Published in final edited form as:
Signal Processing. 2012 August 1; 92(8): 1902–1915. doi:10.1016/j.sigpro.2011.11.028.

Regulatory component analysis: a semi-blind extraction approach to infer gene regulatory networks with imperfect biological knowledge

Chen Wang[a], **Jianhua Xuan**[*,a], **Ie-Ming Shih**[b], **Robert Clarke**[c], and **Yue Wang**[a]
[a]Bradley Dept. of Electrical and Computer Engineering, Virginia Tech, Arlington, VA 22203, USA

[b]Dept. of Pathology, Johns Hopkins University, Baltimore, MD 21231, USA

[c]Lombardi Comprehensive Cancer Center and Department of Oncology, Physiology and Biophysics, Georgetown University, Washington, DC 20057, USA

Abstract

With the advent of high-throughput biotechnology capable of monitoring genomic signals, it becomes increasingly promising to understand molecular cellular mechanisms through systems biology approaches. One of the active research topics in systems biology is to infer gene transcriptional regulatory networks using various genomic data; this inference problem can be formulated as a linear model with latent signals associated with some regulatory proteins called transcription factors (TFs). As common statistical assumptions may not hold for genomic signals, typical latent variable algorithms such as independent component analysis (ICA) are incapable to reveal underlying true regulatory signals. Liao et al. [1] proposed to perform inference using an approach named network component analysis (NCA), the optimization of which is achieved by a least-squares fitting approach with biological knowledge constraints. However, the incompleteness of biological knowledge and its inconsistency with gene expression data are not considered in the original NCA solution, which could greatly affect the inference accuracy. To overcome these limitations, we propose a linear extraction scheme, namely regulatory component analysis (RCA), to infer underlying regulatory signals even with partial biological knowledge. Numerical simulations show a significant improvement of our proposed RCA over NCA, not only when signal-to-noise-ratio (SNR) is low, but also when the given biological knowledge is incomplete and inconsistent to gene expression data. Furthermore, real biological experiments on E. coli are performed for regulatory network inference in comparison with several typical linear latent variable methods, which again demonstrates the effectiveness and improved performance of the proposed algorithm.

Keywords

Transcriptional regulatory network inference; Source extraction; Gene expression; Genomic signal processing

[*]Corresponding author: xuan@vt.edu.

1. Introduction

With advancement of biotechnologies, various types of genomic data provide researchers with great opportunities to study cellular systems in a global perspective, facilitating the understanding of biological functions and disease progression causes [2–4]. To fully exploit the information from genomic data, numerous of machine learning and signal processing techniques have been proposed and developed to model genetic systems in a quantitative way [5, 6]. Among them, linear statistical latent variable algorithms such as principal component analysis (PCA) and independent component analysis (ICA), which have been applied successfully in many biomedical applications [7, 8], are also adopted to analyze distinct multi-dimensional genomic signals, including metabolic data [9], DNA polymorphism data [10], and gene expression data [11–15].

Specifically, for gene expression data describing the messenger RNA (mRNA) dynamics of genes, PCA and ICA serve as useful computational tools in various applications. For examples, in [13], both PCA and ICA were applied to perform dimension reduction of gene expression data, and it was shown that statistical assumption based linear transformation can lead to biologically meaningful components; Lee et al. [12] systemically compared PCA and ICA for gene clustering applications on six different datasets, showing that ICA approaches based on higher-order statistics are more capable for detecting functional enriched gene groups than PCA; in [15], PCA and ICA were used to generate feature patterns from an endometrial cancer dataset consisting of benign and malignant samples, and it was reported that ICA was superior to PCA in characterizing expression signatures of malignant samples; based on the features constructed by ICA, an improved tumor classification rate was also achieved by a regularized regression scheme in [11].

Despite the initial success of applying statistical linear latent variable methods for gene expression data analysis, several limitations of these totally “blind” approaches still exist: firstly, the underlying true dimension is hard to determine computationally, and an over-/under-estimation of the signal source number will lead to misinterpretation of gene expression data; secondly, expression level measurements are acquired through sophisticated microarray biotechnologies such as hybridization where large amounts of errors and noises exist in data [16], pure data-driven approaches often suffer from the problems of low reproducibility [17] and over-fitting [18]; finally and most importantly, statistics-based “blind” approaches can only produce abstract components without concrete biological implications. Although estimated source signals could be related to underlying biological processes after some post-processing [13, 12], it remains unclear which molecule(s) correspond to each source. This makes it difficult to design biological experiments validating computational findings. A general requirement, worth noting for next-generation computational approaches in the field of bioinformatics and systems biology, is that computational approaches should lead to biologically testable hypotheses [4, 19].

Network component analysis (NCA) [1, 20, 21], which explicitly incorporates biological knowledge into modeling, establishes a solid link between latent variables and underlying biological regulatory signals. Through integration of biological knowledge, the linear model in NCA has a clear biological implication by using a bipartite network for regulatory network modeling. For the NCA model, if we denote gene expression matrix $\mathbf{X}$ as a linear mixture model, $\mathbf{X} = \mathbf{AS}$, the mixing matrix $\mathbf{A}$ corresponds to hidden activities of regulatory proteins (transcription factors (TFs)), and the source matrix $\mathbf{S}$ reflects the controlling weight from TFs to their target genes. Therefore, computational results based on the NCA model are biologically interpretable. This could lead to hypotheses that can be tested through further experimental studies. Arguably, the statistical assumptions of un-correlatedness and

independence are inappropriate in describing real hidden biological signals. Multiple TFs could potentially work coordinately to control expressions of target genes, and therefore dependence among them cannot be simply ignored [44]. With no such assumptions, the NCA solution is simply achieved by minimizing the fitting error of matrix decomposition following a structure constraint derived from biological knowledge.

Even with several algorithmic improvements [20, 21] and biological studies (e.g., yeast cell cycle studies [1]) conducted based on the original NCA model, several major issues still hinder wider NCA applications. The issues are mostly linked to the availability and quality of biological knowledge[1]: 1) Knowledge incompleteness: biological knowledge is generally incomplete, especially for organisms other than some simple model systems such as yeast and E. coli, whereas NCA application usually assumes that full biological knowledge is available. 2) Knowledge-data inconsistency: biological knowledge is accumulated through scientific literature or experiments, thus it also contains a significant amount of noise and errors. Moreover, biological systems largely behave in a condition-specific manner. The knowledge generated from one experiment does not necessarily reflect the truth in other experiments. Therefore, biological knowledge has been found inconsistent with gene expression data when the NCA model is directly applied [22, 23].

With the awareness of imperfect biological knowledge, we propose and develop a semi-blind extraction algorithm called regulatory component analysis (RCA). The algorithm aims to estimate hidden regulatory components, or equivalently, infer quantitative configurations of transcriptional regulatory networks. The proposed scheme differs from the matrix decomposition optimization in NCA that requires full knowledge of all regulatory components; it can be applied even with partial knowledge of one regulatory component. The RCA criterion is designed to maximize the consistency of extracted components with knowledge, rather than fully follow the given knowledge that may be inconsistent to gene expression data. Thus, RCA is less affected by false-positives (FPs) and false-negatives (FNs) within biological knowledge. With simulations, statistical assumption-based methods (e.g., ICA and PCA) and knowledge guided methods (e.g., NCA and RCA) are fairly compared, to the best of our knowledge, for the first time. In reality, the given biological knowledge could be incomplete and inconsistent to gene expression data. Therefore, we design our comparison experiments to reflect this reality. Furthermore, real biological expression data with ground truth collected from knowledge database are also used to compare performance of all the methods. Therefore, our comparison results would also serve as a reference for other researchers in the field of signal processing and bioinformatics to further develop other improved approaches.

2. Problem formulation and methodology

2.1. General linear latent model of genomic signals

First, we briefly review a general interpretation of linear latent model for genomic signals. Given a high-dimensional data matrix $\mathbf{X} = [\mathbf{x}[1], \cdots, \mathbf{x}[N]] \in \mathbb{R}^{M \times N}$, which can be seen as N realizations of random vector $\mathbf{x} \in \mathbb{R}^{M}$, the purpose of statistical latent variable algorithms such as PCA and ICA is to find a linear transformation $\mathbf{W} \in \mathbb{R}^{L \times M}$, through which the transformed components of $\mathbf{y} = \mathbf{W}\mathbf{x} = (y_1, \cdots, y_L)^T$ are statistically uncorrelated (PCA) or independent (ICA). When observed data can be assumed as linear mixtures of underlying regulatory components or sources: $\mathbf{x}[n] = \mathbf{A}\mathbf{s}[n]$, where components of sources $\mathbf{s} \in \mathbb{R}^{L}$ are non-Gaussian distributed and independent, ICA can be used to perform blind separation of sources; its estimates correspond to underlying sources up to some scale and order

[1]In the present study, biological knowledge mainly refers to the connectivity pattern between TFs and their target genes in the context of regulatory network inference, while general biological knowledge is a much broader concept.

ambiguities even without the exact distribution form of latent variables [24, 25]. ICA algorithms have been applied to some biomedical problems where the assumption of source independence holds [7, 8].

Recently, PCA and ICA models have also been found useful for linear representation of genomic signals. A common biological interpretation of latent variable model is shown in Fig. 1. $\mathbf{X}$ is a genomic signal matrix of M measurements by N 'genomic instances'. These instances could be transcripts of genes [12, 13], metabolisms [9], or genome loci [10]. Realizations of the l-th latent component $[s_l[1], \cdots, s_l[N]]^T$ are generally assumed to reflect the genomic influence of some underlying biological processes to all the genomic instances. Given that cellular systems are energy efficient, each biological process is further assumed to only affect the activities of small portions of genomic instances. Therefore, a super-Gaussian distribution of each component s_l can be assumed approximately. This assumption is supported by previous comparison studies between ICA and PCA for microarray analysis [12, 13].

The applications of statistical latent variable algorithms are mainly limited in exploratory analysis of genomic data. However, computational results are too general to be interpreted in a specific biological context. Therefore, focus is given on gene expression analysis for transcriptional regulatory network inference, where a clear generative model can be formulated with biological implications. The details will be discussed in the following sections.

2.2. Gene expression and NCA model

2.2.1. Gene expression and transcription model—Gene expression generally refers to an information conversion process from DNA sequence of one gene to its mRNA, which will be further translated to corresponding protein(s). Therefore, mRNA molecular concentrations of genes are generally called gene expression levels or expression data. Expression data are acquired through a series of biochemistry-photo transformation, providing the parallel mRNA measurement of thousands of genes in a single microarray chip. Gene expression is one of the genomic data types received the most intensive research attention. This is not only due to its relatively low acquisition cost, but also attributed to its ability in reflecting genetic dynamics of cellular systems [26].

Having M microarray measurements with N genes, gene expression data can be denoted as a matrix $\mathbf{E} \in \mathbb{R}^{+M \times N}$, where e_{mn} reflects the concentration of the n-th gene in the m-th microarray measurement. We denote normal concentration of the n-th gene as $e_n^{(0)}$, which is usually generated in baseline condition as a reference signal. It is known that transcription rate of genes are determined by concentrations of some special proteins called transcription factors (TFs) [1]. The following transcription rate equation can be obtained through approximation of a series of differential equations under equilibrium assumptions:

$$\frac{e_{mn}}{e_n^{(0)}} = \prod_{l=1}^{L} \left(\frac{p_{ml}}{p_l^{(0)}} \right)^{s_{ln}}, \tag{1}$$

where p_{ml} and $p_l^{(0)}$ are concentrations of the l-th TF in the m-th microarray measurement and under baseline condition, respectively. The exponential item s_{ln} reflects how the l-th TF regulates the transcription rate of the n-th gene, with $s_{nl} = 0$ as no regulation, $s_{nl} > 0$ as transcription promotion (or up-regulation), and $s_{nl} < 0$ as transcription suppression (or down-regulation). It should be noticed that only the expression concentration e_{mn} and $e_n^{(0)}$ are directly observable, whereas p_{ml}, $p_l^{(0)}$ and s_{nl} are all hidden variables.

By denoting

$$x_{mn}=\log\frac{e_{mn}}{e_n^{(0)}} \tag{2}$$

and

$$a_{ml}=\log\frac{p_{ml}}{p_l^{(0)}}, \tag{3}$$

Equation (1) can be expressed as

$$x_{mn}=\sum_{l=1}^{L} a_{ml}s_{ln}, \tag{4}$$

or in a matrix multiplication form with an additive noise matrix $\mathbf{\Gamma}\in\mathbb{R}^{M\times N}$

$$\mathbf{X}=\mathbf{AS}+\mathbf{\Gamma}. \tag{5}$$

Equation (5) can be further written in the form of latent variable model with respect to gene index n:

$$\mathbf{x}[n]=\mathbf{A}\mathbf{s}[n]+\gamma[n], \tag{6}$$

where $\mathbf{x}[n]=[x_{1n},\cdots,x_{Mn}]$, $\mathbf{s}[n]=[s_{1n},\cdots,s_{Ln}]$, and $\boldsymbol{\gamma}[n]=[\gamma_{1n},\cdots,\gamma_{Mn}]$ are the gene expression profile, regulatory component, and noise vectors of the n-th gene, respectively. Equation (5) is called log-linear model, considering the transformations in Equation (2) and (3) [1]. The log-ratio transformation of gene expression data can be fit approximately with Gaussian distribution [13]. Different from the general latent variable analysis, now everything has a clear biological implication; latent factors of the linear model (6) correspond to TFs. The l-th column of matrix $\mathbf{A}=[\mathbf{a}_1,\cdots,\mathbf{a}_L]$ is defined as the l-th transcription factor activity (TFA), which reflects the hidden protein relative activity of the l-th TF. The influence variable of the l-th TF s_l is called the l-th regulatory component (RC). In the present paper, the mixing matrix **A** and source matrix **S** are referred to as TFA matrix and regulatory component matrix (or RC matrix) to highlight their biological implications.

2.2.2. NCA model—Before we discuss regulatory component estimation, we introduce specific biological knowledge that facilitates the estimation. Each regulatory component s_l in Equation (6) corresponds to the regulatory effect of certain TF on gene transcription rates. One TF has to bind to the DNA promoter region of a target gene to regulate gene expression. Such physically-binding relationship can be measured through biological experiments [27] or predicted through computational sequence analysis [28]. Based on the potential binding relationship of TF to gene, the physical binding relationship from TFs to genes is defined as *network connectivity pattern* $\mathbf{B}\in(0,1)^{L\times N}$. This is a binary matrix with element $b_{ln}=1$ indicating potential regulatory relationship from the l-th TF to the n-th gene.

Genes regulated by TFs are typically called *target genes*. Assuming there is no feedback from target genes to TFs, transcriptional regulatory network describing the relationship between TFs and target genes is a bipartite network, where the nodes of latent layer and observed layer are TFs and downstream genes, respectively. Regulatory component matrix **S**

describes weights of bipartite network edges. Therefore, estimation of hidden regulatory components is equivalent to the inference of underlying regulatory network (Fig. 2).

To solve Equation (5) based on available biological knowledge **B**, the original NCA algorithm is designed to estimate **A** and **S** by minimizing the fitting error [1]:

$$(\widehat{\mathbf{A}}, \widehat{\mathbf{S}}) = \arg\min_{(\mathbf{A},\mathbf{S})} \left\| \mathbf{X} - \mathbf{AS} \right\|_2^2, \tag{7}$$

$$\text{s.t. } \mathbf{S} \in \mathbb{Z}_0. \tag{8}$$

In constraint (8), $\mathbb{Z}_0$ is a regulatory matrix set, derived from biological knowledge of connectivity matrix:

$$\mathbb{Z}_0 \triangleq \left\{ \mathbf{S} \in \mathbb{R}^{L \times N} | s_{ln} = 0 \text{ for } b_{ln} = 0 \right\}. \tag{9}$$

Assuming the elements of noise matrix $\mathbf{\Gamma}$ is i.i.d Gaussian distributed, the NCA criterion is equivalent to maximizing the likelihood with respect to noise distribution [21]. The NCA criterion does not incorporate any statistical priori of **A** or **S**. This is motivated by discussions in [1] that statistical assumptions may not fit to biological reality. Therefore, the NCA criterion is simply a least-square fitting with structure constraint on **S**. In the original NCA paper, TFA **A** is regarded as underlying regulatory signals, where regulatory component matrix **S** is treated as a mixing matrix. Actually the definitions of mixing matrix and source matrix are interchangeable for NCA through a matrix transposition, as no statistical properties are assumed according to either matrix. Arguably, **S** is more appropriately assumed as the underlying source than A for applying statistical latent variable methods. This is because non-Gaussianity assumption of each component s_l approximately holds, considering the fact that one TF can only regulate a small portion of genes [1, 21, 20].

As the NCA optimization procedure involves biological knowledge **B**, the structure characteristic of **B** is essential for NCA estimation. This is reflected from identifiability conditions of NCA. In the noiseless case, the identifiability conditions for NCA are proved when the following four assumptions are met [1]:

Identifiability conditions of NCA

Assumption 1: The number of microarray samples (M) should be greater than or equal to the number of TFs (L).

Assumption 2: Different TFAs, $\mathbf{a}_l$, $l = 1, \cdots, L$, are linearly independent.

Assumption 3: For connectivity pattern matrix **B**, if any TF and its associated genes are taken out, the modified connectivity pattern matrix $\tilde{\mathbf{B}}$ should have full row rank (rank = $L - 1$).

Assumption 4: The network connectivity pattern **B** is perfectly known (a priori).

Both Assumption 1 and 2 are almost universal presumptions for linear latent algorithms. Assumption 1 is generally needed to ensure that the problem is not underdetermined. Assumption 2 is also similar to the presumption for PCA/ICA models in that mixing matrix **A** is non-singular.

Assumption 3 indicates that if one TF is determined, the rest $L - 1$ TFAs can still be uniquely determined. By explicitly exploiting the property of Assumption 3, Chang et al. [29] proposed an alternative algorithm fastNCA, which can be several of tens times faster than the original NCA algorithm. Assumption 3 is not always fulfilled for a given connectivity, thus a condition check is usually performed and the connections violating this assumption are pruned [1, 21].

However, effective condition check for Assumption 3 also relies on Assumption 4, assuming that given $\mathbf{B}$ reflects the underlying true relationship $\mathbf{B}_0$. Therefore, the estimation accuracy of both NCA and fastNCA are expected to heavily depend on the availability and quality of given biological knowledge, which will be discussed in Section 2.2.3.

Ambiguities: Although prior biological knowledge eliminates the ordering ambiguity of regulatory components, the scaling of underlying signals is still undetermined. Therefore, even with the fulfillment of all identifiability conditions, the estimated regulatory component $\hat{\mathbf{s}}$ by NCA could still differ from the underlying true signals $\mathbf{s}$ up to some scaling ambiguity $\hat{\mathbf{s}} = \mathbf{Ds}$, where $\mathbf{D}$ is an arbitrary diagonal matrix with non-zero diagonal items. Such ambiguity is usually acceptable in source separation applications as it is "waveform preserved" [30], which means the waveforms of original signals are correctly captured.

2.2.3. Degeneration of biological knowledge—Assumption 4 assumes that the complete biological knowledge-connectivity pattern matrix $\mathbf{B}$ is (a) complete (including all TFs), and (b) accurate (consistent to expression data $\mathbf{X}$). However, biological connection knowledge is often incomplete in reality. This is especially true for species like humans, where only a few transcription factors can be known in advance. Aside from knowledge incompleteness, biological knowledge is also generally inconsistent with gene expression data. Such knowledge-data inconsistency mainly stems from two situations: 1) part of given knowledge is generated from other biological experiments, which may introduce errors; and 2) knowledge is very general and may not be specific to biological conditions under which gene expression data are acquired. Thus, biological knowledge usually contains a considerable amount of FPs and FNs, which should not be ignored for computational modeling.

The incompleteness of biological knowledge and its inconsistency with expression data are summarized as knowledge degeneration (Fig. 3). We denote $\mathbf{B}_0=\left[\mathbf{b}_1^{(0)}, \cdots, \mathbf{b}_L^{(0)}\right]^T$, where $\mathbf{b}_l^{(0)}$ represents the true connectivity pattern for the l-th TF. In Fig. 3, an extreme case is illustrated when only knowledge of the third TF $\mathbf{b}_3$ is available. However, the given $\mathbf{b}_3$ is still different from true $\mathbf{b}_3^{(0)}$ because of FPs and FNs in biological knowledge.

2.3. Regulatory component analysis

With the awareness of degeneration of given biological knowledge, we describe in this section the motivation and criterion of the proposed RCA.

2.3.1. From decomposition to extraction—According to Assumption 2 of NCA, different TFAs are linearly independent so that matrix $\mathbf{A}$ is invertible, a regulatory component estimate $\hat{\mathbf{s}}$ can usually be achieved through a linear projection from expression matrix $\mathbf{X}$:

$$\widehat{\mathbf{s}}=\mathbf{Wx}, \tag{10}$$

where the projection matrix **W** is also called the de-mixing matrix in the blind source separation problem. A perfect **W** should be the pseudo-inverse of mixing matrix up to a scaling ambiguity

$$\mathbf{W}=\mathbf{D}\mathbf{A}^{\dagger}. \tag{11}$$

In Equation (11), † is the notation for pseudo-inverse operator. Whereas the goal of PCA or ICA is to find a projection matrix $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_L]^T$ so that the resulting components are statistically uncorrelated or independent, the purpose of NCA projection is to find the source matrix exactly following the given connectivity knowledge and minimizing the fitting error. Instead of using matrix decomposition, PCA and ICA solutions can also be achieved in an extraction manner by maximizing the variance and non-Gaussianity of estimated components, respectively. Extraction is usually implemented using a linear projection:

$$\mathbf{y}=\mathbf{w}^T\mathbf{X}, \tag{12}$$

where a good extraction filter $\mathbf{w} \in \mathbb{R}^M$ should correspond to one row of the de-mixing matrix in Equation (11). When only certain sources are of interest, blind extraction appears to be a more efficient scheme than fully blind separation. Typical blind extraction algorithms are designed to recover components of interest by maximizing certain desired characteristics of extracted components:

$$\widehat{\mathbf{w}}=\arg\max_{\mathbf{w}} J(\mathbf{y})=\arg\max_{\mathbf{w}} J(\mathbf{w}^T\mathbf{X}), \tag{13}$$

where function forms of $J(\cdot)$ are generally designed according to some properties of underlying source signals. These properties include non-Gaussianity, temporal continuity, etc. [31]. The linear extraction scheme also avoids the dimension determination problem for latent components if only a few of components are of interests.

Based on the discussions in Section 2.2.3, an extraction scheme is very attractive for gene regulatory network inference, especially when only partial knowledge is available. However, an extraction scheme is not immediately clear for the NCA scheme. The NCA criterion of minimizing the fitting error requires all regulatory components to be estimated in parallel. Given the limitations of NCA and inspired by the extraction framework originated from ICA, we propose a linear extraction algorithm for regulatory network inference capable of incorporating partial biological knowledge.

2.3.2. Regulatory component analysis formulation—Assuming only one column of **B** is given, say the l-th column $\mathbf{b}_l$, we propose a scheme to extract the corresponding regulatory component. First, according to $\mathbf{b}_l$, the column vectors of matrix **X** are divided into two non-overlapped sets:

$$\mathbb{X}_{+}^{(l)}=\{\mathbf{x}_i | b_{li}=1\} \tag{14}$$

and

$$\mathbb{X}_{-}^{(l)}=\{\mathbf{x}_j | b_{lj}=0\}. \tag{15}$$

The number of members in $\mathbb{X}_{+}^{(l)}$ and $\mathbb{X}_{-}^{(l)}$ is denoted as N_+ and N_-, respectively ($N_+ + N_- = N$). Regulatory component analysis is designed to find a linear projection maximizing the following cost function:

$$J_0(\mathbf{X}, \mathbf{b}_l, \mathbf{w}) = \frac{\frac{1}{N_+}\sum_{\mathbf{x}_i \in \mathbb{X}_+^{(l)}} (\mathbf{w}^T \mathbf{x}_i)^2}{\frac{1}{N_-}\sum_{\mathbf{x}_j \in \mathbb{X}_-^{(l)}} (\mathbf{w}^T \mathbf{x}_j)^2}. \quad (16)$$

The function value of $J_0(\mathbf{X},\mathbf{b}_l,\mathbf{w})$ has intuitive explanation with data-knowledge consistency, reflecting how well the estimated regulatory component is supported by given biological knowledge. The interpretation is given as follows: the regulatory component **y** of the *l*-th TF describes the regulation relationship from this TF to all the genes. Considering that one TF could up- or down-regulate its target genes, we simply define regulatory influence from this TF to the *n*-th gene as y_n^2 to cancel the sign. If estimated component well corresponds to biological knowledge of the *l*-th TF, the averaged regulatory influence on its target genes defined by knowledge ($b_{ln} = 1$) should be much larger than the average regulatory influence on the remaining genes ($b_{ln} = 0$). The numerator and denominator of Equation (16) are the averaged regulatory influence of target genes and non-target genes of the *l*-th TF defined by knowledge, respectively. Therefore, the larger function value of $J_0(\cdot)$, the more consistent estimated component $\mathbf{y} = \mathbf{w}^T\mathbf{X}$ with given knowledge vector $\mathbf{b}_l$. In the noiseless case and with perfectly given biological knowledge, the average regulatory influence on target genes is non-zero and average influence on remaining non-target genes is zero: $J_0 \to \infty$. With function value equals to 1, it suggests that estimated regulatory component is not consistent with biological knowledge, as the averaged regulatory influence of potential target genes is the same with of non-target genes.

We further stack the members of each set to form two matrices $\mathbf{X}_+^{(l)}$ and $\mathbf{X}_-^{(l)}$, which correspond to $\mathbb{X}_1^{(l)}$ and $\mathbb{X}_0^{(l)}$, respectively. The criterion function is rewritten as

$$J_0(\mathbf{X}, \mathbf{b}_l, \mathbf{w}) = \frac{N_-}{N_+} \frac{\mathbf{w}^T \mathbf{X}_+^{(l)} \left(\mathbf{X}_+^{(l)}\right)^T \mathbf{w}}{\mathbf{w}^T \mathbf{X}_-^{(l)} \left(\mathbf{X}_-^{(l)}\right)^T \mathbf{w}}. \quad (17)$$

Equation (17) has a Rayleigh quotient form so that through some mathematical manipulations (Appendix A), the following equation can be obtained by maximizing RCA criterion function $J_0(\cdot)$:

$$\mathbf{w}^T \mathbf{X}_+^{(l)} \left(\mathbf{X}_+^{(l)}\right)^T = \lambda \mathbf{w}^T \mathbf{X}_-^{(l)} \left(\mathbf{X}_-^{(l)}\right)^T, \quad (18)$$

which can be effectively solved using generalized eigenvalue decomposition between $\mathbf{X}_+^{(l)} \left(\mathbf{X}_+^{(l)}\right)^T$ and $\mathbf{X}_-^{(l)} \left(\mathbf{X}_-^{(l)}\right)^T$. The RCA estimated extraction filter $\hat{\mathbf{w}}$ will be the eigenvector associated with the maximum generalized eigenvalue of Equation (18).

The proposed RCA criterion has several advantages over traditional NCA approaches [1, 20, 21, 29]:

1. Instead of requiring the complete prior knowledge of all TFs for pursuing a constrained least-square solution, RCA can incorporate incomplete knowledge to estimate individual regulatory component by maximizing a knowledge-data consistency criterion.
2. Rather than strictly following given biological knowledge, the RCA criterion function allows mismatch between estimated regulatory component and biological

knowledge. As a result, estimated regulatory weight y_n could be any value, regardless if there is existing knowledge to support it or not (b_{ln} = 1 or 0). According to estimated regulatory component of the l-th TF, if a large absolute value of y_n is observed with no existing biological support (b_{ln} = 0), the regulatory relationship from the l-th TF to the n-th gene could be a false negative in given knowledge. On the contrary, a small absolute value of y_n associated with b_{ln} = 1 may reflect a false positive in given knowledge. Therefore, this feature enables the detection of FPs and FNs of biological knowledge, with the information from expression data.

3. The Rayleigh ratio function form of RCA criterion reduces computation burden with an efficient optimization using generalized eigenvalue decomposition. Moreover, incorporating other regularization items with the form of $J_r(\mathbf{X}, \mathbf{b}_l, \mathbf{w}) = \frac{\mathbf{w}^T F(\mathbf{X}, \mathbf{b})\mathbf{w}}{\mathbf{w}^T \mathbf{X}_-^{(l)} \left(\mathbf{X}_-^{(l)}\right)^T \mathbf{w}}$ is more convenient if extra prior knowledge is known. The extended criterion function $J(\mathbf{w}) = J_0(\mathbf{w}) + \alpha J_r(\mathbf{w})$ can be efficiently solved using generalized eigenvalue decomposition, where α is some trade-off parameter. Notice that generalized eigenvalue decomposition has been widely used in various pattern recognition applications [32], as well as statistical criterion-based blind separation problems [33]. This suggests that the proposed RCA has the potential to be extended with other prior property function terms, which is a topic in our future investigation. A priori property function can be designed to reflect the prior information of underlying components, such as “non-Gaussianity”.

Identifiability condition of RCA: We accept Assumption 1 and 2, which are common assumptions for linear latent model. In noiseless case with perfect given knowledge of the l-th TF ($\mathbf{b}_l = \mathbf{b}_l^{(0)}$), the estimated regulatory component by maximizing RCA criterion function will only differ from true signal s_l with some non-zero scaling factor c:

$$y[n] = \widehat{\mathbf{w}}^T \mathbf{x}[n] = c s_l[n], n = 1, \cdots, N \tag{19}$$

if remaining sources are *linearly independent*, i.e., rank($\mathbf{S}^{(\backslash l)}$) = $L - 1$, where $\mathbf{S}^{(\backslash l)} = [\mathbf{s}_1, \cdots, \mathbf{s}_{l-1}, \mathbf{s}_{l+1}, \cdots, \mathbf{s}_L]^T$.

The proof is presented in Appendix B. This condition is much more relaxed than original NCA Assumption 3. It suggests that in ideal case the perfect recovery of one regulatory component only requires the corresponding perfect knowledge, and the statistical independence of sources is not required. For non-ideal cases with noises and contaminated knowledge, we will investigate RCA performance through following simulations.

3. Simulation

3.1. Simulation description

Following the characteristics of true regulatory network, connectivity matrix $\mathbf{B}_0$ is generated with sparse property. It is known that transcription regulation can be involved with synergistic mechanism (one gene can be regulated through the collaboration of two or more TFs) so that regulatory components are dependent with each other. Dependent regulatory component with an average pair-wise correlation around 0.1 is generated. In evaluating the impact of biological knowledge to estimation, two simulated scenarios are considered:

1. Perfect connectivity pattern is given ($\mathbf{B} = \mathbf{B}_0$).

2. Imperfect connectivity pattern is given ($\mathbf{B} \neq \mathbf{B}_0$). In simulating the real situation where biological knowledge is incomplete and inconsistent, the given $\mathbf{B}$ input to algorithms is generated in two steps. First, only some row vectors of true $\mathbf{B}_0$ are given. Second, the given partial $\mathbf{B}_0$ is corrupted with FPs and FNs.

In each scenario, the estimation performance of multiple algorithms (PCA, fastICA, JADE, NCA, fastNCA and proposed RCA) are tested under various signal-to-noise-ratio (SNR) conditions. Based on Equation (5), SNR is defined as

$$\mathrm{SNR}=10\log_{10}\frac{\mathrm{Power}_{\mathrm{signal}}}{\mathrm{Power}_{\mathrm{noise}}}=10\log_{10}\frac{\sum_{n=1}^{N}\sum_{m=1}^{M}(x_{mn}-\gamma_{mn})^2}{\sum_{n=1}^{N}\sum_{m=1}^{M}\gamma_{mn}^2}. \quad (20)$$

As the regulatory component estimation problem is also equivalent to inference of transcriptional regulatory network, we define two performance evaluation functions for $\hat{s}_l$ estimated by each algorithm:

The Averaged pair-wise absolute correlation (APAC)

$$APAC=\frac{1}{L}\sum_{l=1}^{L}|corr(\widehat{s_l}, s_l)|, \quad (21)$$

and the Averaged Area Under precision-recall Curve (AAUC)

$$AAUC=\frac{1}{L}\sum_{l=1}^{L}AUC\left(\widehat{s_l}, b_l^{(0)}\right). \quad (22)$$

In Equation (22), $b_l^{(0)}$ is the true biological knowledge of the l-th TF, which corresponds to the l-th row of true connectivity pattern matrix $\mathbf{B}^{(0)}$. $AUC(\cdot,\cdot)$ is a function calculating the value of area under precision-recall curve (see Appendix C), which describes how well the estimated component can reveal the true target genes of corresponding TF. APAC has clear implications for signal estimation accuracy, whereas AAUC is more suitable for evaluating biological ground truth when quantitative regulatory component is usually not available.

3.2. Regulatory component estimation

PCA and ICA: After the components $\mathbf{y}_l, l = 1, \cdots, L$ are estimated, correspondence relationships need to be established with true components $\mathbf{s}_l$ for performance evaluation. Given that NCA, fastNCA, and RCA approaches are implicitly incorporated with biological knowledge, the correspondence is simple: $\hat{\mathbf{s}}_l = \mathbf{y}_l$. However, ordering ambiguities still exist for PCA and ICA. Therefore, $\mathbf{y}_l$ is designed to correspond to $\hat{\mathbf{s}}_{l'}$, knowledge vector $\mathbf{b}_{l'}$ of which has the highest similarity with $\mathbf{y}_l$ in terms of absolute correlation value. Two popular ICA algorithms were adopted in simulation studies: JADE [34, 35], which is based on algebra criterion to jointly diagonalize a set of higher-order statistics matrices; and fastICA [36], which is based on information theory-derived criterion to maximize negative-entropy or the distance with Gaussian distribution.

NCA and fastNCA: PCA, ICA, and RCA allow $\hat{s}_{ln}$ to be an arbitrary value even with no biological support $b_{ln} = 0$, whereas NCA and fastNCA explicitly require $\hat{s}_{ln} = 0, \forall b_{ln} = 0$. As one of our purposes in simulation is to detect with false knowledge how well the underlying true regulatory component can still be recovered, we develop a natural extension for NCA and fastNCA. Assuming the non-singularity of mixing matrix **A** (based on Assumption 2 in Section 2.2.2), $\hat{\mathbf{s}} = \hat{\mathbf{A}}^{\dagger}\mathbf{x}$ is used as the estimate for regulatory components, in which $\hat{\mathbf{A}}$ is the estimate of TFA matrix from the NCA or fastNCA algorithm. Through this simple transformation, $\hat{s}_{ln}$ can be of any value even for $b_{ln} = 0$. All methods can then be fairly compared.

3.3. Simulation results

3.3.1. Biological knowledge is perfectly given ($\mathbf{B} = \mathbf{B_0}$)—To obtain a full spectrum of comparison, we tested all the methods under SNR conditions from −1 dB to 15 dB. For each SNR condition, experiments were conducted 50 times to calculate the average performance value. A transcriptional regulatory network consisting of 300 genes regulated by 15 TFs was randomly constructed. Based on the generated network, simulated expression data with 35 samples were produced according to Equation (5) ($M = 35, N = 300, L = 15$). As shown in Fig. 4, two performance evaluations display quite consistent pictures. In general, RCA and NCA exhibited better performance than the two ICA algorithms JADE and fastICA. PCA remains the worst. This observation is understandable as the implicit utilization of knowledge gives the advantages to NCA and RCA. However, fastNCA showed similar performance with both NCA and RCA in high SNR region, but underwent a dramatic degradation in low SNR region. This performance occurred because fastNCA is derived differently from least-squares solution of NCA and is based on a signal sub-space approach based on Assumption 3. Thus, the accurate estimation of sub-space is essential for its estimation accuracy. While in the high SNR conditions the sub-space estimation was generally reliable, fastNCA performance tended to degrade in low SNR conditions. In contrast, although matrix decomposition-based NCA was more computationally costly than fastNCA, its performance was more robust.

3.3.2. Biological knowledge is imperfectly given ($\mathbf{B} \neq \mathbf{B_0}$)—While keeping all the other simulation configuration parameters unchanged, we modified the quality of input biological knowledge **B**. This scenario was designed to evaluate the effect of imperfect biological knowledge on regulatory component estimation by only providing 10 TFs information out of underlying 15 TFs. Moreover, the given knowledge of these 10 TFs were contaminated with moderate FP and FN (FP rate = 1% and FN rate = 10%) to simulate a real biological study. Given that the estimation of regulatory component was equivalent to regulatory network inference, two performance evaluations exhibited consistent comparison orderings: RCA > NCA > (JADE and fastICA) > (fastNCA and PCA), shown in Fig. 5. Noticeably, fastNCA performed miserably with performance sometimes even worse than that of PCA. fastNCA depends heavily on Assumption 3 and 4, which were severely violated in this simulation case. Moreover, although least-squares-based NCA maintained a relatively robust performance, it is apparently inferior to the proposed RCA algorithm. In both simulations, two ICA algorithms consistently outperformed PCA because the non-Gaussianity property used by ICA is well matched with the sparse regulation relationship of regulatory components, even when independence assumption was violated.

To illustrate the estimation difference, some regulatory component estimation results produced in single simulation running when SNR= 3dB are presented in Fig. 6, with corresponding precision-recall curves. It can be observed that RCA generated the most similar waveform with underlying true regulatory component. As a result, its precision-recall curve has larger area-under-curve(AUC) than AUC of all the other methods.

4. Real biological experiments

In previous section, simulation data verified the effectiveness and illustrated the superior performance of the proposed RCA algorithm. We were also willing to proceed to real biological data analysis. However, the revealing of real transcriptional regulation network for human beings is still ongoing, and many related mechanisms remain unclear. Hence, we tested all the algorithms on inferring transcriptional regulatory network on E. coli, a simple bacterium well studied as the model system for various biological studies. We extracted biological knowledge of TFs from a knowledge database called RegulonDB (http://regulondb.ccg.unam.mx) with recently updated version 7.0 [37]. The RegulonDB database contains a collection of TF-target relationships that have been experimentally verified in E. coli. Out of 169 TFs recorded in RegulonDB, 30 TFs with at least 15 experimental validated target genes were selected to form the initial connectivity pattern matrix. This selection criterion was based on the considerations for reliable precision-recall curve estimation and performance evaluation. The target genes of the 30 selected TFs overlap with a huge expression compendium [38], which contains 445 E. coli microarray samples under distinct biological conditions. Subsequently, a network connectivity pattern matrix with 1193 target genes and 30 TFs was obtained. Moderate amounts of FPs and FNs (FP rate = 1% and FN rate = 10%) were added to the connectivity pattern matrix to test how well the regulatory components could be estimated with incomplete and inconsistent knowledge.

As there is no quantitative ground truth for true regulatory component, the AAUC criterion was used to evaluate the performance. In addition, we observed that AAUC is highly correlated with APAC in our previous simulation studies, so AAUC should serve as a reasonable performance evaluation here. Each time, 100 microarray samples were randomly selected from 445 total microarray samples to estimate the regulatory components for all the methods. We obtained 50 random selections to calculate the performance evaluation of AAUC. As shown in Fig. 7, RCA significantly outperformed all the other methods in retrieving the true target genes regulated by corresponding TFs. To illustrate further the retrieval performance of different methods, we present the precision-recall curves for two TFs ArgR and LexA as examples in Fig. 8. Comparing to the simulation studies, the performance of all the methods dropped. It suggests that the estimation of regulatory components in the real dataset would be more difficult. Nevertheless, RCA still achieved very robust performance, much better than all the other methods.

5. Discussions and conclusions

Linear latent variable models are widely used in biomedical applications for identifying or extracting underlying biological signals corrupted by artifacts or undesired signals. Statistical assumptions such as un-correlatedness and independence are readily accepted in many of these applications, such as analyses of ECG, EEG, and MEG data [7, 8]. However, when these statistical tools are applied to analyze complicated genomic data, the results become very difficult to interpret. Instead of enforcing strong statistical assumptions, NCA incorporates biological knowledge into the optimization process of a linear latent model for gene expression data analysis. This leads to biologically interpretable sources, which are called regulatory components in the present paper. Noticeably, this linear model is also equivalent to a bipartite regulatory network describing the regulatory relationship between TFs and their target genes. However, optimization of NCA is performed based on a least-squares fitting with biological knowledge constrained. Thus, NCA estimation is largely dependent on available TF-gene binding knowledge, as well as the quality of given knowledge. Unfortunately, real biological knowledge is generally incomplete and inconsistent with gene expression data under study.

Given the aforementioned pitfalls in biological knowledge, we have proposed a linear extraction-based framework called RCA. RCA explicitly finds a linear projection that maximizes the coincidence with given partial biological knowledge. The linear extraction scheme also allows RCA to detect FPs and FNs of biological knowledge, which is inconsistent with gene expression data. The contributions of our present study are multi-folded. First, from the perspective of general linear latent model for genomic signals, the equivalence of the network inference problem with linear latent variable model is reviewed, which could serve as a useful reference for signal processing researchers interested in genomic signal processing. Second, for the first time, a linear extraction scheme is formulated to infer transcriptional regulatory networks, taking into account incomplete but informative biological knowledge. The proposed scheme shows a significant performance improvement over traditional NCA methods in both simulations and real biological experiments (E. coli). Third, simulation studies show that it is not a trivial problem to integrate biological knowledge effectively and efficiently, considering that given biological knowledge is usually incomplete and inconsistent to available data. An inappropriate incorporation of biological knowledge to the computational methods may result in worse performances than those without using biological knowledge at all.

Notice that the RCA criterion defined in (17) has a similar Rayleigh quotient function form with those of linear discriminate analysis (LDA) [39] and Locality Preserving Projections (LPP) [40]. However, these two methods were designed with different implications and used to achieve distinct goals. LDA, LPP and the proposed RCA are all linear dimension reduction schemes by applying a projection **w**on high-dimensional data **X**, and optimization criterion functions all follow the same Rayleigh quotient form: $R(\mathbf{w})=\frac{\mathbf{w}^T\mathbf{M}_1\mathbf{w}}{\mathbf{w}^T\mathbf{M}_2\mathbf{w}}$. However, the implication of each criterion function is different. LDA solves **w** by maximizing the discrimination function $R_{LDA}(\mathbf{w})=\frac{\mathbf{w}^T\Sigma_{between}\mathbf{w}}{\mathbf{w}^T\Sigma_{within}\mathbf{w}}$, which is the ratio of between-class variance and within-class variance. LPP computes **w** to preserve the local closeness of data-points on a low dimensional manifold while keeping the scale of data point the same; the solution of LPP is through minimization of criterion function $R_{LPP}(\mathbf{w})=\frac{\mathbf{w}^T(\mathbf{D}-\mathbf{C})\mathbf{w}}{\mathbf{w}^T\mathbf{D}\mathbf{w}}$, where $\mathbf{C}\in\mathbb{R}^{M\times M}$ is a symmetric matrix defining the adjacency of data points on manifold and $\mathbf{D}\in\mathbb{R}^{M\times M}$ is a diagonal matrix with $d_{ii}=\sum_{j=1}^{M}c_{ij}$. Guided by biological knowledge of TFs, RCA maximizes the criterion function $R_{RCA}(\mathbf{w})=\frac{\mathbf{w}^T\frac{\mathbf{x}_+^{(l)}\left(\mathbf{x}_+^{(l)}\right)^T}{N_+}\mathbf{w}}{\mathbf{w}^T\frac{\mathbf{x}_-^{(l)}\left(\mathbf{x}_-^{(l)}\right)^T}{N_-}\mathbf{w}}$, where the numerator reflects average regulation influence of target genes of the *l*-th TF defined by knowledge and denominator is the average regulation influence of genes without knowledge support. To further understand the difference between LDA, LPP and RCA, we can look into the implications of the corresponding projection $y_n = \mathbf{w}^T\mathbf{x}_n$ for the *n*-th data vector $\mathbf{x}_n$. For LDA, y_n is used for classification decision by checking whether its value falls above or below some threshold; y_n in LPP is a coordinate of *n*-th data vector in low dimensional manifold; y_n in RCA has biological meaning to describe the regulation influence from the *l*-th TF to the *n*-th gene, given $\mathbf{b}_l$ as its knowledge guidance.

For the future research, it would be very meaningful to apply and extend RCA to analyze different microarray datasets, such as time course dataset, to further understand its usefulness and limitation in real biological studies. Since biological knowledge plays an essential role in the proposed scheme, some consistency check between knowledge and data could also be performed to ensure the quality of given knowledge. For example, a qualitative way to filter out inconsistent biological knowledge with gene expression data has

been proposed in [41], which could serve as a useful pre-processing step to refine initial biological knowledge set of RCA. It is also worthy to notice that the currently proposed scheme is mainly based on linear approximation. The computational modeling of nonlinear interactions among genes has also been extensively studied, for examples, mutual information has been employed to address pair-wise nonlinear gene-gene interactions [42]; tree-based ensemble regression has also been shown as an effective approach to reveal combinatorial and nonlinear regulation relationships [43]. As one of future research directions, it would be very important to incorporate certain nonlinearity into the modeling of transcriptional regulation. With increasing accumulated biological knowledge, Bayesian technique could be a promising alternative to incorporate prior information through a probabilistic formulation [45], instead of enforcing biological knowledge directly in the matrix decomposition. Therefore, another potential research direction would be how to extend RCA approach by using Bayesian techniques.

Acknowledgments

This study is supported in part by the National Institutes of Health under Grants (CA139246, CA149653, CA149147 and NS29525).

References

1. Liao JC, Boscolo R, Yang YL, Tran LM, Sabatti C, Roychowdhury VP. Network component analysis: reconstruction of regulatory signals in biological systems. Proc Natl Acad Sci U S A. 2003; 100(26):15522–7. [PubMed: 14673099]
2. Huttenhower C, Hofmann O. A quick guide to large-scale genomic data mining. PLoS Comput Biol. 2010; 6(5):e1000779. [PubMed: 20523745]
3. Joyce A, Palsson B. The model organism as a system: integrating ‘omics’ data sets. Nature Reviews Molecular Cell Biology. 2006; 7(3):198–210.
4. Clarke R, Ressom HW, Wang A, Xuan J, Liu MC, Gehan EA, Wang Y. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. Nature Reviews Cancer. 2008; 8(1):37–49.
5. Dougherty ER, Datta A, Sima C. Research issues in genomic signal processing. Signal Processing Magazine, IEEE. 2005; 22(6):46–68.
6. Jie C, Huai L, Kaihua S, Kim B. How will bioinformatics impact signal processing research? Signal Processing Magazine, IEEE. 2003; 20(6):106–206.
7. Jung TP, Makeig S, Westerfield M, Townsend J, Courchesne E, Sejnowski T. Removal of eye activity artifacts from visual event-related potentials in normal and clinical subjects. Clinical Neurophysiology. 2000; 111(10):1745–1758. [PubMed: 11018488]
8. Vigario R, Sarela J, Jousmaki V, Hamalainen M, Oja E. Independent component approach to the analysis of eeg and meg recordings. IEEE Trans Biomed Eng. 2000; 47(5):589–93. [PubMed: 10851802]
9. Scholz M, Gatzek S, Sterling A, Fiehn O, Selbig J. Metabolite fingerprinting: detecting biological features by independent component analysis. Bioinformatics. 2004; 20(15):2447–54. [PubMed: 15087312]
10. Dawy Z, Sarkis M, Hagenauer J, Mueller JC. Fine-scale genetic mapping using independent component analysis, Computational Biology and Bioinformatics. IEEE/ACM Transactions on. 2008; 5(3):448–460.
11. Huang DS, Zheng CH. Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. Bioinformatics. 2006; 22(15):1855–62. [PubMed: 16709589]
12. Lee SI, Batzoglou S. Application of independent component analysis to microarrays. Genome Biol. 2003; 4(11):R76. [PubMed: 14611662]
13. Liebermeister W. Linear modes of gene expression determined by independent component analysis. Bioinformatics. 2002; 18(1):51–60. [PubMed: 11836211]

14. Raychaudhuri S, Stuart JM, Altman RB. Principal components analysis to summarize microarray experiments: application to sporulation time series. Pac Symp Biocomput. 2000:455–66. [PubMed: 10902193]
15. Saidi SA, Holland CM, Kreil DP, MacKay DJ, Charnock-Jones DS, Print CG, Smith SK. Independent component analysis of microarray data in the study of endometrial cancer. Oncogene. 2004; 23(39):6677–83. [PubMed: 15247901]
16. Klebanov L, Yakovlev A. How high is the level of technical noise in microarray data? Biol Direct. 2007; 2:9. [PubMed: 17428341]
17. Kreil DP, MacKay DJ. Reproducibility assessment of independent component analysis of expression ratios from dna microarrays. Comp Funct Genomics. 2003; 4(3):300–17. [PubMed: 18629283]
18. Särelä J, Vigário R. Overlearning in marginal distribution-based ica: analysis and solutions. Journal of Machine Learning Research. 2003; 4:1447–1469.
19. Lander AD. The edges of understanding. BMC Biol. 2010; 8:40. [PubMed: 20385033]
20. Tran LM, Brynildsen MP, Kao KC, Suen JK, Liao JC. gnca: a framework for determining transcription factor activity based on transcriptome: identifiability and numerical implementation. Metab Eng. 2005; 7(2):128–41. [PubMed: 15781421]
21. Boscolo R, Sabatti C, Liao JC, Roychowdhury VP. A generalized framework for network component analysis, Computational Biology and Bioinformatics. IEEE/ACM Transactions on. 2005; 2(4):289–301.
22. Brynildsen MP, Tran LM, Liao JC. A gibbs sampler for the identification of gene expression and network connectivity consistency. Bioinformatics. 2006; 22(24):3040–6. [PubMed: 17060361]
23. Wang C, Xuan J, Chen L, Zhao P, Wang Y, Clarke R, Hoffman E. Motif-directed network component analysis for regulatory network inference. BMC Bioinformatics. 2008; 9(Suppl 1):S21. [PubMed: 18315853]
24. Cardoso JF. Blind signal separation: statistical principles. Proceedings of the IEEE. 1998; 86(10): 2009–2025.
25. Lee TW, Girolami M, Bell AJ, Sejnowski TJ. A unifying information-theoretic framework for independent component analysis. Computers and Mathematics with Applications. 2000; 39(11):1–21.
26. Stafford P, Yidong C. Expression technology - a review of the performance and interpretation of expression microarrays. Signal Processing Magazine, IEEE. 2007; 24(1):18–26.
27. Wu J, Smith LT, Plass C, Huang TH. Chip-chip comes of age for genome-wide functional analysis. Cancer Res. 2006; 66(14):6899–902. [PubMed: 16849531]
28. Ji H, Wong WH. Computational biology: toward deciphering gene regulatory information in mammalian genomes. Biometrics. 2006; 62(3):645–63. [PubMed: 16984301]
29. Chang C, Ding Z, Hung YS, Fung PC. Fast network component analysis (fastnca) for gene regulatory network reconstruction from microarray data. Bioinformatics. 2008; 24(11):1349–58. [PubMed: 18400771]
30. Tong L, Liu, Soon VC, Huang YF. Indeterminacy and identifiability of blind identification, Circuits and Systems. IEEE Transactions on. 1991; 38(5):499–509.
31. Cruces-Alvarez SA, Cichocki A, Amari S. From blind signal extraction to blind instantaneous signal separation: criteria, algorithms, and stability, Neural Networks. IEEE Transactions on. 2004; 15(4):859–873.
32. De Bie, T.; Cristianini, N.; Rosipal, R.; Corrochano, B. Handbook of Geometric Computing : Applications in Pattern Recognition, Computer Vision, Neuralcomputing, and Robotics. Springer-Verlag; 2005. Eigenproblems in pattern recognition; p. 129-170.
33. Parra L, Sajda P. Blind source separation via generalized eigenvalue decomposition. J Mach Learn Res. 2003; 4:1261–1269.
34. Cardoso JF, Souloumiac A. Blind beamforming for non-gaussian signals, Radar and Signal Processing. IEE Proceedings F. 1993; 140(6):362–370.
35. Cardoso JF. High-order contrasts for independent component analysis. Neural Computation. 1999; 11:157–192. [PubMed: 9950728]

36. Hyvärinen A. Fast and robust fixed-point algorithms for independent component analysis. IEEE Transactions on Neural Networks. 1999; 10(3):626–634. [PubMed: 18252563]
37. Gama-Castro S, Salgado H, Peralta-Gil M, Santos-Zavaleta A, Muniz-Rascado L, Solano-Lira H, Jimenez-Jacinto V, Weiss V, Garcia-Sotelo JS, Lopez-Fuentes A, Porron-Sotelo L, Alquicira-Hernandez S, Medina-Rivera A, Martinez-Flores I, Alquicira-Hernandez K, Martinez-Adame R, Bonavides-Martinez C, Miranda-Rios J, Huerta AM, Mendoza-Vargas A, Collado-Torres L, Taboada B, Vega-Alvarado L, Olvera M, Olvera L, Grande R, Morett E, Collado-Vides J. Regulondb version 7.0: transcriptional regulation of escherichia coli k-12 integrated within genetic sensory response units (gensor units). Nucleic Acids Res. 2010; 39(Database issue):D98–105. [PubMed: 21051347]
38. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS. Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol. 2007; 5(1):e8. [PubMed: 17214507]
39. Fisher RA. The use of multiple measurements in taxonomic problems. Annals of Eugenics. 1936; 7(2):179–188.
40. He, X.; Niyogi, P. Locality preserving projections. In: Thrun, S.; Saul, L.; Scholkopf, B., editors. Advances in Neural Information Processing Systems. Vol. 16. MIT Press; 2004.
41. Guziolowski C, Veber P, Le Borgne M, Radulescu O, Siegel A. Checking consistency between expression data and large scale regulatory networks: A case study. The Journal of Biological Physics and Chemistry. 2007; 7:37–43.
42. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics. 2006; 7(Suppl 1):S7. [PubMed: 16723010]
43. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring regulatory networks from expression data using tree-based methods. PLoS One. 2010; 5(9):e12776. [PubMed: 20927193]
44. Bhardwaj N, Carson MB, Abyzov A, Yan K, Lu H, Gerstein MB. Analysis of combinatorial regulation: scaling of partnerships between regulators with the number of governed targets. PLoS computational biology. 2010; 6(5):e1000755. [PubMed: 20523742]
45. Sabatti C, James GM. Bayesian sparse hidden components analysis for transcription regulation networks. Bioinformatics. 22(6):739–746. [PubMed: 16368767]

Appendix A. Optimization of RCA criterion function

To estimate the linear extraction filter according to RCA criterion, we have

$$\widehat{\mathbf{w}}=\arg\max_{\mathbf{w}}\frac{N_-}{N_+}\frac{\mathbf{w}^T\mathbf{X}_+^{(l)}\left(\mathbf{X}_+^{(l)}\right)^T\mathbf{w}}{\mathbf{w}^T\mathbf{X}_-^{(l)}\left(\mathbf{X}_-^{(l)}\right)^T\mathbf{w}}=\arg\max_{\mathbf{w}}\frac{\mathbf{w}^T\mathbf{X}_+^{(l)}\left(\mathbf{X}_+^{(l)}\right)^T\mathbf{w}}{\mathbf{w}^T\mathbf{X}_-^{(l)}\left(\mathbf{X}_-^{(l)}\right)^T\mathbf{w}}, \tag{A.1}$$

which can be equivalently expressed as:

$$\widehat{\mathbf{w}}=\arg\max_{\mathbf{w}^T}\quad \mathbf{w}^T\mathbf{X}_+^{(l)}\left(\mathbf{X}_+^{(l)}\right)^T\mathbf{w}, \tag{A.2}$$

$$\text{s.t. } \mathbf{w}^T\mathbf{X}_-^{(l)}\left(\mathbf{X}_-^{(l)}\right)^T\mathbf{w}=1 \tag{A.3}$$

We convert (A.2) with constraint (A.3) using the Lagrange method as follows:

$$\Lambda(\mathbf{w},\lambda)=\mathbf{w}^T\mathbf{X}_+^{(l)}\left(\mathbf{X}_+^{(l)}\right)^T\mathbf{w}+\lambda\left(1-\mathbf{w}^T\mathbf{X}_-^{(l)}\left(\mathbf{X}_-^{(l)}\right)^T\mathbf{w}\right). \tag{A.4}$$

The partial derivate of Lagrange function (A.4) with respect to **w** leads to following equation:

$$\frac{\partial \Lambda(\mathbf{w}, \lambda)}{\partial \mathbf{w}^T} = \mathbf{w}^T \mathbf{X}_+^{(l)} \left(\mathbf{X}_+^{(l)}\right)^T - \lambda \mathbf{w}^T \mathbf{X}_-^{(l)} \left(\mathbf{X}_-^{(l)}\right)^T = 0 \tag{A.5}$$

or

$$\mathbf{w}^T \mathbf{X}_+^{(l)} \left(\mathbf{X}_+^{(l)}\right)^T = \lambda \mathbf{w}^T \mathbf{X}_-^{(l)} \left(\mathbf{X}_-^{(l)}\right)^T, \tag{A.6}$$

which is a generalized eigenvalue equation between $\mathbf{X}_+^{(l)} \left(\mathbf{X}_+^{(l)}\right)^T$ and $\mathbf{X}_-^{(l)} \left(\mathbf{X}_-^{(l)}\right)^T$. $\hat{\mathbf{w}}$ is the eigenvector associated with the maximum generalized eigenvalue of the above equation.

Appendix B. Proof of identifiability condition of RCA

To simplify discussions of extraction of the *l*-th component, we denote

$$K(\mathbf{w}) = \frac{1}{J_0(\mathbf{X}, \mathbf{b}_l, \mathbf{w})} = \frac{\mathbf{w}^T \mathbf{X}_-^{(l)} \left(\mathbf{X}_-^{(l)}\right)^T \mathbf{w}}{\mathbf{w}^T \mathbf{X}_+^{(l)} \left(\mathbf{X}_+^{(l)}\right)^T \mathbf{w}} \tag{B.1}$$

so that RCA estimation is equivalently expressed as follows:

$$\widehat{\mathbf{w}} = \arg \min_{\mathbf{w}} K(\mathbf{w}). \tag{B.2}$$

Denote perfect de-mixing matrix as the pseudo inverse of TFA matrix $\mathbf{A}: \mathbf{W}^{(0)} = \mathbf{A}^{\dagger} = \left[\mathbf{w}_1^{(0)}, \cdots, \mathbf{w}_L^{(0)}\right]^T$. Since noiseless case is assumed, $\left(\mathbf{w}_l^{(0)}\right)^T \mathbf{x}[n] = s_l[n]$. Additionally, since knowledge of the *l*-th TF is assumed to be perfectly given, we will have following equation for any non-zero factor *c*:

$$K(c\mathbf{w}_l^{(0)}) = 0. \tag{B.3}$$

Since the function domain of $K(\cdot)$ is $[0, \infty)$, apparently $c\mathbf{w}_l^{(0)}$ $(\forall c \neq 0)$ are optimization solutions for (B.2). This is because $s_l[n] = s_{ln} = 0$, $\forall b_{ln} = 0$. Equation (B.3) is equivalent with

$$\left(\mathbf{w}_l^{(0)}\right)^T \mathbf{X}_-^{(l)} = \mathbf{0}^T. \tag{B.4}$$

Equation (B.4) suggests optimal extraction filter $\mathbf{w}_l^{(0)}$ is the null vector of $\mathbf{X}_-^{(l)}$ space. Now we need to prove there are no any other equivalently good extraction filters $\widehat{\mathbf{w}}' \neq c\mathbf{w}_l^{(0)}$ also making $K(\hat{\mathbf{w}}') = 0$. We can prove this by contradiction:

Let us assume the rank($\mathbf{S}^{(l)}$) = $L-1$ and we have $\widehat{\mathbf{w}}' \neq c\mathbf{w}_l^{(0)}$ with $K(\hat{\mathbf{w}}') = 0$. It suggests that there is another null vector of $\mathbf{X}_-^{(l)}$ space different from $\mathbf{w}_l^{(0)}$, leading to the conclusion that

the rank of $\mathbf{X}_{-}^{(l)}$ is less or equal to $L-2$, which contradicts our assumption since $\text{rank}(\mathbf{X}_{-}^{(l)})=\text{rank}(\mathbf{A}^{(\backslash l)}\mathbf{S}^{(\backslash l)})=L-1$.

Therefore, $c\mathbf{w}_l^{(0)}$ are the only solutions so that estimated component will only differ from the true signals upon some scaling ambiguity.

Appendix C. Calculation of precision and recall values

Having the estimated value of the l-th regulatory component $[\hat{s}_l[1], \cdots, \hat{s}_l[N]]$, we define a function Descend_Sort $([|\hat{s}_l[1]|, \cdots, |\hat{s}_l[N]|], K)$, output of which is a gene index set associated with top K absolute regulatory component value. We are interested in evaluating its capability to retrieve the genes truly affected by the l-th TF according to a gene index set $\mathbb{G}_{l,\text{True}}$, which defines truly affected genes:

$$\mathbb{G}_{l,\text{True}}=\{i|b_{li}^{(0)}=1\}. \tag{C.1}$$

Precision and recall of top K genes sorted by estimated regulatory component are defined as follows, respectively:

$$Precision=\frac{\#(\mathbb{G}_{l,\text{True}} \cap \text{Descend\_ Sort}([|\widehat{s_l}[1]|, \cdots, |\widehat{s_l}[N]|], K))}{K} \tag{C.2}$$

and

$$Recall=\frac{\#(\mathbb{G}_{l,\text{True}} \cap \text{Descend\_ Sort}([|\widehat{s_l}[1]|, \cdots, |\widehat{s_l}[N]|], K))}{\#(\mathbb{G}_{l,\text{True}})}, \tag{C.3}$$

where #(.) is the operator to count number of members.

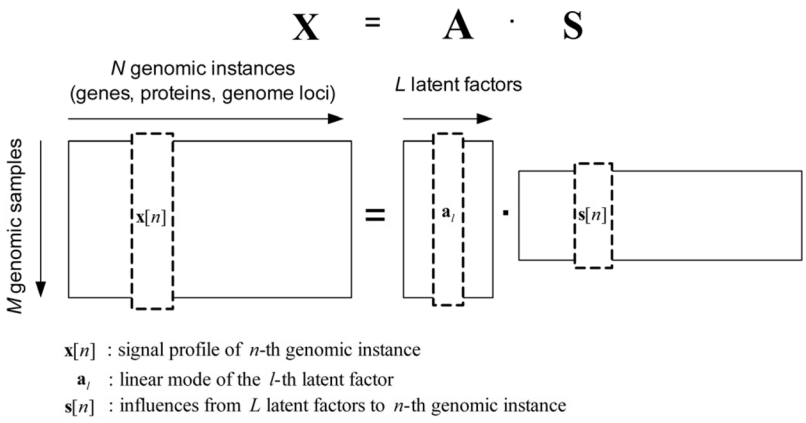

Figure 1.
General linear latent variable model for genomic signals.

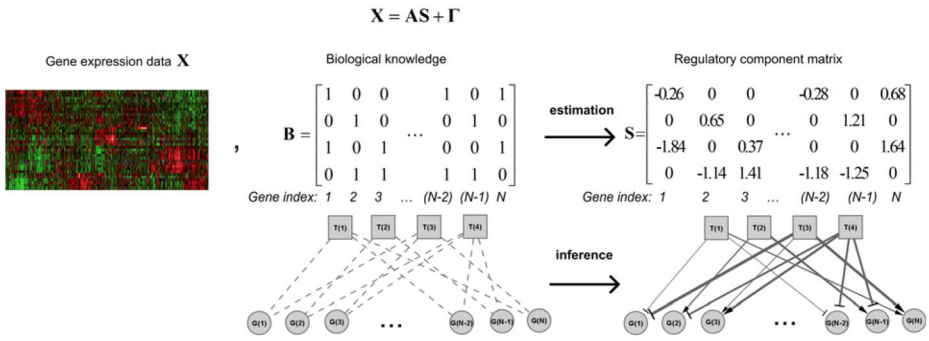

Figure 2.

Equivalence relationship of regulatory component estimation and regulatory network inference; Left side is the illustrative heatmap of gene expression data **X**; the network in the center with dashed connections represents given biological knowledge - initial network connectivity **B**; the network on the right side represents the inferred regulatory network components **S** based on given expression data and biological knowledge, where the width of edge is proportional to the absolute value of corresponding regulatory component elements, and the arrow shape of edge indicates the sign of regulatory component elements

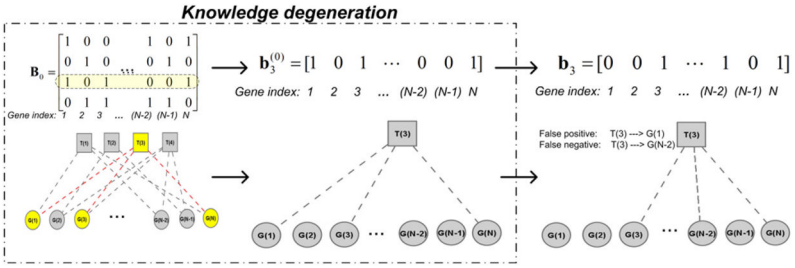

Figure 3.
Illustration of biological knowledge degeneration. The left arrows indicate incompleteness of biological knowledge, and the arrows in the center indicate that false positives and false negatives could contaminate the final knowledge we obtained.

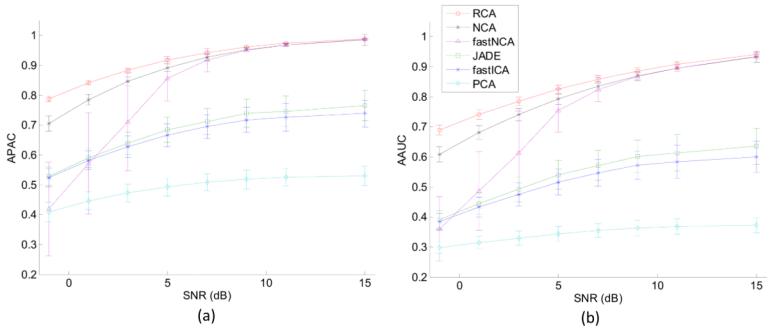

Figure 4.

Estimation performance curves for all the methods in scenario 1, where biological knowledge is perfectly given ($\mathbf{B} = \mathbf{B}_0$). (a) corresponds to the performance evaluation in averaged pair-wise absolute correlation (APAC) and (b) corresponds to the performance evaluation in Averaged Area-Under-precision-recall-Curve (APAC).

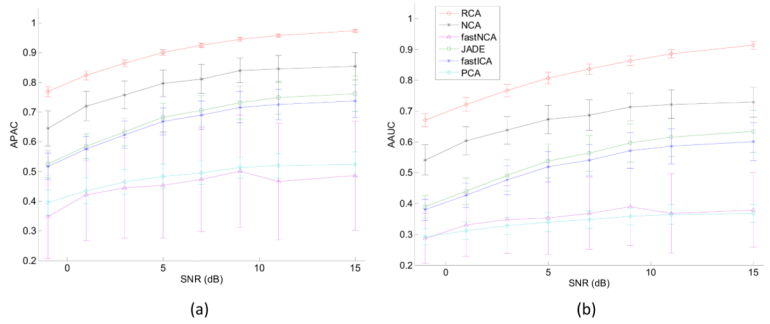

Figure 5.

Estimation performance curves for all the methods in scenario 2, where biological knowledge is imperfectly given ($\mathbf{B} \neq \mathbf{B}_0$). (a) corresponds to the performance evaluation in averaged pair-wise absolute correlation (APAC) and (b) corresponds to the performance evaluation in Averaged Area-Under-precision-recall-Curve (APAC).

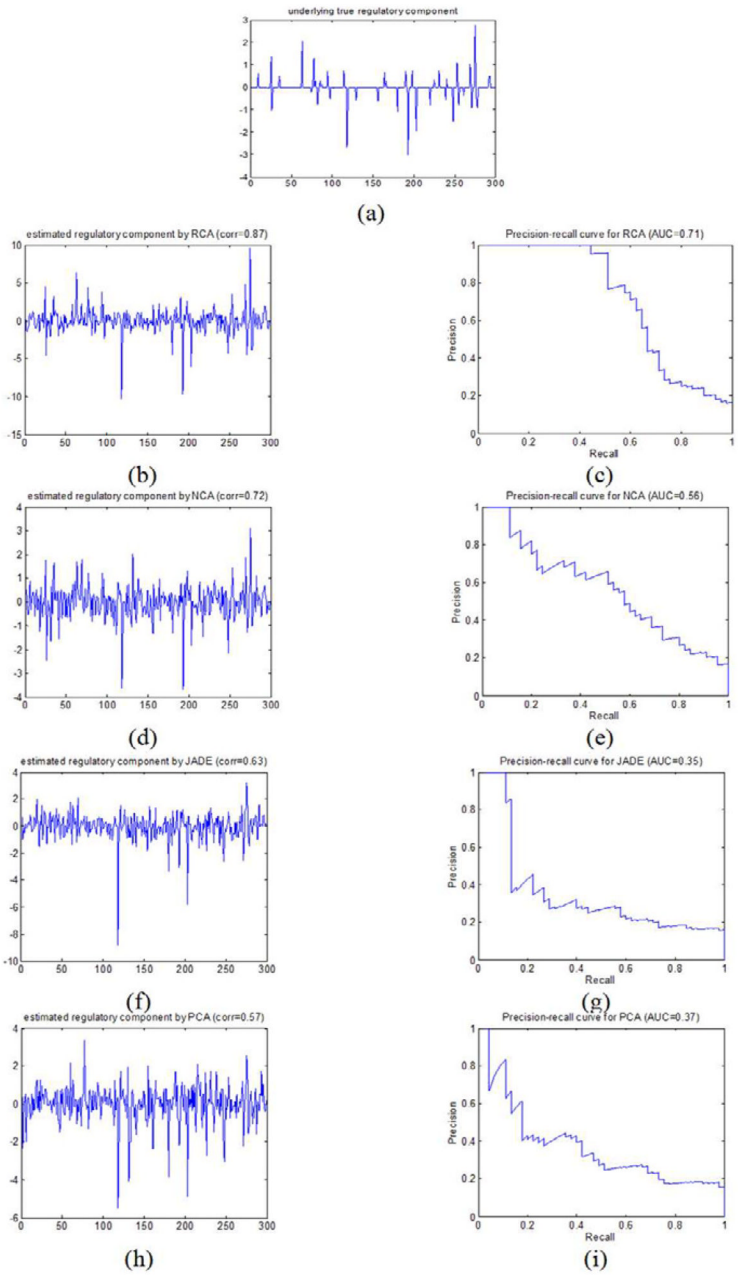

Figure 6.

Estimated regulatory component profiles and associated precision-recalling curves for retrieving the genes truly affected by corresponding TFs. (a) is the underlying true regulatory component profile; (b), (d), (f) and (h) are estimated regulatory component profiles according to RCA, NCA, JADE and PCA, respectively. (c), (e), (g), and (i) are precision-recalling curves for retrieving the genes truly affected by corresponding TF, according to RCA, NCA, JADE and PCA, respectively.

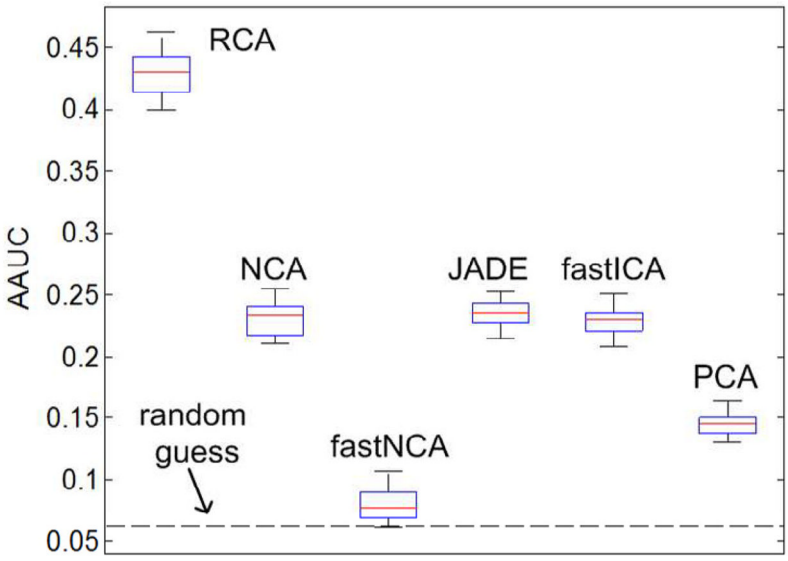

Figure 7.
Boxplots for Averaged Area-Under precision-recall Curve (AAUC). Where the red-line of each boxplot corresponds to median of all AAUC values, and top and bottom of boxplot corresponds to 75% and 25% Quantile of all AAUC values.

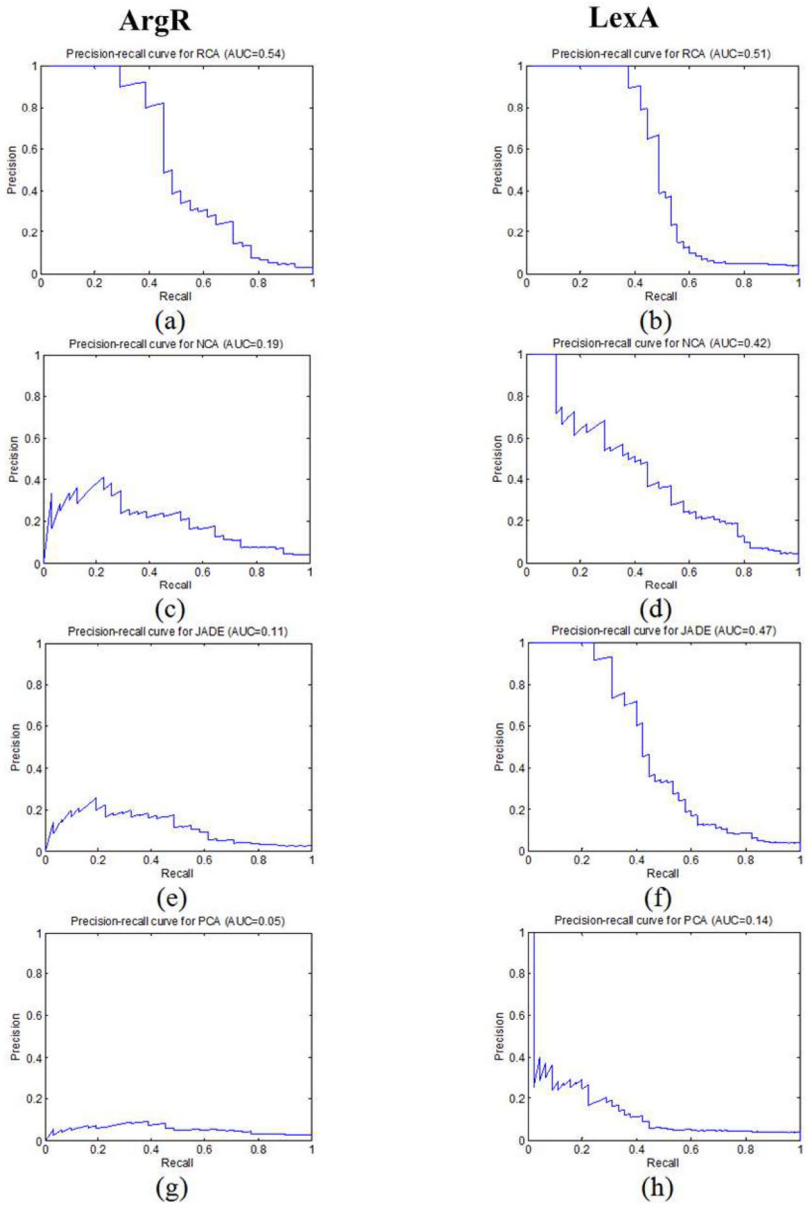

Figure 8.

Precision-recalling curves for retrieving the genes truly affected by corresponding TFs, in E.coli experiments. (a), (c), (e) and (g) are curves according to TF ArgR, by using RCA, NCA, JADE and PCA, respectively. (b), (d), (f) and (h) are curves according to TF LexA, by using RCA, NCA, JADE and PCA, respectively.