3-18-2016

Bottom-up GGM algorithm for constructing multilayered hierarchical gene regulatory networks that govern biological pathways or processes

Sapna Kupari
Michigan Technological University

Wenping Deng
Michigan Technological University

Chathura J. Gunasekara
Michigan Technological University, cjgunase@mtu.edu

Vincent Chiang
North Carolina State University at Raleigh

Huann-sheng Chen
National Institutes of Health

Kumari et al. BMC Bioinformatics (2016) 17:132
DOI 10.1186/s12859-016-0981-1

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Bottom-up GGM algorithm for constructing multilayered hierarchical gene regulatory networks that govern biological pathways or processes

Sapna Kumari[1], Wenping Deng[1], Chathura Gunasekara[1], Vincent Chiang[2], Huann-sheng Chen[3], Hao Ma[4], Xin Davis[2] and Hairong Wei[1*]

Abstract

Background: Multilayered hierarchical gene regulatory networks (ML-hGRNs) are very important for understanding genetics regulation of biological pathways. However, there are currently no computational algorithms available for directly building ML-hGRNs that regulate biological pathways.

Results: A bottom-up graphic Gaussian model (GGM) algorithm was developed for constructing ML-hGRN operating above a biological pathway using small- to medium-sized microarray or RNA-seq data sets. The algorithm first placed genes of a pathway at the bottom layer and began to construct a ML-hGRN by evaluating all combined triple genes: two pathway genes and one regulatory gene. The algorithm retained all triple genes where a regulatory gene significantly interfered two paired pathway genes. The regulatory genes with highest interference frequency were kept as the second layer and the number kept is based on an optimization function. Thereafter, the algorithm was used recursively to build a ML-hGRN in layer-by-layer fashion until the defined number of layers was obtained or terminated automatically.

Conclusions: We validated the algorithm and demonstrated its high efficiency in constructing ML-hGRNs governing biological pathways. The algorithm is instrumental for biologists to learn the hierarchical regulators associated with a given biological pathway from even small-sized microarray or RNA-seq data sets.

Keywords: Multilayered gene regulatory network, Pathway, Microarray or RNA-seq data

Background

Present knowledge indicates that genes in genomes operate in multilayered hierarchical gene regulatory networks (ML-hGRNs) to control biological processes and pathways [1–6]. A typical hGRN contains a few high hierarchical regulators, some middle-level regulators and many terminal/structural genes at bottom layer. Studies have shown high hierarchical regulators seem to be global modulators that respond to various cellular signals [7, 8] and environmental cues [3, 9]. The middle-level genes play the manager-like roles, through which the commands from high hierarchical regulators at upper layers are synthesized and then passed down to terminal genes at bottom layer for execution [3, 10]. In general, the high hierarchical regulators at the top levels have more pleiotropic effect while terminal or structural genes have more specific functions. Genes involved in various metabolic or canonical pathways are regulated by ML-hGRNs [11, 12]. To understand how pathways are regulated, we should develop methods for constructing ML-hGRNs that operate above biological pathways and processes. This kind of ML-hGRNs can provide a hierarchy in addition to connectivity of regulators, which are essential in understanding wired complex regulation on metabolic or canonical pathways through multiple chains-of-command [13].

Despite the critical importance of ML-hGRNs, there is a lack of computational algorithms for directly building

* Correspondence: hairong@mtu.edu
[1]School of Forest Resources and Environmental Science, Michigan Technological University, Houghton, MI 49931, USA
Full list of author information is available at the end of the article

ML-hGRNs from gene expression data. Reverse engineering of ML-hGRN governing a biological pathway or process remains challenging, and the algorithms that are well suited for building ML-hGRNs have not yet been established. Although several algorithms have been developed for reverse-engineering GRNs, they are not specifically tailored for constructing the ML-hGRNs that mimic the hierarchical regulation [10]. Currently, the majority of gene expression data in public repositories are static data, namely, non-time-course data or time course data with large time intervals that vary from a few hours to even several days [14]. These types of data often miss some regulatory events and interactions between two adjacent time intervals, nullifying dynamic methods that include differential equation [15], finite state [16], dynamic bayesian network [17], control logic [18], boolean network [19], and stochastic networks [20]. In general, the methods that are of more useful to static data include various static methods that comprise GGM [21], mutual information based-RN [22], Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE) [23], Context Likelihood of Relatedness (CLR) [24], C3NET [25], MI3 [26], and probabilistic based–Bayesian network [27]. Bayesian network, a probabilistic graphic model that represents a set of variables and their conditional dependencies via a directed acyclic graph, has been used to infer the "optimal" structure of a set of genes. However, due to astronomical number of possible structures and the computational complexity, approximate inference methods, such as Gibbs sampling, Metropolis–Hastings algorithms and other variant methods, instead of the original probability method of Bayesian network, were used to approximate the inference of the possible structures of GRNs [27], and these methods are capable of producing a local interacting dependence variables containing causality relationships. ARACNE is the other widely used algorithm for building graphic dependency network using mutual information-based method. It can identify both linear and non-linear dependence relationships between genes while eliminating potential indirect gene–gene interactions through implementing data processing inequality (DPI). Another information-theory-based algorithm is the Context Likelihood of Relatedness (CLR), which scores all possible pairwise interactions based on mutual information and compared that to an interaction specific background distribution. Although computationally less complicated, most of the methods that evaluate pairwise relationships among genes can easily lead to spurious relationships when the number of RNA-seq or microarray data sets is small. Finally, GGM uses the partial correlation as a measure of dependency and interaction between any pair of genes by removing effect of third one [28], allowing to distinguish direct from indirect interactions [4]. Recently, GGM with a limited-order partial correlation function, which estimates correlations conditional on one or two, but not all other genes, and has been used to infer gene networks from *Arabidopsis* [29] and yeast [30] transcript profiles. However, as aforementioned, these methods are not specifically designed and tuned for inferring ML-hGRNs.

Although mathematical models are critically important in capturing the causality relationships for reconstructing gene networks, what is also very important is the biological regulatory models, which, when integrated into mathematical models, can empower efficiency of the network construction algorithms substantially. Biological regulatory model, in this study, refers to the defined regulatory structure to which the input genes can be functionally fitted in and then evaluated as a building block during network construction process. The integration of a biologically valid regulatory model into an algorithm demands seamless design that can enhance the recognition of the causal regulatory relationships of input genes. In this study, we developed a novel algorithm, named as bottom-up GGM algorithm, specifically for reverse engineering of ML-hGRNs that govern a biological pathway or process through integration of GGM algorithm with an authentic biological regulatory model. The input files for bottom-up GGM algorithm include: (1) the transcriptomic profiles of differentially expressed genes (DEGs) involved in a known metabolic pathway, or a canonical pathway defined by a gene ontology (2) the transcriptomic profiles of differentially expressed transcription factors (TFs) or all TFs under experimental condition. We evaluated bottom-up GGM algorithm for several pathways or biological processes and found it in general performs well for constructing ML-hGRNs. We believe the algorithm can meet the great needs for constructing ML-hGRNs using small- to medium-sized gene expression data sets, and the ML-hGRNs built will be instrumental for us to understand the hierarchical regulation of many biological processes and pathways.

Results

Selection of genes for bottom-layer and top regulatory layers

We applied our bottom-up GGM algorithm to multiple pathways for reverse-engineering ML-hGRNs, each is considered to govern a given pathway or biological process. For each pathway, there were two files: one contained the gene expression profiles of pathway genes of interest, and the other file contains the expression profiles of all TF genes. The genes involved in a pathway are generally non-regulatory genes,, which can be obtained from existing annotation of metabolic pathways. Certainly, genes involved in a biological process, for example, as defined by a gene ontology that is enriched in differentially expressed genes, can be treated as a canonical pathway, and used to replace the pathway genes.

Reverse-engineering of ML-hGRN governing lignocellulosic pathway

The plant lignocellulosic pathway controls the biosynthesis of wide variety of secondary metabolic compounds including cellulose and lignin [31]. In addition to their roles in the structure and protection of the plants, cellulose and lignin have important roles in the structural integrity of plant cell walls, and the stiffness and strength of stems [32, 33].

The inputs for our bottom-up GGM algorithm include the profiles of 25 pathway genes, and 1622 TFs extracted from the 128 pooled *Arabidopsis* microarray data sets under short-day condition that is known to induce secondary wall biosynthesis. We constructed a ML-hGRN using these 25 pathway genes used as bottom layer and these TFs as candidates for top layers. The construction of ML-hGRN was a dynamic process in our pipeline software with the parameters predetermined by users. These parameters include the number of layers, significant levels of p values for correlation and partial correlation and their differences, and a percentage of genes to be kept for each layer above the bottom one. The pipeline first built the second layer immediately above the bottom (or first) layer, and then used the second layer as the bottom layer, and repeated the above procedure to obtain the third layer and so on. For this ML-hGRN (Fig. 1a), we obtained 14, 16, and 18 TFs for second, third and fourth layer respectively above the bottom layer. The detailed results that include correlations and p-values for all the interfering TFs are given in (Additional file 1: Table S1). There are 14 TFs in the second layer (Fig. 1), out of which 10 TFs (*GATA12*, *SND1*, 2, and 3, *MYB85*, *NST1 and 2*, *MYB103*, *MYB46* and *MYB58*) were positive TFs known to regulate lignocellulosic biosynthesis [34]. NAC domain proteins: NST1 [35], NST2 [36], and SND1 (also called NST3) [35, 37] are key regulators of secondary wall biosynthesis. NST1, NST2 and NST3 are key regulators involved in wall thickenings in various tissues when expressed ectopically [36, 38]. The expressions of *SND2, SND3, MYB103*, and *MYB46* are regulated by SND1 and all are developmentally associated with cells undergoing secondary wall thickening [39–41]. The SND2 regulates genes involved in secondary cell wall development in *Arabidopsis* fibres, and increases fibre cell area in *Eucalyptus* [42]. MYB46 regulates the biosynthesis pathways of cellulose, xylan, and lignin [40] and *GATA12* [43] controls xylem vessel differentiation. In the third layer, there are 6 positive TFs (MYB43, *MYB92, MYB61, MYB63, MYB86, GRF3*). The MYB63 is known to be involved in the activation of lignin biosynthetic pathway during secondary wall formation in *Arabidopsis*. The TF MYB61 controls stomatal aperture in *Arabidopsis* [44] and is required for mucilage deposition and extrusion in

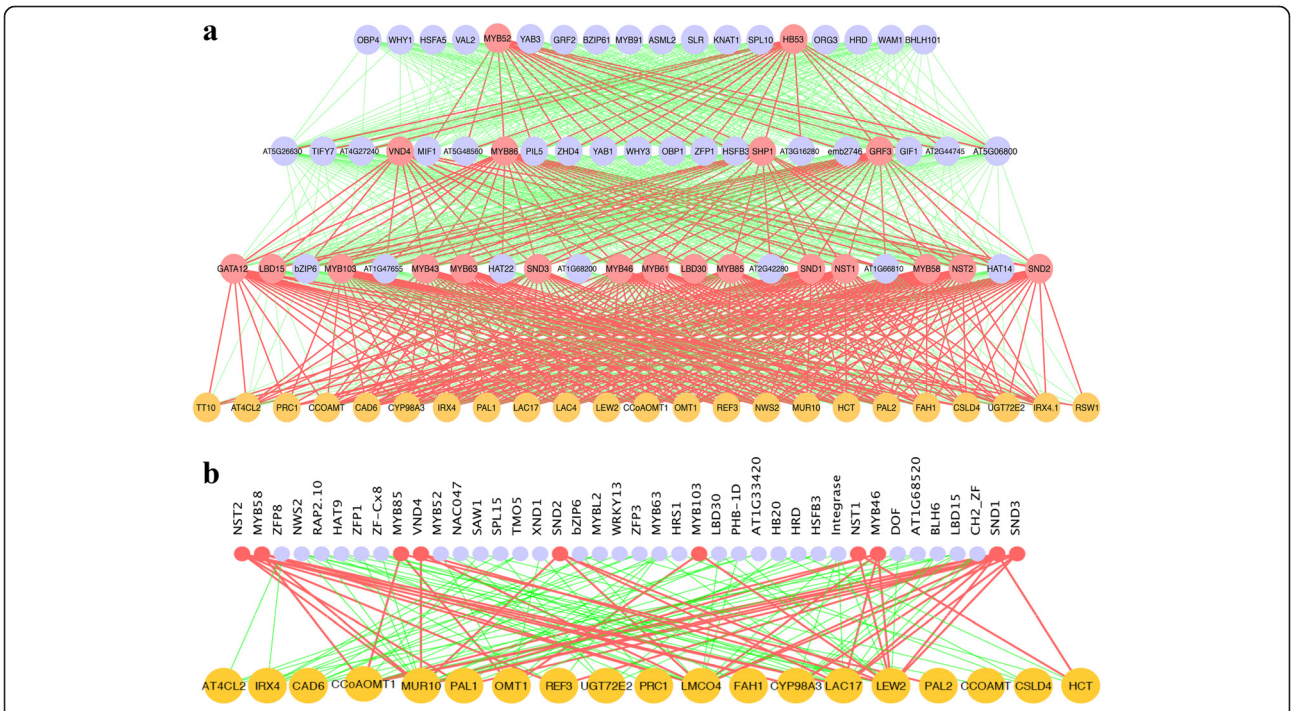

Fig. 1 Construction of ML-hGRN for lignocellulosic pathway. **a**. Four-layered hGRN built with GGM bottom-up algorithm. The gene IDs represented by each symbol can be found in Additional file 1: Table S1. **b**. The gene association network of lignocellulosic pathway built with ARACNE algorithm that identifies expression profile-dependent genes. The input files for both bottom-up GGM algorithm and ARANCE include the profiles of 1622 transcription factors and 25 lignocellulosic pathway genes. The nodes with coral (red) color highlight in both networks are known regulatory TFs for lignocellulosic pathway in existing knowledgebase. The gene IDs represented by each symbol can be found in Additional file 1: Table S2

the seed coat [45]. MYB84 regulates the accumulation of the UV-protectant compound sinapoylmalate by repressing the transcription of the gene encoding the lignocellulosic enzyme cinnamate 4-hydroxylase [46], and the *MYB43* regulates the thickening of secondary wall of cells [47]. In fourth layer, there are 18 TFs. Most appear to be the high hierarchical regulators. Some of them are known to be responsive to various environmental cues and intercellular cues, for example, SLR [48] and HB53 [49] are auxin-inducible whereas ERF38 [50] and OBP3 [51] are responsive to ethylene and salicylic acid, respectively. In this layer, the SHP1, also referred as to be SHATTERPROOF1, is known to control the differentiation of the dehiscence zone where it promotes the lignification of adjacent cells [52] while HB53 boosts vascular development in meristem [53].

Although we have indicated there are currently no computational algorithms for directly building ML-hGRNs from gene expression data, ARANCE can take the same inputs as our bottom-up GGM algorithm and obtain TFs that have dependency on pathway genes. These TFs can serve as rough controls and allow us to obtain some idea of the performance of our bottom-up GGM algorithm. We input these expression profiles of 25 pathway genes and 1622 TFs to ARANCE, and obtained the TFs that have dependency with at least two pathway genes with mutual information > 0.25 (Additional file 1: Table S2). Of the 40 TFs obtained, there are 13 positive TFs (Fig. 1B). Ten of them are common with the ML-hGRN built with bottom-up GGM algorithm. Although bottom-up GGM failed to capture three TFs, *VND4, MYB52* and *XND1*, which were recognized by ARANCE, it identified eight more positive TFs that included *GATA12, MYB86, MYB43, MYB61, GRF3, MYB92, SHP1*, and *HB53*.

Construction of ML-hGRN that controls human embryonic pluripotency renewal

There are three master transcription factors, NANOG, POU5F1, and SOX2, which are known to govern the pluripotentcy renewal in human embryonic stem cells [54]. Early studies have identified some target genes that can be bound by the above three transcription factors using ChIP-seq experiments [54, 55]. We assume that these target genes belong to a canonical pathway that plays key roles in pluripotency renewal. We would like to test if we could infer these three master transcription factors by building a one-layered hGRN using our bottom-up GGM algorithm. The 189 microarray data sets [56, 57] for human stem cells was collected from 17 experiments in which hES cells were treated with various differentiation reagents. Therefore, these datasets include states involved in many regulatory events underpinning pluripotency, such as ES maintenance, exiting the pluripotent state, and differentiation. We used 19 target genes as bottom-layer, all TFs in human as inputs, and then used our bottom-up GGM algorithm to build one regulatory layer above these 19 pathway genes. The network we obtained is shown in Fig. 2a, with 25 top genes shown in second layer. All above three transcription

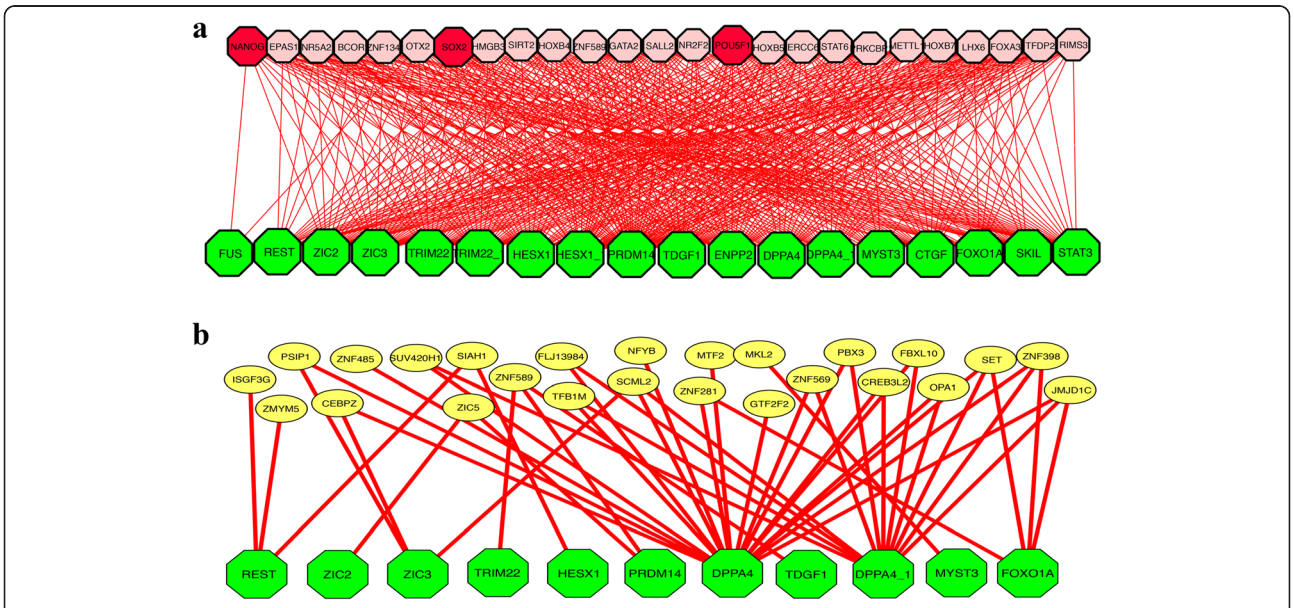

Fig. 2 Construction of ML-hGRN for putative pluripotency renewal pathway in human embryonic stem cells. **a**. Two-layered hGRN built with bottom-up GGM algorithm. The gene IDs represented by each symbol can be found in Additional file 1: Table S3. **b**. The network built with ARACNE algorithm. The input files for both bottom-up GGM algorithm and ARANCE include the expression profiles (microarrays) of 2189 transcription factors and 19 putative pathway genes identified from literature [54]. Only one layer was built and red nodes in the second layer represent the three master transcription factors known to control the pluripotency renewal. The gene IDs represented by each symbol can be found in Additional file 1: Table S4

factors were shown up in top 25 TFs captured (2λ was used). When the same inputs were used for ARANCE, and the network obtained is shown in Fig. 2b. We also kept top 25 TFs based on the mutual information (MI) on the second layer but none of above three TFs was present in these 25 genes. We searched the rankings of above three TFs in the ARANCE output sorted by MI, and found SOX2 and NANOG ranked at 68 and 154, respectively.

Implementation of the bottom-up GGM to synthetic yeast data

We generated a set of synthetic compendium gene expression data sets using the SynTReN software [58] and the regulatory network models based upon yeast experimental data as original seeds. There are 200 genes each with 100 expression values. From these 200 genes, we randomly selected 48 non-regulatory genes, which are presumably the pathway genes, and were used as bottom layer. The remaining genes are regulatory genes that contained 26 positive TFs. The results are shown in Additional file 1: Table S5. We constructed only one layer in order to make a better comparison with the result generated with ARCNE method. The list of top 50 TFs yielded from bottom-up GGM algorithm contains 25 the positive TFs, whereas, the top 50 TFs yielded from ARANCE contain 15 positive TFs, indicating the evaluation of triplet genes for causal relationships may have some advantages over pairwise evaluation used by ARANCE.

Performance comparison between the networks constructed by bottom-up GGM and ARACNE

We ran bottom-up GGM and ARANCE algorithm with the 25 wood formation genes as bottom layer, and five groups of 322 TFs as candidates for regulatory layer. Based on the number of positive TFs that were built-in regulatory layer, we calculated the sensitivities and specificities for eight different numbers of TF cutoffs, namely, the number of TFs retained in the second layer. For ARCNE, we just kept those TFs that were directly dependent on wood formation genes and counted their frequencies. After the same TF cut-offs as bottom-up GGM algorithm were applied, we calculated the sensitivities and specificities. For both bottom-up GGM and ARACNE, we plotted ROC curves (sensitivity vs (1-specificity)) (Fig. 3a and b). The dashed curves shown in Fig. 3 (a, and b) correspond to the five groups of TF inputs and the solid curve is

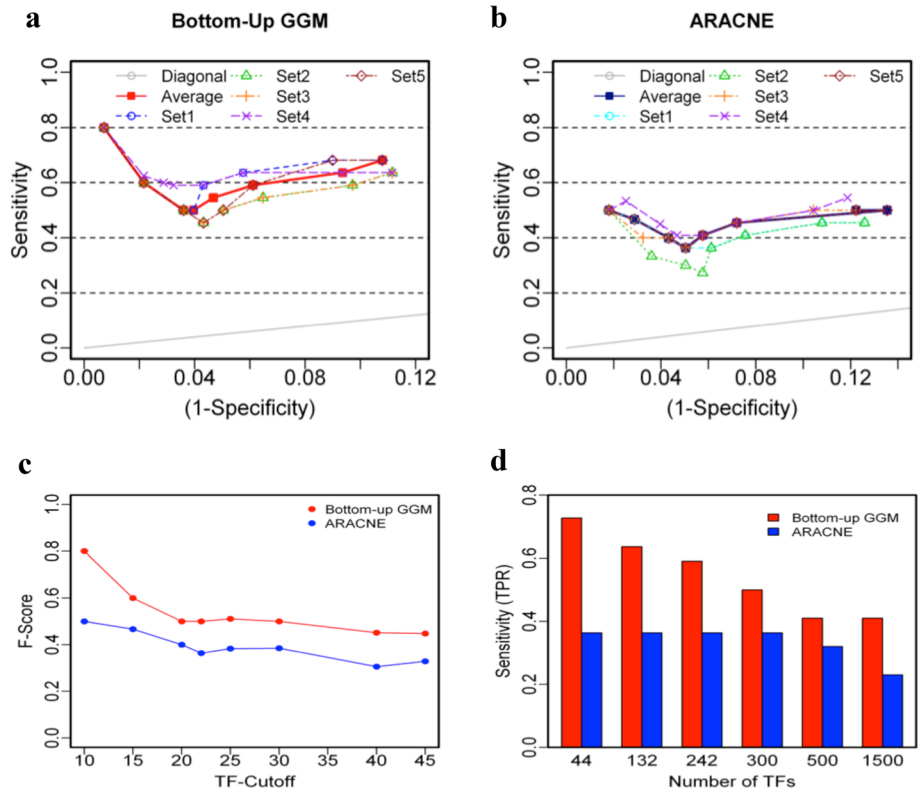

Fig. 3 The efficiency of bottom-up GGM algorithm. **a**. ROC curves of bottom-up GGM algorithm resulted from five testing data sets, each contains 300 TFs, and 25 pathway genes. **b**. ROC curves of ARACNE resulted from five testing data sets, each contains 300 TFs, and 25 pathway genes. **c**. F scores of bottom-up GGM and ARACNE in terms of different TF-cutoffs. **d**. The relationship between true positive rates (TPR) and different numbers of TFs as inputs. The TPR of ARACNE is uniform because it captured just one positive for various numbers of TF inputs varying from 44 to 1500

the average of the five curves. The cohesion of dashed curves suggests that the performance of our algorithms was persistent and did not change much with different groups of TFs. We also calculated the F-scores from the averaged curves and plotted against the different TF cutoffs for both bottom-up GGM algorithm and ARANCE (Fig. 3c). Higher F-score represents better performance. To see how the performance of the bottom-up GGM algorithm changes with the size of the data set (number of genes in the input), we calculated and plotted the True Positive Rates (TPR) against different number of TFs in the input for same TF cutoffs (Fig. 3d). We used 22 TF cutoff to find the TPR in Fig. 3d as there are 22 positive TFs in the input file. It is obvious that the bottom-up GGM has significantly higher TPR than ARACNE.

Discussion

Like GGM, we have used the first order partial correlation but adopted a biological and mathematical model integrated approach to infer ML-hGRNs. Our approach is based on a biological theory that when a TF exerted its control over a pair of genes, the correlation between these two genes will be either enhanced or impaired. The change in correlation is represented by a significant difference between the correlation of the pair of genes (in the presence of a regulatory TF, z), namely, r_{xy}, and the correlation of the same pair of genes at the absence of this TF, namely $r_{xy|z}$. We implemented a multivariate delta method [59] to test the significance of difference $d = r_{xy} - r_{xy|z}$ in four different circumstances where the null hypothesis of zero difference. The higher accuracy of method can be at least partially ascribed to the integration of biological triplet gene regulatory model, which played significantly roles in the following aspects: (1) gene (variable) reduction. By fitting differentially expressed regulatory genes and a group of pathway genes into the model based on their annotation, the dimensionality of gene space can be significantly reduced. (2) noise reduction. By filtering out the irrelevant genes that do not fit the biological model, we significantly reduced the noise, enabling the true regulatory relationships to be easily emerged. (3) reduction in gene dimensionality and explicitly defined roles of input genes in turn empowered mathematical modeling for capturing true regulatory relationships. In addition to the integration of biological regulatory model that enhanced the efficiency of the approach, the association of two paired genes under same regulatory mechanism was achieved by Spearman method, a non-parametric method that measures the strength of association between two ranked variables, which was demonstrated to be extremely well-suited for associating functionally relevant genes as compared to seven other gene association methods [57]. Spearman correlation method makes no assumption about the distribution of the data, and Spearman rho measures the strength of linear/non-linear monotonic relationship between two variables [57]. In addition, the bottom-up GGM algorithm performs exhaustive combinations of genes, providing numerous evaluation opportunities for positive genes to be eventually emerged from large number of candidates. Finally, we integrated weighted sparse canonical correlation analysis method (WSCCA) to determine the number of genes being kept for each layer, which in general could produce ideal number of regulators for each layer. To identify the TFs with higher interference frequency to pathway genes, one can double the λ value.

The efficiency of triplet gene model utilized in this study was experimentally validated in our earlier study [4]. In that study, we used a probability model and triple gene model combined approach to build a SND1-mediated two-layered hGRN in a top-down fashion with 12 RNA-seq libraries as input. Up to 97 % regulatory relationships in the built network were successfully proven by ChIP-PCR using SND1 antibodies and again verified by RT-PCR in stable transgenic lines where these targets were activated by overexpressed SND1 [4], indicating the efficiency of triplet gene regulatory model. In addition, the early version of this bottom-up GGM algorithm was implemented to nine RNA-seq libraries generated from nine independent poplar overexpression transgenic lines of microRNA397a [5]. microRNA397a targets mRNAs species of different laccase genes whose proteins catalyze polymerization reactions of S, G, and H monomers during lignin biosynthesis [5]. We constructed a three-layered hGRNs with 14 differentially expressed laccase genes as bottom-layer. The algorithm successfully identified microRNA397a from 1208 regulatory genes and constructed it into the secondary layer, where it directly regulated 12 laccase genes at the bottom layer. We chose seven laccase genes to validate using 5'RACE, and five of them were proven to be down-regulated in the microRNA397a transgenic lines. In the network built, several transcription factors that were known to govern lignin biosynthesis were recognized by our triple gene model and built into the regulatory layers. These experimental results implicate that the networks built from our bottom-up GGM algorithm are highly accurate and trustworthy.

Finally, the bottom-up GGM algorithm we developed can be used to obtain a hierarchy of TFs that are coordinated to pathway genes in expression profiles at the bottom layer. We believe these hierarchical TFs contain significantly enriched positive genes that govern the underlying pathway either directly or indirectly, allowing biologists to initiate the experimental validation. Our method is compliant with the present knowledge that when the gene expression profile of a transcription factor and the profile of target gene are correlated, it is more likely that the target genes are authentic targets [60, 61]. We employed robust Spearman-rank correlation that has

been proven to be efficiently in associating loosely and functionally coordinated genes as shown earlier [57] to augment the recognition of real hierarchical TFs that collectively control underlying pathway genes. The computational validation with test data sets has shown the positive regulatory genes can be significantly enriched in the built ML-hGRNs. We believe the algorithm is instrumental for constructing the hGRNs that govern pathways or biological processes.

Conclusions

A bottom-up graphic Gaussian model (GGM) algorithm was developed for constructing a multilayered hierarchical gene regulatory network that operates above a given metabolic or canonical pathway using small- to medium-sized microarray or RNA-seq data sets. The algorithm was validated with both synthetic and real gene expression data sets, leading to the networks that were dominated with significantly enriched known positive regulatory genes in most of the cases. We believe the algorithm is in particular instrumental for biologists to identify the hierarchical regulators associated with a given biological pathway of interest for experimental validation.

Methods

Arabidopsis and human microarray data sets

The *Arabidopsis* gene expression data used in this study were downloaded from public repository. The wood formation compendium data set contains the 128 microarrays pooled from six experiments, which have the accession identifiers of GSE607, GSE6153, GSE18985, GSE2000, GSE24781, and GSE5633, in NCBI Gene Expression Omnibus(GEO) (http://www.ncbi.nlm.nih.gov/geo/). These data sets were obtained from hypocotyledonous stems under short-day that can induce secondary wood formation [62]. The salt stress compendium microarray data set contains108 microarrays from 6 experiments with the accession identifiers of GSE5620, GSE6153, GSE24781, GSE5633, GSE6151, GSE18985) in the NCBI GEO database. This salt compendium data were used in our earlier research [56, 63]. All data sets mentioned above were derived from hybridization of Affymetrix 25 k ATH1 microarrays. The original CEL files were downloaded and processed by the robust multiarray analysis (RMA) algorithm using the Bioconductor package. For quality control we used the methods that were previously described [64]. The 189 human microarray data sets were introduced in detail in our previous publication [56].

Biological model for reverse-engingeering ML-hGRNs

Based on the fact that genes with similar patterns of transcriptomic expression are likely to be regulated via same mechanism [65–67], we proposed that when a regulatory gene in layer $i+1$ of a hGRN controls a pair of genes in layer i, *where* $i \geq 0$, the presence of this regulatory gene either significantly enhances or impairs the correlation of the paired genes at one level below, we consider this regulatory gene interfere with the paired genes. This model was integrated into the bottom-up GGM algorithm to evaluate the building blocks of combined triple genes, namely, a regulatory gene in layer $i+1$ *and* a pair of genes in layer i, during the construction of ML-hGRN using GGM.

Biological model based graphic Gaussian model (GGM) for construction of ML-hGRNs

We developed a bottom-up GGM algorithm that contains multi-step mathematical procedure to construct ML-hGRN operating above a biological process or pathway. The algorithm integrated the GGM with the biological model as described earlier. Initially, the pathway genes were placed at bottom layer, namely layer $i=1$, while regulatory genes like TFs were used as candidate genes for layer $i+1$. The algorithm evaluated each combination of triplet genes, namely one regulatory gene, and two pathway genes, to determine if the regulatory gene significantly interfered with the two pathway genes. The interference of regulatory gene on two pathway genes could enhance or impair the coordination of two pathway genes. The significance of interference was tested by examining if the significant difference existed between the correlation of two pathway genes in the presence of the regulatory gene, namely, r_{xy}, and the correlation of the same pair of genes at the absence of this TF, namely $r_{xy|z}$. We implemented a multivariate delta method [59] to test the significance of difference $d = r_{xy} - r_{xy|z}$ in four different circumstances where the null hypothesis of zero difference. The four circumstances shown in Fig. 4 are: (1) The correlation is significant but the partial correlation is not significant, indicating that the presence of the TF make the paired pathway genes more coordinated. Therefore, the TF interfered with the pathway genes pair. (2) Both the correlations are significant. The pathway genes in the pair are correlated before and after removing the effect of the transcription factor. To find if TF has significant effect on the relation of the two pathway genes, we need to determine if the difference between the correlations and partial correlation are significant. (3) The correlation is not significant but partial correlation is significant, which implies the TF is interfering. (4) Both correlation and partial correlation are not significant. In this case, we discarded the triplet and moved to the next triplet genes. Both correlation and partial correlation were calculated using Spearman Rank Correlation method. Spearman method is one of the most effective method to identify functionally associated genes

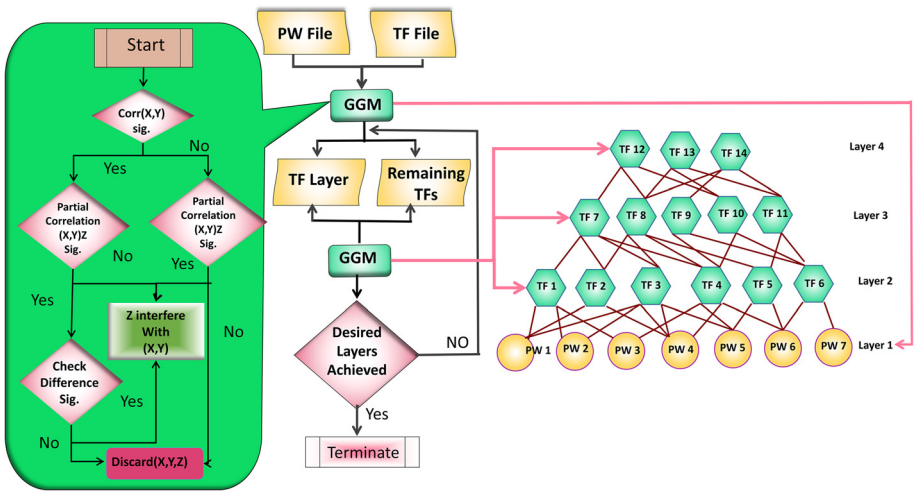

Fig. 4 The flowchart illustrating the procedure of bottom-up GGM algorithm using a group of pathway genes and all regulatory genes or a subset of regulatory genes that are significantly altered under experimental condition

using microarray data [57]. The pseudo code for the bottom-up GGM-algorithm is shown below.

Pseudo code for regulatory model based statistical model

1 Input: (G, R) [$G \rightarrow$ BPP genes, $R \rightarrow$ TFs], where BPP represents biological pathways or processes
2 For each pair (x, y), $[(x, y) \in G]$
  Calculate pairwise correlation of (x, y), $r_{xy} = cor(x, y)$, and partial correlation $r_{xy|z} = pcor(x, y|z)$ (partial correlation of x, y given z$\in Z$).
3 Perform the significance test for the correlation and the partial correlation and find the corresponding p-values.
4 Find the difference of the correlations and perform the significance test:
  Difference of correlations: $D = r_{xy} - \frac{r_{xy} - r_{zx} r_{zy}}{\sqrt{(1-r_{zx}^2)(1-r_{zy}^2)}}$
  Standard error of $D \triangleq D_{se}$
  Find the test statistics, $z_1 = \frac{D}{D_{se}}$, which follows a standard normal distribution approximately.

The difference of correlation, D, is a function of three correlations, r_{xy}, r_{yz} and r_{xz}, and denote $D = f(r_{xy}, r_{yz}, r_{xz})$. Let

$$\sum_r = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix}$$

be the covariance matrix of (r_{xy}, r_{yz}, r_{xz}).

Let (pd_1, pd_2, pd_3) denote the partial derivatives $\left(\frac{\partial D}{\partial r_{xy}}, \frac{\partial D}{\partial r_{zy}}, \frac{\partial D}{\partial r_{xz}}\right)$ of D with respect to (r_{xy}, r_{yz}, r_{xz}), then according the multivariate Delta method [59], the variance of D, is approximated by

$$var(D) \approx \sum_{i=1}^{3} \sum_{j=1}^{3} pd_i \sigma_{ij} pd_j.$$

The standard error is taken as the square root of $var(D)$. That is, $D_{se} = \sqrt{var(D)}$.

The formulae to calculate the variance and covariance among the correlations were based on asymptotic theory [68].

$$\sigma_{11} = var(r_{xy}) = \frac{(1-r_{xy}^2)^2}{n}$$

$$\sigma_{22} = var(r_{yz}) = \frac{(1-r_{yz}^2)^2}{n}$$

$$\sigma_{33} = var(r_{xz}) = \frac{(1-r_{xz}^2)^2}{n}$$

$$\sigma_{12} = cov(r_{xz}, r_{xy}) = \frac{(2r_{zy} - r_{xz}r_{xy})(1-r_{zy}^2 - r_{xz}^2 - r_{xy}^2) + r_{zy}^3}{2n}$$

$$\sigma_{13} = cov(r_{xz}, r_{zy}) = \frac{(2r_{xy} - r_{xz}r_{zy})(1-r_{zy}^2 - r_{xz}^2 - r_{xy}^2) + r_{xy}^3}{2n}$$

$$\sigma_{23} = cov(r_{zy}, r_{xy}) = \frac{(2r_{zx} - r_{zy}r_{xy})(1-r_{zy}^2 - r_{xz}^2 - r_{xy}^2) + r_{zx}^3}{2n}$$

The partial derivatives of the difference are

$$pd_1 = 1 - \frac{1}{\sqrt{1-r_{zy}^2} * \sqrt{1-r_{xz}^2}}$$

$$pd_2 = \frac{\left(r_{xz} - r_{xy} * r_{zy}\right)}{\left(\left(\sqrt{1-r_{xz}^2}\right) * \left(1-r_{zy}^2\right)^{1.5}\right)}$$

$$pd_3 = \frac{\left(r_{zy} - r_{xy} * r_{zx}\right)}{\left(\left(\sqrt{1-r_{zy}^2}\right) * \left(1-r_{zx}^2\right)^{1.5}\right)}$$

Bottom-up GGM algorithm for reconstruction of ML-hGRNs controlling a biological pathway or biological process

Based on the regulatory model and mathematical model as described, we developed the following integrated procedure to reconstruct ML-hGRNs using two inputs: the profiles of pathway genes, and the profiles of TFs. These pathway genes and TFs are in general from differentially expressed gene sets under the experimental condition in which profiles were obtained.

1. Initially, the algorithm set a group of pathway genes to be the bottom layer, and built regulatory layer immediately above the bottom layer.
2. Using the aforementioned GGM method as shown in Pseudo code for regulatory model based statistical model, we evaluated the relationships between each paired pathway gene pair from the bottom layer with a TF from TF pool. When a TF interfered (either enhanced or impaired) the coordination of two paired genes with significant p-values, keep them. Discard this TF otherwise. Continue until all TFs are evaluated.
3. When a given pathway gene pair against all TF are completed, pick up a new pair of genes from bottom layer. Repeat Step 2 and 3 until all pair of bottom-layer genes against all TFs are completed.
4. Correct the interfering p-values for multiple testing and keep the triple genes in which the TFs and two paired bottom-layer genes have corrected p-values less than the given significance level.
5. Count the total number of paired genes each TF interferes with.
6. Sort the TFs by the number of significantly interfered genes from the largest to smallest and determine the number of TFs of interest to be retained in current layer using the method given below.
7. Remove the TFs kept in current layer from the TF input file, and then set the current layer as bottom layer, and then repeat step 1-6 with the remaining TFs to obtain a new layer.
8. We repeat the above steps from 2-7 until the desired number of layers is obtained.

This framework enhances the discovery of regulatory layers operating over a pathway or a biological process by recursively evaluating and identifying candidate regulatory genes with strongest interference on layer in a layer-by-layer fashion.

Determination of TF interference on paired target genes

In the step 3 of above-mentioned procedure, we need to determine how TF interferes two paired genes. Let p_1 and p_2 be the p-values of the significance tests of correlation and partial correlation respectively. If r_{xy} and $r_{xy|z}$, are not significant i.e., the p-values p_1 and p_2 are greater than the significance level, discard (x, y, z). If r_{xy} is significant and $r_{xy|z}$ is not significant i.e., p_1 is less and p_2 is greater than the given significance level then z interferes with x and y. If both, r_{xy} and $r_{xy|z}$ are significant then test the significance of the difference between r_{xy} and $r_{xy|z}$. Let p_3 to be the p-value of the test. If there is significance difference between the correlations i.e., if p_3 is less than the pre-specified significance level, then z interferes with x and y. There is no interference of z otherwise. If p_1 is greater but p_2 is less than the given significance level, then, z interferes with x and y.

Determination of number of TFs to be kept in each layer

In order to determine how many TFs should be included in each layer, we designed a weighted sparse canonical correlation analysis method (WSCCA), which is similar to Witten's sparse canonical correlation analysis method (SCCA) [69]. For the construction of first layer network, let $X_{n \times p}$ be the pathway gene expression matrix, where n denotes the number of samples and p denotes the number of pathway genes. Let $Y_{n \times q}$ be the expression matrix of TFs we get in step 6 of above given algorithm, where q denotes the number of TFs. The matrices, X and Y are centered and scaled. Canonical correlation analysis (CCA), developed by Hetelling [70], involves finding vectors u and v that maximize $cor(Xu, Yv)$, that is

$$maximize_{u,v} u^T X^T Y v \;\; subject\ to\ u^T X^T X u = 1,\; v^T Y^T Y v = 1$$

There is a closed-form solution for canonical vectors u and v. However, u and v are not sparse, and these vectors are not unique if p or q exceeds n. SCCA use l_1 penalization for high-dimensional problems to get sparse u and v. So the optimization problem now is

$$maximize_{u,v} u^T X^T Y v + \lambda_1 ||u||_1 + \lambda_2 ||v||_1$$
$$subject\ to\ \ ||u||_2^2 = 1, ||v||_2^2 = 1$$

In our setting, as the number of pathway genes is not big, there is no need to impose l_1 penalization on vector u. We used the number of interference of each TF as weight to penalize vector v, our optimization model is

$$maximize_{u,v} u^T X^T Y v + \lambda || W^T v||_1$$
$$subject\ to\ ||u||_2^2 = 1,\ ||v||_2^2 = 1$$

Where the i^{th} element of weight vector W is maximal number of interference minus the number of interference of TF_i. Then the above model becomes a biconvex problem. We designed the following algorithm to solve it.

1. Initialize v to have l_2 norm 1.
2. Iterate until convergence:
   a. Fix v, solve $u \leftarrow argmax_u u^T X^T Y v$ *subject to* $||u||_2^2 = 1$.
   b. Fix u, solve $v \leftarrow argmax_v u^T X^T Y v + \lambda || W^T v||_1$ *subject to* $||v||_2^2 = 1$.

The tuning parameter λ is determined by cross-validation. We keep the TFs selected by this algorithm in the current layer.

Testing efficiency of the methods

We performed sensitivity, specificity and F-score analyses to assess the efficiency and quality of our algorithm. First, we built a test data set with the 129 microarray data sets, and 108 microarray data sets we used earlier [63]. The 25 pathway genes we used for initiating the ML-hGRN construction was the same as those used for building ML-hGRN governing lignocellulosic pathway and to construct the upper layers, 1500 TFs including 22 true positive TFs that are known to govern lignocellulosic pathway were used. We classified the 1500 TFs into five groups for testing the efficiency of the bottom-up GGM algorithm. Each group contained 278 randomly selected TFs from 1500 TFs. We then added 22 positive TFs to each group. These 300 TFs, together with the 25 wood formation genes, were used as inputs for building one layer of hGRN. We built one layer of hGRN and calculated the sensitivity specificity and F-score using the formulae as shown below. ARACNE was used for comparison.

$$Sensitivity(TPR) = \frac{TP}{TP + FN} \quad (1)$$

$$Specificity(SPC) = \frac{TN}{FP + TN} \quad (2)$$

$$F\text{-}score = \frac{2 \times Precision \times Recall}{(Precision + Recall)} \quad (3)$$

Where,

$$Recall = Sensitivity \quad (4)$$

$$Precision(PPV) = \frac{TP}{TP + FP} \quad (5)$$

TP, *FP*, *TN*, and *FN* are true positive, false positive, true negative, and false negative respectively.

Additional file

Additional file 1: Table S1. The output of ML-hGRN operating over lignocellulosic pathway built with the bottom-up GGM algorithm. **Table S2.** The output of regulatory layer over lignocellulosic pathway identified by ARACNE algorithm. **Table S3.** The ML-hGRN operating over human pluripotency renewal pathway built with the bottom-up GGM algorithm. **Table S4.** The output of regulatory layer over human stem cell renewal pathway identified by ARACNE algorithm. **Table S5.** Comparison of bottom-up GGM algorithm with ARANCE using yeast synthetic data. (XLSX 232 kb)

Competing interests

The authors declare that they have no competing interests.

Authors' contribution

SK and HW developed bottom-up GGM algorithm, SK wrote the methods, WD developed WSCCA, CG performed the data analysis and produced the networks, VC and XD and HM were involved in data collection, preparation and analysis, HSC helped SK in algorithm development and participated data analysis, and HW developed the project and wrote the manuscript. All authors read and approved the final manuscript.

Authors' information

SK (Research Assistant Professor), WD (Ph.D. candidate), CG (Ph.D. candidate) and HW (Associate Professor), School of Forest Resources and Environmental Science, Michigan Technological University, Houghton, MI 49931, USA. VC (Professor) and XD (Postdoctoral researcher), Forest Biotechnology Group, Department of Forestry and Environmental Resources, North Carolina State University, Raleigh, NC 27695. HSC (Mathematical Statistician), Statistical Methodology and Applications Branch, Division of Cancer Control and Population Sciences, National Cancer Institute, National Institutes of Health, Maryland, United States of America. Rockville, MD 20850, HM (Research Molecular Biologist), NCCWA, USDA ARS, Kearneysville, WV 25430

Author details

[1]School of Forest Resources and Environmental Science, Michigan Technological University, Houghton, MI 49931, USA. [2]Department of Forestry and Environmental Resources, North Carolina State University, Raleigh, NC 27695, USA. [3]Statistical Methodology and Applications Branch, Division of Cancer Control and Population Sciences, National Cancer Institute, National Institutes of Health, Rockville, MD 20850, USA. [4]NCCWA, USDA ARS, Kearneysville, WV 25430, USA.

Received: 13 August 2015 Accepted: 9 March 2016
Published online: 18 March 2016

References

1. Jin Y, Guo H, Meng Y. A hierarchical gene regulatory network for adaptive multirobot pattern formation. IEEE Trans Syst Man Cybern B Cybern. 2012; 42(3):805–16.
2. Wei H et al. Genetic networks involved in poplar root response to low nitrogen. Plant Signal Behav. 2013;8(11):e27211.

3. Wei H et al. Nitrogen deprivation promotes Populus root growth through global transcriptome reprogramming and activation of hierarchical genetic networks. New Phytol. 2013;200(2):483–97.
4. Lin YC et al. SND1 transcription factor-directed quantitative functional hierarchical genetic regulatory network in wood formation in Populus trichocarpa. Plant Cell. 2013;25(11):4324–41.
5. Lu S et al. Ptr-miR397a is a negative regulator of laccase genes affecting lignin content in Populus trichocarpa. Proc Natl Acad Sci U S A. 2013;110(26):10848–53.
6. Ma HW et al. An extended transcriptional regulatory network of Escherichia coli and analysis of its hierarchical structure and network motifs. Nucleic Acids Res. 2004;32(22):6643–9.
7. Erwin DH, Davidson EH. The evolution of hierarchical gene regulatory networks. Nat Rev Genet. 2009; 10(2):141–8.
8. Martínez-Antonio A. Escherichia coli transcriptional regulatory network. Network Biology 1. 2011;1(1):21–33.
9. Balazsi G, Barabasi AL, Oltvai ZN. Topological units of environmental signal processing in the transcriptional regulatory network of Escherichia coli. Proc Natl Acad Sci U S A. 2005;102(22):7841–6.
10. Yu H, Gerstein M. Genomic analysis of the hierarchical structure of regulatory networks. Proc Natl Acad Sci U S A. 2006;103(40):14724–31.
11. Blum A. Drought resistance - is it really a complex trait? Functional & Plant Biology. 2011;38:753–7.
12. Hasegawa PM et al. Plant Cellular and Molecular Responses to High Salinity. Annu Rev Plant Physiol Plant Mol Biol. 2000;51:463–99.
13. Bhardwaj N, Kim PM, Gerstein MB. Rewiring of transcriptional regulatory networks: hierarchy, rather than connectivity, better reflects the importance of regulators. Sci Signal. 2010;3(146):ra79.
14. Yang C and Wei H. Designing Microarray and RNA-seq Experiments for Greater Systems Biology Discovery in Modern Plant Genomics. Mol Plant 2014.
15. Chen T, He HL, and Church GM. Modeling gene expression with differential equations. Pac Symp Biocomput. 1999: p. 29-40.
16. Ruklisa D, Brazma A, Viksna J. Reconstruction of gene regulatory networks under the finite state linear model. Genome Inform. 2005;16(2):225–36.
17. Dojer N et al. Applying dynamic Bayesian networks to perturbed gene expression data. BMC bioinformatics. 2006;7:249.
18. Louis M, Becskei A. Binary and graded response in gene networks. Sci STKE. 2002;43:PE33.
19. Kauffman S. Homeostasis and differentiation in random genetic control networks. Nature. 1969;224(5215):177–8.
20. Chen BS et al. Robust model matching design methodology for a stochastic synthetic gene network. Math Biosci. 2011;230(1):23–36.
21. Schafer J, Strimmer K. An empirical Bayes approach to inferring large-scale gene association networks. Bioinformatics. 2005;21(6):754–64.
22. Butte A, Kohane I. Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements. Proc Pac Symp Biocomput. 2000;5:415–26.
23. Margolin AA et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics. 2006;7 Suppl 1:S7.
24. Faith JJ et al. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol. 2007;5(1):e8.
25. Altay G, Emmert-Streib F. Inferring the conservative causal core of gene regulatory networks. BMC Syst Biol. 2010;4:132.
26. Luo W, Hankenson KD, Woolf PJ. Learning transcriptional regulatory networks from high throughput gene expression data using continuous three-way mutual information. BMC bioinformatics. 2008;9:467.
27. Friedman N et al. Using bayesian networks to analyze expression Data. J Comput Biol. 2000;7(3/4):601–20.
28. Whittaker J. Graphical models in applied multivariate statistics. New York: John Wiley; 1990.
29. Magwene PM, Kim J. Estimating genomic coexpression networks using first-order conditional independence. Genome Biol. 2004;5(12):R100.
30. Wille A et al. Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana. Genome Biol. 2004;5(11):R92.
31. Dixon RA, Paiva NL. Stress-induced phenylpropanoid metabolism. Plant Cell. 1995;7(7):1085.
32. Chabannes M, Ruel K, Yoshinaga A, Chabbert B, Jauneau A, Joseleau JP, Boudet AM. In situ analysis of lignins in transgenic tobacco reveals a differential impact of individual transformations on the spatial patterns of lignin deposition at the cellular and subcellular levels. The Plant journal for cell and molecular biology. 2001;28:271–282.
33. Donaldson LA. Lignification and lignin topochemistry - an ultrastructural view. Phytochemistry. 2001;57:859–873.
34. Zhong R et al. A battery of transcription factors involved in the regulation of secondary cell wall biosynthesis in Arabidopsis. Plant Cell. 2008; 20(10):2763–82.
35. Mitsuda N, Ohme-Takagi M. NAC transcription factors NST1 and NST3 regulate pod shattering in a partially redundant manner by promoting secondary wall formation after the establishment of tissue identity. Plant J. 2008;56(5):768–78.
36. Mitsuda N et al. The NAC transcription factors NST1 and NST2 of Arabidopsis regulate secondary wall thickenings and are required for anther dehiscence. Plant Cell. 2005;17(11):2993–3006.
37. Kim WC et al. Transcription factor MYB46 is an obligate component of the transcriptional regulatory complex for functional expression of secondary wall-associated cellulose synthases in Arabidopsis thaliana. J Plant Physiol. 2013;170(15):1374–8.
38. Mitsuda N et al. NAC transcription factors, NST1 and NST3, are key regulators of the formation of secondary walls in woody tissues of Arabidopsis. Plant Cell. 2007;19(1):270–80.
39. Zhong R, Demura T, Ye Z-H. SND1, a NAC domain transcription factor, is a key regulator of secondary wall synthesis in fibers of Arabidopsis. The Plant Cell Online. 2006;18(11):3158–70.
40. Zhong R, Richardson EA, Ye Z-H. The MYB46 transcription factor is a direct target of SND1 and regulates secondary wall biosynthesis in Arabidopsis. The Plant Cell Online. 2007;19(9):2776–92.
41. Zhong R, Richardson EA, Ye Z-H. Two NAC domain transcription factors, SND1 and NST1, function redundantly in regulation of secondary wall synthesis in fibers of Arabidopsis. Planta. 2007;225(6):1603–11.
42. Hussey SG, et al. SND2, a NAC transcription factor gene, regulates genes involved in secondary cell wall development in Arabidopsis fibres and increases fibre cell area in Eucalyptus. BMC Plant Biology. 11(1): p. 173.
43. Endo H et al. Multiple Classes of Transcription Factors Regulate the Expression of VASCULAR-RELATED NAC-DOMAIN7, a Master Switch of Xylem Vessel Differentiation. Plant Cell Physiol. 2015;56(2):242–54.
44. Liang Y.K, et al. AtMYB61, an R2R3-MYB transcription factor controlling stomatal aperture in Arabidopsis thaliana. Curr Biol. 2005;15(13):1201–6.
45. Penfield S et al. MYB61 is required for mucilage deposition and extrusion in the Arabidopsis seed coat. The Plant Cell Online. 2001;13(12):2777–91.
46. Hemm MR, Herrmann KM, Chapple C. AtMYB4: a transcription factor general in the battle against UV. Trends Plant Sci. 2001;6(4):135–6.
47. Wang HZ, Dixon RA. On-off switches for secondary cell wall biosynthesis. Mol Plant. 2012;5(2):297–303.
48. De Smet I. Multimodular auxin response controls lateral root development in Arabidopsis. Plant Signal Behav. 2010;5(5):580–2.
49. Son O et al. Induction of a homeodomain-leucine zipper gene by auxin is inhibited by cytokinin in Arabidopsis roots. Biochem Biophys Res Commun. 2005;326(1):203–9.
50. Lasserre E et al. AtERF38 (At2g35700), an AP2/ERF family transcription factor gene from Arabidopsis thaliana, is expressed in specific cell types of roots, stems and seeds that undergo suberization. Plant Physiol Biochem. 2008; 46(12):1051–61.
51. Kang HG, Singh KB. Characterization of salicylic acid-responsive, arabidopsis Dof domain proteins: overexpression of OBP3 leads to growth defects. Plant J. 2000;21(4):329–39.
52. Colombo M et al. A new role for the SHATTERPROOF genes during Arabidopsis gynoecium development. Dev Biol. 2010;337(2):294–302.
53. Baima S et al. The arabidopsis ATHB-8 HD-zip protein acts as a differentiation-promoting transcription factor of the vascular meristems. Plant Physiol. 2001;126(2):643–55.
54. Boyer LA et al. Core transcriptional regulatory circuitry in human embryonic stem cells. Cell. 2005;122(6):947–56.
55. Sharov AA et al. Identification of Pou5f1, Sox2, and Nanog downstream target genes with statistical confidence by applying a novel algorithm to time course microarray and genome-wide chromatin immunoprecipitation data. BMC Genomics. 2008;9:269.
56. Nie J et al. TF-Cluster: a pipeline for identifying functionally coordinated transcription factors via network decomposition of the shared coexpression connectivity matrix (SCCM). BMC Syst Biol. 2011;5:53.
57. Kumari S et al. Evaluation of gene association methods for coexpression network construction and biological knowledge discovery. PLoS One. 2012; 7(11):e50411.

58. Van den Bulcke T et al. SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. BMC Bioinformatics. 2006;7:43.
59. MacKinnon DP et al. A comparison of methods to test mediation and other intervening variable effects. Psychol Methods. 2002;7(1):83.
60. Kim H, Hu W, Kluger Y. Unraveling condition specific gene transcriptional regulatory networks in Saccharomyces cerevisiae. BMC Bioinformatics. 2006;7:165.
61. Karczewski KJ et al. Coherent functional modules improve transcription factor target identification, cooperativity prediction, and disease association. PLoS Genet. 2014;10(2):e1004122.
62. Chaffey N et al. Secondary xylem development in Arabidopsis: a model for wood formation. Physiol Plant. 2002;114(4):594–600.
63. Cui X et al. TF-finder: a software package for identifying transcription factors involved in biological processes using microarray data and existing knowledge base. BMC Bioinformatics. 2010;11:425.
64. Persson S et al. Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. Proc Natl Acad Sci U S A. 2005;102(24):8633–8.
65. Clements M et al. Integration of known transcription factor binding site information and gene expression data to advance from co-expression to co-regulation. Genomics Proteomics Bioinformatics. 2007;5(2):86–101.
66. Yeung KY, Medvedovic M, Bumgarner RE. From co-expression to co-regulation: how many microarray experiments do we need? Genome Biol. 2004;5(7):R48.
67. Allocco DJ, Kohane IS, Butte AJ. Quantifying the relationship between co-expression, co-regulation and gene function. BMC Bioinformatics. 2004;5:18.
68. Olkin I, Siotani M. Asymptotic distribution of functions of a correlation matrix. In: Ikeda S, editor. Essays in Probability and Statistics Chapter 16, vol. MR0603847. Tokyo: Shinko Tsusho; 1976. p. 235–51.
69. Witten DM, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics. 2009;10(3):515–34.
70. Hotelling H. Relations between two sets of variates. Biometrika. 1936;28:321–77.