Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 138410, 10 pages
http://dx.doi.org/10.1155/2014/138410

Research Article

msiDBN: A Method of Identifying Critical Proteins in Dynamic PPI Networks

Yuan Zhang,[1] Nan Du,[2] Kang Li,[2] Jinchao Feng,[1] Kebin Jia,[1] and Aidong Zhang[2]

[1] *College of Electronic Information and Control Engineering, Beijing University of Technology, Beijing 100124, China*
[2] *Department of Computer Science and Engineering, State University of New York at Buffalo, Buffalo, NY 14260, USA*

Correspondence should be addressed to Yuan Zhang; zhangyuan1012@gmail.com

Received 29 January 2014; Accepted 9 March 2014; Published 2 April 2014

Academic Editor: FangXiang Wu

Dynamics of protein-protein interactions (PPIs) reveals the recondite principles of biological processes inside a cell. Shown in a wealth of study, just a small group of proteins, rather than the majority, play more essential roles at crucial points of biological processes. This present work focuses on identifying these critical proteins exhibiting dramatic structural changes in dynamic PPI networks. First, a comprehensive way of modeling the dynamic PPIs is presented which simultaneously analyzes the activity of proteins and assembles the dynamic coregulation correlation between proteins at each time point. Second, a novel method is proposed, named msiDBN, which models a common representation of multiple PPI networks using a deep belief network framework and analyzes the reconstruction errors and the variabilities across the time courses in the biological process. Experiments were implemented on data of yeast cell cycles. We evaluated our network construction method by comparing the functional representations of the derived networks with two other traditional construction methods. The ranking results of critical proteins in msiDBN were compared with the results from the baseline methods. The results of comparison showed that msiDBN had better reconstruction rate and identified more proteins of critical value to yeast cell cycle process.

1. Introduction

A biological process is a complexity of spatial and temporal interactions among innumerable molecules. Understanding dynamic biological processes and revealing the mechanisms behind dynamic systems are of great value for a wide variety of important biological and medical issues, such as understanding aging, cancers, and other perplexing diseases. Dynamic biological network mining has attracted increasing attention from biologists in the past few years [1–3]. For example, we want to know which gene or protein is of critical effect to disease development. In this domain, microarray gene expression data offers useful dynamic information and is generally exploited to locate differentially expressed genes that may be related to specific abnormal conditions. A few tools are available for finding differentially expressed genes under varying conditions among which statistical methods are widely accepted, including methods based on t-test and SAM [4, 5]. However, change level of gene expression is such a representation that is far from satisfaction to explain the complex dynamic mechanism, considering that it is not capable of investigating the dynamic changes of relationships of proteins in consecutive protein-protein interaction networks (PPINs). For instance, methods based on sole gene expression analysis cannot capture genes with medium expression, but in contrast more accurate and complete understanding can be achieved by putting them into the PPIN.

There are mainly two challenges in dynamic network analysis. The first one is to construct the dynamic networks that accurately model the dynamic processes. Proteins perform their functions at specific times under distinguished conditions, which we call them in their active forms. Dynamic PPINs reveal the instant relationships of functional proteins. However, the publicly available PPI datasets are mostly aggregates of all possible interactions obtained under different examined conditions or time points [6] and are oblivious to the temporal changes of these networks. Dynamic analysis involves extracting dynamic PPI networks from these known datasets and the methods mainly fall into two directions. The first way is based on the differential coexpression correlations. Studies [7] have shown that highly positive coexpressed proteins tend to form the most static modules appearing at all times and at the center of which there are some hubs with high degree being referred to as "party" hubs. Further, some less positive coexpressed proteins interact at particular time points, the hubs therefore being referred to as "date" hubs that are believed to cause dynamic interactions and plausibly induce aberrant pathways and molecular disorders. Taylor et al. [8] also observed multimodal distribution of correlation coefficients of gene expression using curated sources from the literature. The second way to construct dynamic PPINs is based on expression variance [9] by determining the peak time points of expression for each protein. Thus, if a protein is at its peak point, it is considered to be in its active form, the status at which a protein can interact with its active neighbors. This assumption allows computing scored gene expression activity using a single threshold [10] or a systematical threshold [11]. In the present work, we assert that coexpression correlation may describe only the possible coregulated relationships of proteins, while existence of a specific interaction at a certain time point would depend on the activities of the two associated proteins. Hence, the integration of these two aspects becomes necessary in the construction of dynamic networks and a comprehensive way of defining the existence of dynamic PPIs is needed. In addition, some researchers argue that the gene expression data contain far more noise that will induce unauthentic factors. For example, the genes are sent into a filter that defines a criterion for genes of being dynamic or stable in Xiao et al.'s paper [12], and the stable ones are left out of the subsequent construction of dynamic networks. However, the definition of dynamic networks is slightly different from Xiao et al.'s. In our case, the stable active proteins are impartially included in the dynamic networks.

The second challenge of dynamic network analysis is to identify the most critical proteins out of a series of dynamic networks. As discussed above, Han et al. [7] concluded that hubs can be divided into two categories: the "party" and "date" hubs, among which the latter ones are more essential to global connectivity and functions that cells process. In this paper, the proteins exhibiting dramatic structural changes in the set of consecutive networks are defined as critical proteins which serve a compensation of the definition of "date" hubs to some extent. The intuition is that a series of networks in the same biological process should share a certain degree of consistence in structure. By extracting the consistence and reserving the structural difference of dynamic networks, we are able to find the critical proteins that are extremely important for the dynamic process. To this end, the consistent and varying properties of local structures in dynamic PPINs are studied in this work and a critical node detection method based on integration of multiple deep belief networks are proposed to identify the most critical proteins that are responsible for dynamic changes during a certain time period, and specifically, the case of yeast cell cycle processes is studied in our present work. There are several comparative methods of extracting consistent information from multiple graphs, such as the most straightforward average network and the joint nonnegative matrix factorization (JNMF) [13]. NMF tries to decompose the original graph to linear combination of basis vectors and is usually used in clustering problems, graph partition problems, and so on. In this work the hierarchical fashion of PPIN structures is taken into consideration by building a multisource integrated deep belief network (msiDBN) as a joint multilayer model that extracts the common higher levels of structural units. Our network is based on the previous work in [14]; however, in this msiDBN model, we decipher the structural varieties of nodes at different time points as the residuals of reconstruction and believe that the nodes with dramatic changes of structure in the networks play an important role in the progression of the cell cycle.

The framework of this present work is shown in Figure 1. It is assumed that a small part of proteins in the network is associated with the changing of dynamic processes, marked by red circles in Figure 1 as an example. The changing of local structure is studied through our msiDBN method and in summary this work contributes in there ways.

(i) A new method for constructing dynamic coregulated PPINs has been proposed and it gets better representation of dynamic process by comparing to other construction methods.

(ii) A multisource integrated deep belief network (msiDBN) is developed to extract the common representation of multiple networks, reconstruct the dynamic networks, analyze the residual of reconstruction, and identify the critical proteins in the yeast cell cycle processes.

(iii) Experimental results show that our strategy of dynamic network construction is superior to the other baseline methods and the msiDBN is able to reconstruct the dynamic networks with the lowest root mean square error (RMSE) than the comparative methods for it extracts the consistent hierarchical structures while others do not have any deep insight of the networks.

The rest of the paper is organized as follows: the proposed dynamic PPIN construction method is described in Section 2; Section 3 defines the critical protein identification problem in an anomaly detection fashion and introduces the msiDBN method and the critical node ranking criteria; and the proposed methods are evaluated in Section 4 from different aspects. Finally, the conclusion of this work is given in Section 5.

2. Dynamic PPI Network Construction

In the traditional dynamic network construction methods, the instant interactions between proteins are determined using either coexpression correlation or gene expression level shift. However, we construct the dynamic networks by integrating both assumptions in the present work.

2.1. Activity Determination. Different peak time points of gene expressions may represent the dynamic changes in

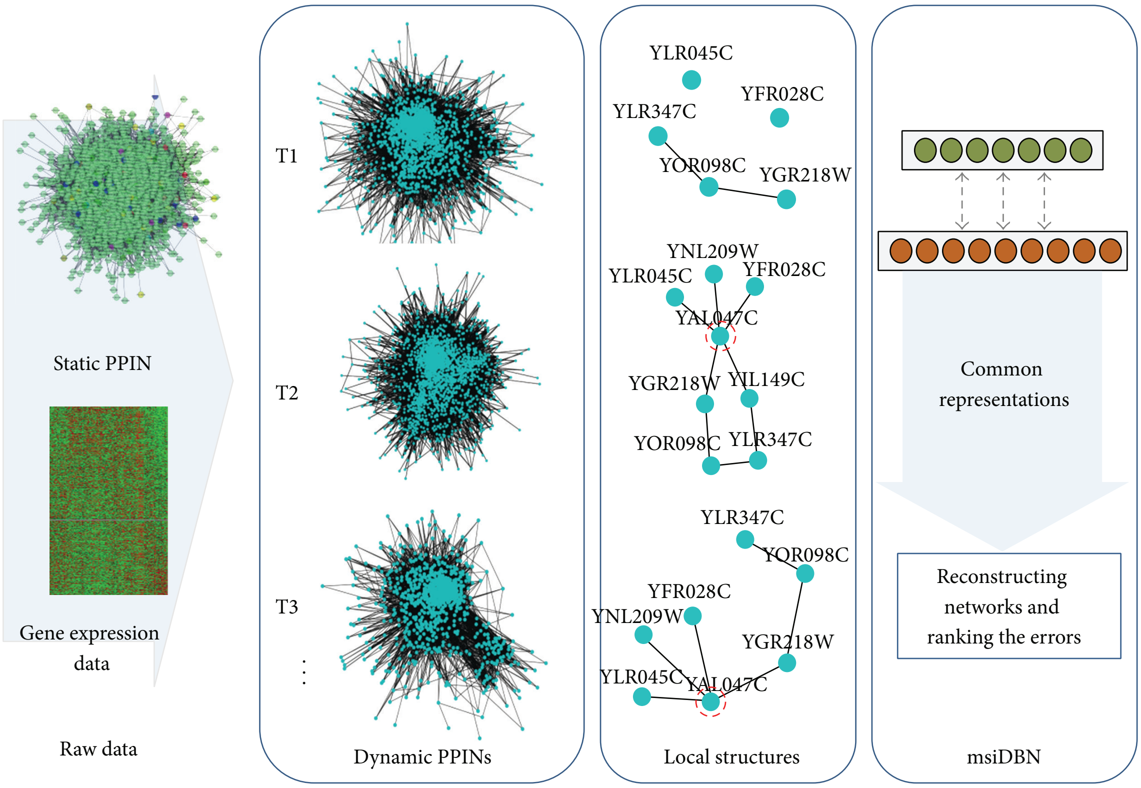


Figure 1: The framework of this paper.

Table 1: Active proteins and their interactions in different dynamic networks.

Methods	$T1$	$T2$	$T3$	$T4$	$T5$	$T6$	$T7$	$T8$	$T9$	$T10$	$T11$	$T12$
Our, AP	1068	1010	1019	882	853	828	898	1057	1195	1031	1086	1036
Our, Ins	10489	9890	10733	8100	6975	6609	7662	11584	15906	11850	12857	11040
Th = 0.7, AP	1071	1080	1062	962	902	842	843	1066	1162	991	1051	885
Th = 0.7, Ins	12267	13034	13653	10463	9345	8370	8133	13263	16302	13586	13640	10010
3segma, AP	531	545	505	393	364	343	361	603	688	473	545	449
3segma Ins	3003	3325	3140	1841	1462	1278	1588	4352	6380	3654	4123	2611

protein activities. Here we have assumed that proteins are active at their most peak points of gene expressions as discussed in Wang et al.’s work [11] and a similar threshold is set for the expression of each gene that is collected under continuous conditions. The active score is determined by

$$\text{AcScore}(p) = \text{thr}_1(p) \times F(p) + \text{thr}_2 \times (1 - F(p)), \quad (1)$$

where $\text{thr}_1(p)$ is the mean of the gene expression of protein p, which is also denoted as $\mu(p)$, $\text{thr}_2(p) = \mu(p) \times \sigma(p)$, where $\sigma(p)$ is the standard deviation of the gene expression of protein p, and $F(p) = 1/(1+\sigma(p))$. As seen from (1), $F(p)$ is a weight function of $\sigma(p)$ and occurs in the range of $(0, 1)$. An empirical parameter α was set for maintaining the active score AcScore within the range of $(\mu(p), \mu(p) + \alpha(\sigma(p))^3/(1 + \sigma(p)^2)))$. The performances of different empirical α have been discussed in the experimental section.

By setting such an active score threshold, the activity PPI networks Act_t were built for each timestamp:

$$\text{Act}_t = \delta_t \delta_t^T, \quad (2)$$

where δ_t is a column vector representing the activity of proteins at time t and δ_t^T is the transpose of the column vector. Each element in δ_t is determined by the binary threshold function as shown below:

$$\delta_t(p) = \begin{cases} 1 & \text{if } g_t(p) \geq \text{AcScore}(p), \\ 0 & \text{if otherwise.} \end{cases} \quad (3)$$

2.2. Combining with Coexpression Correlation and Static PPIN. It has been demonstrated previously that functionally related genes are frequently coexpressed across multiple conditions and different organisms [15]. Coexpression correlation coefficient is used as a measure of coexpressed genes having the same expression variance patterns across different conditions [16]. We have used the Pearson correlation coefficient [17] (normalized to the range of 0 to 1) to calculate the coexpression correlation and build coregulation protein networks. Since the computation of correlation coefficient requires expression data that cover a period of time, a time window was set on the original expression dataset which covered the time points from $t-1$ to $t+1$, where t is the current time point. The correlation coefficient matrix at time t is denoted as CoE_t. Combining the static PPIN and the activity PPIN provides the dynamic coregulation protein network at each time point:

$$A_t = \mathrm{CoE}_t \circ \mathrm{Act}_t \circ Ppi, \tag{4}$$

where Ppi denotes the static PPI network adjacency matrix and $\circ$ represents element-wise multiplication.

Given the adjacency matrices of networks, the structural difference of networks can be studied in many different ways. The most important point is to incorporate the changes induced by neighbours' behaviors. Hence, we use higher order of the adjacency matrices to mimic random walks on these networks while keeping the nonnegative property at the same time.

3. Critical Node Detection Based on Multisource Integrated Deep Belief Network

3.1. Definition of Critical Node Detection Problem in Dynamic Networks. Given a set of PPINs $\{A_1, A_2, \ldots, A_T\}$ under T time points, they are naturally evolving all along the biological process. The structure of network is represented by high order of adjacency matrix in this paper, which can be considered as the reachability of one node to the other in certain steps of rand walk. PPINs exhibit hierarchical structure and the trigger of changes in biological processes can be a small but complex set of molecules [18]. At a certain time point, the proteins in the PPIN are taken as nodes and each row of the high order adjacency matrix that a node corresponds to is seen as its feature at that time. There are T sources about the N nodes. To rank the most critical proteins, our intuition is that a node will receive low score if its topological structures of neighborhoods are consistent across the evolving networks and vice versa. It is impossible to directly compare the network structures because of the noise, sparsity, and indirect paths issues. However, because of the hierarchy of PPINs, we can extract hierarchical latent layers hidden in the networks that explain the evolution of network structures and protein functions. In other words, the hidden layers can be seen as the implied reasons of dynamic changes, by which the proteins fall into different groups of different characters of structural changes.

The flow of msiDBN is shown in Figure 2 where T matrices of evolving networks are fed in as inputs. The msiDBN model tries to find the latent layers $H_l^{(t)}$, representing the hidden variables of the lth layer for the tth network and the symmetrical weighted connections between input layers and hidden layers, that is, $W_l^{(t)}$. As shown in Figure 2, multiple inputs are trained separately at the beginning and then are all combined to extract the common factors in the top layer.

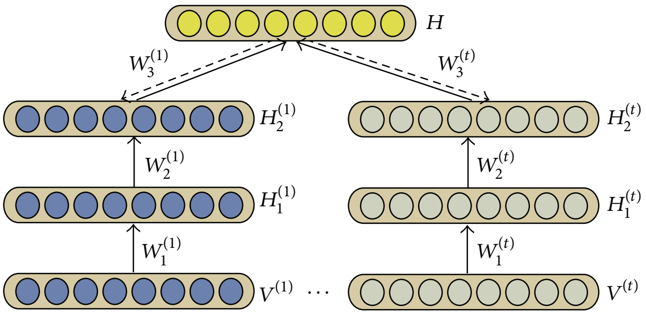


Figure 2: The flow of msiDBN.

3.2. DBN in Critical Protein Detection. To explain the framework of DBN, we should first go through the concept of restricted Boltzmann machines (RBMs), which are stacked one on top of each other to compose the DBNs [19]. RBM is defined as a network of symmetrically coupled binary random variables or units. As shown in Figure 3, these units can be divided into two groups: the visible variables, $v \in \{0,1\}^{|v|}$, and the hidden variables, $h \in \{0,1\}^{|h|}$ ($|\cdot|$ gets the dimension of the object inside it). The visible variables can be the original input or the transformed results from last layer according to the position of current RBM in the whole DBN model. The hidden variables imply the dependencies among the visible variables through their mutual interactional relationships as mimicked by the weighted matrix of W. In RBM, the interactions among visible-to-visible variables and among hidden-to-hidden ones are ignored [20]. Hence, we get a bipartite graph with completed connections.

The RBM defines an energy function between the visible and hidden layer variables:

$$E(v,h) = h^T W v + d^T h + b^T v, \tag{5}$$

where h and v are row vectors in H and V, respectively, b and d are the bias to the visible layer and hidden layer, and W is the weights between two layers. In RBM the training purpose is to learn the weights and biases between adjacent layers so that the energy function achieves its lowest level. The joint probability distribution of RBM with a normalization factor Z is

$$P(v,h) = \frac{1}{Z} E(v,h). \tag{6}$$

With the restricted conditions, the hidden variables are independent given the visible variables and this property factorizes the individual activation probabilities of a hidden variable as follows:

$$P\left(h_j = 1 \mid v\right) = \mathrm{sigmoid}\left(d_j + \sum_i W_{ij} v_i\right). \tag{7}$$

Likewise, we have the individual activation probabilities of a visible variable as

$$P\left(v_i = 1 \mid h\right) = \mathrm{sigmoid}\left(b_i + \sum_j W_{ij} h_j\right), \tag{8}$$

where the sigmoid represents the logistic sigmoid function.

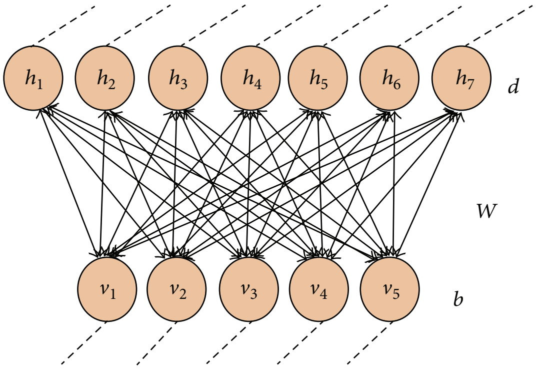


Figure 3: RBM in the DBN model.

To train the probabilistic models, we typically adapt and find the best parameters that maximize the likelihood of the training data. The most straightforward way is to maximize the likelihood following the log-likelihood gradient. However, in the gradient of the log-likelihood, there are terms that are intractable, that is, the ones that compute the expectations over the joints of variables v and h. There are several ways of dealing with this problem, like the contrastive divergence (CD) [21] which uses a very short Gibbs chain to estimate the expectation of the joints of v and h. The reliability of CD has been proved by different groups of researchers [22–24].

The RBM model extracts the latent variables hidden in the training data and several RBMs are stacked one on top of others, using the hidden variables derived from lower models as the input, to get deeper layer variables that explain the hierarchical factorizations of PPINs. Given l layers of RBMs, the joint distribution is

$$\begin{aligned} & P(v, h_1, h_2, \ldots, h_l) \\ & \quad = P(v \mid h_1) P(h_1 \mid h_2) \cdots P(h_{l-2} \mid h_{l-1}) P(h_{l-1} \mid h_l). \end{aligned} \tag{9}$$

As the variables inside each layer are independent and considering the biases for each layer, we get

$$P(h_{li} = 1 \mid h_{l+1}) = \text{sigmoid}\left(d_{li} + \sum_j W_{lij} h_{(l+1)j} \right). \tag{10}$$

3.3. Critical Protein Detection from Reconstruction of msiDBN. The msiDBN model to detect the critical proteins is built upon the assumption that most proteins have similar behavior patterns across the time courses while the most critical proteins that are responsible for the progression of the yeast cell cycle exhibit different expression levels and more importantly they engage in different interactions with contemporary neighbors. The other intuition is that the integrated deep belief networks extract the common features at the top layer (Figure 2) which represent the hidden deeper reasons for which the interactions change at different time points. Although we get merely J hidden variables at the top level, the feature space it can represent is scale of 2^J which is a much larger space than the common NMF clustering method gets. The joint probability of msiDBN is given as follows:

$$P\left(v^{(1)}, \ldots, v^{(t)}, h\right) = P\left(h_2^{(1)}, \ldots, h_2^{(t)}, h\right) \prod_t P\left(v^{(t)}, h_1^{(t)}, h_2^{(t)}\right), \tag{11}$$

where $P(h_2^{(1)}, \ldots, h_2^{(t)}, h)$ is

$$\begin{aligned} P\left(h_2^{(1)}, \ldots, h_2^{(t)}, h\right) \propto \exp\Biggl(& \sum_t \sum_i d_t h_{2i}^{(t)} + \sum_k c_k h_k \\ & + \sum_t \sum_{i,k} h_{2i}^{(t)} W_3^{(t)} h_k \Biggr), \end{aligned} \tag{12}$$

where d_t is the bias variable of $h^{(t)}$ and c_k is the bias of the top hidden variable h. As discussed above, the conditional distributions can be derived according to the independency conditions as follows:

$$\begin{aligned} P\left(h_2^{(t)} \mid h\right) &= \text{sigmoid}\left(d_t + \sum_i h_i W_{3i}^{(t)} \right), \\ P\left(h \mid h_2^{(1)}, \ldots, h_2^{(t)}\right) &= \text{sigmoid}\left(c + \sum_t \sum_i W_{3i}^{(t)} h_{2i}^{(t)} \right). \end{aligned} \tag{13}$$

The parameters of msiDBN can be learned approximately by greedy layer-wise training using CD. Therefore, with the common hidden variables and the trained weight matrices we are able to build an auto-encoder machine to reconstruct the input dynamic networks. The pseudo code in Algorithm 1 shows how to train the msiDBN model. With the common representation of multiple networks, we reconstruct them using (13) for the sampled data from $P(h_2^{(t)} \mid h)$ which can be viewed as the approximation of original data.

We quantify the reconstruction error using root mean square error (RMSE) which is denoted by Er and calculated as follows:

$$\text{Er}_i^{(t)} = \sqrt{\frac{1}{N} \sum_{j=1}^{N} \left(A_{ij}^{(t)} - A_{ij}^{(R)} \right)^2}, \tag{14}$$

where $A_{ij}^{(R)}$ denotes the reconstructed network and $\text{Er}_i \in \mathbb{R}^T$ is a vector representing the RMSE of protein i between the original data and reconstructed data across time 0 to T. The dispersion of Er_i is rated based on the relative standard error (RSD) which is RSD $= \sigma/\mu$, with μ denoting the mean and σ denoting the standard deviation of Er_i. The lower RSD scores correspond to the proteins that are well recovered by the model and also show average smoothness across time courses, while the higher RSD scores reveal the ones that are more likely having varying structures at different time points and are expected to play important roles during research like drug targets design.

```
Input: 2nd order of adjacency matrices of dynamic networks
    $A^{(1)}, \ldots, A^{(T)}$ and learning rate $\varepsilon$;
Output: Weight matrices $A^{(1)\prime}, \ldots, A^{(T)\prime}$, bias for each network
    $d_t$ and bias for top common layer $c$;
Initialize Weight matrices $A^{(1)\prime}, \ldots, A^{(T)\prime}$, $d_t$ and $c$
repeat
    repeat
        pick up the samples $v^{(t)}$ of proteins $s$ at all time points;
        for all hidden units $i$ do
            compute $P\left(h^{(1i)} = 1v^{(1)}, \ldots, v^{(T)}\right)$ using (14);
            sample $h^{(1i)}$ from $P\left(h^{(t)} = 1v^{(1)}, \ldots, v^{(T)}\right)$;
        end for
        for all $t$ in $[1 : T]$ do
            compute $P\left(v^{(t)'} h^{(1)}\right)$ using (13);
            sample $v^{(t)\prime}$ from $P\left(v^{(t)'} h^{(1)}\right)$;
        end for
        for all hidden units $i$ do
            compute $P\left(h^{(2i)} = 1v^{(1)'}, \ldots, v^{(T)'}\right)$ using (14);
        end for
    until for all samples in $A$
    $A^{(t)\prime} \leftarrow A^{(t)\prime} + \varepsilon\left(h^{(1)}v^{(t)'} - P(h^{(2)} = 1 \mid v^{(t)})v^{(t)'}\right)$;
    $c \leftarrow c + \varepsilon\left(h^{(1)} - P\left(h^{(2)} = 1v^{(1)'}, \ldots, v^{(T)'}\right)\right)$;
    $d^{(t)} \leftarrow d^{(t)} + \varepsilon\left(v^{(t)} - v^{(T)'}\right)$;
until all parameters are converged
```

Algorithm 1: Multisource integrated deep belief nets (msiDBNs).

4. Experiments and Results

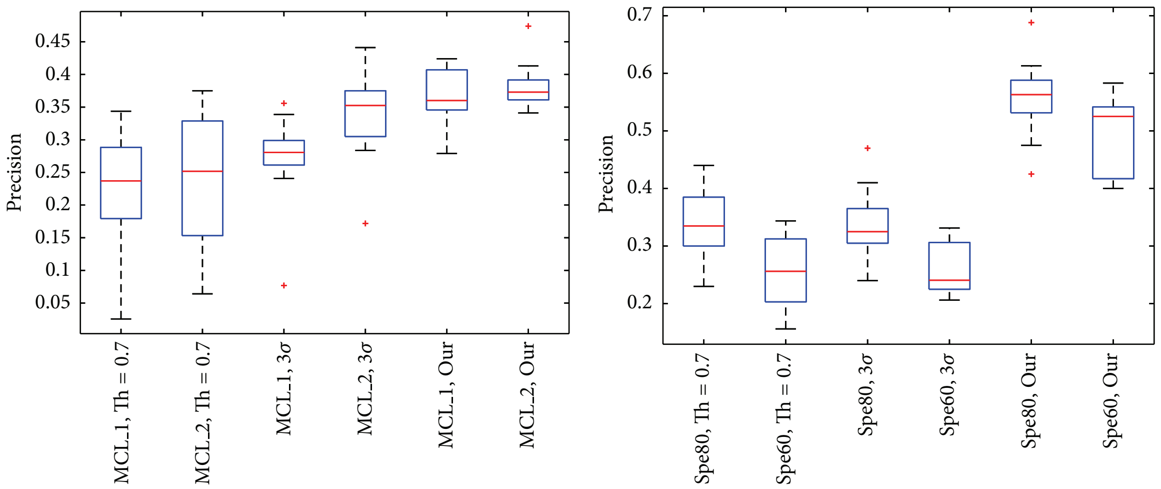
4.1. Datasets. The gene expression data from GSE3431 [25] was used as the time course data to construct time course PPINs. GSE3431 is an expression profiling of yeast over three successive metabolic cycles. The overall design of this expression experiment is 12 time intervals per cycle, and approximately 25 minutes per time interval. Thus, for each gene there are 12 expression values at 12 time points in each cycle. In order to calculate the instant coexpression correlation coefficient, we choose $t-1$, t, and $t+1$ as three time points in a snapshot and at each time point there are three successive expression values serving as replicate samples. Particularly, for the first time point of the cell cycle, the last time point was chosen as its previous time point, and vice versa. Further, we also adopted another reference cell cycle gene expression data for yeast indexed by GSE7645 to alleviate the bias of expression in the calculation of mean and variance for each gene. In the experiment generating GSE7645, *S. cerevisiae* was cultured under oxidative stress induced by cumene hydroperoxide (CHP) and the transcriptional profile is collected at $t = 0$ (immediately before adding CHP) and at 3, 6, 12, 20, 40, 70, and 120 minutes after adding the oxidant.

The static PPIN of yeast was collected from BioGRID dataset for yeast and the cell cycle regulated protein dataset was downloaded from http://nar.oxfordjournals.org/content/38/suppl_1/D699 which will serve as the golden data in validation. We also constructed the cell cycle related static PPIN based on these proteins and their first neighbors in BioGRID PPIN. Finally we get a static PPIN with 2069 proteins and 43462 interactions between them.

The function of detected modules was validated by adopting the CYC2008 human-curated complex dataset as benchmark data [26]. CYC2008 is a comprehensive catalogue of manually curated 408 heteromeric protein complexes in *S. cerevisiae* reliably backed by small-scale experiments from the literature.

4.2. Evaluation of Dynamic Network Construction. We compare the proposed dynamic network construction method with traditional methods, that is, methods from the work of Tang et al. [10] and Wang et al. [11]. It is widely believed that the dynamic network reveals more accurate functional interactions between proteins than static PPIN and also a better dynamic network construction method should achieve better functional module analysis results. Hence, we run two traditional clustering methods on the different sets of dynamic networks and compare the precision of module detection results. Known complexes in the CYC2008 dataset served as a gold-standard data to evaluate the experimental results.

It is expected for a module detection method that the predicted clusters (P_c) and the reference complexes (R_c)



(a) MCL performance with different parameter setting on three dynamic networks

(b) Spectral clustering performance with different community number on three dynamic networks

Figure 4: Different dynamic network construction methods.

match as much as possible. The overlapping scores $\mathrm{OL}(P_c, R_c)$ are used to find the matched complexes:

$$\mathrm{OL}\left(P_c, R_c\right) = \frac{\left|V_{P_c} \cap C_{R_c}\right|}{\left|V_{P_c}\right| \times \left|V_{R_c}\right|}, \tag{15}$$

where $|V_{P_c}|$ is the size of the predicted cluster, $|V_{R_c}|$ is the size of the known complex, and $|V_{P_c} \cap V_{R_c}|$ is the number of the intersections of the predicted cluster and the known complex. P_c and R_c are considered to be matched if their OL score is larger than a threshold σ, which is typically chosen as 0.2 [27, 28]. Precision is defined as Prec = TP/(TP + FP), where TP (true positive) is the number of the predicted clusters matched with known complexes by OL $> \sigma$, and FP (False Positive) is the number of the unmatched known complexes in the predicted clusters.

4.2.1. Comparison of Different Dynamic Networks Construction Methods. In the experiments, a fixed threshold (Th = 0.7) was set to Tang's method, and the second construction method is described in Wang et al.'s paper [11] which uses a three-sigma threshold to determine the status of proteins. Table 1 shows the numbers of active proteins derived from three different threshold setting methods. In our method, the parameter α was chosen to be 1.5. The performance of MCL and spectral clustering method on different dynamic PPI construction methods (Figure 4) indicates that our integrative method is more effective in constructing dynamic networks, and both functional module detection methods benefit from it by achieving the highest precisions compared to the other two dynamic PPIN construction methods.

4.2.2. Parameter Setting. Initially, we analyzed the effect of change in parameters on our dynamic network construction method. Thus, we performed the spectral method to detect dynamic functional modules at 12 time points and compared the results with the CYC2008 dataset. The Precs of the results under different parameter settings have been compared as shown in Table 2 and the mean and variance of Prec are shown in Figure 5. From the results of comparison, it was obvious that on fixing α at 1.5 the precision of the functional module detection achieved the highest score. Thus, in the following comparisons with other module detection methods, this prime parameter setting has been used.

4.3. Critical Proteins Identification with msiDBN. In the msiDBN model, the visible input variables were chosen as the high order of the adjacency matrices; in this case the 2nd-order was used, and the self-transmissions were ignored which meant that the diagonal elements in each matrix were set to 0. We built the separate training DBNs as 2-layer models and combined different sources on the third layer as shown in Figure 2. We compared our method with two basic reconstruction methods and also with the original DBN to verify the effectiveness of our method.

An online resource of cell-cycle-related gene list from *Cyclebase* database was used as the golden data which contains 150 genes that are proved to be related to yeast cell cycle process. The evaluation metric of this experiment was also Prec with the same definition as the one introduced above.

The baseline methods include joint NMF (JNMF) method, the straightforward average network, and the original DBN method. The JNMF method learns a common base matrix from different sources that best approximates the original sources. It is often used in clustering problems and dimension reduction problems. In our experiments the prior low dimension of JNMF was set as 500 by which the approximation to the original data generally achieved the best position. And the method which adopts the average network,

Table 2: Parameter settings, $\alpha \in (0.5{\sim}3)$.

α	$T1$	$T2$	$T3$	$T4$	$T5$	$T6$	$T7$	$T8$	$T9$	$T10$	$T11$	$T12$	Average
0.5	0.270	0.350	0.325	0.325	0.375	0.350	0.200	0.375	0.225	0.400	0.275	0.325	0.316
1	0.340	0.483	0.483	0.317	0.350	0.417	0.283	0.450	0.400	0.333	0.333	0.317	0.376
1.5	**0.533**	**0.517**	**0.483**	**0.400**	**0.417**	**0.417**	**0.400**	**0.550**	**0.583**	**0.533**	**0.567**	**0.533**	**0.494**
2	0.330	0.400	0.480	0.350	0.320	0.320	0.380	0.460	0.490	0.490	0.460	0.500	0.415
2.5	0.470	0.387	0.445	0.328	0.345	0.312	0.320	0.478	0.495	0.545	0.478	0.470	0.423
3	0.423	0.394	0.437	0.280	0.287	0.316	0.316	0.451	0.473	0.501	0.423	0.416	0.393
3.5	0.390	0.371	0.421	0.315	0.290	0.265	0.303	0.421	0.528	0.453	0.415	0.415	0.382

Table 3: GO enrichment of unmatched proteins in the top 150 list.

Protein	GO-ID	Term description
YGL016W	GO:0016021	Integral to membrane
	GO:0006606	Protein import into nucleus
YKL203C	GO:0000080	G1 phase of mitotic cell cycle
	GO:0007049	Cell cycle
	GO:0030037	Actin filament reorganization involved in cell cycle
YBR078W	GO:0031505	Fungal-type cell wall organization
	GO:0031225	anchored to membrane
	GO:0005618	cell wall
YLL031C	GO:0015867	ATP transport
	GO:0009277	fungal-type cell wall
YBR122C	GO:0032543	Mitochondrial translation
	GO:0005762	mitochondrial Large ribosomal subunit
	GO:0005739	Mitochondrion

denoted as AVG in the following content, simply extracted the average of the 2nd order adjacencies of the series of dynamic networks. Comparing with our msiDBN, the DBN method just processes our 12 networks through one straightforward deep structure of three layers to get the common representation and derive the reconstruction errors. By comparing the RMSEs, it is easy to see in Figure 6 that the msiDBN method obtains the best reconstruction while the AVG gets the worst in all of the four methods. The property of msiDBN is to extract a hierarchy of hidden features which naturally meets the characteristics of PPIN. The JNMF is analogous to one layer feature extraction model that does not fit in to its best within this scenario. In addition, our method surpasses the traditional DBN which considers all the networks identically and shows the promising results of the framework of multisource integrated deep belief networks.

As we know, RSD is a measure that quantifies whether a set of variables are constant or have more variabilities. A high RSD number indicates that the data is more varied. The RSD scores of each protein were calculated and ranked to get the top 150 proteins in these three methods. The proteins in the top 150 RSD lists, derived from these three methods, were compared with the golden standard list from *Cyclebase* database. We ran msiDBN and JNMF 100 times separately to get the average performance and the comparison of precision results is shown in Figure 7. The matched proteins that are truly associated with cell cycle process in msiDBN were around 75, which was much higher than JNMF and AVG method. Among those unmatched proteins inside our top 150 list, we saw an interesting thing: according to the gene ontology annotations, a few of the unmatched proteins are relative to the true cell-cycle-related proteins, as shown in Table 3 where just a part of the unmatched proteins are listed due to limitations on space. We also checked in the static PPI network and discovered that most of the unmatched proteins are directly linked to the cell-cycle-related proteins or within short length of link distance. This phenomenon is consistent with our structural changing assumption about critical proteins in which for a critical protein, which has been varied structure during the cell cycle progression, the structure of its neighbors must change along with it.

5. Conclusion

In this paper, the structural variability of dynamic PPINs was studied to identify the critical proteins in the yeast cell cycle process. A comprehensive method of constructing dynamic active PPI networks was proposed which simultaneously modeled the activity of proteins and assembled the dynamic coregulation protein network at each time point. And then a critical node detection method that integrated multiple networks into deep belief network model was developed, in which the reconstruction results were ranked by the variabilities of the reconstruction errors across time courses and finally the top proteins in rank order were selected to be the critical ones that may play important roles in

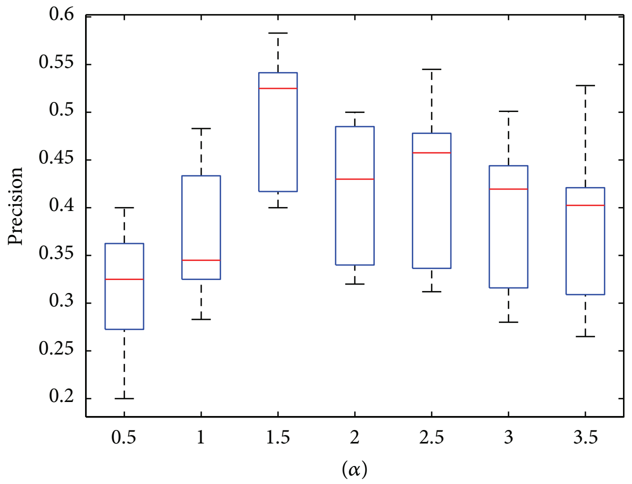


FIGURE 5: The distribution of precision under different parameter settings.

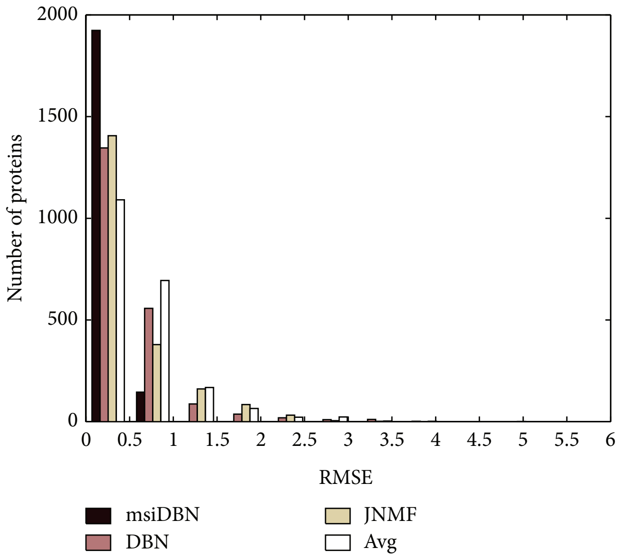


FIGURE 6: Comparison of RSME.

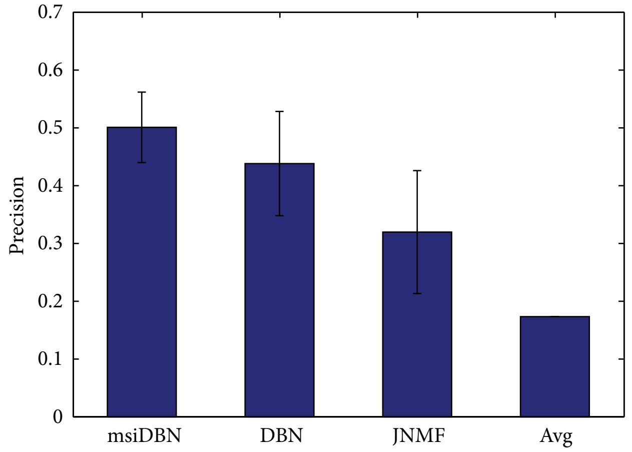


FIGURE 7: Precision comparison of different methods.

dynamic mechanisms. We evaluated our network construction method by comparing the functional representations of proteins in the derived networks with that from two traditional construction methods and our method achieved superior function analysis results. The critical protein ranking results from msiDBN were compared with results from JNMF reconstruction method and the comparison of results showed that msiDBN had better reconstruction rate and identified more proteins of critical value to yeast cell cycle process.

The fact that a few proteins among the unmatched protein lists are truly relevant to the cell cycle process inspires an interesting idea that the system analysis of dynamic networks should be done to reveal groups of critical proteins with the same or relative functional roles in the dynamic mechanism. We will focus more on a system level study of the dynamic networks in the future research.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant no.s 81000624 and 81370038, the National Natural Science Foundation of Beijing under Grant no. 7142012, the Scientific Research Project of Beijing Educational Committee under Grant no. km201410005003, and the Rixin Fund of Beijing University of Technology under Grant no. 2013-RX-L04.

References

[1] X. Chang, T. Xu, Y. Li, and K. Wang, “Dynamic modular architecture of protein-protein interaction networks beyond the dichotomy of “date” and “party” hubs,” *Scientific Reports*, vol. 3, 2013.

[2] K. Komurov and M. White, “Revealing static and dynamic modular architecture of the eukaryotic protein interaction network,” *Molecular Systems Biology*, vol. 3, no. 1, pp. 1–11, 2007.

[3] U. De Lichtenberg, L. J. Jensen, S. Brunak, and P. Bork, “Dynamic complex formation during the yeast cell cycle,” *Science*, vol. 307, no. 5710, pp. 724–727, 2005.

[4] V. G. Tusher, R. Tibshirani, and G. Chu, “Significance analysis of microarrays applied to the ionizing radiation response,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 9, pp. 5116–5121, 2001.

[5] W. Pan, “A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments,” *Bioinformatics*, vol. 18, no. 4, pp. 546–554, 2002.

[6] N. Du, Y. Zhang, K. Li et al., “Evolutionary analysis of functional modules in dynamic ppi networks,” in *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, pp. 250–257, ACM, 2012.

[7] J.-D. J. Han, N. Bertin, T. Hao et al., “Evidence for dynamically organized modularity in the yeast protein–protein interaction network,” *Nature*, vol. 430, no. 6995, pp. 88–93, 2004.

[8] I. W. Taylor, R. Linding, D. Warde-Farley et al., “Dynamic modularity in protein interaction networks predicts breast

cancer outcome," *Nature Biotechnology*, vol. 27, no. 2, pp. 199–204, 2009.

[9] K. Tarassov, V. Messier, C. R. Landry et al., "An in vivo map of the yeast protein interactome," *Science*, vol. 320, no. 5882, pp. 1465–1470, 2008.

[10] X. Tang, J. Wang, B. Liu, M. Li, G. Chen, and Y. Pan, "A comparison of the functional modules identified from time course and static PPI network data," *BMC Bioinformatics*, vol. 12, article 339, 2011.

[11] J. Wang, X. Peng, M. Li, and Y. Pan, "Construction and application of dynamic protein interaction network based on time course gene expression data," *Proteomics*, vol. 13, no. 2, pp. 301–312, 2013.

[12] Q. Xiao, J. Wang, X. Peng, and F. X. Wu, "Detecting protein complexes from active protein interaction networks constructed with dynamic gene expression profiles," *Proteome Science*, vol. 11, supplement 1, p. S20, 2013.

[13] L. Ge, J. Gao, X. Yu, W. Fan, and A. Zhang, "Estimating local information trustworthiness via multisource joint matrix factorization," in *Proceedings of the IEEE 12th International Conference on Data Mining (ICDM '12)*, pp. 876–881, 2012.

[14] L. Ge, J. Gao, X. Li, and A. Zhang, "Multi-source deep learning for information trustworthiness estimation," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '13)*, pp. 766–774, ACM, New York, NY, USA, 2013.

[15] H. K. Lee, A. K. Hsu, J. Sajdak, J. Qin, and P. Pavlidis, "Coexpresion analysis of human genes across many microarray data sets," *Genome Research*, vol. 14, no. 6, pp. 1085–1094, 2004.

[16] K. Basso, A. A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera, and A. Califano, "Reverse engineering of regulatory networks in human B cells," *Nature Genetics*, vol. 37, no. 4, pp. 382–390, 2005.

[17] N. Bhardwaj and H. Lu, "Correlation between gene expression profiles and protein-protein interactions within and across genomes," *Bioinformatics*, vol. 21, no. 11, pp. 2730–2738, 2005.

[18] V. Arnau, S. Mars, and I. Marín, "Iterative cluster analysis of protein interaction data," *Bioinformatics*, vol. 21, no. 3, pp. 364–378, 2005.

[19] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[20] P. Smolensky, "Information processing in dynamical systems: foundations of harmony theory," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, D. E. Rumelhart, J. L. McClelland, and C. PDP Research Group, Eds., vol. 1, pp. 194–281, MIT Press, Cambridge, Mass, USA, 1986, http://dl.acm.org/citation.cfm?id=104279.104290.

[21] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[22] A. Yuille, "A comment on contrastive divergence," Tech. Rep., Proc. of NIPS, 2004.

[23] Y. Bengio and O. Delalleau, "Justifying and generalizing contrastive divergence," *Neural Computation*, vol. 21, no. 6, pp. 1601–1621, 2009.

[24] I. Sutskever and T. Tieleman, "On the convergence properties of contrastive divergence," in *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS '10)*, 2010.

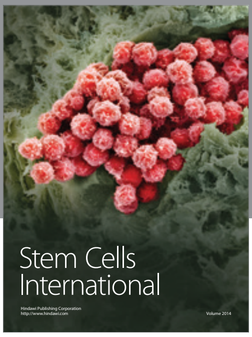
[25] B. P. Tu, A. Kudlicki, M. Rowicka, and S. L. McKnight, "Cell biology: logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes," *Science*, vol. 310, no. 5751, pp. 1152–1158, 2005.

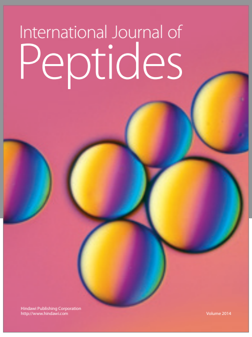
[26] S. Pu, J. Wong, B. Turner, E. Cho, and S. J. Wodak, "Up-to-date catalogues of yeast protein complexes," *Nucleic Acids Research*, vol. 37, no. 3, pp. 825–831, 2009.

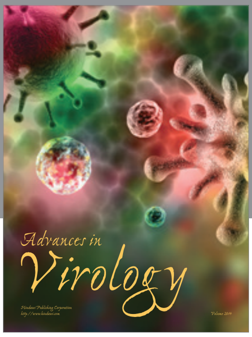
[27] F. Markowetz and O. G. Troyanskaya, "Computational identification of cellular networks and pathways," *Molecular BioSystems*, vol. 3, no. 7, pp. 478–482, 2007.

[28] Y.-R. Cho, L. Shi, and A. Zhang, "FlowNet: flow-based approach for efficient analysis of complex biological networks," in *Proceedings of the 9th IEEE International Conference on Data Mining (ICDM '09)*, pp. 91–100, December 2009.

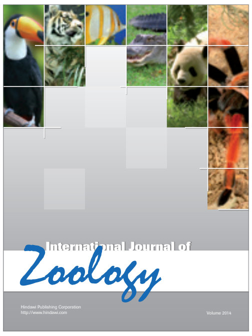
BioMed
Research International

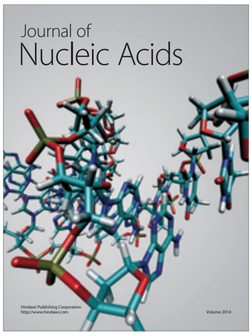

Stem Cells
International


International Journal of
Peptides


Advances in
Virology

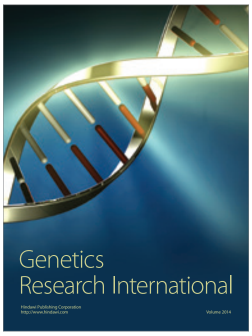
International Journal of
Genomics

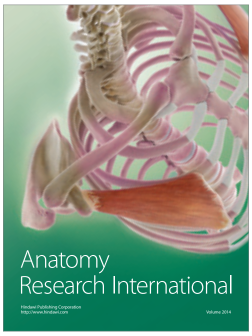

International Journal of
Zoology


Journal of
Nucleic Acids

Journal of
Signal Transduction

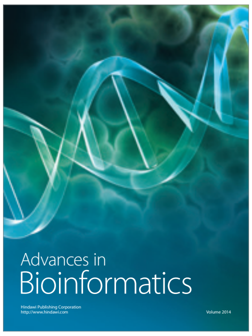
The Scientific
World Journal

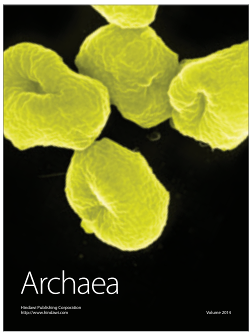

Genetics
Research International

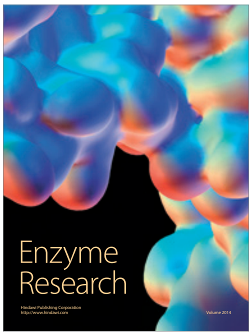

Anatomy
Research International

International Journal of
Microbiology

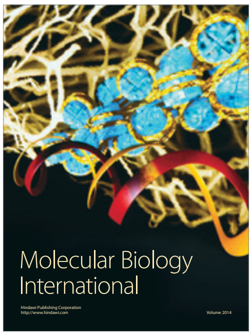
Biochemistry
Research International


Advances in
Bioinformatics


Archaea


Enzyme
Research

International Journal of
Evolutionary Biology


Molecular Biology
International

Journal of
Marine Biology