BMC Bioinformatics

Research article

Modeling human cancer-related regulatory modules by GA-RNN hybrid algorithms

Jung-Hsien Chiang*† and Shih-Yi Chao†

Address: Department of Computer Science and Information Engineering, National Cheng Kung University, Tanian, Taiwan

Email: Jung-Hsien Chiang* - jchiang@mail.ncku.edu.tw; Shih-Yi Chao - chaosy@cad.csie.ncku.edu.tw

* Corresponding author †Equal contributors

Published: 14 March 2007

BMC Bioinformatics 2007, **8**:91 doi:10.1186/1471-2105-8-91

Received: 31 October 2006

Accepted: 14 March 2007

This article is available from: http://www.biomedcentral.com/1471-2105/8/91

Abstract

Background: Modeling cancer-related regulatory modules from gene expression profiling of cancer tissues is expected to contribute to our understanding of cancer biology as well as developments of new diagnose and therapies. Several mathematical models have been used to explore the phenomena of transcriptional regulatory mechanisms in *Saccharomyces cerevisiae*. However, the contemplating on controlling of feed-forward and feedback loops in transcriptional regulatory mechanisms is not resolved adequately in *Saccharomyces cerevisiae*, nor is in human cancer cells.

Results: In this study, we introduce a Genetic Algorithm-Recurrent Neural Network (GA-RNN) hybrid method for finding feed-forward regulated genes when given some transcription factors to construct cancer-related regulatory modules in human cancer microarray data. This hybrid approach focuses on the construction of various kinds of regulatory modules, that is, Recurrent Neural Network has the capability of controlling feed-forward and feedback loops in regulatory modules and Genetic Algorithms provide the ability of global searching of common regulated genes. This approach unravels new feed-forward connections in regulatory models by modified multi-layer RNN architectures. We also validate our approach by demonstrating that the connections in our cancer-related regulatory modules have been most identified and verified by previously-published biological documents.

Conclusion: The major contribution provided by this approach is regarding the chain influences upon a set of genes sequentially. In addition, this inverse modeling correctly identifies known oncogenes and their interaction genes in a purely data-driven way.

Background

A regulatory module is a set of genes that is regulated or co-regulated by one or more common transcription factors (TFs). A TF is a protein that binds to a cis-regulatory element (e.g. an enhancer, a TATA box) and thereby, directly or indirectly, positively or negatively affects the initiation of transcription of regulated genes. A cancer-related regulatory module is a set of genes (oncogenes or tumor suppressor genes) that is regulated by one ore more common TFs. Modeling the cancer-related regulatory modules of the cell division cycle in human cells is a critical and fundamental step toward understanding cancers. The aim of this paper is not only to drive cancer-related regulatory modules, but also to identify the relationships of regulations between genes that fit the feed-forward or feedback influences. A feed-forward regulatory module,

contains a TF that controls a second TF at later time points and has the additional feature that both TFs bind to common target genes. Therefore, the major contribution of this study is regarding the chain influences upon a set of genes sequentially. That is, to construct a simple cancer-related regulatory pathway with feedback loop and feed-forward controlled relationships achieved by modified Recurrent Neural Network (RNN) architecture [1]. Combining modified multi-layer RNN with the global searching ability of Genetic Algorithms (GA) [2], this approach can efficiently select regulated target genes as well. We also provide the solution of analysis time-course gene expression data. For example, one particular TF expressed highly in S/G1 phase may regulate its target genes expressed highly in the M (mitotic) phase. That is, our modified GA-RNN hybrid algorithm has the capability of finding target regulated genes at a later time point (e.g. $t + 2$) when given a TF at an earlier time point (e.g. t).

Machine learning approaches to microarray analysis

There are many types of gene transcriptional regulatory related approaches which have been proposed in the past. Their nature and composition are categorized by several factors: considering gene expression values [3,4], the causal relationship between genes, e.g. with Bayesian analysis or Dynamic Bayesian Networks[5,6], and the time domain e.g. discrete or continuous time [7-10]. The genome-wide transcriptional program during the cell cycle has been investigated in a wide range of organisms, including yeast [11], bacteria [12], primary human fibroblasts [13,14], and human HeLa cells [15]. However, consideration of feedback and feed-forward control in regulatory modules is also important. That is, some genes have unique characteristics, for instance, they regulate themselves or they regulate genes in the following further time points. Unfortunately, constructing regulatory modules with feedback and feed-forward controls is not mentioned by [3-6]. Moreover, genes may have one or more activators or inhibitors which co-regulate the transcription levels of genes in regulatory modules. Lots of the cell's activities are organized as sets of genes co-regulated by some particular TFs to respond to different conditions. Therefore, the present challenge is to understand how transcription factors control global gene expression programs, i.e., specific gene expression programs involve regulated transcription of many other genes in different time points or involve regulated transcription of themselves. Our approach aims to provide a system to construct regulatory modules with feedback and feed-forward control mechanisms that illustrate cancer-related genes and their cause-effect relations to other genes.

Recent approaches to gene expression of human cancers

Gene expression profiling has been widely used for cancer research. The results provided by [16] and [17] have shown that co-expression of gene pairs in multiple data sets are correlated with functional relatedness. According to [16], they seek pairs of genes based on the correlation of their expression profiles in multiple data sets, and define these pairs of genes reliably co-expressed to establish a high-confidence network of more than 8,000 genes connected by co-expression links that are observed in the data sets. However, global co-expression patterns have not been determined for cancer, and it is still unknown what are the key genes or gene groups that have been causing or stabilizing the observed global cancer-related patterns. Likewise, it is of interest to know which genes are the factors that initiate the regulation to another possibly pathological state. Hence, we try to deal with this problem by identifying genes whose regulatory functions have interventions in the global cancer-related gene expression profiles.

Results

Results of human cancer data

The human cell cycle data set consists of almost 30,000 genes and over 44 time points for each of the experimental data. As a result, the number of gene combinations is more significant than the yeast data set. We list some experimental results in Table 4. In these experiments, the number of epochs of RNN is 200, and the number of generations of GA is varied while recording the error rates for the training data. The minimum value of RMSE decreases as the number of GA generation increases.

Notice that the minimum RMSE for GA generations 2000 and 2500 are 0.17 and 0.16, respectively. Comparing these two experimental results, although the number of GA generations is supplementary, there is no conspicuous diminishing for the value of RMSE. Moreover, the maximum number of GA generations for the yeast data set is less than half of the human data set. Searching for "good" combinations of regulated target genes from the human data set yields much more permutations than yeast.

We also demonstrate some relationships of cancer-related regulation in Figure 4. The E2F1 gene has been biologically proven a key regulator of the cell cycle. As noted by

Table 1: The encoded numbers of amino acids

Amino acids	A	G	C	T	P	R	S	W	D	E	F	I	K	L	N	Q	V	H	Y	M
Encoded number	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	1

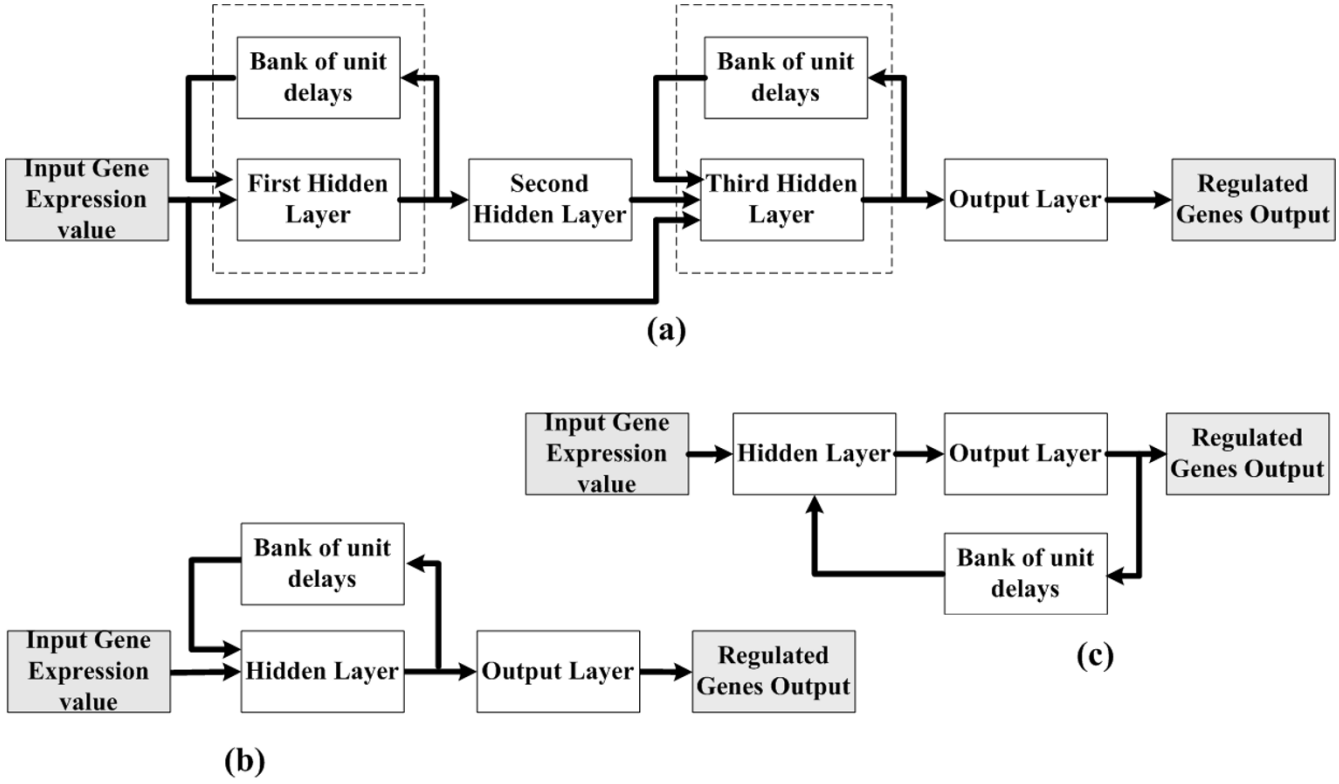

Figure 1
RNN architectures for each different kind of regulatory module module used by this approach. (a) is represented for gene feed-forward regulatory modules, (b) is represented for auto-regulate modules, and (c) is represented for single-input and multi-input regulatory modules.

Stanelle *et al*. [18], the p16/RB/E2F regulatory pathway, which controls transit through the G1 restriction point of the cell cycle, is one of the most frequent targets of genetic alterations in human cancer. Likewise, Figure 4 shows the same idea of the E2F1 and RB regulatory relationship, and also proves the ability of our system. It is also known that, E2F1 participates in the progress of apoptosis by regulating p53, and this is indicated in Figure 4 as well. P53 is a tumor suppressor whose inactivation is observed in most human cancers [19]. P53 also plays a central role in regulating cell growth, particularly in response to various forms of stress, including DNA damage and viral infection, because p53 sits on a critical node of signal transduction networks that control cell growth and death. Figure 4 illustrates that PCAF is a co-activator of p53 transcription identified by our system, which is biologically supported by Zhao *et al*. [20]. The same as CDC6, PCAF is identified as an auto-regulated TF, moreover, it up-regulates the NICD, CBF1, and BRCA2 target genes. More results and biological evidences are provided in additional file 1.

Compare human cancer data results to yeast cell cycle

An example of periodically expressed gene in human cells but not in yeast is the human CDK7 (homologous of *S. cerevisiae* KIN28). Yeast KIN28 is not found as a TF at the transcriptional level, but the CDK7 is a TF that regulates MO15. Human gene CDC6 is homologous of *S. cerevisiae* CDC6, which is identified by our system, and is an auto-regulated and crucial TF to GCN5, pRB, and NICD genes while yeast CDC6 is not. The most interesting and complicated gene of the human data set that system came out is E2F1. It is likely that some of the periodically expressed genes in human cell that do not have periodically expressed correspondents in yeast are subject to multiple layers of regulation. It is also reasonable that, multi-layer regulation in human cell, such as the phosphorylation and proteolysis, are known for some well-studied cell cycle genes. Therefore, the regulatory module starting on E2F1 TF shows the chain processes of the regulatory mechanism, which implies the complication of human cancer-related regulatory manipulations. One known gene, p53,

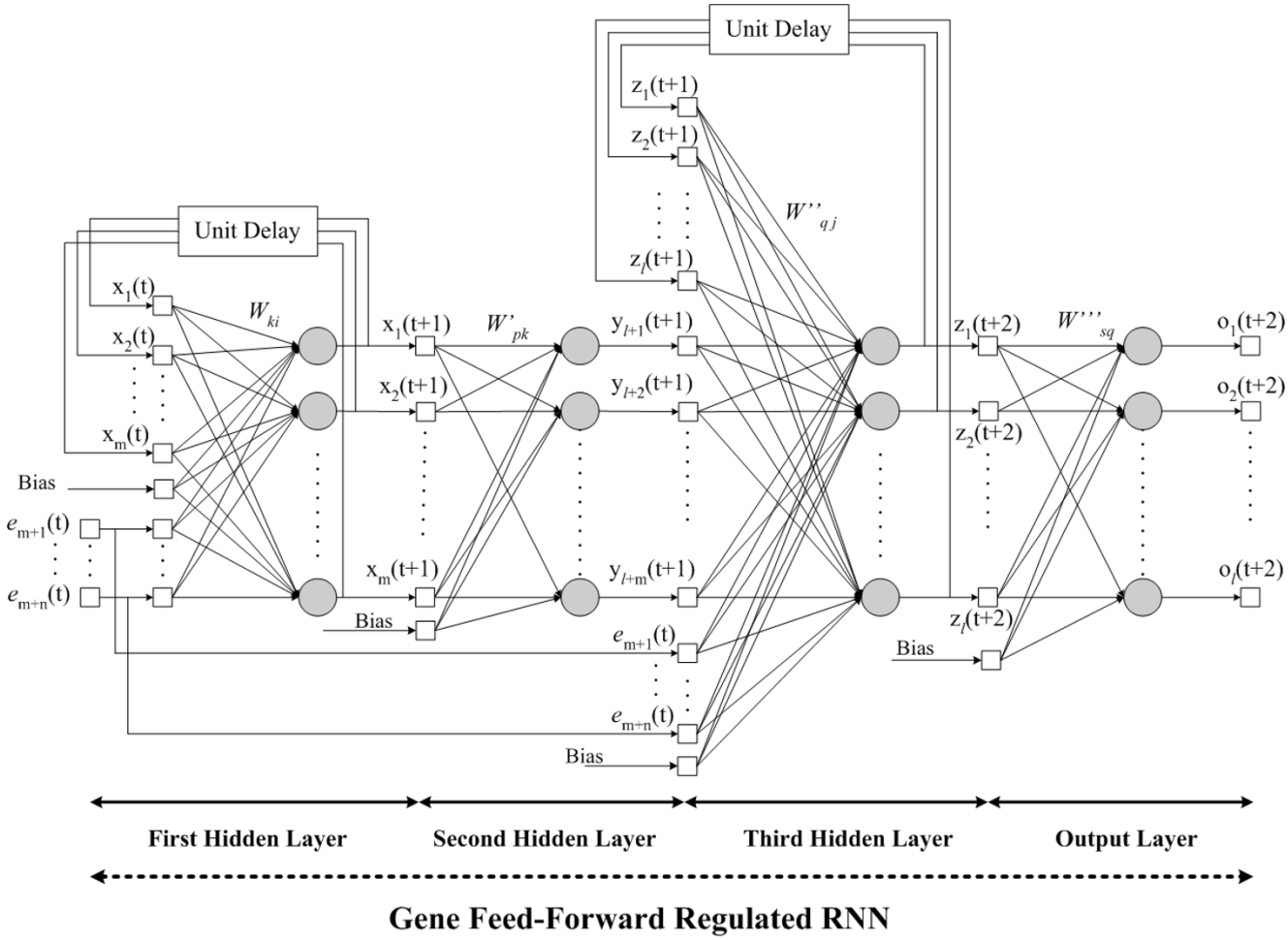

Figure 2
The RNN architecture for constructing feed-forward module.

also represents an intriguing consequence. After inputting HDAC3 TF to the GA-RNN algorithm, some negative values are reported by the neural network weight matrix, which indicates inhibitor targets. It appears that there is a similar case, the GATA2, down-regulated by HDAC3 as well, and both are confirmed by Juan *et al*. [21] and Ozawa *et al*. [22]. These listed genes are strongly expressed in proliferative tumors, and have regulatory relationships with other genes. This may further prove to be a useful source of additional drug targets of this kind. On the other hand, most of the genes identified by our system periodically expressed in both species are involved in DNA replication, DNA repair, DNA metabolism and mitosis. In the viewpoint of biology, these genes involved in mitosis or DNA replication, in all probability, have connections with cancer diseases.

We also show some experimental results with G1/S/G2/M phases in figure 5. The green lines shown in figure 5 indicate regulatory connections predicted by this approach. The red lines indicate regulatory connections that predicted by this approach and also are confirmed by biological experiments. The red dotted lines represent the negative regulatory controlling, which are also confirmed by biological experiments. The blue lines shown in figure 5 demonstrate the feed-forward controlling between E2F1, RB and the target gene cycA. It is clear that the E2F1 is a start TF and controls the second TF, RB. Both E2F1 and RB regulate the target gene, cycA, which form a feed-forward regulatory motif.

Discussion

The construction of cancer-related regulatory modules from temporal expression data is one of the most important problems in computational biology. It is acknowledged that the causes of heterogeneity genetic-related circumstances, such as the cell cycle, or cancer diseases, are products of complex interactions between genes over

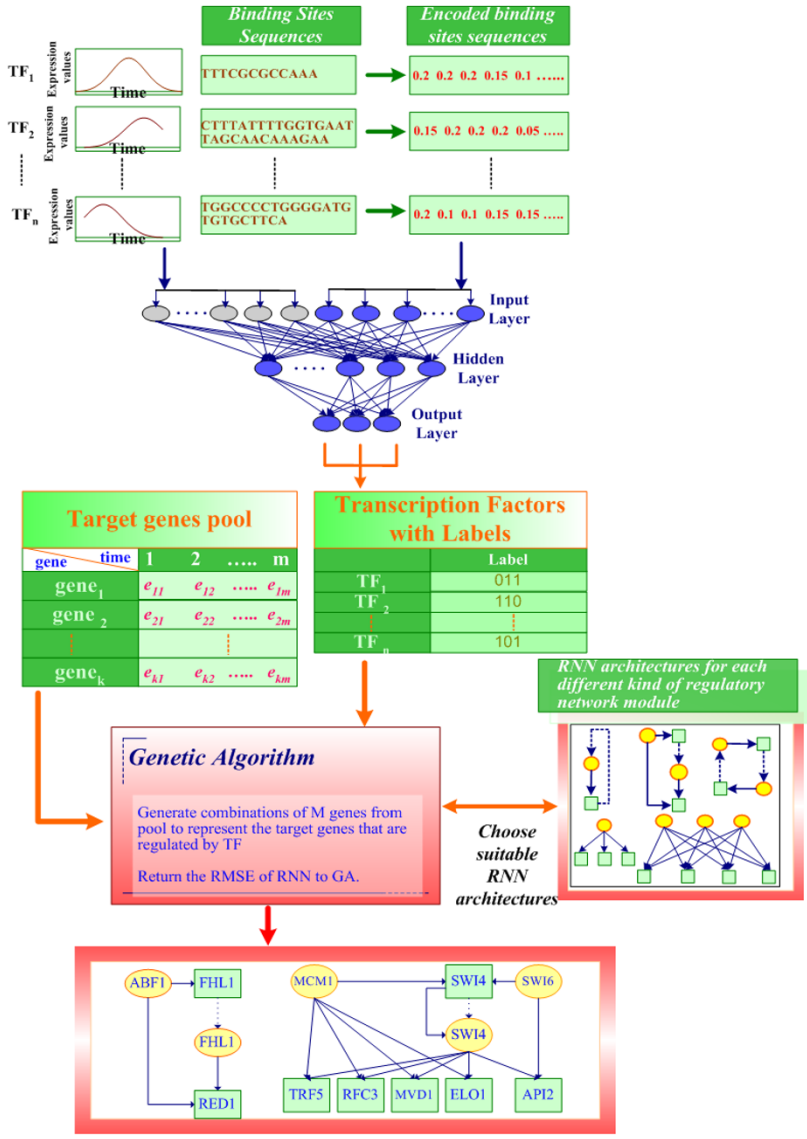

Figure 3

The graphic algorithm of system procedures. The whole system is implemented by Matlab 6.5. Notice that the RBF classifier is not designed for predicting novel TFs for yeast or human species. The main purpose of the RBF classifier is to group the kinds of categories that TFs belong to. As Figure shows, we input TFs into the system and the trained RBF classifier then can decide on the category according to the microarray expression values and transcription factor binding sites sequences of the TFs. This step also indicates that what kinds of RNN architecture are used in the following steps. The GA uses a standard random mutation and a standard binary representation with one point crossover. One TF may not only regulate one target gene, but may regulate several target genes simultaneously. To acquire the "good" combinations of target genes that are regulated by one or more common TFs, we select appropriate GA mutation and crossover operators to alter the chromosomes. One chromosome of the GA represents a number of genes taken from the full set of genes and is used by RNN to check how "good" the expression values of this combination of genes affected by particular transcription regulators are. On the other hand, the GA consists of populations of such chromosomes, and each chromosome is evaluated by the RNN for its fitted value to the given TFs. The choice of RNN architecture is according to the labels assigned by the RBF classifier. Furthermore, the final returned RNN output error (RMSE) is treated as a fitted value for some particular combinations of target genes. The stopping criterion includes not only the fitted value fit for some criteria but also the determination of the RNN selecting steps. In other words, the GA never stops until all appropriate RNN architectures are executed. In that case, each TF can choose suitable RNN architecture more than once, and find out a dissimilar set of target genes. After all TFs are run by this system procedures and output regulatory modules, the GA is then complete.

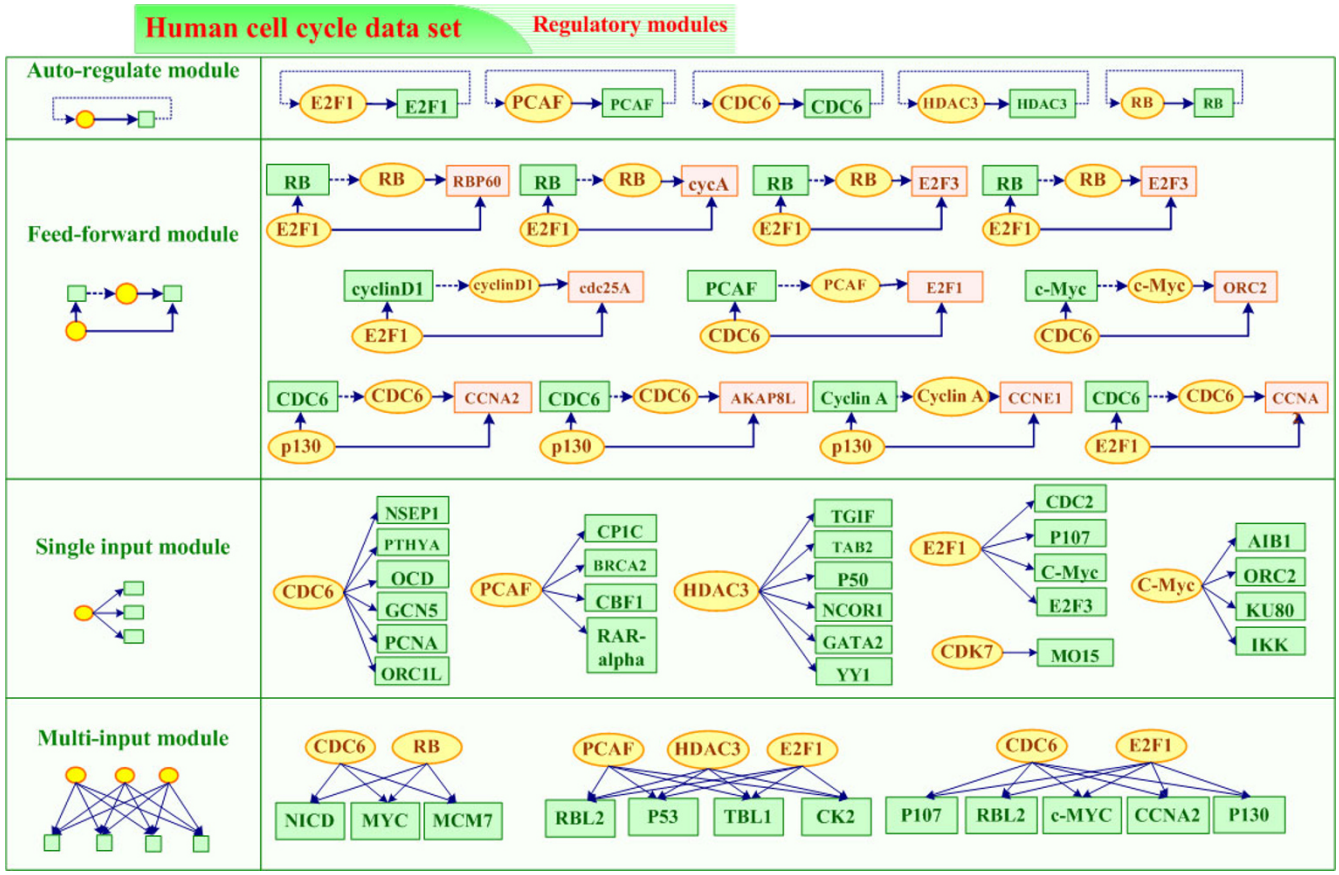

Figure 4

Some experimental results of regulatory modules for human cell cycle data set. Circles indicate TFs and squares represent genes. Solid arrows indicate regulation relationships between TFs and their target genes.

time. The analysis of cancer-related gene expression data will thus become increasingly widespread. When appraising approaches for discovery of cancer-related regulatory modules, the amount and type of sources of data must be taken into account. Besides, the approach must be capable of handling noisy and high dimensional gene expression data. The approach described here has been shown to be effective with real-world expression data. The stochastic nature of GA means that the same results can not be expected from each run of the algorithm, and the GA is run for a fixed number of generations for each output of regulatory modules. However, to increase the number of genes that the GA can select from, it could require more GA generations. As a result, increasing the GA generations also increases the computational time, although it does show that results on microarray data can be discovered correctly by the GA used in our approach. In addition, this approach builds modules "piece by piece", that is, regulatory module by regulatory module. Imagine that the network motif described in section 2 is one of the transcriptional regulatory mechanism units with a specific set of genes, including the influences and the targets. We discover all the formed units one by one and eventually join these units by their simultaneously existing TFs. The above-mentioned contents are the advantages of generating smaller but more precise regulatory modules, in that each of the paths or the units (or genes) in the modules can be seen without being masked by other connections. It is not the same as traditional complicated regulatory relationships, which are too many to visualize as a network to yield useful information in a digestible format for biologists.

Compare to related researches

The ordering of regulatory processing can be displayed faithfully by this approach, especially the feed-forward motif, which represents a very simple regulator pathway. The phenomenon of this biology mechanism is often seen but ignored when constructing regulatory modules. Bayesian Networks have been proven to be an efficient methodology to reveal the cause-effect relationships from microarray experimental data, but may be deficient in dealing with the control of feedback or feed-forward issues. A major contribution provided by this approach is

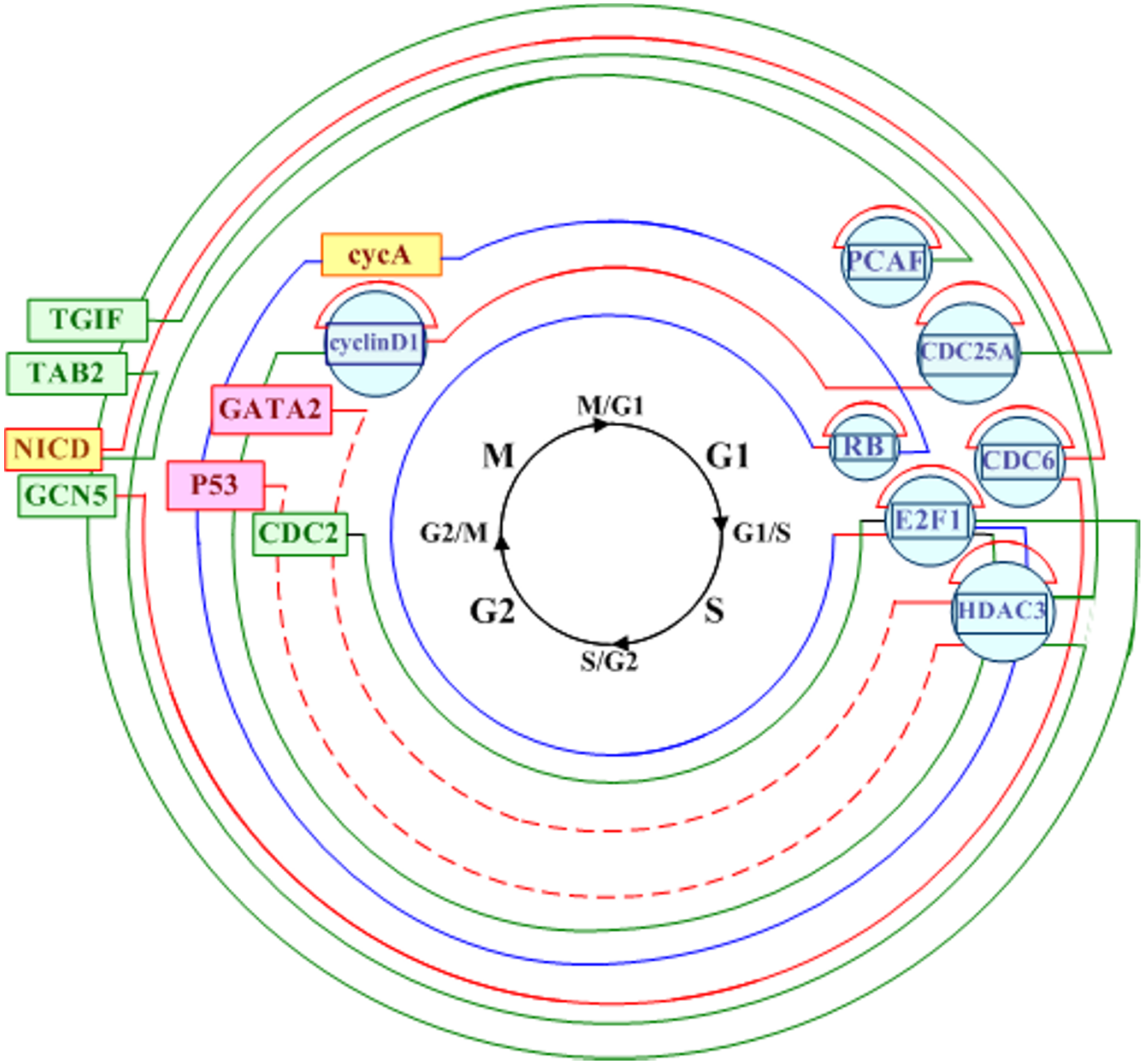

Figure 5
Some experimental results of regulatory modules for human cell cycle data set. The green lines indicate regulatory connections predicted by this approach. The red lines indicate regulatory connections that predicted by this approach and also are confirmed by biological experiments. The red dotted lines represent the negative regulatory controlling. The blue lines shown in this graph demonstrate the feed-forward controlling between E2F1, RB and the target gene cycA. All TFs are represented by blue circles and the squares indicate the target genes.

the discovery of not only the cause-effect relationships between genes but also the chain influences upon a set of genes sequentially. The mechanism of chain influences is achieved by modified RNN architecture, which includes the nonlinear mapping capability and the ability of time delay in constructing regulatory modules. The other advantage of this study is the power of global searching for regulated target genes provided by GA. Those selected target genes are judged whether they are "true target genes" or not according to the RMSE provided by RNN, and the results substantiate the capability of our hybrid GA-RNN method. Compared to [5] (see additional file 2), this paper limits potential regulators to those genes with either earlier or simultaneous expression changes (up- or

down-regulation) in relation to their target genes, which we also achieve the results. Moreover, we provide feedback and feed-forward regulation control relations between transcription factors and their targets, that is, we contribute more complicated regulatory relationships of regulatory modules.

Compared to Keedwell *et al*. [3], which uses a supervised single-layer artificial neural network (ANN) to construct the regulatory connections between genes, the first improvement of our approach is that we use modified multi-layer RNN to complete the feed-forward, feedback, auto-regulate, and multi-input regulatory modules, which means we show clearly the transcriptional regulatory mechanisms. Secondly, Keedwell *et al*. [3] provided most of the significant connections in the network by repeated GA. In our approach, we intersect the regulatory modules that output from different GA generations. Take Table 4 as an example, the GA generations for human cell cycle data are 100, 500, 1000, 1500, 2000, and 2500, respectively. We collect the repeated regulatory modules that are appeared in 1000, 1500, 2000, and 2500 generations. It means that the significant regulatory modules are aggregated. Finally, Keedwell *et al*. construct their regulatory connections by microarray expression values; nevertheless, we also consider the transcription factor binding sites sequences, which consider more biological factors while constructing regulatory modules. We also provide more biological evidences for human cell cycle data, and the results compared with [7] are listed in additional file 1.

Conclusion

We combine the GA and RNN computing approaches to construct the cancer-related regulatory modules in *silico*. Upon the microarray data and the sequences of transcription factor binding sites, the approach has been shown to be able to accurately fit the data on which it is trained. We also observe that some TFs play critical roles in various motifs. In other words, some functions of TFs are fit for several kinds of regulatory modules. We then adopt these characteristics by training the RBF classifier [1] for categorizing TFs. Additionally, the experimental results have proven that the GA-RNN hybrid algorithm has the capability of constructing the feedback and feed-forward regulatory modules. RNNs with diversified architectures indicate varied regulatory mechanisms to construct complete regulatory modules with feedback and feed-forward controls. Combining modified RNN with GA, it provides the global searching capacities to find proper target regulated genes for some TFs. The chromosomes that the GA used are combinations of target genes and the crossover and mutation operators used by GA on all chromosomes alter the choice of output gene combinations. This approach is on the basis of both gene expression data and sequences data, so it is time significant and binding region significant data analysis. Summing up, since this method has been previously shown to also classify TFs as well and then construct regulatory modules, it can be considered a candidate multipurpose tool for microarray expression data analysis.

Methods

Human cell cycle microarray data

Microarray time course measurement of genome-wide mRNA expression levels allows genome-wide prediction of cell division cycle regulated genes. In each cell division cycle, cells pass through four phases, namely, M-G2-S-G1, in a fixed order. Each cell division cycle regulates gene expressions in one of these phases, and results in a rise in the possibility that the cell division cycle regulated genes would reveal periodic expressions if they are studied for more than one cycle. This phenomenon is the basis for detecting genes with oscillated expressions in synchronized cell culture to discover the cell division cycle regulated genes. Data from Whitfield *et al*. [23] is downloaded from the reference web site. The genome-wide program of gene expression during the cell division cycle in a human cancer cell line (HeLa) is characterized using cDNA microarrays. The goal of human cell cycle analysis from Whitfield *et al*. is to identify >850 genes periodically expressed during the cell cycle, and to show that most of these genes have been previously associated with the proliferation of tumors during the human cell division cycle as well. The data in this report provide a comprehensive catalog of cell cycle regulated genes that can serve as a starting point for functional discovery. We adopt this data set to construct cancer-related regulatory modules with feedback or feed-forward controlled target genes which are regulated by some specific TFs.

Yeast cell cycle data

The data of Spellman *et al*. [24] is downloaded from the reference web site. We use this data set to construct regulatory modules and target genes which are also regulated by some specific TFs. Since this data set has been used to construct regulatory modules by various approaches in the past, we regard it as the testing data set to prove the efficiency of our approach. The experimental results and biological supports are provided in additional file 2 and additional file 3.

Transcription factor binding sites

Our computational method attempts to integrate gene expression data and sequence data. Transcription factors bind short DNA motifs, namely, the transcription factor binding site, which plays a central role in recruiting the transcriptional mechanism at the promoters of genes and leading to the initiation of their transcription. As a result, we collect known transcription factor binding sites sequences from ENSEMBL [25], TRANSFAC [26], SGD

[27] and YPD [28] for human and yeast, which are listed in additional file 4.

The network motifs

According to Lee *et al*. [29], there are five network motifs for transcriptional regulatory modules. We provide the detail descriptions in additional file 5.

Pre-processing

With all the materials that described above, we next perform the data pre-processing procedure. For instance, microarray data with missing values that occurred at certain time points for some genes have been deleted from the data set, since we can not appraise what the real values are. Besides, all microarray numeric data are re-scaled between 0 and 1 by divided the maximum value for each gene. All the transcription factor binding sites queried from databases are originally represented by sequences of amino acids and encoded into numeric codes which are listed in Table 1.

Categorize human cell cycle-related transcription factors by RBF

Whitfield *et al*. [23] have identified 874 genes that show periodic expression across the human cell cycle in a well-studied cancer cell line (HeLa). However, not all the 874 genes are TFs; our approach is intended to reveal the target genes that are regulated by one particular TF. To achieve this, we search GO (Gene Ontology) terms for 874 genes from the GO web site [30]. Only genes with GO annotation terms, such as "transcription factor activity", "transcription factor complex", "regulation of cell cycle" and so on, are kept. Genes without related transcriptional function GO terms are left behind. In an analogous manner, we regroup TFs for Homo sapiens. Take transcription factor E2F1 as an example, a list of reactions which go out from E2F1 that representing E2F1 is served as a signal donor. These downstream reactions contain E2F1 itself, which cause E2F1 to fit the definition of an auto-regulating TF. What is more, the expression of most E2F1-dependent genes, such as P107 and RB1, peaks at the G1/S boundary. Additionally, E2F1 is also involved in regulating genes that control other phases of the cell cycle. Such as cycA and cdc2, whose expression remain high throughout the S-phase and into the G2 phase, which characterizes E2F1 as a single-input, multi-input and feed-forward motif factor [31]. As a result, we regroup human sapiens TFs into several catalogs according to biological documents and list the classified samples in Table 2.

Under this approach, the RBF network architecture is one input layer with two kinds of input data, one hidden layer, and one output layer, which is illustrated in additional file 6.

Construction of regulatory modules

RNNs are neural networks with one or more feedback loops. Given a multilayer perceptron as the basic building block, we may have feedback from the output neurons of the multilayer perceptron to the input layer. Similarly, the architectures of gene regulatory modules also have feedback from the target genes to TFs, which represent positive or negative effects on those genes that influence themselves. When the multilayer perceptron has two or more hidden layers, the possible forms of global feedback expand even further, such as the feed-forward motif. We train different RNN architectural layouts for various network modules, as described in the material section.

Figure 1 demonstrates RNN architectures for various different kinds of regulatory modules used in our approach. The diagram shown in Figure 1(a) stands for the feed-forward motif, which contains the starter TF that controls the second TF in later time points and has the additional feature that both TFs have common target genes. It is of interest that the second TF is regulated by the principal TF at first and then controls other target genes together with the principal TF. This framework has an advantage of building simple regulatory pathways with the ordering of transcriptional influences and the time delay of this mechanism. In this approach, the modified RNN is altered from [1] to adjust to the nature of feed-forward transcriptional regulatory mechanisms. Figure 2 shows the detailed RNN architecture for constructing feed-forward regulatory modules. In this diagram, the expression level of a gene at a certain time point can be calculated by the weighted sum of the expression levels of all potential TFs in the network at a previous time point. For a time delay system, the models can be represented as:

$$x_k(t+1) = \varphi\left(\sum_{i=1}^{m} w_{ki} x_i(t) + \sum_{i=m+1}^{m+n} w_{ki} e_i(t) + w_{kb} B\right) \quad (3)$$

Table 2: Some examples of regrouped TFs for Homo sapiens

Gene	Single Input	Multi Input	Feed-forward	Auto-regulate	Encoded catalog number of RBF classifier
E2F1	√	√	√	√	00
CDC6	√	√	√	√	00
HDAC3	√	√		√	10
CDK7	√	√			11

Table 3: GA and RNN parameter settings

Yeast Cell Cycle Data set		Human Cell Cycle Data set	
GA parameters	**Values**	**GA parameters**	**Values**
Crossover	One Point, crossover rate (0.9)	Crossover	One Point, crossover rate (0.8)
Mutation	Random, mutation rate (0.05)	Mutation	Random, mutation rate (0.1)
Selector	Roulette Wheel	Selector	Roulette Wheel
Population Size	50 ~ 250	Population Size	100 ~ 250
Generations	100 ~ 1000	Generations	100 ~ 2500
RNN parameters	**Values**	**RNN parameters**	**Values**
Epochs	50 ~ 100	Epochs	100 ~ 200
Gradient descent	Standard	Gradient descent	Standard
Weight Update	Online	Weight Update	Online

where the $e_i(t)$ is the gene expression level for the i^{th} gene ($m + 1 \le i \le m + n$, n is the number of beginning TFs), w_{ki} ($m + 1 \le i \le m + n$) represents the effect of the k^{th} gene on the i^{th} gene, and B is the bias. The $x_i(t)$ represents the m neurons in the hidden layer that are connected to the feedback nodes in the input layer, and the matrix w_{ki} ($1 \le i \le m$) represents the synaptic weights of the m neurons. A negative value of w_{ki} represents the inhibition of the k^{th} gene on the i^{th} gene, while a positive value indicates the activation controls. The $\varphi()$, is a nonlinear sigmoidal function, usually in the form of $\varphi(z) = 1/(1 + \exp^{-z})$. The symbol B represents the bias, and the matrix w_{kb} contains of the weights that represent bias terms applied to neuron 1 to k. The $x_k(t + 1)$ represents the first hidden layer in the gene feed-forward regulated RNN, which stands for biological second ordered TFs (controlled by the starter TFs); the second and third hidden layers are listed as follows, respectively:

$$\gamma_p(t+1) = \varphi\left(\sum_{k=1}^{m} w'_{pk} x_k(t+1) + w'_{pb} B'\right) \qquad (4)$$

$$z_q(t+2) = \varphi\left(\sum_{j=1}^{n} w''_{qj} e_j(t) + \sum_{j=n+1}^{n+m} w''_{qj} \gamma_j(t+1) + \sum_{j=n+m+1}^{n+m+l} w''_{qj} z_j(t+1) + w''_{qb} B''\right) \qquad (5)$$

where the w'_{pk} is a weight matrix, for $1 \le k \le m$, and B' is the bias for this hidden layer. As in Figure 2, the third hidden layer contains $e_i(t)$ and $\gamma_p(t + 1)$ to demonstrate the expression values of the starter TFs and TFs controlled by themselves, respectively, and to co-regulate their common target genes described as:

$$O_s(t+2) = \varphi\left(\sum_{q=1}^{l} w'''_{sq} z_q(t+2) + w'''_{sb} B'''\right) \qquad (6)$$

It is logical to assume that the time stamp of target genes is marked as $t + 2$, $t + 1$ for controlled TFs, and time stamp t for the starter TFs. In transcriptional progresses, regulatory pathway follows the prescribed order to "turn-on" or "turn-off" some target genes, and that is why we design a gene feed-forward regulated RNN architecture to fit the characteristic of the regulatory module. In the problem of regulatory module inference, the goal is to recover the regulatory interactions w_{ki}, w'_{pk}, w''_{qj} and w'''_{sq}. The instantaneous sum of squared errors at time t is defined in terms of $E(t)$ by

$$E(t) = \frac{1}{2}\sum_{s=1}^{l} (O_s(t) - d_s(t))^2 \qquad (7)$$

The objective of the learning process is to minimize a cost function obtained by summing $E(t)$ over all time t; that is,

$$E_{Total} = \sum_{t} E(t) \qquad (8)$$

which measures the deviation of network output $O(t)$ from the measurement (the target) $d(t)$. We use gradient

Table 4: The experimental results of GA with RNN for human cell cycle data

GA generations	Average RMSE	The minimum RMSE
100	5.34	2.78
500	3.47	1.86
1000	1.96	0.55
1500	1.31	0.38
2000	0.84	0.17
2500	0.81	0.16

descent to determine the weights of the network and the weights correction of this gene feed-forward regulated RNN in the training phase are as follows:

$$\Delta w_{ki} = -\eta \frac{\partial E}{\partial w_{ki}} = -\eta \cdot \left[\sum_t \sum_s (O_s - d_s) \right] \cdot \frac{\partial O_s}{\partial w_{ki}} \quad (9)$$

$$\Delta w'_{pk} = -\eta' \frac{\partial E}{\partial w'_{pk}} = -\eta' \cdot \left[\sum_t \sum_s (O_s - d_s) \right] \cdot \frac{\partial O_s}{\partial w'_{pk}} \quad (10)$$

$$\Delta w''_{qj} = -\eta'' \frac{\partial E}{\partial w''_{qj}} = -\eta'' \cdot \left[\sum_t \sum_s (O_s - d_s) \right] \cdot \frac{\partial O_s}{\partial w''_{qj}} \quad (11)$$

$$\Delta w'''_{sq} = -\eta''' \frac{\partial E}{\partial w'''_{sq}} = -\eta''' \cdot \left[\sum_t \sum_s (O_s - d_s) \right] \cdot \frac{\partial O_s}{\partial w'''_{sq}} \quad (12)$$

where the η, η', η'' and η''' represent the learning rate. The first derivative of O_s with respect to w_{ki}, w'_{pk}, w_{qj}'' and w_{sq}''' are listed as follows:

$$\frac{\partial O_s}{\partial w'''_{sq}}(t+2) = \varphi(\Omega) \cdot (1 - \varphi(\Omega)) \cdot z_q(t+2) \quad (13)$$

$$\frac{\partial O_s}{\partial w''_{qj}}(t+2) = \frac{\partial O_s}{\partial z_q} \cdot \frac{\partial z_q}{\partial w''_{qj}}(t+2)$$
$$= \begin{cases} \varphi(\Omega) \cdot (1 - \varphi(\Omega)) \cdot \varphi(\Psi) \cdot \varphi(1 - \varphi(\Psi)) \cdot e_j(t), & 1 \le j \le n \\ \varphi(\Omega) \cdot (1 - \varphi(\Omega)) \cdot \varphi(\Psi) \cdot \varphi(1 - \varphi(\Psi)) \cdot y_j(t+1), & n+1 \le j \le n+m \\ \varphi(\Omega) \cdot (1 - \varphi(\Omega)) \cdot \varphi(\Psi) \cdot \varphi(1 - \varphi(\Psi)) \cdot z_j(t+1), & n+m+1 \le n+m+l \end{cases} \quad (14)$$

$$\frac{\partial O_s}{\partial w'_{pk}}(t+1) = \frac{\partial O_s}{\partial z_q} \cdot \frac{\partial z_q}{\partial y_j} \cdot \frac{\partial y_j}{\partial w'_{pk}}(t+1)$$
$$= \varphi(\Omega) \cdot (1 - \varphi(\Omega)) \cdot \varphi(\Psi) \cdot (1 - \varphi(\Psi)) \cdot \varphi(\Lambda) \cdot (1 - \varphi(\Lambda)) \cdot \varphi(\Gamma) \cdot (1 - \varphi(\Gamma)) \cdot x_i(t+1) \quad (15)$$

$$\frac{\partial O_s}{\partial w_{ki}}(t) = \frac{\partial O_s}{\partial z_q} \cdot \frac{\partial z_q}{\partial y_j} \cdot \frac{\partial y_j}{\partial x_k} \cdot \frac{\partial x_k}{\partial w_{ki}}(t)$$
$$= \begin{cases} \varphi(\Omega) \cdot (1 - \varphi(\Omega)) \cdot \varphi(\Psi) \cdot (1 - \varphi(\Psi)) \cdot \varphi(\Lambda) \cdot (1 - \varphi(\Lambda)) \cdot \varphi(\Gamma) \cdot (1 - \varphi(\Gamma)) \cdot x_i(t), & 1 \le i \le n \\ \varphi(\Omega) \cdot (1 - \varphi(\Omega)) \cdot \varphi(\Psi) \cdot (1 - \varphi(\Psi)) \cdot \varphi(\Lambda) \cdot (1 - \varphi(\Lambda)) \cdot \varphi(\Gamma) \cdot (1 - \varphi(\Gamma)) \cdot e_i(t), & n+1 \le i \le n+m \end{cases} \quad (16)$$

where

$$\Omega = \sum_{q=1}^{l} w'''_{sq} z_q(t+2) \quad (17)$$

$$\Psi = \sum_{j=1}^{n} w''_{qj} e_i(t) + \sum_{j=n+1}^{n+m} w''_{qj} y_j(t+1) + \sum_{j=n+m+1}^{n+m+l} w''_{qj} z_j(t+1) \quad (18)$$

$$\Lambda = \sum_{k=1}^{m} w'_{pk} x_k(t+1) \quad (19)$$

$$\Gamma = \sum_{i=1}^{m} w_{ki} x_i(t) + \sum_{i=m+1}^{m+n} w_{ki} e_i(t) \quad (20)$$

Figures 1(b) and 1(c) demonstrate the RNN architectures for auto-regulate and multi-input motifs, respectively. The major difference between figure 1(b) and 1(c) is that the feedback comes from the output neurons or the hidden neurons of the multilayer perceptron to the input layer. Especially for figure 1(c), the single-input motif can be considered a subclass of the multi-input motif, and that is the reason we use the same RNN architecture to represent single-input and multi-input motifs.

As described in previous section, we have various cell cycle-regulated TFs classified by RBF, and the next process is to find out the target genes. To resolve this problem, we must perform a global search for the optimal target genes which the GA produces. Hence, the neural computing method adopted in this paper combines the GA with the RNN architecture to form a hybrid system. We demonstrate the graphic algorithm in figure 3, and the procedures work as follows:

Input: various transcription factors

Output: regulatory modules with feedback and feed-forward control

Procedure:

For each transcription factor

BEGIN:

1. Randomly choose one TF *A* as the "input" gene, $gene_A$.

2. If $gene_A$ is labeled as "010", then run steps 3 ~ 9 three times, one for the single-input motif RNN architecture, another for the multi-input motif RNN architecture, and the other for the feed-forward motif RNN architecture. It infers that, we have designed several RNN architectures for different motifs that are described in above section.

3. Use the genetic algorithm (GA) to generate combinations of M genes ($gene_1$, $gene_2$, ..., $gene_m$) to represent the target genes that are regulated by $gene_A$. Each combination is a chromosome. The initial set of combinations is composed of the initial population of chromosomes.

4. The training set, including the initial population of chromosomes and $gene_A$, will consist of the microarray expression values for all the time points, and the initial population of chromosomes will be the target output for the RNN.

5. Execute the gradient descent algorithm on this training data via the RNN to determine the weights between the

input genes and the output genes until stopping criterion is met.

6. Return the RMSE of RNN to GA. This is the fitted value for a particular chromosome.

7. Repeat 3, 4, and 5 for each chromosome.

8. Repeat steps 3~6 as a GA run, using crossover and mutation operators on all chromosomes to alter the choice of output gene combinations.

9. When some stopping criterion is met, the GA stops. Record the best chromosome and the weights of the RNN.

END

Change another TF (i.e. $gene_B$), until no TFs are left.

When all the steps described above are completed, regulatory module structures can be derived to comprise the connections of the TFs and their target regulated genes. The parameters for the GA operators and RNN parameters are shown individually in Table 3.

List of abbreviations

Genetic Algorithm (GA), Recurrent Neural Network (RNN), Transcription Factors (TFs), Radial Basis Function (RBF), Dynamic Bayesian Network (DBN)

Authors' contributions

JHC and SYC participated in the design of this approach, analyzed the experimental results, and writing of the manuscript. SYC participated in the coding of the experiments and JHC participated in preparing the final draft of the manuscript. Both authors read and approved the final manuscript.

Additional material

Additional file 1

Experimental results for human cell cycle and the biological support of the gene regulations. *Experimental results for human cell cycle and the biological support evidences of the gene regulations are listed in this file.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-8-91-S1.pdf]

Additional file 2

The precision for yeast cell cycle. *The precision (TP/(TP+FP)) for yeast cell cycle data.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-8-91-S2.pdf]

Additional file 3

Experiment for yeast cell cycle and the biological supports of the gene regulations. *Additional supporting analyses of yeast cell cycle and the biological supports for the article.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-8-91-S3.pdf]

Additional file 4

Examples of TFs and their specific binding sites sequences. *The supplementary Tables A and B. We list examples of TFs and their binding sites sequences for human and yeast.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-8-91-S4.pdf]

Additional file 5

The network motifs. *The network motifs used by this approach are described in additional file 5, including the auto-regulatory, feed-forward, single-input and multiple-input regulatory motifs.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-8-91-S5.pdf]

Additional file 6

The architecture of RBF classifier. *The architecture of RBF classifier used by this approach is illustrated in this additional file.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-8-91-S6.pdf]

Acknowledgements

This research work was supported in part by Research Grant NSC95-2221-E-006-321 from the National Science Council, Taiwan.

References

1. Haykin S: **Neural Networks: a comprehensive foundation.** In *Dynamically Driven Recurrent Networks* Tom Robbins. New Jersey; 1999:732-778.
2. Michalewicz Z: *Genetic Algorithms + Data Structures = Evolution Programs* Third, Revised and Extended edition. Springer-Verlag. New York; 1999.
3. Keedwell E, Narayanan A: **Discovering Gene Networks with a Neural-Genetic Hybrid.** *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2005, **2:**231-242.
4. Shmulevich I, Dougherty ER, Zhang W: **From Boolean to Probabilistic Boolean Networks as Models of Genetic Regulatory Networks.** *Proceedings of the IEEE* 2002, **90:**1778-1790.
5. Zou M, Conzen SD: **A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data.** *Bioinformatics* 2005, **21:**71-79.
6. Husmeier D: **Sensitivity and Specificity of Inferring Genetic Regulatory Interactions from Microarray Experiments with Dynamic Bayesian Networks.** *Bioinformatics* 2003, **19:**2271-2282.
7. Li X, Rao S, Jiang W, Li C, Xiao Y, Guo Z, Zhang Q, Wang L, Du L, Li J, Li L, Zhang T, Wand QK: **Discovery of time-delayed gene regulatory networks based on temporal gene expression profiling.** *BMC Bioinformatics* 2006, **7:**26-46.
8. He F, Zeng AP: **In search of functional association from time-series microarray data based on the change trend and level of gene expression.** *BMC Bioinformatics* 2006, **7:**69-84.
9. Filkov V, Skiena S, Zhi J: **Analysis techniques for microarray time-series data.** *J Comput Boil* 2002, **9:**317-331.

10. Qian J, Dolled-Filhart M, Lin J, Yu H, Gerstein M: **Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions.** *J Mol Biol* 2001, **314:**1053-1066.
11. Kim H, Hu W, Kluger Y: **Unraveling condition specific gene transcriptional regulatory networks in Saccharomyces cerevisiae.** *BMC Bioinformatics* 2006, **7:**165-182.
12. Laub MT, McAdams HH, Feldblyum T, Fraser CM, Shapiro L: **Global analysis of the genetic network controlling a bacterial cell cycle.** *Science* 2000, **290:**2144-2148.
13. Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, Lee JC, Trent JM, Staudt LM, Hudson J Jr, Boguski MS, Lashkari D, Shalon D, Botstein D, Brown PO: **The transcriptional program in the response of human fibroblasts to serum.** *Science* 1999, **283:**83-87.
14. Ishida S, Huang E, Zuzan H, Spang R, Leone G, West M, Nevins JR: **Role of E2F in control of both DNA replication and mitotic functions as revealed from DNA microarray analysis.** *Mol Cell Biol* 2001, **21:**4684-4699.
15. Crawford DF, Piwnica-Worms H: **The G2 DNA damage checkpoint delays expression of genes encoding mitotic regulators.** *J Biol Chem* 2001, **276:**37166-37177.
16. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P: **Coexpression analysis of human genes across many microarray data sets.** *Genome Res* 2004, **14:**1085-1094.
17. Stuart JM, Segal E, Koller D, Kim SK: **A gene coexpression network for global discovery of conserved genetic modules.** *Science* 2003, **302:**249-255.
18. Stanelle J, Stiewe T, Theseling CC, Peter M, Putzer BM: **Gene Expression changes in response to E2F1 activation.** *Nucleic Acids Res* 2002, **30:**1859-1867.
19. Levine AJ: **p53, the Cellular Gatekeeper for Growth and Division.** *Cell* 1997, **88:**323-331.
20. Zhao LY, Liu Y, Bertos NR, Yang XJ, Liao D: **PCAF is a coactivator for p73-mediated transactivation.** *Oncogene* 2003, **22:**8316-8329.
21. Juan LJ, Shia WJ, Chen MH, Yang WM, Seto E, Lin YS, Wu CW: **Histone deacetylases specifically down-regulate p53-dependent gene activation.** *Journal of Biological Chemistry* 2000, **275:**20436-20443.
22. Ozawa Y, Towatari M, Tsuzuki S, Hayakawa F, Maeda T, Miyata Y, Tanimoto M, Saito H: **Histone deacetylase 3 associates with and represses the transcription factor GATA-2.** *Blood* 2001, **98:**2116-2123.
23. Whitfield ML, Sherlock G, Saldanha AJ, Murray JI, Ball CA, Alexander KE, Matese JC, Perou CM, Hurt MM, Brown PO, Botstein D: **Identification of Genes Periodically Expressed in the Human Cell Cycle and Their Expression in Tumors.** *Mol Biol Cell* 2002, **13:**1977-2000.
24. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: **Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces Cerevisiae by Microarray hybridization.** *Mol Biol cell* 1998, **9:**3273-3297.
25. Hubbard T, Andrews D, Caccamo M, Cameron G, Chen Y, Clamp M, Clarke , Coates G, Cox T, Cunningham F, Curwen V, Cutts T, Down T, Durbin R, Fernandez-Suarez XM, Gilbert J, Hammond M, Herrero J, Hotz H, Howe K, Iyer V, Jekosch K, Kahari A, Kasprzyk A, Keefe D, Keenan S, Kokocinsci F, London D, Longden I, McVicker G, Melsopp C, Meidl P, Potter S, Proctor G, Rae M, Rios D, Schuster M, Searle S, Severin J, Slater G, Smedley D, Smith J, Spooner W, Stabenau A, Stalker J, Storey R, Trevanion S, Ureta-Vidal A, Vogel J, White S, Woodwark C, Birney E: **Ensembl 2005.** *Nucleic Acids Res* 2005, **33:**447-453 [http://www.ensembl.org/index.html].
26. Wingender E, Dietze P, Karas H, Knuppel R: **TRANSFAC: a database on transcription factors and their DNA binding sites.** *Nucleic Acids Res* 1996, **24:**238-241 [http://www.gene-regulation.com/pub/databases.html].
27. Christie KR, Weng S, Balakrishnan R, Costanzo MC, Dolinski K, Dwight SS, Engel SR, Feierbach B, Fisk DG, Hirschman JE, Hong EL, Issel-Tarver L, Nash R, Sethuraman A, Starr B, Theesfeld CL, Andrada R, Binkley G, Dong Q, Lane C, Schroeder M, Botstein D, Cherry JM: **Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms.** *Nucleic Acids Res* 2004, **32:**311-314 [http://www.yeastgenome.org].
28. Hodges PE, McKee AH, Davis BP, Payne WE, Garrels JI: **Yeast proteome database(YPD): a model for the organization and presentation of genome-wide functional data.** *Nucleic Acids Res* 1999, **27:**69-73.
29. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett M, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel , Gifford DK, Young RA: **Transcriptional Regulatory Networks in Saccharomyces cerevisiae.** *Science* 2002, **298:**799-804.
30. **Gene Ontology Consortium, 1999–2006** [http://www.geneontology.org/]
31. Kel AE, Kel-Margoulis OV, Farnham PJ, Bartley SM, Wingender E, Zhang MQ: **Computer-assisted identification of cell cycle-related genes: new targets for E2F transcription factors.** *Molecular Biology* 2001, **309:**99-120.