responses (DR-cAMP) and time–course (TC-IS) microarray data (Fig. 3A). We used the qualitative regulation recorded in RegulonDB and the default parameter settings for OptNCA in reconstructing the quantitative GRNs. The regulation between CRP and AscG was initialized as an unknown regulation. The estimation of gene expression values over time and dosage were globally optimized through the cooperation of the estimated CSs and TFAs using OptNCA. OptNCA was used to conduct 30 independent runs of OSA to obtain two CRP-mediated GRNs (Supplementary Fig. S2). The averaged mean squared errors were 0.021 and 0.223, respectively, in approximating the gene expression profiles of DR-cAMP and TC-IS. Considering that the expression values of DR-cAMP and TC-IS were, respectively, in the range of [−2.79, 6.60] and [−4.77, 4.70], OptNCA performed well in fitting the expression profiles. Consequently, we want to further validate the estimated TFAs.

For experimental validation, we randomly chose a small CRP-regulated from all the sub-networks with the smallest regulation by outer TFs on the target genes in the sub-network (Supplementary Fig. S3). The sub-network consists of two genes (*ascF* and *ybiT*) and their upstream TFs (CRP and AscG) from the qualitative CRP-mediated GRN (Fig. 3A). The TFAs of CRP and AscG were estimated from the expression profiles of their downstream genes that CRP regulates 432 genes and AscG regulates three genes (*ascG*, *ascF* and *ybiT*). By observing the expression levels of *ascG*, the self-regulation of AscG was found to have occurred in the TC-IS but not DR-cAMP dataset (Supplementary Dataset S2). The averaged standard deviations of the estimated TFAs over 30 independent runs of OptNCA were very small (0.0060 for CRP and 0.0092 for AscG) from DR-cAMP dataset (Supplementary Fig. S4A). The expression profiles of *ascF* and *ybiT* were perfectly estimated, and the trends of real and estimated profiles of *ascG* were very similar (Supplementary Fig. S4B). In the GRN inferred from TC-IS dataset, the averaged standard deviations of the estimated TFAs over 12 growth times were 0.0077 for CRP and 0.1059 for AscG (Supplementary Fig. S5A). The TFA of AscG featured tolerable variance and the estimated expression profile of *ascG* was less accurate due to the self-regulation (Supplementary Fig. S5B). The gene expression of self-regulation is affected by many factors such as post-transcriptional regulation, transcriptional regulation and other mechanisms. Furthermore, the NCA model deals mainly with the transcriptional regulation only. However, the estimated expression ratios of *ascF* and *ybiT* were still accurate.

Consequently, we verified the results of those fold changes by knocking out these two TFs for three concentrations (0.01, 0.1 and 1 mM) in DR-cAMP and five growth times in TC-IS by averaging from three replicates of quantitative RT-PCR (qRT-PCR, Supplementary Materials). The quantitative results of this sub-network obtained using GeNOSA on the qualitative GRN CRP-ybiT are shown along with the experimental validation results (Figs. 3B and C). These two figures show the estimated TFAs of CRP and AscG indicating that CRP activates AscG in both the DR-cAMP and TC-IS microarray data. Furthermore, the PCC test indicated high correlations of relative gene expressions between the qRT-PCR and GeNOSA results for both the dose-responses and time-course datasets. The correlation coefficients (the two-tailed *t*-test *P* value) for DR-cAMP and TC-IS are 0.85 ($P = 6.9 \times 10^{-5}$) and 0.88 ($P = 5.59 \times 10^{-12}$), respectively (Supplementary Tables S3 and S4). The strength of association between the qRT-PCR and GeNOSA results is highly related and statistically significant from the null hypothesis ($P < 0.0001$). Consequently, we can say that GeNOSA can explain the coefficients of determination (72% in DR-cAMP and 77% in TC-IS) of variance in the result of qRT-PCR.

3.5 Experimental validations for the regulation of CRP on AscG

The estimation results described in Section 3.3 show that CRP exerts a strong regulation on AscG, shown in Figures 3B and C. Experimental validations were used to confirm the novel regulation of CRP on AscG. The *in vitro* electrophoretic mobility shift assays (EMSA) for the CRP binding sites on *ascG*, *ascF* and *ybiT* genes were shown in Figure 4A. The quantitative results obtained using Zero-D software (Scanalytics, Billerica, MA) indicate that CRP clearly retarded the migration of the *ascF*, *ascG* and *ybiT* oligonucleotides (Fig. 4B) and CRP was responsible for the promoters of *ascF* and *ybiT* in agreement with the regulation of CRP (Ishida *et al.*, 2009; Raghavan *et al.*, 2011). Moreover, the CRP binding site of *ascG* was recognized near positions ranging from −63 to −78 upstream of *ascG* with TGTGAAACCGGTCACC for the first time. Therefore, a novel regulation between CRP and AscG is validated.

To clarify the effects of CRP regulation on *ascG*, *ascF* and *ybiT*, wild-type and mutant strains (such as Δ*crp*, Δ*ascG* and Δ*crp*Δ*ascG*) were grown to an exponential phase to detect the gene expressions using qRT-PCR. The data from each mutant strain were normalized against those for the 16 S rRNA as a positive control. The results show that the expressions of *ascF* and *ascG* were significantly repressed in the Δ*crp* strain and slightly reduced in *ybiT* (Fig. 4C). This result indicates that cAMP-CRP is an activator for *ascF*, *ascG* and *ybiT*. In contrast, the gene expression of *ascF* was depressed in

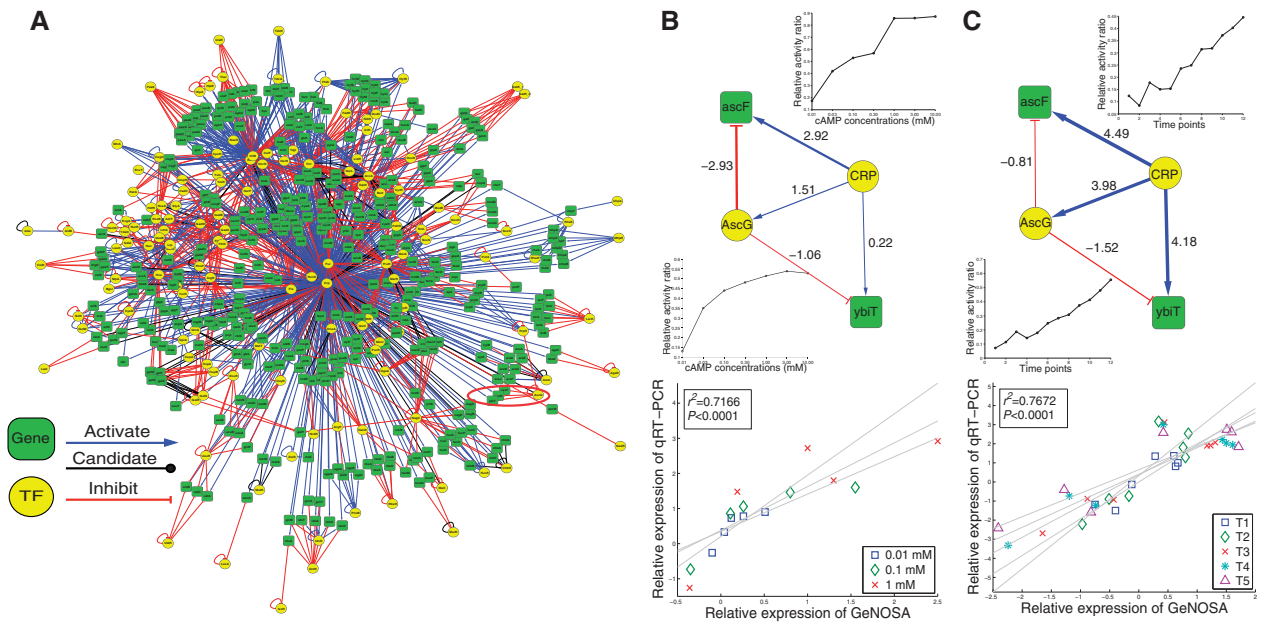

Fig. 3. A CRP-regulated network (CRP-ybiT) for a new regulation and their quantitative results. We had reconstructed a qualitative GRN as shown in (**A**). We choose an independent sub-network including an unknown regulation between CRP and AscG from this GRN and design biological experiments to perform quantitative analysis for GeNOSA. The quantitative and correlation results for (**B**) DR-cAMP (five mutant experiments for each of the three concentrations) and (**C**) TC-IS (seven mutant experiments for each of the five time points). The quantitative values of CS of DR-cAMP and TC-IS were illustrated by the line width and the activities of TFs are also shown to demonstrate the dynamics of the reconstructed GRN (Upper). The correlation between results of qRT-PCR and GeNOSA are plotted (Lower, see Supplementary Tables S3 and S4 in detail). Each expression is the average of three replicates of qRT-PCR. The Pearson's correlation coefficients are 0.85 and 0.88 for DR-cAMP and TC-IS, respectively

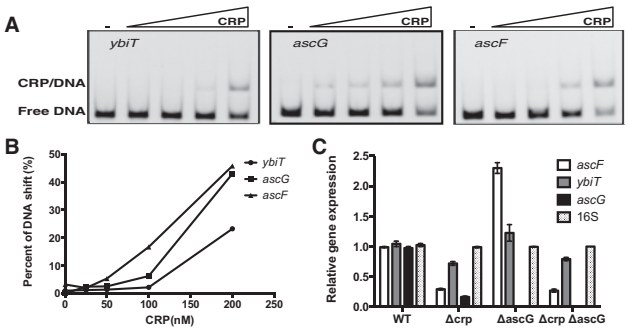

Fig. 4. Experimental validations on the quantitative sub-network. (**A**) The binding affinities of CRP to the promoters of *ascF*, *ascG* and *ybiT* are determined by EMSA, (**B**) quantified by measuring the intensity of retard bands in subfigure **A** to retain bands with Zero-D scan software and (**C**) the binding affinities of CRP to the promoters of *ascF*, *ascG* and *ybiT* are determined by qRT-PCR

the Δ*ascG* strain and slightly increased in *ybiT*. The Δ*crp*Δ*ascG* double-mutant strain was further used to interpret the effects of the two transcription regulators on gene expression of *ascF* and *ybiT*. The result indicates that the expression patterns of *ascF* and *ybiT* in the Δ*crp*Δ*ascG* double-mutant strain were similar to that of the Δ*crp* strain. This finding suggests that AscG exhibits a suppressive effect and CRP is the dominant regulator for *ascF* and *ybiT* in response to activation of gene expression.

4 Conclusion

This work proposes a complete framework GeNOSA to infer large-scale, qGRNs using OptNCA based on the model of NCA. Using an OSA algorithm, the proposed OptNCA can efficiently yield a good solution to the decomposition problem without performing matrix reduction. To obey the three identifiability criteria, existing NCA-derived algorithms except OptNCA usually perform matrix reduction and ignore some genes and transcription factors of interest. However, with prior knowledge of connectivity information from the literature and existing databases, OptNCA can retain this reliable connectivity information in inferring GRNs. The general-purpose GeNOSA can quantitatively estimate CS and TFA as well as discover novel regulations for specific conditions and applications. Estimating quantitative TFAs improves understanding of how the physical states of cells respond to environmental changes and elucidate the transcriptional architectures for various applications such as drug design.

OptNCA performs well compared to existing NCA-derived algorithms in terms of reliable connectivity information and reconstruction accuracy using synthetic and real *E. coli* datasets. Experimental results indicate that GeNOSA can infer CRP-regulated GRNs to quantitatively deduce condition-dependent and undiscovered regulations. In the NCA model, the quantity of the activity of a TF mainly depends on its target genes. By investigating the stability of a TFA that targets on multiple genes, GeNOSA provides a solution to elucidate relative CS of TF-mediated genes and the dynamics of TFAs (upper parts of Figs. 3B and C) for various dosages and time points.

While existing NCA-derived algorithms focus on approximating expression profiles for GRN reconstruction with qualitative regulation or quantitative regulation, GeNOSA aims to minimize a profile approximation error while accurately identifying a topological structure with quantitative regulation for efficiently establishing large-scale GRNs. The framework GeNOSA consists of tools, including generation of input files for OptNCA and optimization of GRNs, to make the best use of connectivity information from the literature and regulation databases for time-course or dose–response

microarray datasets. We believe that GeNOSA could serve as a fundamental tool for estimating TFAs in the reverse engineering of inferring quantitative GRNs and lead to the development of a rich set of applications in the field of systems biology.

Acknowledgement

The authors thank the members of Applied Microbiology and Bioengineering Laboratory in NCTU for valuable discussions: Yi-Pei Chen, Yen-Hua Chen and Shin-Wen Wei.

Funding

This work was funded by National Science Council of Taiwan under the contract number NSC-103-2221-E-009-117-, and 'Center for Bioinformatics Research of Aiming for the Top University Program' of the National Chiao Tung University and Ministry of Education, Taiwan, ROC for the project 103W962. This work was also supported in part by the UST-UCSD International Center of Excellence in Advanced Bioengineering sponsored by the Taiwan National Science Council I-RiCE Program under Grant Number: NSC-102-2911-I-009-101.

Conflict of Interest: none declared.

References

Bar-Joseph,Z. *et al.* (2012) Studying and modelling dynamic biological processes using time-series gene expression data. *Nat. Rev. Genet.,* **13**, 552–564.

Bourdon,J. *et al.* (2011) Integrating quantitative knowledge into a qualitative gene regulatory network. *PLoS Comput. Biol.,* **7**, e1002157.

Boussaid,I. *et al.* (2013) A survey on optimization metaheuristics. *Inf. Sci.,* **237**, 82–117.

Brynildsen,M.P. *et al.* (2006) A gibbs sampler for the identification of gene expression and network connectivity consistency. *Bioinformatics,* **22**, 3040–3046.

Chang,C. *et al.* (2008) Fast network component analysis (fastnca) for gene regulatory network reconstruction from microarray data. *Bioinformatics,* **24**, 1349–1358.

Cooke,E.J. *et al.* (2009) Computational approaches to the integration of gene expression, chip-chip and sequence data in the inference of gene regulatory networks. *Semin. Cell Dev. Biol.,* **20**, 863–868.

di Bernardo,D. *et al.* (2005) Chemogenomic profiling on a genomewide scale using reverse-engineered gene networks. *Nat. Biotechnol.,* **23**, 377–383.

Dybas,J.M. *et al.* (2008) Computational analysis and experimental validation of gene predictions in *Toxoplasma gondii*. *PLoS ONE,* **3**, e3899.

Galbraith,S.J. *et al.* (2006) Transcriptome network component analysis with limited microarray data. *Bioinformatics,* **22**, 1886–1894.

Gama-Castro,S. *et al.* (2011) Regulondb version 7.0: transcriptional regulation of *Escherichia coli* k-12 integrated within genetic sensory response units (gensor units). *Nucleic Acids Res.,* **39**, D98–D105.

Geeven,G. *et al.* (2012) Identification of context-specific gene regulatory networks with gemula-gene expression modeling using lasso. *Bioinformatics,* **28**, 214–221.

Harbison,C.T. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature,* **431**, 99–104.

Ho,S.J. *et al.* (2006) Optimizing fuzzy neural networks, for tuning pid controllers using an orthogonal simulated annealing, algorithm osa. *IEEE Trans. Fuzzy Syst.,* **14**, 421–434.

Ho,S.Y. *et al.* (2004) Intelligent evolutionary algorithms for large parameter optimization problems. *IEEE Trans. Evolut. Comput.,* **8**, 522–541.

Ho,S.Y. *et al.* (2008) Opso: orthogonal particle swarm optimization and its application to task assignment problems. *IEEE Trans. Syst. Man Cybern. Part A,* **38**, 288–298.

Huynh-Thu,V.A. *et al.* (2010) Inferring regulatory networks from expression data using tree-based methods. *PLos One,* **5**, e12776.

Ishida,Y. *et al.* (2009) Participation of regulator ascg of the beta-glucoside utilization operon in regulation of the propionate catabolism operon. *J. Bacteriol.,* **191**, 6136–6144.

Jacklin,N. *et al.* (2012) Noniterative convex optimization methods for network component analysis. *IEEE/ACM Trans. Comput. Biol. Bioinf.,* **9**, 1472–1481.

Kao,K.C. *et al.* (2004) Transcriptome-based determination of multiple transcription regulator activities in *Escherichia coli* by using network component analysis. *Proc. Natl. Acad. Sci. USA,* **101**, 641–646.

Keseler,I.M. *et al.* (2013) Ecocyc: fusing model organism databases with systems biology. *Nucleic Acids Res.,* **41**, D605–D612.

Liao,J.C. *et al.* (2003) Network component analysis: reconstruction of regulatory signals in biological systems. *Proc. Natl. Acad. Sci. USA,* **100**, 15522–15527.

Marbach,D. *et al.* (2010) Revealing strengths and weaknesses of methods for gene network inference. *Proc. Natl. Acad. Sci. USA,* **107**, 6286–6291.

Marbach,D. *et al.* (2012) Wisdom of crowds for robust gene network inference. *Nat. Methods,* **9**, 796–804.

Margolin,A.A. *et al.* (2006) Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics,* 7(Suppl 1), S7.

Misra,A. and Sriram,G. (2013) Network component analysis provides quantitative insights on an arabidopsis transcription factor-gene regulatory network. *BMC Syst. Biol.,* **7**, 126.

Noor,A. *et al.* (2013) Robnca: robust network component analysis for recovering transcription factor activities. *Bioinformatics,* **29**, 2410–2418.

Pan,Y. *et al.* (2007) Connecting quantitative regulatory-network models to the genome. *Bioinformatics,* **23**, I367–I376.

Raghavan,R. *et al.* (2011) Genome-wide identification of transcription start sites yields a novel thermosensing rna and new cyclic amp receptor protein-regulated genes in *Escherichia coli*. *J. Bacteriol.,* **193**, 2871–2874.

Shao,L.Y. *et al.* (2012) Dynamic network of transcription and pathway crosstalk to reveal molecular mechanism of mgd-treated human lung cancer cells. *PLoS ONE,* **7**, e31984.

Tikhonov,A.N. and Arsenin,V.Y. (1977) *Solutions of ill-posed problems*. Winston, Washington, DC.

Tran,L.M. *et al.* (2005) gnca: a framework for determining transcription factor activity based on transcriptome: identifiability and numerical implementation. *Metab. Eng.,* **7**, 128–141.

Tran,L.M. *et al.* (2010) Trimming of mammalian transcriptional networks using network component analysis. *BMC Bioinformatics,* **11**, 511.

Tran,L.M. *et al.* (2012) Determining pten functional status by network component deduced transcription factor activities. *PLoS ONE,* **7**, e31053.

Wang,Y. *et al.* (2006) Inferring gene regulatory networks from multiple microarray datasets. *Bioinformatics,* **22**, 2413–2420.

Yang,Y.L. *et al.* (2005) Inferring yeast cell cycle regulators and interactions using transcription factor activities. *BMC Genomics,* **6**, 90.

Ye,C. *et al.* (2009) Using network component analysis to dissect regulatory networks mediated by transcription factors in yeast. *PLoS Comput. Biol.,* **5**, e1000311.

Zhang,X.J. *et al.* (2013) Narromi: a noise and redundancy reduction technique improves accuracy of gene regulatory network inference. *Bioinformatics,* **29**, 106–113.