Journal of Process Control 60 (2017) 24–33

Contents lists available at ScienceDirect

Journal of Process Control

journal homepage: www.elsevier.com/locate/jprocont

ISNCA: A new iterative approach for constrained matrix factorization methods☆

Naresh Doni Jayavelu[a,b,*], Nadav Bar[c,**]

[a] Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, WA 98195-7720, USA
[b] Turku Centre for Biotechnology, Turku, Finland
[c] Department of Chemical Engineering, Norwegian University of Science and Technology, Trondheim NO-7491, Norway

ARTICLE INFO

Article history:
Received 17 November 2016
Received in revised form 5 August 2017
Accepted 12 August 2017
Available online 31 August 2017

MSC:
00-01
99-00

Keywords:
Data analysis
High-dimensional data
Gene regulatory networks
Microarray
RNA-seq
MicroRNAs
Transcription factors
NCA
ISNCA

ABSTRACT

High-dimensional space of data is abundant in many fields, including medicine, machine learning, computer imaging, financial data, internet and data mining. These datasets usually suffer from large number of components but low sample sizes. One particular datasets are gene regulatory networks (GRNs) in systems biology. They are complex and involve thousands of components but they are seldom measured by more than a few dozens samples. High-dimensional analysis methods that attempt to extract hidden regulatory signals from such data are based on statistical models that often impose restrictions on a network topology and size. These restrictions often omit key components and therefore provide predictions that are less feasible from a biological perspective. To relax these restrictions, we developed iterative sub-network component analysis (ISNCA) that solves two or more sub-networks with joint components at one iteration and then updates solution at next iteration. It does so by subtracting the contribution of shared components from each sub-networks. Our approach of network division and update can analyze large networks that do not satisfy the restrictions of standard analysis algorithms, such as network component analysis. In this work, we generalized the ISNCA to include both target genes (TGs) and regulators, i.e. transcription factors (TFs) or microRNAs (miRNAs) as shared components and studied predictions of ISNCA to a new type of networks, miRNAs–TGs networks. Furthermore, we tested performance of the ISNCA with several new expression data obtained from different and independent platforms, and several new *a priori* knowledge databases. The generalized ISNCA can be used as a chassis to relax restrictions on network structure of other data analysis methods.

1. Introduction

Living organisms are composed of intrinsic elements such as DNA, RNA, proteins and cells. The complex interactions between these elements make it hard to understand their working principles and behaviors. Therefore, it is important to study a biological system in an holistic manner rather than by its individual elements. Systems biology is relatively new and interdisciplinary field developed to study and understand the complex behavior at the system level and it is achieved by integrating experimental data and known systems knowledge in an iterative manner [1].

Gene regulatory networks (GRNs) represent interactions between regulators (TFs, miRNAs) and TGs. TFs are proteins that bind to promoter regions of TGs and regulate their expression. TFs are known to have dual regulatory roles, i.e. a TF can either activate (as an activator) or repress (as a repressor) gene expression. MiRNAs are small (~22 nucleotides in length) non-coding RNA molecules that bind to TGs and repress gene expression. These regulators control gene expression in a spatiotemporal manner. These GRNs are key to understand biological systems and how cells function and express their output in response to signaling cues and perturbations [2], but they face the *curse of dimensionality*, the apparent intractability of systematically searching through a high-dimensional space and of accurately approximating a general high-dimensional function. These networks are complex in nature and contain thousands of components (TGs and their regulators TFs, miRNAs) with their inter-connections. Cellular response (in terms of expression levels of TGs) are obtained by high-throughput experimental platforms such as microarray and

☆ A preliminary version of this work has been presented in the 13th Symposium on Computer Application in Biotechnology (CAB) 2016, Trondheim, Norway.

* Corresponding author at: Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, WA 98195-7720, USA.
** Corresponding author.
E-mail addresses: ndjuw@uw.edu (N.D. Jayavelu), nadi.bar@ntnu.no (N. Bar).

http://dx.doi.org/10.1016/j.jprocont.2017.08.006

RNA-sequencing (RNA-seq), but the underlying GRNs are largely unknown. Therefore, computing the activity levels of TFs and miRNAs by decomposing expression data in order to construct the GRNs is a common practice in systems biology [3,4].

Several statistical models and computational methods were developed in the last two decades in order to solve this inverse problem but the majority of these suffer from imposed restrictions on network topology and size. Decomposition methods such as singular value decomposition (SVD) [5], principal component analysis (PCA) [6], independent component analysis (ICA) [7] and partial least squares regression [8] have been applied with assumptions of statistical independence and orthogonality on reconstructed regulatory signals. However, these mathematical assumptions are usually not valid in biological systems. In contrast, network component analysis (NCA) [9] combines *a priori* known biological knowledge in terms of TF–TG or miRNA–TG interactions with expression data for reconstruction. The NCA predicts the temporal activity of regulators (TFs or miRNAs) in GRNs using the following linear model:

$$E = AP + \epsilon \tag{1}$$

where, $E \in \mathbb{R}^{n\times m}$ is gene expression measurement data, $A \in \mathbb{R}^{n\times l}$ is initial connectivity data represented as adjacency matrix, defining how each of l regulators (TFs or miRNAs) and n TGs are connected in network. $P \in \mathbb{R}^{l\times m}$ is reconstructed activity levels of regulators (TFs or miRNAs). The index m is number of experimental conditions across which gene expression data is measured. The decomposition of E into A and P is achieved by performing a two step alternating least squares optimization problem under three criteria termed as NCA criteria: (a) The matrix A must be a full-column rank, (b) If a regulatory node is removed from A along with their connected TGs the resulting reduced matrix still should be a full-column rank and (c) Activity matrix P must be a full-row rank. If NCA criteria is satisfied, then the problem reduces to minimizing a bi-linear objective function, preserving the non-zero pattern of connectivity matrix.

$$||E - AP||_F^2 \text{ s.t. } A \in Z_0 \tag{2}$$

where $||\cdot||_F$ is Frobenius norm and Z_0 is the topology induced by the initial network.

The initial NCA algorithm suffered from unstable solutions resulting from ill-conditioned matrices and multiple local solutions. Subsequently, several algorithms were developed, mainly focusing on accurate, stable solutions and minimizing the computational complexity in terms of time. However, none of these focused on reconstructing larger networks and retaining key components of the system [10–14]. Moreover, these algorithms failed to capture well known biological features of redundancy and co-regulation (of TFs or miRNAs) [15]. Recently, we introduced the ISNCA algorithm which solved the NCA compliant sub-networks and predicted the solution to a larger NCA incompliant network, by using a predict-update strategy [16,17]. Briefly, ISNCA predicts the solution of the first sub-network and updates the shared components expression of a second sub-network. Then, ISNCA solves the second sub-network and updates the first sub-network, continuing in an iterative manner until the solution of combined network converges to a minimum threshold. In our previous work, we demonstrated ISNCA by including only TGs as shared components. We used small networks and real expression datasets obtained from microarray platform only.

In this work, we present generalization of the ISNCA to include both TGs and TFs as shared components. Then, we analyze the effect of network partition type on the ISNCA performance, i.e. only shared TGs, versus shared TGs and TFs. We also present an algorithm for network partition, that can modify the structure of large networks to be solved with the ISNCA approach. We also extend the application of ISNCA to analyze miRNA–TG networks, and to analyze expression datasets obtained also from RNA-seq technologies thus making it suitable to work with any kind of regulatory networks and technology. Finally, we compared the performance of ISNCA with standard NCA solvers in retaining key TGs of system using machine learning measures of precision and recall. The rest of the paper is organized as follows: In Section 2, we present the ISNCA algorithm along with the datasets used. The results are systematically presented in Section 3, which is followed by discussion and conclusions in Section 4.

2. Methods

2.1. Network component analysis

Network component analysis algorithms decompose gene expression data matrix into a weighted topology TF–TG (or miRNA–TG) matrix and the temporal profile matrix of the TFs (or miRNAs). The model can be represented in the matrix form as follows:

$$E = AP + \epsilon \tag{3}$$

where, $E \in \mathbb{R}^{n\times m}$ is gene expression measurement data, $A \in \mathbb{R}^{n\times l}$ is initial connectivity data represented as adjacency matrix, defining how each of l regulators (TFs or miRNAs) and n TGs are connected in network. $P \in \mathbb{R}^{l\times m}$ is reconstructed activity levels of regulators (TFs or miRNAs). The index m is number of experimental conditions across which gene expression data is measured. The decomposition of E into A and P is achieved by performing a two step alternating least squares optimization problem under three criteria termed as NCA criteria: (a) The matrix A must be a full-column rank (b) If a regulatory node is removed from A along with their connected TGs the resulting reduced matrix still should be a full-column rank (c) Activity matrix P must be a full-row rank. If NCA criteria is satisfied, then the problem reduces to minimizing the bi-linear objective function preserving the non-zero pattern of connectivity matrix.

$$||E - AP||_F^2 \text{ s.t. } A \in Z_0 \tag{4}$$

where $||\cdot||_F$ is Frobenius norm and Z_0 is the topology induced by initial network.

2.2. Iterative sub-network component analysis

In ISNCA approach, we first divide the NCA incompliant network (network that does not satisfy NCA criteria) into two NCA compliant sub-networks (networks that satisfy NCA criteria), i ($i \in 1, 2$) with shared components either only TGs or both TGs and TFs (we present a pseudo algorithm later in this section). Let the subscripts u and c denote the unique and common components of each sub-network. Then, the matrices E, A and P in Eq. (1) for each sub-network i can be divided by the following:

$$E_1 = A_1P_1 = \begin{bmatrix} E_{u1} \\ E_c \end{bmatrix} = \begin{bmatrix} A_{u1u1} & A_{u1c} \\ A_{cu1} & A_{cc} \end{bmatrix} \begin{bmatrix} P_{u1} \\ P_c \end{bmatrix} \tag{5a}$$

$$E_2 = A_2P_2 = \begin{bmatrix} E_{u2} \\ E_c \end{bmatrix} = \begin{bmatrix} A_{u2u2} & A_{u2c} \\ A_{cu2} & A_{cc} \end{bmatrix} \begin{bmatrix} P_{u2} \\ P_c \end{bmatrix} \tag{5b}$$

where $E_{ui} \in \mathbb{R}^{n_{ui}\times m}$ and $E_c \in \mathbb{R}^{n_c\times m}$ denote the expression matrices of sub-networks $i = 1, 2$, and $n_{u1} + n_{u2} + n_c = n$. $A_{uiui} \in \mathbb{R}^{n_{ui}} \times l_{ui}$, $A_{uic} \in \mathbb{R}^{n_{ui}} \times l_c$, $A_{cui} \in \mathbb{R}^{n_c} \times l_{ui}$ and $A_{cc} \in \mathbb{R}^{n_c} \times l_c$ denote partition matrices of A of sub-networks, i and $l_{u1} + l_{u2} + l_c = l$. In all the following, when we write A_i, E_i or P_i, we refer to matrices of the entire sub-network i, including both its unique and common components.

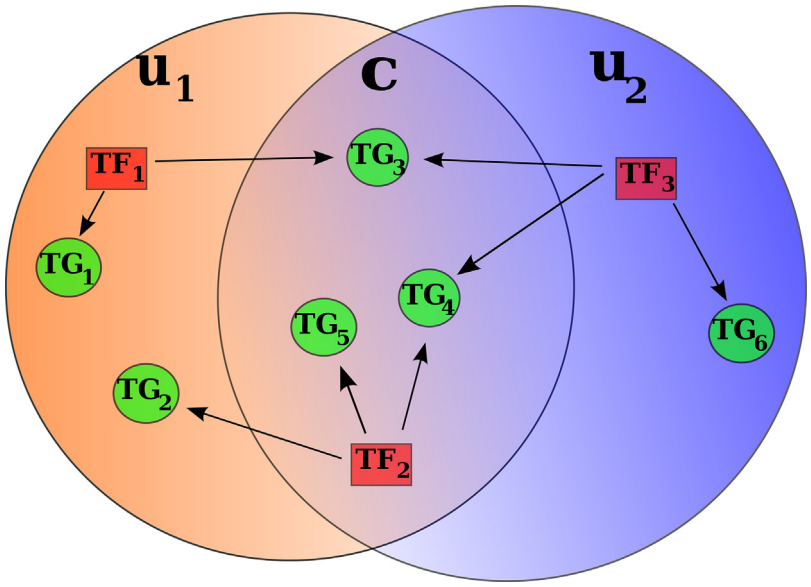

Fig. 1. Network division in example 1, its unique sub-networks components u_1 (left), u_2 (right) and common components c (middle). TG_i (green circles) and TF_i denote target genes and transcription factors. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The two sub-networks, i ($i \in 1, 2$) can be united to one joined network representation:

$$E = AP = \begin{bmatrix} E_{u1} \\ E_c \\ E_{u2} \end{bmatrix} = \begin{bmatrix} A_{u1u1} & A_{u1c} & \mathbf{O}_2 \\ A_{cu1} & A_{cc} & A_{cu2} \\ \mathbf{O}_1 & A_{u2c} & A_{u2u2} \end{bmatrix} \begin{bmatrix} P_{u1} \\ P_c \\ P_{u2} \end{bmatrix} \tag{6}$$

The matrices $\mathbf{O}_1 \in \mathbb{R}^{n_{u2} \times l_{u1}}$ and $\mathbf{O}_2 \in \mathbb{R}^{n_{u1} \times l_{u2}}$ denote zero matrices.

In case only the TGs are shared, Eq. (6) can be represented by the following:

$$E = AP = \begin{bmatrix} E_{u1} \\ E_c \\ E_{u2} \end{bmatrix} = \begin{bmatrix} A_{u1u1} & \mathbf{O}_2 \\ A_{cu1} & A_{cu2} \\ \mathbf{O}_1 & A_{u2u2} \end{bmatrix} \begin{bmatrix} P_{u1} \\ P_{u2} \end{bmatrix} \tag{7}$$

Example 1. Network division to sub-networks: Consider the example network we presented in Fig. 1. We can divide its connectivity matrix A to unique and common components as:

$$A = \begin{bmatrix} A_{u1u1} & A_{u1c} & \mathbf{O}_2 \\ A_{cu1} & A_{cc} & A_{cu2} \\ \mathbf{O}_1 & A_{u2c} & A_{u2u2} \end{bmatrix} = \begin{array}{c} \\ TG_1 \\ TG_2 \\ TG_3 \\ TG_4 \\ TG_5 \\ TG_6 \end{array} \left[\begin{array}{c|c|c} TF_1 & TF_2 & TF_3 \\ \hline 1 & 0 & 0 \\ 0 & 1 & 0 \\ \hline 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \\ \hline 0 & 0 & 1 \end{array}\right] \tag{8}$$

2.3. ISNCA algorithm

At the first iteration ($k = 0$) of the ISNCA algorithm, the network is divided as follows: The known expression matrix E is divided to initial $E_i(0)$, $i = 1, 2$ and network matrix, A to A_i according to Eq. (5). Note that the zero elements of A and A_i are known by initial connectivity data, and the non-zero elements are randomly chosen. At the start of each iteration k, we compute a unique solution to $\|E_i(k) - A_i P_i\|$, separately for sub-networks 1 and 2 using any standard NCA algorithm, and obtain $\hat{A}_i(k)$ and $\hat{P}_i(k)$. Then we construct $\hat{A}(k)$ and $\hat{P}(k)$ as described in Eqs. (6) and (7), as

$$\hat{A}(k) = \begin{bmatrix} \hat{A}_{u1u1} & \hat{A}_{u1c} & \mathbf{O}_2 \\ \hat{A}_{cu1} & \hat{A}_{cc} & \hat{A}_{cu2} \\ \mathbf{O}_1 & \hat{A}_{u2c} & \hat{A}_{u2u2} \end{bmatrix}, \ \hat{P}(k) = \begin{bmatrix} \hat{P}_{u1} \\ \hat{P}_c \\ \hat{P}_{u2} \end{bmatrix} \tag{9}$$

for both TGs and TFs shared case, and

$$\hat{A}(k) = \begin{bmatrix} \hat{A}_{u1u1} & \mathbf{O}_2 \\ \hat{A}_{cu1} & \hat{A}_{cu2} \\ \mathbf{O}_1 & \hat{A}_{u2u2} \end{bmatrix}, \ \hat{P}(k) = \begin{bmatrix} \hat{P}_{u1}(k) \\ \hat{P}_{u2}(k) \end{bmatrix} \tag{10}$$

for the TGs shared case. Then, we compute the error of the entire network at iteration k,

$$e(k) = ||E - \hat{A}\hat{P}||_F \tag{11}$$

where E is the expression matrix (initial data matrix). If the error is sufficiently small, for instance by

$$e(k+1) - e(k) < \epsilon \tag{12}$$

then we exit by the ISNCA algorithm with $\hat{A}$ and $\hat{P}$. In our simulations, we set ϵ to be 10^{-5} and maximum number of iterations to 100.

If the error does not converge, we proceed to update the sub-networks: Let $T_i(k)$ be the common TGs contribution from sub-network i, that is,

$$T_1(k) = \begin{bmatrix} \hat{A}_{cu1} & \hat{A}_{cc} \end{bmatrix} \begin{bmatrix} \hat{P}_{u1} \\ \hat{P}_c \end{bmatrix} \tag{13a}$$

$$T_2(k) = \begin{bmatrix} \hat{A}_{cu2} & \hat{A}_{cc} \end{bmatrix} \begin{bmatrix} \hat{P}_{u2} \\ \hat{P}_c \end{bmatrix} \tag{13b}$$

for both TGs and TFs shared case, whereas for the only TGs shared case it is

$$T_1(k) = \hat{A}_{cu1}\hat{P}_{u1} \tag{14a}$$

$$T_2(k) = \hat{A}_{cu2}\hat{P}_{u2} \tag{14b}$$

Note that $\hat{A}_{cc}$ is computed from both sub-networks, so we take the mean value. We then update the matrices E_1 and E_2 for the next iteration $k+1$ from Eq. (15) by subtracting the common TGs contribution from the other sub-network:

$$E_1(k+1) = \begin{bmatrix} E_{u1} \\ E_c - \delta \cdot T_2(k) \end{bmatrix} \tag{15a}$$

$$E_2(k+1) = \begin{bmatrix} E_{u2} \\ E_c - \delta \cdot T_1(k) \end{bmatrix} \tag{15b}$$

where, $\delta \in [0, 1]$ denotes the damping coefficient. Note that E_c and E_{ui} do not change from iteration to iteration as they represent the original expression matrices. We then advance to the next iteration and predict the solution to the expression $\|E_i(k) - A_i P_i\|$ using any standard NCA algorithms.

2.3.1. Pseudo algorithm: generalized ISNCA

Algorithm 1. ISNCA.

input : Network with topology A, measurement E

INIT divide A to two overlapping, NCA compliant, sub-networks with shared component(s) ($E_1(0)$, $A_1(0)$ and $E_2(0)$, $A_2(0)$) using algorithm 2;
SET $e(0) \leftarrow 0$, $T_i(0) \leftarrow 0$, choose δ, ϵ;
for *K:=1 to number of iterations* **do**
  Predict:;
  CALL standard NCA solver to obtain ($\hat{A}_i(k)$, $\hat{P}_i(k)$) $\longleftarrow NCA(E_i(k-1), A_i(k-1))$; (eq. 1)
  CALCULATE common TGs contribution, $T_1(k)$ and $T_2(k)$ (either eq. 13 or eq. 14);
  Exit condition:;
  DETERMINE $\hat{A}(k)$, $\hat{P}(k)$ (either eq. 9 or eq. 10);
  CALCULATE error, $e(k)$ (eq. 11);
  if $\|e(k) - e(k-1)\| < \epsilon$ **then**
    Exit the algorithm with $\hat{A}(k)$, $\hat{P}(k)$;
  end
  Update:;
  CALCULATE $E_1(k)$ and $E_2(k)$ (eq. 15);
end
output: $\hat{A}(k)$, $\hat{P}(k)$

Remark 1. The network division must contain shared components, and each of the sub-network must contain the mutual exclusive components (Fig. 1).

Remark 2. The two sub-networks that were partitioned at the initiation step must each separately satisfy the NCA criteria.

Remark 3. Only the sub-networks expression matrices $E_i(k)$ are passed on to the next iteration by Eq. (15). The matrix E of the entire network used in Eq. (11) does not change at any step, whereas A_i, P_i, $\hat{A}_i(k)$ and $\hat{P}_i(k)$ are re-computed at each iteration.

2.3.2. Pseudo algorithm: network division to sub-networks

Here, we present the work flow to divide the initial NCA incompliant network into two overlapping NCA compliant sub-networks (Algorithm 2). First, we find the largest possible NCA compliant sub-network. Then we find the second sub-network by excluding only the TFs or miRNAs (regulators) and not TGs present in the first sub-network. Once the two NCA compliant sub-networks are identified it is easy to define the unique and shared components for each sub-network, and we can feed these to the INIT section of Algorithm 1.

Algorithm 2. Network division.

input : initial network with topology A_0, measurement E_0
START: (N,L,M) $\longleftarrow$ CALCULATE size of A_0 and E_0
CALCULATE *row sum* of A_0
$A^1 \longleftarrow$ remove rows of A_0, IF *row sum* > M (checking NCA criteria III)
CALCULATE *column sum* of A^1
$A^2 \longleftarrow$ remove columns of A^1, IF *column sum* < 3
DETERMINE dependent columns of A^2
$A^3 \longleftarrow$ remove dependent columns of A^2, IF any (checking NCA criteria I)
SET *flag* $\leftarrow 0$ and $A_{sub1} \leftarrow A^3$ (checking NCA criteria II)
while *flag* $\neq$ *1* **do**
  DETERMINE ranks of reduced matrices, Ar_i for A_{sub1}
  L $\longleftarrow$ number of columns in A_{sub1}
  R $\longleftarrow$ cumulative sum of ranks of Ar_i
  if *R == L(L-1)* **then**
    flag $\leftarrow 1$
    $A_{sub1} \leftarrow A_{sub1}$
  end
  else
    flag $\leftarrow 0$
    badTFs $\longleftarrow$ IF any $Ar_i <$ L-1
    $A_{sub1} \longleftarrow$ remove bad columns (TFs) of A_{sub1}
  end
end
(E_{sub1}, A_{sub1}) $\longleftarrow$ match TGs in E_0 and A_{sub1}
DEFINE new A_0 and E_0 for finding second NCA compliant sub-network
(E_0, A_0) $\longleftarrow$ remove TFs present in A_{sub1} from A_0
go to START

2.4. NCA solvers used

We used FastNCA and ROBNCA solvers inside ISNCA algorithm for solving individual sub-networks. Both solvers are non-iterative and their computational load is minimal. Although FastNCA is faster than ROBNCA, it is limited to solving smaller networks. The accuracy of FastNCA depends on an accurate estimation of signal sub-space of data. The ROBNCA explicitly models the outliers of the expression data and is therefore more suitable for noisy data and larger networks. The complete details of both solvers and other NCA solvers are available elsewhere [18,11,14].

2.5. Computation of precision–recall measures

We downloaded the list of TGs that are known to be involved in breast cancer from the cancer gene database [19] and it serves as a breast cancer reference gene set. Next, we generated 500 miRNA–TG networks of different sizes by randomly choosing TGs and reconstructed networks using ISNCA and NCA. Then, we computed precision and recall measures as defined below for each reconstructed network by ISNCA and NCA. Higher values of precision and recall suggests better performance of the algorithm.

$$Precision = \frac{\#(\text{TGs in network} \cap \text{Breast cancer TGs reference set})}{\#(\text{TGs in network})}$$

$$Recall = \frac{\#(\text{TGs in network} \cap \text{Breast cancer TGs reference set})}{\#(\text{Breast cancer TGs reference set})}$$

where $\#(\cdot)$ denotes number of TGs in the set and $\cap$ denotes intersection between sets.

2.6. Expression datasets used

2.6.1. Synthetic expression data for case study

We generated 100 different expression matrices E of size 15×10 (TGs, $n = 15$ and measurement points $m = 10$) for each case study. We used Gaussian distribution to generate random elements of E (including both positive and negative values). We used Matlab function "*randn*" for this purpose. The initial connectivity matrices A of size 15×10 (TGs, $n = 15$ and TFs, $l = 10$) are also generated in similar manner while preserving its connectivity structure. Then the TF activity matrices were computed using ISNCA approach and the calculated error e of the reconstruction were calculated using Eq. (11).

2.6.2. Human breast cancer cells microarray data

We downloaded the microarray data of human breast cancer cells treated with heregulin (HRG) at 12 time points from the GEO database with accession number: GSE13009 [20]. We applied Loess normalization within time points and quantile normalization across time points. The expression values were averaged over replicate measurements. The differentially expressed genes (DEGs) with fold change (FC) >1.5 and P-value < 0.05 at minimum 2 time points were selected for further analysis.

2.6.3. Mouse T cells microarray data

The microarray data of mouse T cells treated with interleukin-2 (IL-2) at 10 time points over a period of 24 h was downloaded from GEO database with accession number: GSE6085 [21]. The data was processed using limma package in R/Bioconductor [22]. The expression values were averaged over replicated measurements. The DEGs were identified with fold change (FC) >1.5 and adjusted P-value < 0.05 at minimum two time points.

2.6.4. Human developmental stages RNA-seq data

The single cell RNA-seq data for human developmental stages is acquired from Yan et al. [23]. We conducted a statistical test for each developmental stage to find the TGs that have a differential expression compared to the remaining developmental stages. For this purpose we used a moderated F-test which was implemented in the limma package in R/Bioconductor, comparing each stage against the remaining stages [22]. The DEGs at each developmental stage were selected with fold change (FC) >2 and adjusted P-value < 0.05.

2.7. Network datasets used

2.7.1. miR–TG network data

We downloaded a manually curated and experimentally verified miR–TG interaction data from miRTarBase database [24] as an adjacency list, and then converted to an adjacency matrix to use with ISNCA and NCA algorithms.

2.7.2. TF–TG network data

We obtained the TF–TG interaction data from TFactS [25], HTRIdb [26] and TRRUST [27] databases as adjacency lists. We filtered for only experimentally verified and manually curated interactions. We found that few entries (TF–TG) were present in more than one database. Therefore, we made a union of three data sources, removed duplicated entries (present in 2 or 3 sources) and retained only unique TF–TG entries.

3. Results

We generalized the ISNCA approach in which the sub-networks are shared with both TGs and TFs (or miRNAs, see Methods). We then presented an approach to divide the initial NCA incompliant network to NCA compliant sub-networks by providing a pseudo algorithm (see Methods).

3.1. Effect of network division on ISNCA performance

We tested how network division to sub-networks with both TGs and TFs as shared components versus the case in which only the TGs are shared affects the ISNCA approach, in terms of accuracy and the size of the reconstructed networks. For this purpose, we constructed a synthetic network consisting of 10 TFs, 15 TGs and 36 interactions (Fig. 2A). This network is not satisfying the NCA criteria (NCA incompliant network) and therefore does not have a unique solution with standard NCA algorithms. We solved this network with the ISNCA approach by dividing it into two NCA compliant sub-networks (each satisfies the three NCA criteria). We divided this initial network to two sub-networks in two different manners: (1) only TGs are shared between sub-networks (Case 1, Fig. 2B), and (2) both TGs and TFs are shared between the sub-networks (Case 2, Fig. 2C). Then we examined the network reconstruction accuracy (e using Eq. (11)) of the ISNCA approach, layered with two NCA solvers, the FastNCA and the ROBNCA. The reason we used two NCA solvers was to find the best solver for the ISNCA. For this purpose, we generated 100 random expression matrices, E of size, 15×10 (TGs, $n = 15$ and measurement points $m = 10$) and with initial connectivity matrix, A for two cases (Fig. 2B and C), we computed the error of reconstruction for the entire network, e (see Methods). The errors of these 100 expression matrices for the network division in Case 1 were distributed over a small range, compared to the network division presented in Case 2, when the ISNCA was layered with FastNCA. Although the mean error over 100 simulations was smaller for Case 1 than for Case 2, this difference was not statistically significant ($P = 0.138$, Wilcoxon rank sum test) (Fig. 2D). However, when we used ISNCA layered with ROBNCA, the mean error for Case 1 was significantly ($P = 0.0036$, Wilcoxon rank sum test) smaller than for Case 2. The accuracy of the FastNCA is dependent on an accurate estimation of signal subspace of data and is therefore more suitable for smaller networks (see Methods). The ROBNCA explicitly models the outliers and is therefore more suitable for noisy expression data, and for larger networks. The error convergence rates were smooth for both network division cases, and for both solver (Fig. 2F and G). The initial error convergence rate was slightly sharp at first two iterations for Case 1 than for Case 2, but converged for both cases to the same value in next few iterations. Although we found that both network partitions performed similarly, convergence of Case 1 was better with the ROBNCA solver. Therefore, we included only TGs in shared components between sub-networks in subsequent analyses.

3.2. Performance of ISNCA on miRNA–TG networks

Next, we present an extended application of ISNCA to a real case study on microarray expression data obtained from human breast cancer cells and miRNA–TG networks. The miRNA–TG network consists of 20 miRNAs, 47 TGs and 60 interactions (Fig. 3A). The ISNCA algorithm solved the whole network by dividing it into two smaller sub-networks. In comparison, traditional NCA algorithms we tested could only reconstruct a network size of 14 miRNAs, 39 TGs and 42 interactions, since they pruned 6 miRNAs and their connected TGs (Fig. 3B). However, these pruned miRNAs were shown to have key roles in breast cancer. The temporal activities of the miRNA predicted by the ISNCA were scaled and sorted in order to reveal the sequential activation of miRNAs (Fig. 3C). The miRNAs miR-302c and miR-1181 were activated early 15 and 30 min after the treatment with interleukin-2 (IL-2). Several miRNAs, including miR-100 and miR-130b were activated after 1 h, whereas miR-141 and miR-320a were activated 6–8 h after the IL-2 treatment. The miR-105 and miR-449a were activated late, after 12–24 h. We found that these miRNAs are key regulators, and are vital for the system dynamics (Table 1). Furthermore, ISNCA predicted the actual connectivity strength of each miRNA to their connected TGs (Fig. 3D).

3.3. Network size comparison reconstructed by ISNCA approach and NCA algorithms

We also focused on network sizes reconstructed by ISNCA and compared to those reconstructed by traditional NCA algorithms. For this purpose, we generated 100 initial networks, randomly chosen by a set of TGs and their connected regulators (TFs or miRNAs). The network sizes in terms of the number of miRNAs and TGs reconstructed by ISNCA and NCA are summarized in Fig. 4A and

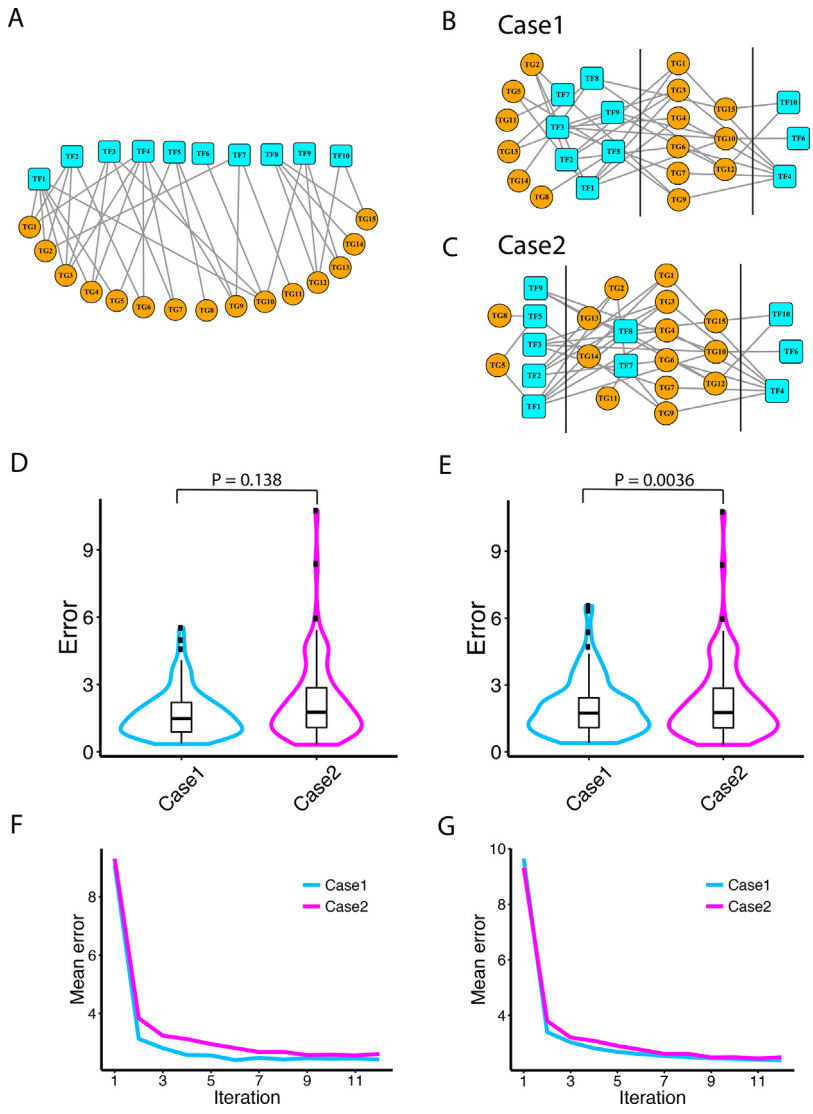

Fig. 2. Network division comparison. (A) Initial NCA incompliant network. (B) Division of network to sub-networks with only TGs shared (Case 1). Solid vertical lines partition the sub-network's unique and common components. (C) Division of network to sub-networks with both TGs and TFs shared (Case 2). Solid vertical lines partition the sub-network's unique and common components. (D) Comparison of error distributions of Case 1 and Case 2 networks solved with ISNCA with FastNCA as solver. *P*-values for statistical significance are computed using Wilcoxon rank sum tests. (E) Comparison of error distributions of Case 1 and Case 2 networks solved with ISNCA with ROBNCA as solver. *P*-values for statistical significance are computed using Wilcoxon rank sum tests. (F) Error convergence of Case 1 and Case 2 networks with each iteration solved with ISNCA with FastNCA as solver. (G) Error convergence of Case 1 and Case 2 networks with each iteration solved with ISNCA with ROBNCA as solver.

Table 1
The list of key miRNAs (all reported to be associated with breast cancer), that were pruned by the NCA algorithm but retained by the ISNCA. % ISNCA denotes number of times a particular miRNA was retained by the ISNCA and % NCA denotes number of times it was retained by the NCA algorithm.

miR name	% ISNCA	% NCA	Reference
miR-299	72	0	[28]
miR-15b	70	0	[29]
miR-625	70	0	[30]
miR-302d	68	0	[31]
miR-520a	68	0	[32]
miR-302c	67	0	[31]
miR-448	67	0	[33]
miR-130b	66	0	[34]
miR-18b	66	0	[35]
miR-19b	65	0	[36]
miR-106b	64	0	[37]
miR-503	35	3	[38]
miR-149	34	4	[39]
miR-218	34	3	[40]

B, respectively. The largest possible network that was modeled by the ISNCA consisted of 185 miRNAs, 403 TGs and 830 interactions, whereas the NCA reconstructed a network no larger than 115 miRNAs, 318 TGs and 546 interactions. On average, ISNCA solved 74%, 73% and 81% larger networks than the NCA, in terms of miRNAs, TGs and interactions respectively on 100 tested networks. Moreover, these differences were significant ($P < 10^{-8}$) and were not observed by random networks.

We are interested to find whether the ISNCA feature of reconstructing larger networks was dependent on the specific expression data and connectivity datasets or it was a robust feature valid across any type of datasets. To validate this, we analyzed additional two expression datasets (1) microarray data of mouse T-cells, (2) RNA-seq data of human developmental stages. We repeated the similar analysis of comparing network sizes reconstructed by ISNCA and NCA on 100 randomly selected networks from each dataset. We used the TF–TG interaction data from literature for initial network construction. The ISNCA consistently showed superior

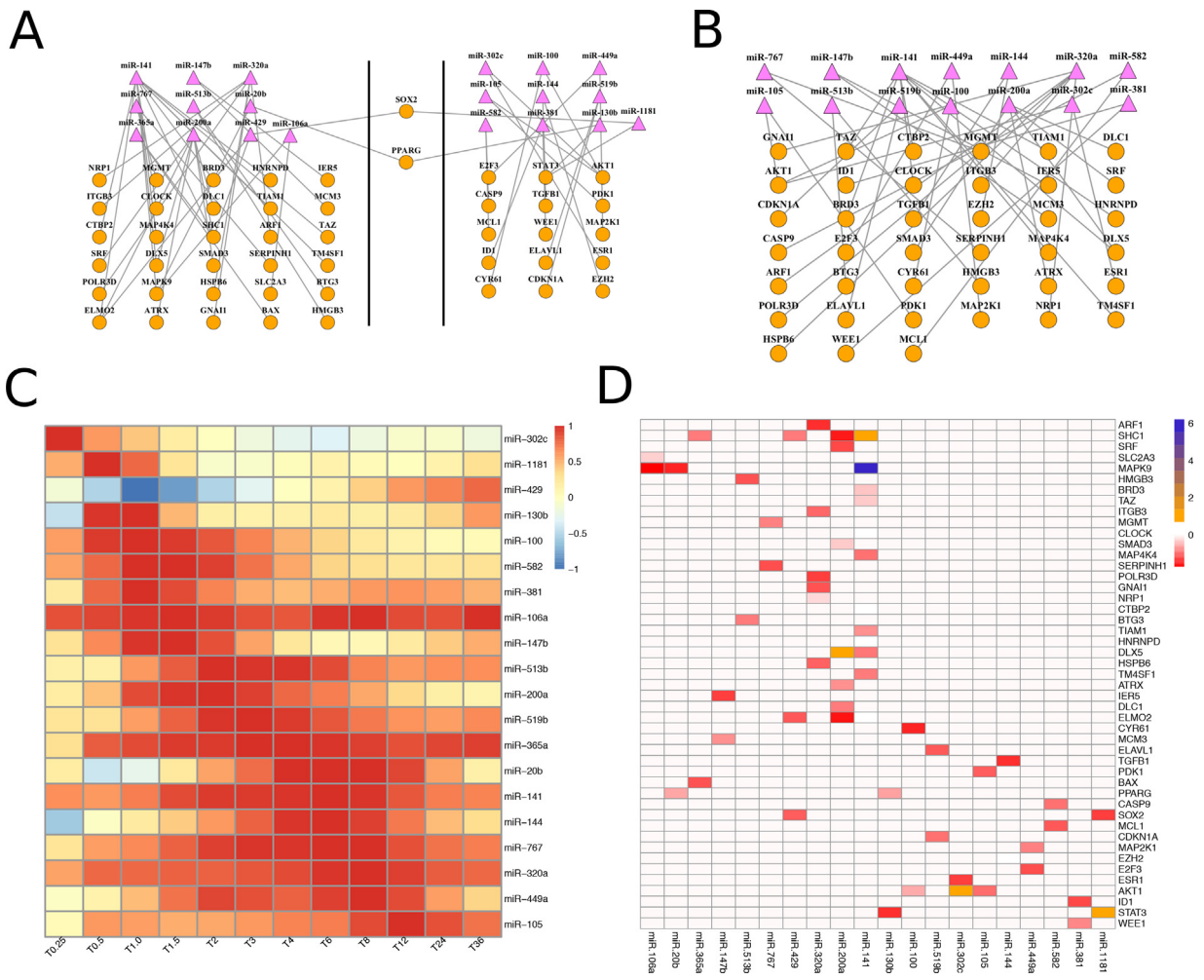

Fig. 3. miRNA–TG network. (A) ISNCA network. Solid vertical lines partition the sub-network's unique and common components. (B) NCA network. (C) Heatmap of computed miRNA temporal activities by ISNCA. (D) Heatmap of predicted connectivity strength matrix by ISNCA.

performance to the traditional NCA methods (Fig. 4C–F) irrespective of the regulatory networks and expression data.

3.4. Performance of ISNCA in retaining key components of system

Next, we tested the performance of ISNCA in retaining key network components. For this purpose, we downloaded a list of TGs, known to be involved in breast cancer [19]. Then, compared the two well-known machine learning performance measures, precision and recall for ISNCA and NCA (see Methods). The precision measure provides fraction of relevant TGs (breast cancer) among the retained TGs in a network, and the recall provides the fraction of relevant TGs that were retained over the total relevant TGs. We generated 500 miRNA–TG networks of different sizes by randomly choosing TGs and reconstructed networks using ISNCA and NCA. Then, we computed precision and recall measures for each reconstructed network by ISNCA and NCA (Fig. 5). Higher values of precision and recall suggest better performance of the algorithm. The majority of the networks reconstructed by ISNCA exhibited significantly ($P < 2.2e^{-16}$) higher precision values than by the NCA. Recall that values were also higher for ISNCA reconstructed networks than for the NCA ones for all the tested networks ($n = 500$). This analysis demonstrated superior performance of ISNCA in not only reconstructing larger networks, but also in retaining key components of the system.

We then tested and compared the ability of ISNCA and NCA to retain key miRNAs, which are essential to understand the cellular dynamics of breast cancers. Because the information related to miRNAs which are involved in breast cancer was limited, we were unable to compute the machine learning precision and recall measures (that require large datasets). Instead, we simulated 100 random networks and compared the results to manually curated data from the literature. Table 1 lists the key miRNAs in our network that are known to be relevant for breast cancer (manually curated, their association in breast cancer is described in mentioned references). Over 78% (11/14) of the key miRNAs that are known to be associated with breast cancer were retained by at least 64% of the random networks we tested by ISNCA. Yet those were never retained by the standard NCA, demonstrating loss of important information. This suggests that ISNCA approach is able to capture key components of the miRNA regulatory network that standard NCA algorithms fail to retain, because they violated the NCA criteria.

4. Discussion and conclusions

Previously [17] we showed that the ISNCA converges with real measured biological data consisting of TFs and TGs, with topological information obtained from TFacts database [25]. Here, we extended this analysis in several directions. (i) We used new biological measurements and additional independent TF–TG topological

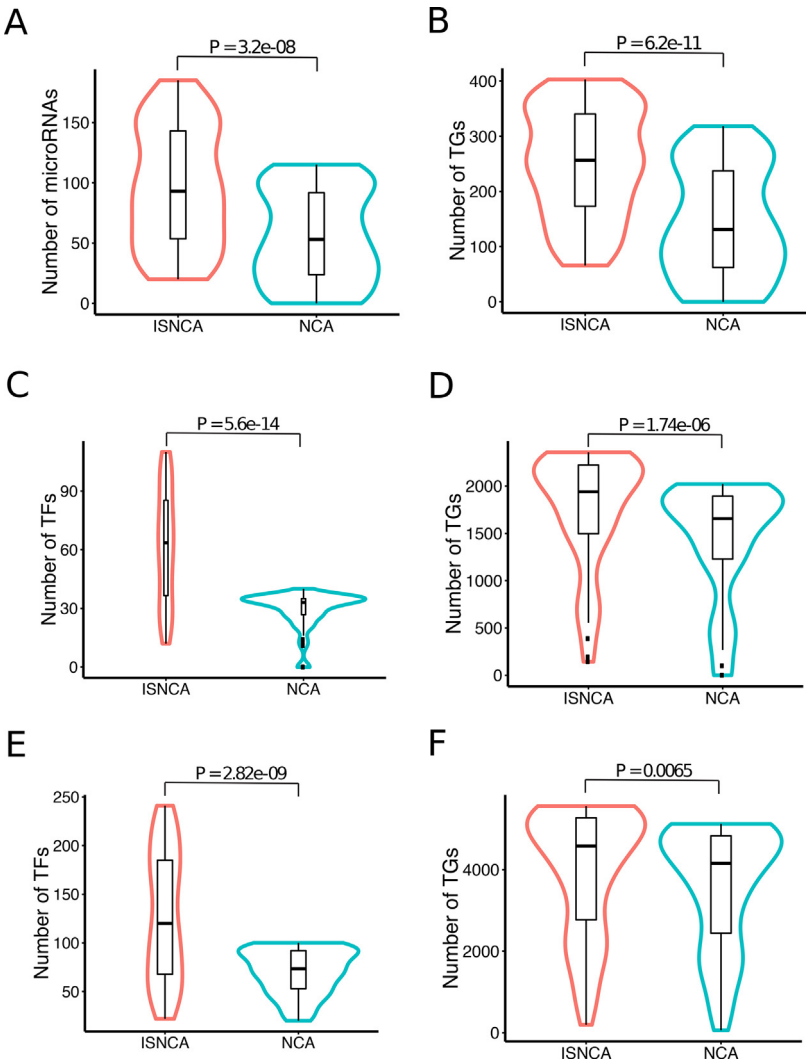

Fig. 4. Network size comparisons on more datasets. Distribution of number of miRNAs (A) and TGs (B) reconstructed by ISNCA and NCA on 100 tested expression datasets obtained from breast cancer cells microarray data. Distribution of number of TFs (C) and TGs (D) reconstructed by ISNCA and NCA on 100 tested expression datasets obtained from mouse T cells microarray data. Distribution of number of TFs (E) and TGs (F) reconstructed by ISNCA and NCA on 100 tested expression datasets obtained from human developmental RNA-seq data. *P*-values for statistical significance are computed using Wilcoxon rank sum tests.

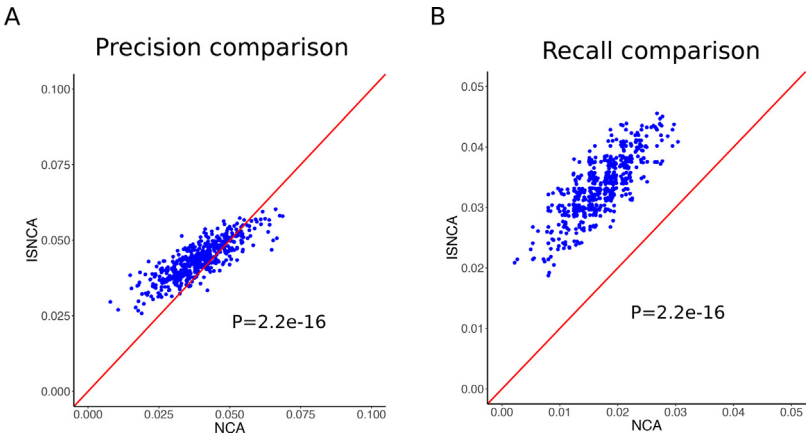

Fig. 5. Precision–recall measures. Comparison of precision (A) and recall (B) measures for ISNCA and NCA. *P*-values for statistical significance are computed using Wilcoxon rank sum tests.

information obtained from HTRIdb and TRRUST, and showed that the ISNCA predicts accurately also in different *a priori* studied databases. (ii) More importantly, we tested whether the ISNCA manages to preserve information and enlarge the size of a new, independent biological networks consisting of microRNAs (not TFs). This regulatory network is a different form of regulation than

we previously tested [16,17]. (iii) Since information on the involvement of few miRNAs in breast cancer is already known, we could verify our predictions and evaluate the performance of ISNCA more accurately than before. ISNCA retained miRNAs miR-299, miR-15b and miR-625 over 70 times of 100 tested networks. These miRNAs were shown to be key regulators of breast cancer. The miR-299 was shown to be a potential biomarker for detection and classification of breast cancer stages [28]. The miR-15b controls the migration and invasion of breast cancer cells by regulating the target gene MTSS1 [41]. The miR-625 suppresses cell proliferation and migration by regulating HMGA1 and proven to be a promising prognostic biomarker and a potential therapeutic target for breast cancer [30]. Overall, we showed that the ISNCA retained all the key components we tested, in contrast to the traditional network analysis methods (NCA, FastNCA, ROBNCA), that removed these components in order to satisfy the factorization conditions. (iv) The ISNCA consistently improved the predictions over traditional NCA methods in all our case studies, demonstrating that the ISNCA is not restricted to specific datasets, measurements, network type or to a specific curated *a priori* known information. (v) The ISNCA exhibited good performance in terms of well-known machine learning measures of precision and recall in retaining potentially important breast cancer TGs.

In the context of high dimensional data, ISNCA can serve as a platform to relax restrictions set by network structure matrices, and thus be able to reconstruct larger networks. Here we demonstrated that the restrictions on A (Eq. (1)) could be relaxed by solving iteratively smaller network sizes that do not violate the rank conditions (in A) of the NCA [9]. This important feature of the ISNCA not only retains key network components, but also better exploits the scarcely available measurements, and extract more information on temporal regulatory behavior of the network. We demonstrated the ISNCA platform on FastNCA and ROBNCA, but we propose that the ISNCA can be employed for other methods than NCA based, extending the big data analysis to other network analysis methods with restrictions on the network topology, such as partial least squares regression methods or functional PCA. The ISNCA, coupled with locally optimal network division approach (e.g. Branch and bounds) can serve as a powerful tool for high dimensional data analysis.

The identification and characterization of cis-regulatory elements (putative enhancers and silencers) have been reported in many human and mouse cell types [42]. These regulatory elements also control the gene expression. Recently, few research groups have developed the experimental technology to study 3D genome architecture, i.e. finding the putative cis-regulatory elements interactions with promoters of TGs [43]. When these datasets will be available in sufficient amounts, we can predict the activity of these cis-regulatory elements and study their role in regulating the gene expression. Thus, the proposed ISNCA approach will serve as unique universal tool to study all kinds of gene regulatory networks.

Acknowledgements

The authors would like to thank L. Aasgaard for his effort during the early stages of the method development.

References

[1] H. Kitano, Computational systems biology, Nature 420 (2002) 206–210.
[2] L.T. MacNeil, A.J.M. Walhout, Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression, Genome Res. 21 (2011) 645–657.
[3] M.L. Arrieta-Ortiz, C. Hafemeister, A.R. Bate, T. Chu, A. Greenfield, B. Shuster, S.N. Barry, M. Gallitto, B. Liu, T. Kacmarczyk, F. Santoriello, J. Chen, C.D. Rodrigues, T. Sato, D.Z. Rudner, A. Driks, R. Bonneau, P. Eichenberger, An experimentally supported model of the *Bacillus subtilis* global transcriptional regulatory network, Mol. Syst. Biol. 11 (2015).
[4] T.A. Long, S.M. Brady, P.N. Benfey, Systems approaches to identifying gene regulatory networks in plants, Annu. Rev. Cell Dev. Biol. 24 (2008) 81–103.
[5] M.E. Wall, P.A. Dyck, T.S. Brettin, SVDMAN – singular value decomposition analysis of microarray data, Bioinformatics 17 (2001) 566–568.
[6] S. Raychaudhuri, J.M. Stuart, R.B. Altman, R.B. Altman, Principal components analysis to summarize microarray experiments: application to sporulation time series, Pac. Symp. Biocomput. (2000) 452–463.
[7] W. Liebermeister, Linear modes of gene expression determined by independent component analysis, Bioinformatics 18 (2002) 51–60.
[8] A.L. Boulesteix, K. Strimmer, Predicting transcription factor activities from combined analysis of microarray and chip data: a partial least squares approach, Theor. Biol. Med. Model. 2 (2005) 23.
[9] J.C. Liao, R. Boscolo, Y.L. Yang, L.M. Tran, C. Sabatti, V.P. Roychowdhury, Network component analysis: reconstruction of regulatory signals in biological systems, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 15522–15527.
[10] S.J. Galbraith, L.M. Tran, J.C. Liao, Transcriptome network component analysis with limited microarray data, Bioinformatics 22 (2006) 1886–1894.
[11] C. Chang, Z. Ding, Y.S. Hung, P.C. Fung, Fast network component analysis (fastNCA) for gene regulatory network reconstruction from microarray data, Bioinformatics 24 (2008) 1349–1358.
[12] L.M. Tran, M.P. Brynildsen, K.C. Kao, J.K. Suen, J.C. Liao, gNCA: a framework for determining transcription factor activity based on transcriptome: identifiability and numerical implementation, Metab. Eng. 7 (2005) 128–141.
[13] C. Wang, J. Xuan, I.-M. Shih, R. Clarke, Y. Wang, Regulatory component analysis: a semi-blind extraction approach to infer gene regulatory networks with imperfect biological knowledge, Signal Process. 92 (2012) 1902–1915.
[14] A. Noor, A. Ahmad, E. Serpedin, M. Nounou, H. Nounou, ROBNCA: robust network component analysis for recovering transcription factor activities, Bioinformatics 29 (2013) 2410–2418, http://dx.doi.org/10.1093/bioinformatics/btt433.
[15] U. Alon, An Introduction to Systems Biology: Design Principles of Biological Circuits, Chapman and Hall/CRC (Taylor and Francis Group), 2007.
[16] N.D. Jayavelu, L.S. Aasgaard, N. Bar, Iterative sub-network component analysis enables reconstruction of large scale genetic networks, BMC Bioinform. 16 (2015) 366.
[17] N. Bar, N. Jayavelu, New iterative approach (ISNCA) for constrained matrix factorization methods, IFAC-PapersOnLine 49 (2016) 472–477.
[18] X. Wang, M. Alshawaqfeh, X. Dang, B. Wajid, A. Noor, M. Qaraqe, E. Serpedin, An overview of NCA-based algorithms for transcriptional regulatory network inference, Microarrays 4 (2015) 596–617.
[19] O. An, G.M. Dall'Olio, T.P. Mourikis, F.D. Ciccarelli, NCG 5.0: updates of a manually curated repository of cancer genes and associated properties from cancer mutational screenings, Nucleic Acids Res. 44 (2016) D992–D999.
[20] Y. Saeki, T. Endo, K. Ide, T. Nagashima, N. Yumoto, T. Toyoda, H. Suzuki, Y. Hayashizaki, Y. Sakaki, M. Okada-Hatakeyama, Ligand-specific sequential regulation of transcription factors for differentiation of MCF-7 cells, BMC Genomics 10 (2009) 1–16.
[21] Z. Zhang, A. Martino, J.-L. Faulon, Identification of expression patterns of IL-2-responsive genes in the murine T cell line CTLL-2, J. Interferon Cytokine Res. 27 (2007) 991–996.
[22] M.E. Ritchie, B. Phipson, D. Wu, Y. Hu, C.W. Law, W. Shi, G.K. Smyth, Limma powers differential expression analyses for RNA-sequencing and microarray studies, Nucleic Acids Res. 43 (2015), e47.
[23] L. Yan, M. Yang, H. Guo, L. Yang, J. Wu, R. Li, P. Liu, Y. Lian, X. Zheng, J. Yan, J. Huang, M. Li, X. Wu, L. Wen, K. Lao, R. Li, J. Qiao, F. Tang, Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells, Nat. Struct. Mol. Biol. 20 (2013) 1131–1139.
[24] C.-H. Chou, N.-W. Chang, S. Shrestha, S.-D. Hsu, Y.-L. Lin, W.-H. Lee, C.-D. Yang, H.-C. Hong, T.-Y. Wei, S.-J. Tu, T.-R. Tsai, S.-Y. Ho, T.-Y. Jian, H.-Y. Wu, P.-R. Chen, N.-C. Lin, H.-T. Huang, T.-L. Yang, C.-Y. Pai, C.-S. Tai, W.-L. Chen, C.-Y. Huang, C.-C. Liu, S.-L. Weng, K.-W. Liao, W.-L. Hsu, H.-D. Huang, miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database, Nucleic Acids Res. 44 (2016) D239–D247.
[25] A. Essaghir, F. Toffalini, L. Knoops, A. Kallin, J. Helden, J.B. Demoulin, Transcription factor regulation can be accurately predicted from the presence of target gene signatures in micro array gene expression data, Nucleic Acids Res. 38 (2010) e120.
[26] L. Bovolenta, M. Acencio, N. Lemke, HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions, BMC Genomics 13 (2012) 405.
[27] H. Han, H. Shim, D. Shin, J.E. Shim, Y. Ko, J. Shin, H. Kim, A. Cho, E. Kim, T. Lee, H. Kim, K. Kim, S. Yang, D. Bae, A. Yun, S. Kim, C.Y. Kim, H.J. Cho, B. Kang, S. Shin, I. Lee, TRRUST: a reference database of human transcriptional regulatory interactions, Sci. Rep. 5 (2015) 11432.
[28] E. van Schooneveld, M.C. Wouters, I. Van der Auwera, D.J. Peeters, H. Wildiers, P.A. Van Dam, I. Vergote, P.B. Vermeulen, L.Y. Dirix, S.J. Van Laere, Expression profiling of cancerous and normal breast tissues identifies microRNAs that are differentially expressed in serum from patients with (metastatic) breast cancer and healthy volunteers, Breast Cancer Res. 14 (2012) R34.
[29] M. Kedmi, N. Ben-Chetrit, C. Körner, M. Mancini, N.B. Ben-Moshe, M. Lauriola, S. Lavi, F. Biagioni, S. Carvalho, H. Cohen-Dvashi, F. Schmitt, S. Wiemann, G. Blandino, Y. Yarden, EGF induces microRNAs that target suppressors of cell migration: miR-15b targets MTSS1 in breast cancer, Sci. Signal. 8 (2015) ra29.

[30] W. Bin Zhou, C. Neng Zhong, X. Peng Luo, Y. Yuan Zhang, G. Ying Zhang, D. Xian Zhou, L. Ping Liu, miR-625 suppresses cell proliferation and migration by targeting HMGA1 in breast cancer, Biochem. Biophys. Res. Commun. 470 (2016) 838–844.
[31] L. Zhao, Y. Wang, L. Jiang, M. He, X. Bai, L. Yu, M. Wei, MiR-302a/b/c/d cooperatively sensitizes breast cancer cells to adriamycin via suppressing P-glycoprotein (P-gp) by targeting MAP/ERK kinase kinase 1 (MEKK1), J. Exp. Clin. Cancer Res. 35 (2016) 25.
[32] I. Keklikoglou, C. Koerner, C. Schmidt, J.D. Zhang, D. Heckmann, A. Shavinskaya, H. Allgayer, B. Guckel, T. Fehm, A. Schneeweiss, O. Sahin, S. Wiemann, U. Tschulena, MicroRNA-520/373 family functions as a tumor suppressor in estrogen receptor negative breast cancer by targeting NF-[kappa]B and TGF-[beta] signaling pathways, Oncogene 31 (2012) 4150–4163.
[33] O.A. Bamodu, W.-C. Huang, W.-H. Lee, A. Wu, L.S. Wang, M. Hsiao, C.-T. Yeh, T.-Y. Chao, Aberrant KDM5B expression promotes aggressive breast cancer through MALAT1 overexpression and downregulation of hsa-miR-448, BMC Cancer 16 (2016) 160.
[34] Y.-Y. Chang, W.-H. Kuo, J.-H. Hung, C.-Y. Lee, Y.-H. Lee, Y.-C. Chang, W.-C. Lin, C.-Y. Shen, C.-S. Huang, F.-J. Hsieh, L.-C. Lai, M.-H. Tsai, K.-J. Chang, E.Y. Chuang, Deregulated microRNAs in triple-negative breast cancer revealed by deep sequencing, Mol. Cancer 14 (2015) 36.
[35] M.A. Fonseca-Sanchez, C. Perez-Plasencia, J. Fernandez-Retana, E. Arechaga-Ocampo, L.A. Marchat, S. Rodriguez-Cuevas, V. Bautista-Pina, Z.E. Arellano-Anaya, A. Flores-Perez, J. Diaz-Chavez, C. Lopez-Camarillo, microRNA-18b is upregulated in breast cancer and modulates genes involved in cell migration, Oncol. Rep. 30 (2013) 2399–2410.
[36] M. Liu, R. Yang, U. Urrehman, C. Ye, X. Yan, S. Cui, Y. Hong, Y. Gu, Y. Liu, C. Zhao, L. Yan, C.-Y. Zhang, H. Liang, X. Chen, MiR-19b suppresses PTPRG to promote breast tumorigenesis, Oncotarget 7 (2016) 64100–64108.
[37] X. Ni, T. Xia, Y. Zhao, W. Zhou, N. Wu, X. Liu, Q. Ding, X. Zha, J. Sha, S. Wang, Downregulation of miR-106b induced breast cancer cell invasion and motility in association with overexpression of matrix metalloproteinase 2, Cancer Sci. 105 (2013) 18–25.
[38] J. Long, C. Ou, H. Xia, Y. Zhu, D. Liu, MiR-503 inhibited cell proliferation of human breast cancer cells by suppressing CCND1 expression, Tumor Biol. 36 (2015) 8697–8702.
[39] A. Bischoff, B. Huck, B. Keller, M. Strotbek, S. Schmid, M. Boerries, H. Busch, D. Müller, M.A. Olayioye, miR149 functions as a tumor suppressor by controlling breast epithelial cell migration and invasion, Cancer Res. 74 (2014) 5256–5265.
[40] B. Liu, Y. Tian, F. Li, Z. Zhao, C. Jiang, X. Zhai, X. Han, L. Zhang, Tumor-suppressing roles of miR-214 and miR-218 in breast cancer, Oncol. Rep. 35 (2016) 3178–3184.
[41] M. Kedmi, N. Ben-Chetrit, C. Körner, M. Mancini, N.B. Ben-Moshe, M. Lauriola, S. Lavi, F. Biagioni, S. Carvalho, H. Cohen-Dvashi, F. Schmitt, S. Wiemann, G. Blandino, Y. Yarden, EGF induces microRNAs that target suppressors of cell migration: miR-15b targets MTSS1 in breast cancer, Sci. Signal. 8 (2015) ra29.
[42] An integrated encyclopedia of DNA elements in the human genome, Nature 489 (2012) 57–74.
[43] B.M. Javierre, O.S. Burren, S.P. Wilder, R. Kreuzhuber, S.M. Hill, S. Sewitz, J. Cairns, S.W. Wingett, C. Várnai, M.J. Thiecke, F. Burden, S. Farrow, A.J. Cutler, K. Rehnström, K. Downes, L. Grassi, M. Kostadima, P. Freire-Pritchett, F. Wang, J.H. Martens, B. Kim, N. Sharifi, E.M. Janssen-Megens, M.-L. Yaspo, M. Linser, A. Kovacsovics, L. Clarke, D. Richardson, A. Datta, P. Flicek, H.G. Stunnenberg, J.A. Todd, Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters, Cell 167 (2016), 1369–1384.e19.