The Third International Symposium on Optimization and Systems Biology (OSB'09)
Zhangjiajie, China, September 20–22, 2009

An Information and Combinatorial Theories-based Supervised Learning Framework for Integrative Inference and Analysis of Genetic Regulatory Networks*

Binhua Tang[1] Xuechen Wu[2] Ge Tan[1]
Su-Shing Chen[3] Qing Jing[4] Bairong Shen[5,†]

[1]Department of Bioinformatics, Tongji University, Shanghai, China
[2]Institute of Protein Research, Tongji University, Shanghai, China
[3]CAS-MPG Partner Institute of Computational Biology, Shanghai, China
[4]Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, China
[5]Center for Systems Biology, Soochow University, Suzhou, China

Abstract A supervised learning framework based on information and combinatorial theories is introduced for the inference and analysis of genetic regulatory networks. First, an associativity measure is proposed to quantify the regulatory strength. Next, a phase-shift metric is defined for detecting regulatory orientations among network components. Thus, this framework can solve undirected problems from most current linear/nonlinear relevance methods. For computational redundancy, the size of the classified pair candidates is constrained within a multiobjective combinatorial optimization problem. In comparison with previously reported methods, our flexible approach can be used to efficiently identify a directed biological network that is verified by both synthetic and real-world microarray datasets having different statistical characteristics. Thus, the underlying network-designing mechanisms are deciphered by qualitative and quantitative means.

Keywords Information theory; Signal processing; Combinatorial optimization; Genetic regulatory network

1 Introduction

The phenotypes and functions of cells within multicellular organisms are directly related to the genetic contents decoded from DNA and RNA during the transcriptional and translational processes. Inference of gene regulatory networks or maps for these intracellular processes would provide a better understanding of the underlying genetic regulatory mechanisms. Thus, reconstructing regulatory networks from multisource data measured

*This study was supported by the 973 Program of NSF China (Grant No. 2007CB947002), and the Postgraduate Innovation Fund of Tongji University.
†Corresponding author: bairong.shen@suda.edu.cn

during different cell phases and in different cell types and even species has become one of the most interesting research topics in recent times.

Since simultaneous measurement of multiple expression profiles with increasing accuracy and reasonable costs is possible by using high-throughput microarrays and ChIP assays, learning and inference of regulatory maps and the functionality of these genetic networks is possible and necessary. During the last few decades, manifold inference and learning methods have been proposed that integrate raw data with computational modeling. These include Boolean network (probabilistic or dynamic) methods [1-3], systematic differential/difference equations [4-6], information theory-based modeling [7-9], and graph and control theoretic approaches [1, 10, 11].

Regulatory networks that are currently available are commonly regarded as static descriptions of inherent mechanisms. Once the models and parameters are set, the regulatory processes can be determined. During transcriptional and translational processes, real-world regulatory maps may undergo various perturbations from intercellular and intracellular signals and unknown factors. Therefore, a single model may not be sufficient for characterizing all the possible structures or even the crucial ones for specific analysis purposes. Consequently, more flexible and reasonable models are necessary to improve the present rigid network-inference methods.

In this study, we propose an integrative supervised learning framework based on information and combinatorial theories for inference of regulatory mechanisms. First, we provide brief definitions of correlation and mutual information. We then propose an associative quantity for the dependency measures. Using integrative operations on all pairwise genes from the raw dataset, one may rank the dependency/connectivity among pairwise gene candidates. Based on the signal processing theory [12-14], a phase-shift metric is then introduced for measuring the delay of expression among pairwise candidates. The underlying computational redundancies are reduced by a multiobjective combinatorial optimization (MOCO) approach.

The paper is organized as follows. Section 2 describes the methods used for building the framework. Section 3 presents the application of the proposed methods to network inference problems by theoretical analysis and experimental validation. Finally, Section 4 concludes with remarks on the proposed methods and future directions for the reconstruction of biological networks.

2 A Supervised Learning Framework for Inferring Genetic Regulatory Networks-Methodologies

The supervised learning framework mainly integrates two aspects, *i.e.*, it defines pairwise regulatory strengths and constrains subsequent computational redundancy. This section introduces a dimensionless metric for regulatory strengths and a phase-shift metric for determining regulatory orientations. For biological inference, we propose a MOCO problem for constraining the inference complexities. The framework allows the possibilities of incorporating acquired knowledge and specific analysis for integrative data mining.

2.1 Probability Theory-based Inference of Biological Network Structures

Correlation analysis aims to reveal the strength of a linear relationship between random variables (R.V.); statistical correlation (coefficient) represents the departure of two R.V. from independence. Among the various metrics often used to measure the correlation or association, the *Pearson* product-moment correlation coefficient is applicable to some data of diverse characteristics. Normally, the correlation $\rho_{X,Y}$ is denoted as the covariance of two R.V. divided by the product of their standard deviations, which can be represented as [15]

$$\begin{aligned} \rho_{X,Y} &= \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X \sigma_Y} \\ &= \frac{E(XY)-E(X)E(Y)}{\sqrt{E(X^2)-E^2(Y)}\sqrt{E(Y^2)-E^2(Y)}} \end{aligned} \tag{1}$$

where cov indicates covariance, E is the expected value operator, $\mu_X = E(X)$, and $\sigma_X^2 = E[(X - E(X))^2]$=$E(X^2) - E^2(X)$.

When interpreting the *Pearson* product-moment correlation coefficient, Cohen noted that the proposed interpretative criteria were arbitrary in general and that specific treatments should be adopted for specific cases in fields ranging from physics to social sciences [16]. Apart from the parametric statistic, nonparametric correlation metrics such as the χ^2 test, Spearman's ρ, and Kendall's τ are proposed, and these can be applied to problems with diverse nonnormal distributions [17].

2.2 Information-Theoretic Inference of Biological Network Structures

To quantify the mutual dependence of two R.V., mutual information is frequently adopted as an alternative in information theoretic application, in addition to the above measure. The mutual information of two discrete R.V. can be defined as [18]

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) log \left(\frac{p(x,y)}{p_1(x) p_2(y)} \right) \tag{2}$$

where $p(x, y)$ denotes the joint probability distribution of X and Y, and $p_1(x)$ and $p_2(y)$ represents the marginal probability distributions of X and Y, respectively. The measure normally adopts the well-defined form $I(X, Y,b)$ where b denotes the base. In general, a base of 2 can be specified since that is the common unit of the bit. Thus, for analysis within this context, we consistently use the base of 2.

2.3 Associativity Measure for Describing Regulatory Connectivities

The above-described measures illustrate the correlation and dependence relationships of R.V. Normally, these R.V. characterize different entities within a system structure. The interconnections in the biological network can be weighted by the probability of association between the pairs being investigated [19]. Since the above metrics, *i.e.*, the *Pearson* product-moment correlation coefficient and mutual information are dimensionless vector quantities, we introduce an associativity measure (AM) for illuminating the connectivities

between candidate pairs. Within this uniform measure, the quantities of mutual information and correlation metrics can be projected onto the orthogonal coordinates of a 2D plane. The metric is represented in formal terms as

$$\begin{aligned} AM_i &= w_{i1}\overrightarrow{MI_i} + w_{i2}\overrightarrow{Cor_i} = [w_{i1}MI_i] + j[w_{i2}Cor_i] = |AM_i|\angle\alpha_i \\ &= \sqrt{[w_{i1}MI_i]^2 + [w_{i2}Cor_i]^2}\angle tan^{-1}\left(\frac{w_{i2}Cor_i}{w_{i1}MI_i}\right),\ i \in \mathbf{N} \end{aligned} \tag{3}$$

where MI_i and Cor_i denote the mutual information and correlation quantities, respectively; ω_{i1} and ω_{i2} represent the weights of both quantities; α_i is the phase difference for the ith pair candidate; and N is a set of natural numbers. Note that the weights here aim to leverage any possible asymmetric distribution within the datasets of the above subterms $MI_i$and Cor_i. The weights can be derived from previously acquired knowledge or from a specific theoretical hypothesis, *e.g.*, the respective centroids of datasets.

2.4 Phase-Shift Metric for Determining Regulatory Directions

Currently, most gene expression profiles are discrete time-series data. The data samples are diverse expression densities measured at multiple time points, and the data intervals represent the sampling periods. When n samples are compared, a total of $n(n-1)/2$ pairwise comparisons are obtained. Butte *et al.* utilized a type of signal processing method to cluster and compare the similarity of expression profiles ..[12]. For every potential pairwise regulation, the activities of the investigated genes can be modularized as a subsystem. Their expression patterns might be viewed as input and output signals, as shown in **Fig. 1**.

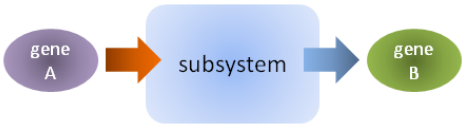

Figure 1: Each pairwise association might be modularized as a subsystem with the expression patterns serving as input and output signals.

For each pair, the coherence, gain, and phase shift might be calculated by discrete Fourier transform (DFT) of the inputs and outputs. The coherence of signals a and b is a function of the power spectral density (PSD) and the cross power spectral density (CPSD) and is defined as

$$Coh_{ab}(f) = \frac{|CPSD_{ab}(f)|^2}{PSD_{aa}(f) \cdot PSD_{bb}(f)} \tag{4}$$

where $PSD_{aa}(f)$, $PSD_{bb}(f)$, and $CPSD_{ab}(f)$ measure the PSD and CPSD of the associated pairwise signals. The symbol f represents a frequency-domain metric. Normally, signals a and b are of the same length. A coherence of 1 represents a scalar multiples relationship between two investigated signals, while 0 indicates that such a relationship is not linearly related. The transfer function (TF) between two associated input/output

signals measures the signal amplification and related time lag/latency properties, which are defined as

$$TF_{ab}(f) = \frac{PSD_{ab}(f)}{PSD_{aa}(f)} \tag{5}$$

The regular transfer functions will be of the complex-valued form, the arctangents of which are the corresponding transfer phases (TP). The absolute values denote the related transfer gains (TG) and are represented as

$$TP_{ab}(f) = \arctan[\frac{PSD_{ab}(f)}{PSD_{aa}(f)}] \tag{6}$$

$$TG_{ab}(f) = abs[\frac{PSD_{ab}(f)}{PSD_{aa}(f)}] \tag{7}$$

Theoretically, the TP illustrates the phase shift between the investigated pairwise signals, *i.e.*, the input and output. The phase shift ranges might be allocated within -π to π, where -π represents a phase lead of half a wavelength and π denotes a phase lag of half a wavelength. Whether the input signals are amplified or not is not illuminated at the output by the transfer gain and determines the related degrees at different frequencies. The larger the ratio, the less energy is lost by the output. Note that at different frequencies, the transfer phase and relative transfer gain may also differ from each other. An effective evaluation criterion for these metrics is the related coherence, *i.e.*, at frequencies where the coherence values are high, the corresponding transfer phases and gains are much more reliable than others.

The advantages of such metrics lie in the flexible and quantitative characteristics of determining the regulatory delay via dynamic thresholding. Factual regulatory mechanisms have multiple possibilities, and inherent regulatory delay effects might vary during biological processes. The phase-shift metric determines such possibilities underlying regulatory mechanisms in a quantitative manner. The advantages include the inherent capabilities of integrating *a priori* biological knowledge. This kind of knowledge-based inference method avoids redundant false-positive connectivities within pair candidates.

Such dynamic thresholding is applicable to the majority of problems facing theoretical and experimental biologists. Since regulatory connectivities underlying pair candidates may differ from each other in various processes or at different sampling times, systematic and quantitative determination of these regulations with empirical and theoretical knowledge will be much more effective than the information generated by most currently available computational approaches [6-8]. Such types of flexible network connectivities and regulations characterize major genetic regulatory processes from the perspective of information and combinatorial theories.

2.5 A MOCO Framework for Constraining Computational Complexities

In the following sections, we extract inherent regulations and decipher network structures by introducing a pairwise gene hierarchy criterion (PGHC) for classifying possible gene pairs into three major groups as follows.

(1) Authentic Pairwise Genes (APGs): These include pairs with mutual information values and correlation coefficients larger than specific thresholds. Moreover, the corresponding *P* value resides in the confidence interval, *i.e.*, it is smaller than 0.05.

(2) Questionable Pairwise Genes (QPGs): These include pairs that do not satisfy both of the thresholds mentioned above. The group contains pairs of two classes. One class has pairs with mutual information larger than specific thresholds but satisfies neither the criteria of correlation coefficients nor *P* values. The other class includes pairs with correlation coefficients larger than specific thresholds and with *P*values residing in the confidence interval but the related mutual information does not satisfy specific thresholds.

(3) Unauthentic Pairwise Genes (UPGs): These include those pair candidates that do not satisfy any criteria of the APGs or QPGs defined above.

The QPGs actually act as a subsidiary candidate pool for the APGs in case the empirical thresholds are set too high to extract structures merely from the APGs. Under such conditions, the QPGs will be ranked according to mutual information values, correlation coefficients, and *P* values. Optimal pairs will then be recruited to the APGs to refine the former network connectivities. The algorithm for the supervised PGHC is shown below.

Algorithm: Pairwise Gene Hierarchy Criterion

Input:
all pairwise gene candidates GPs;
initial MI threshold MIth = MI's centroid;
initial CC threshold CCth = CC's centroid;
increments δ_{mi}, δ_{cc} for MI and CC.
Output:
classified APGs, UPGs and QPGs.
while count(GPs)>0 **do**
1. construct APGs, QPGs using initial MIth, CCth and *P*-value;
2. group the others into UPGs;
if (APGs' undersized) && count(QPGs)>0 **then do**
MIth=MIth-δ_{mi} & CCth=CCth-δ_{cc};
continue Step 1 for QPGs & obtain Δ_{APGs} and Δ_{UPGs};
APG=APGs+Δ_{APGs} & UPGs=UPGs-Δ_{UPGs}.
elseif (APGs' oversized) **then do**
MIth=MIth+δ_{mi} & CCth=CCth+δ_{cc};
continue Step 1 for APGs & obtain Δ_{APGs} and Δ_{UPGs};
APG=APGs-Δ_{APGs} & UPGs=UPGs+Δ_{UPGs}.
endif
end

Thus, network reconstruction might be transformed into a class of MOCO problems [20-22]. The optimization objectives include first reaching suitable thresholds for mutual information and correlation coefficient to maximize the feasible components in the APGs. The inference might be carried out with much more confidence and reliability. The second objective is to maximize the UPGs. The larger the UPGs, the fewer the problems faced during further solution searching. This decreases the feasible solution space for subsequent computations. In addition, the following relative constraints exist. There are nonnegative constraints for the sizes of groups, and the total number of pair candidates is fixed, *i.e.*, the valid combinatorial space is limited. The gain thresholds for guaranteeing valid network connectivities and previously acquired biochemical knowledge and differ-

ent experimental conditions constitute other prominent constraints for the reconstruction process. The MOCO paradigm is described as follows

$$\begin{array}{lll} OBJ & : & F_i = \max_{i \in S_1} \{APGs_i,\ UPGs_i\} \\ \\ s.t. & 1. & APGs_i \geq 0,\ QPGs_i \geq 0,\ UPGs_i \geq 0,\ i \in S_1; \\ & 2. & \sum(APGs_i + QPGs_i + UPGs_i) \in S_2; \\ & 3. & \{GC_i\} \subset S_3; \\ & 4. & \{ABK_i\} \subset S_4. \end{array} \tag{8}$$

where $F_i$is the multiobjective function set; S_1 is the set of feasible group combinations for APGs, QPGs, and UPGs; S_2 is the number set of all gene pairs (S_2= $\{n(n\text{- }1)/2\}$, n is the total number of genes); S_3 is the set of necessary gain constraints (GC); and S_4 is the set of possible constraints from acquired biological knowledge (ABK).

Recently quite a few authors have argued the necessity of incorporating the preferences of decision-maker (DM) into MOCO solution selection .[21, 22]. For the problem under investigation, the DM's preferences mainly stem from the GC (S_3) and ABK (S_4) illustrated above.

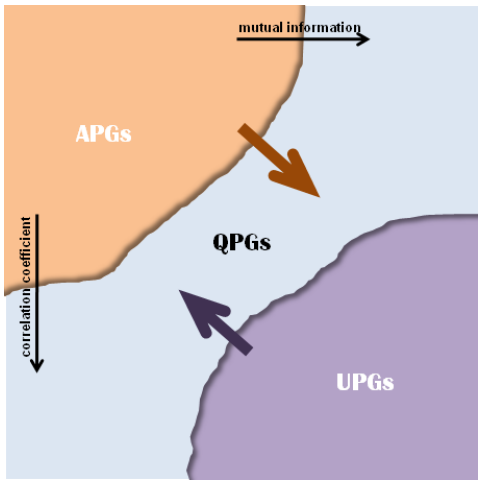

Figure 2: Schematic representation of the MOCO problem by dynamic thresholding of mutual information and correlation metrics. Total pairs are classified into APGs, QPGs, and UPGs. The upper rightward horizontal arrow represents dynamic thresholding by mutual information, and the left descending arrow is for thresholding of the correlation measure.

In cases governed by lower thresholds of mutual information and correlation metrics, APGs will form the group with the maximum components within the total pair candidates. On the other hand, with the heightened thresholds, many more pairs might be grouped into UPGs. This reduces the computational complexity for network reconstruction since APGs have fewer components in such situations. If APGs are classified with above-normal sizes, the reconstructed network will be densely connected and will have much

more redundancies. On the contrary, a sparsely connected structure will be inferred with an undersized candidate group of APGs.

Since biological theoreticians and experimentalists may vary specific mutual information and correlation thresholds to incorporate empirical or concrete knowledge into the reconstruction procedures, the underlying coordination approaches via the MOCO framework might be feasible and significant, especially for those containing pivotal structural connectivities or for specific analysis purposes.

The APGs, QPGs, and UPGs engender the underlying evolutionary mechanisms with respect to dynamic thresholding by the above metrics and related biochemical knowledge, as shown in **Fig. 2**.

3 Experiments and Analysis of a Synthetic Dataset from a Typical Mammalian G_1/S Cell Cycle Transition Network

We validated the proposed methods using three types of datasets. The first is synthesized from a cell cycle network by Swat *et al.* [10, 23], while the other datasets of different statistical properties are collected from the literature. For simplicity, we use the methods illustrated above to describe the essential procedures for analyzing the synthetic dataset, determining pairwise regulatory information, and modeling regulatory networks. Two other datasets are provided in the supplementary file.

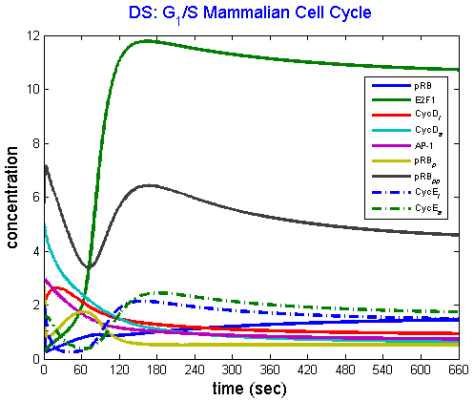

Figure 3: The expression profile of nine genes in the cell cycle transition network. The horizontal coordinate represents sampling time, and the vertical indicates expression concentrations.

The G_1/S transition network consists of nine components, *i.e.*, pRB (retinoblastoma, a tumor suppressor in the pocket protein family), E2F1 (a transcription factor that targets genes that regulate the cell cycle), CycD*i*(the inactive form of the cyclin D/cdk4, 6 complex), CycD*a*(the active form of the cyclin D/cdk4, 6 complex), AP-1 (a family of transcription factors that mediate mitogenic signals), pRB*p*(the phosphorylated form of pRB), pRB*pp* (the double-phosphorylated form of pRB), CycE*i* (the inactive form of the

cyclin E/cdk2 complex), and CycE*a* (the active form of the complex cyclin E/cdk2). The dataset is sampled every 60 s and covers 12 sampling points within a time range of 11 min. The time course is shown in **Fig. 3**, and the nine genes/proteins are listed in the legend. We next calculate the mutual information, correlation coefficients, and *P* values of pairwise genes for constructing the hypothetical cell cycle regulatory network (shown in the supplementary figures).

Through dynamic thresholding of mutual information and correlation coefficient, one may conveniently obtain the global statistical distribution for groups of different sizes using these metrics. Fig. 4 shows the classified groups.

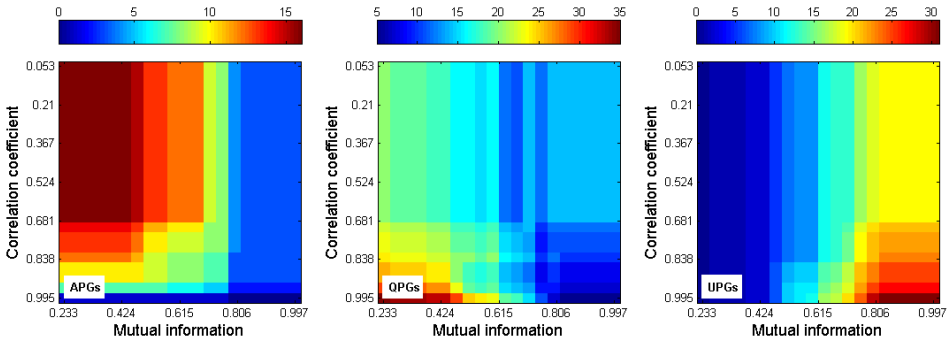

Figure 4: Global statistics for pair candidates via dynamic thresholding of mutual information and correlation coefficient. The *P* value adopts 0.05. In total, 36 pairs are obtained from nine genes/proteins. The horizontal axis represents different mutual information thresholds, while the vertical axis shows the correlation coefficient.

From the global statistics shown above, we may compare different combinatorial optimizations for the groups. For the APGs, we find that the optimal combination for the mutual information and correlation coefficient thresholds is below 0.75 when the size of the APGs is approximately more than 12 (brown zone on the left plot). As shown in the middle one, the QPGs mainly range from 10 to 20. For the UPGs, if the mutual information threshold is below 0.4, the size of the UPGs will be as small as zero or so under any correlation coefficient threshold. However, if the mutual information threshold is above 0.7, the size of the UPGs will increase up to 20 or even 30. Thus, one may conclude the sizes of APGs and UPGs are less sensitive to correlation coefficient thresholds than to mutual information thresholds when the mutual information thresholds operate below 0.3 or above 0.75.

Such information acts as a guide for further supervised inference. First, we set the centroids, *i.e.*, 0.6620 and 0.5957, as the thresholds for mutual information and correlation coefficient, respectively; the increment is 0.05 for both metric values. Next, we determine that the classified APGs are undersized for the reconstruction; therefore, we decrease the thresholds with the given increment.

After several iterative operations and using the inherent biological knowledge, we obtain the relevant APGs, QPGs, and UPGs. The corresponding thresholds for mutual information and correlation coefficient decrease to 0.52 and 0.55, respectively. Supple-

mentary Tables 1 and 2 list the details of APGs and QPGs, and the related gene/protein names are shown in **Fig. 8**. There are 14 candidate pairs in the APGs, 15 in the QPGs, and only 7 in the UPGs.

Once the basic groups are determined, the regulatory direction should be the uppermost issue to challenge further analysis and validation. To determine the systematic phase shifts and time lags for deciding the regulatory orientation, we utilize the signal processing concepts defined above. Without the loss of generality, we consider the pair E2F1-pRBp that has been randomly selected from the APGs. First, we calculate coherence, transfer gain, and phase shift via the 6-point DFT on the related expression profiles. Next, we derive related coherence, transfer gain, and phase-shift metrics at four frequencies, *i.e.* 0, 0.0028, 0.0056, and 0.0083 Hz.

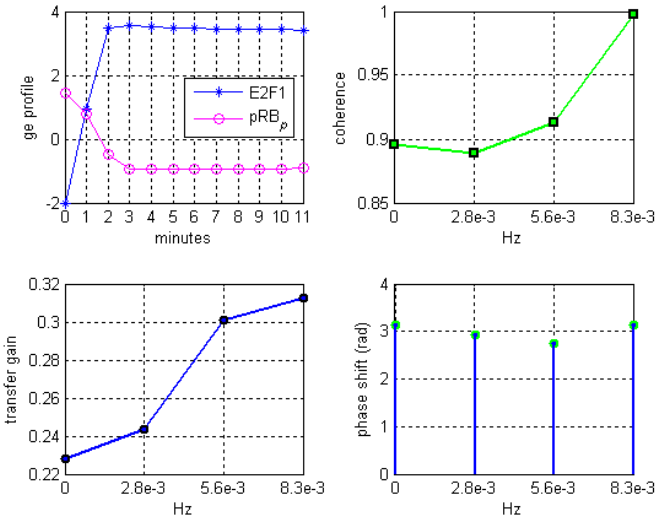

Figure 5: The expression profile of the E2F1-pRBp pair and the related coherence, transfer gain, and phase-shift graphs. The sampling frequency for the profile signals is 1/60 Hz. The metric values are estimated at four frequencies using the 6-point DFT.

On the two lower subgraphs in **Fig. 5**, each phase shift corresponds to a specific transfer gain at the related frequencies. For instance, the phase shift is 2.7496 rads at 0.0056 Hz, and its related transfer gain is 0.3011. If we set 0.3 as the gain threshold, we may acquire two valid phase-shift values at frequencies of 0.0056 Hz and 0.0083 Hz since there are only two gain values at frequencies larger than the threshold. We then average the two phase-shift values and denote the mean value as the corresponding time lag for the underlying regulatory process. Finally, we may assign different signs to each lag (averaged phase shift) derived above, *i.e.*, positive (+) for leading phase shift, undirected (0) for zero phase shift, and negative (-) for lagging phase shift. By means of the systematic phase-shift measure, the regulatory directions might be directly elucidated from the previously determined candidate groups.

Furthermore, from the descriptions given above, we observe that the different gain thresholds play significant roles in determining global pairwise phase shifts and regulatory

orientation. We might define the phase shifts among gene pairs as the functions of the gain thresholds, which we might stipulate according to different experimental situations and empirical knowledge, and then plot the global phase-shift distribution as shown in **Fig. 6**.

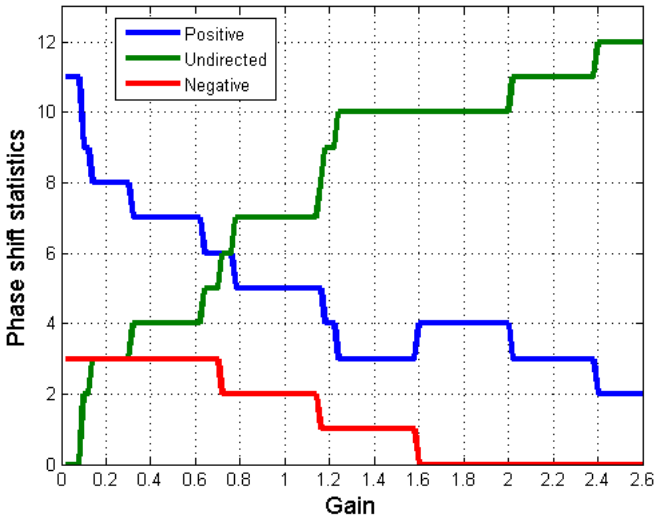

Figure 6: The global phase-shift statistical distribution (with totally14 pairs from the APGs). The phase-shift statistics vary as functions of the gain thresholds. The blue curve represents the integral tendency of pairs with leading phase shifts (positive), red denotes those with lagging phase shifts (negative), and green is for those without any detected phase shift (undirected), *i.e.*, there might be no regulatory activities between corresponding pairs (the same as in the following figures).

From the above-described phase-shift statistics distribution, when one enlarges the gain threshold, the undirected gene pairs will also increase gradually, and the pairs with leading and lagging phase shifts will decrease in jumps and rest at the extreme gain threshold of approximately 2.4. Moreover, there will be no change if the gain is still enlarged. The other extreme gain threshold is 0.1, shown in the statistics distribution illustrated above. Through dynamic gain thresholding, one may easily determine concrete regulatory time lags, directions, and intensities from the quantitative signal processing perspective. See **Fig. 7** for details on the APGs.

We observe certain interesting phenomena in **Fig. 7**. Pairs with both high mutual information and correlation are not necessarily also candidates with strong connectivities under relatively high gains. For example, in pairs 1 to 4 (between E2F1, CycD_a, CycD_a, AP-1, and pRB_p), the phase-shift information permanently changes from +1 to 0 once the gain threshold increases to 0.4. These pairs might be classified as candidates with weak-gain connectivities. In pairs 5 (CycD_i, CycD_a), 8 (CycE_i, CycE_a), 11 (pRB, CycD_i), and 14 (pRB_{pp}, CycE_a), the delay information changes to zero around the gain of 1. In the case of pairs 9 (pRB_p, CycE_a), 10 (CycD_i, CycE_a), 12 (AP-1, pRB_p), and 13 (CycD_a, pRB_p), the delay signs are maintained even when the threshold reaches values as high as 2 or so. We may call such candidate pairs as pairs with strong-gain connectivities.

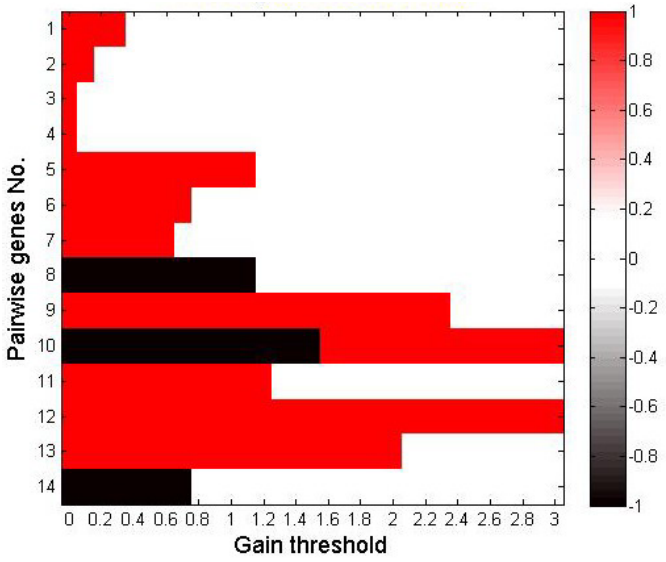

Figure 7: Phase-shift statistics for the APGs (a total of 14 pairwise genes are sorted in a descending order based on mutual information values) calculated on the basis of the signal-processing concepts defined above. The red area (+1) represents the leading phase shift, black (-1) denotes the lagging phase shift, and white shows pairs without any phase shift under specific gain thresholds.

We randomly preset the gain threshold at 0.5. This indicates that during the regulatory process, there is only half transfer loss from each potential signal source to its anticipated destination. Under this condition, E2F1 and pRB_{pp} are isolated from the reconstructed network under the current gain. The acquired knowledge from the literature requires all nine genes to interact within the regulatory network. Thus, to include the two isolated genes into the reconstructed network, we decrease the gain threshold slightly to 0.3. **Fig. 8** shows the reconstructed network.

4 Results and Discussion

We propose an information and combinatorial theories-based learning framework for inference and analysis of genetic networks from microarray datasets. Considering the acquired knowledge, possible preferences of DMs, and practical computational constraints, network inference might be transformed into a type of MOCO problem.

For different kinds of microarray datasets collected from multiple organisms and species, there is still no efficient solution applicable to most problems facing biological theoreticians and experimentalists. In comparison with currently available methods, the associative approaches allow the possibilities of incorporating concrete theoretical and empirical knowledge and thus construct regulatory networks with more reliability and accuracy than ever. Moreover, different regulatory models should focus on specific perspectives and utilities adopted by the builders; thus, inherent complexities from the

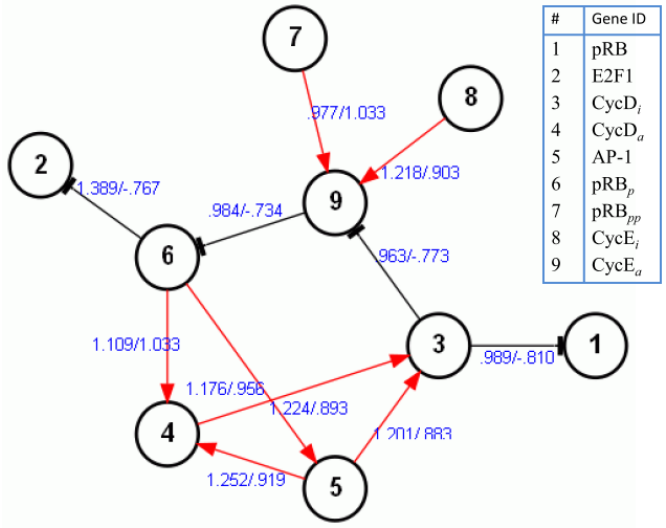

Figure 8: The reconstructed genetic network with a gain threshold at 0.3. Each black-circled node represents the relevant gene/protein, the red arrow denotes activation, while the black tee indicates inhibitory action among bilateral entities. The blue labels along the links describe the respective pairwise associativity measures (see supplementary **Fig. 1-C** for details).

inference procedures and the necessity to optimize the results appeal to such associative relevance metrics and MOCO methods.

Including or excluding specific nodes from the reconstructed networks with sufficient confidence, possible DMs' preferences and previously acquired knowledge provides several design approaches within the proposed framework. Through this study, we can decipher the underlying designing mechanisms of pairwise connectivities by dynamic thresholding of linear/nonlinear relevance metrics. We also determine regulatory orientations among genetic networks with signal processing metrics. With the inference procedure transposed into a MOCO problem, we might constrain the multiobjective iterative searching complexities with reasonable considerations from acquired knowledge, experimental conditions, and other computational limits or from the preferences of DMs.

Finally, we utilize the proposed methods to analyze a synthetic cell cycle dataset and two other microarray datasets of different statistical characteristics. For the sake of simplicity, we validate the approach on a few small-scale datasets. Different clustering and classification methods are beneficial and necessary for preprocessing some large-scale datasets, *e.g.*, those with more than hundreds of gene/proteins. Thus, by qualitative and quantitative means, we reveal the inherent designing mechanisms for genetic networks, facilitating further theoretical analysis and experimental design of specific biochemical purposes.

Availability

Supplementary material is available at http://sites.google.com/site/bhtangsite/.

Acknowledgments

We thank Dr. Li He for his constructive suggestions.

References

[1] Ching, W.K., Zhang, S.Q., Jiao, Y., Akutsu, T., Tsing, N.K., Wong, A.S.: Optimal control policy for probabilistic Boolean networks with hard constraints. IET Systems Biology, **3** (2009) 90-99

[2] Troyanskaya, O.G., Dolinski, K., Owen, A.B., Altman, R.B., Botstein, D.: A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). Proc. Natl. Acad. Sci. USA **100** (2003) 8348-8353

[3] Zou, M., Conzen, S.D.: A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. Bioinformatics **21** (2005) 71-79

[4] Perkins, T.J., Hallett, M., Glass, L.: Inferring models of gene expression dynamics. Journal of Theoretical Biology **230** (2004) 289-299

[5] Tiana, G., Krishna, S., Pigolotti, S., Jensen, M.H., Sneppen, K.: Oscillations and temporal signalling in cells. Physical Biology **4** (2007) R1-R17

[6] Wang, Y., Joshi, T., Zhang, X.-S., Xu, D., Chen, L.: Inferring gene regulatory networks from multiple microarray datasets. Bioinformatics **22** (2006) 2413-2420

[7] Zhao, W., Serpedin, E., Dougherty, E.R.: Inferring connectivity of genetic regulatory networks using information-theoretic criteria. IEEE/ACM Trans. Comput. Biol. Bioinformatics **5** (2008) 262-274

[8] Meyer, P.E., Kontos, K., Lafitte, F., Bontempi, G.: Information-theoretic inference of large transcriptional regulatory networks. EURASIP Journal on Bioinformatics and Systems Biology (2007)

[9] Schneidman, E., Still, S., II, M.J.B., Bialek, W.: Network information and connected correlations. Phys. Rev. Lett. **91** (2003) 238701-238704

[10] Tang, B., He, L., Jing, Q., Shen, B.: Model-based identification & adaptive control of the core module in a typical cell cycle pathway via network and system control theories. Advances in Complex Systems **12** (2009) 21-43

[11] Huber, W., Carey, V., Long, L., Falcon, S., Gentleman, R.: Graphs in molecular biology. BMC Bioinformatics **8** (2007) S8

[12] Butte, A.J., Bao, L., Reis, B.Y., Watkins, T.W., Kohane, I.S.: Comparing the similarity of time-series gene expression using signal processing metrics. Journal of Biomedical Informatics **34** (2001) 396-405

[13] Dougherty, E.R., Shmulevich, I., Bittner, M.L.: Genomic signal processing: the salient issues. EURASIP J. Appl. Signal Process. (2004) 146-153

[14] Candy, J.V.: Model-based signal processing. John Wiley & Sons, Inc., Hoboken, New Jersey (2006)

[15] Papoulis, A.: Probability, random variables, and stochastic processes. McGraw-Hill, New York (1984)

[16] Cohen, J.: Statistical power analysis for the behavioral sciences. Lawrence Erlbaum Associates, Hillsdale, New Jersey (1988)

[17] Simon, M.K.: Probability distributions involving Gaussian random variables. Springer, New York (2002)

[18] Yao, Y.Y.: Information-theoretic measures for knowledge discovery and data mining. In: Karmeshu (ed.): Entropy Measures, Maximum Entropy Principle and Emerging Applications. Springer (2003) 115-136

[19] Forst, C.V., Schulten, K.: Phylogenetic analysis of metabolic pathways. Journal of Molecular Evolution **52** (2001) 471-489

[20] Liefooghe, A., Basseur, M., Jourdan, L., Talbi, E.-G.: Combinatorial optimization of stochastic multi-objective problems: an application to the flow-shop scheduling problem. Evolutionary Multi-Criterion Optimization (2007) 457-471

[21] K*ö*ksalan, M.: Multiobjective combinatorial optimization: some approaches. Journal of Multi-Criteria Decision Analysis (2009)

[22] Jaszkiewicz, A.: Genetic local search for multi-objective combinatorial optimization. European Journal of Operational Research **137** (2002) 50-71

[23] Swat, M., Kel, A., Herzel, H.: Bifurcation analysis of the regulatory modules of the mammalian G_1/S transition. Bioinformatics **20** (2004) 1506-1511