A Hybrid Data and Knowledge Driven Approach for
Gene Clustering and Network Reconstruction

Chung-Hsun Lin   Yu-Ting Hsiao   Wei-Po Lee
Department of Information Management
National Sun Yat-sen University
Kaohsiung, Taiwan

Abstract. Many methods have been developed for reverse engineering gene networks from time series expression data. However, when the number of genes and the complexity of regulation increase, it becomes increasingly difficult to infer gene networks. To tackle this scalability problem, this study presents an approach with two phases: gene clustering and network reconstruction. To perform gene clustering, a hybrid data and knowledge-driven method is developed to calculate both data and semantic similarity between genes. In the network reconstruction procedure, a Boolean network model is inferred from the gene clusters. A series of experiments are conducted to investigate the effect of the hybrid similarity measure in gene clustering and network reconstruction. The results prove the feasibility and effectiveness of the proposed approach.

Keywords: gene network inference; gene clustering; principal component analysis; knowledge ontology; Boolean network

1 Introduction
Gene network construction is one of the most important issues in systems biology research. The procedure involves deriving a model that can describe the observed phenotypic behavior of a target system, which often includes time series expression data. An automated reverse engineering procedure is advocated to save the effort of repeatedly identifying possible interactions in the gene networks and adjusting the network accordingly [1,2]. Many computational methods have been proposed that can successfully infer gene networks [3,4]. However, when the number of genes involved in the interactions increases, network construction becomes more difficult.

To solve the scalability problem for inferring large networks, the decomposition procedure has been proposed. Clustering is a practical technique in genomic studies and is useful for grouping genes [5]. Gene clustering assumes that the genes in a cluster may share some common functions or regulatory elements and can thus be considered and processed together. The goal of the clustering process is to identify genes that have the same functions or regulatory mechanisms. The quantitative expression levels of \( n \) genes under \( d \) different conditions (or time points) can be thought of as \( n \) points in \( d \)-dimensional space. Clustering methods group those points locating closely together in \( d \)-dimensional space. Each gene cluster can be considered a sub-network and inferred separately then gradually combined.

The first step in gene clustering is to determine how to measure the similarity (or dissimilarity) between any two genes; that is, to define the similarity function. All genes in a dataset are then assigned by the clustering algorithms into different clusters.
of similar expression patterns according to a similarity/dissimilarity measure. One popular method to perform such a measurement is to extract data features from the original gene regulatory signals, and the gene distance can then be calculated accordingly. The principal component analysis (PCA) method is a widely used technique that analyzes multivariate data and exploits the data characteristics in order to extract important features for dimension reduction [6,7]. It is a coordinate transformation in which each data is written as a linear sum over basis vectors called principal components (PCs). Usually, only more important PCs are retained to reduce the dimensionality of the data.

The other method to measure data distance is to tackle the problem with a domain knowledge mindset and to consider the gene semantic similarity [8,9]. Biological knowledge can be obtained from scientific literature or public databases, among which gene ontology (GO, [10]) is most popular and prominent biological knowledge source. GO is specifically intended to annotate gene products with these vocabulary terms, based on their functions in the cell. Though GO is a useful method to derive biological similarity between genes, it is notable that the appropriate use of functional similarity measures depends on the applications [11,12]. A given measure can yield good performance for one application but performs poorly for another.

Considering the construction of gene networks, many models have been proposed to address different levels of biological details [3,4]. Abstract models involve less biological detail and display only qualitative dynamic behavior; they are therefore uniquely capable of implementing large-sized networks. In contrast, concrete models describe network dynamics in detail and are closer to biological reality. However, due to their computational complexity, these models can only be applied to very small systems. Because the goal of this work is to construct large-sized networks, we choose to use the most popular abstract model, the Boolean model, in which genes are Boolean variables that only exist in discrete states (i.e., either on or off) [13].

To investigate the effect of using gene clustering to assist the construction of large Boolean networks from gene profiles, we present an approach that includes two major phases. The first phase is gene clustering, in which a hybrid method of data and knowledge-based clustering is developed to extract specific features from the original gene profiles. The PCA method is adopted to derive data features and calculate data similarity, and the GO method is employed to measure the semantic similarity between genes. The two types of similarities are then assembled together and used as the gene-gene distance during the clustering process. The second phase is a network reconstruction procedure that adopts the software BoolNet ([14]) to build networks from the gene clusters. To evaluate the presented approach, a series of experiments have been conducted, and the results show that the presented approach can produce clusters with high interactions between genes, which can lead to better network construction performance. Also, a small network obtained from our two-stage procedure has been analyzed and interpreted for further verification. These results confirm the feasibility and effectiveness of the approach developed herein.

2. The Proposed Method

To reconstruct large networks from gene profiles, in this work we devise a divide-and-conquer approach: the entire set of expression data is divided into strongly
correlated subsets by a clustering method, and small networks are inferred from the subsets of expression data and assembled. To perform gene clustering, we define a criterion of similarity measurement based on the features extracted from the gene data. It is notable that gene networks derived from expression data and gene ontology often disagree with each other, because the information on the time series is insufficient to comprehend the complexity of the GRNs due to the fact that not only mRNA but also other RNA species regulate the gene expression. Therefore, we develop an integrative measurement to take into account both data level and knowledge level information. Two methods are used here for feature extraction and distance measurement: the PCA method to extract features from the expression data; and a knowledge-based method to derive features from gene ontology. Once the gene groups are obtained, a reverse engineering procedure is performed for network reconstruction. Fig. 1 depicts the main phases of our approach, and the following subsections describe how the relevant computational methods, coupled with the integrative gene distance measurement, can be applied to the specific case of gene clustering and network reconstruction.

2.1 Measuring Gene Distance

In addition to the high dimensionality issue mentioned above, another well-recognized challenge in dealing with microarray data is that gene expression could be correlated in a complex way. Many genes interact and highly co-regulate one another; they may share the same molecular function and be involved in the same biological pathway. To tackle these difficulties, we adopt the PCA technique to process time series data. Here, each dataset is considered a data matrix including gene variables as rows and their observations as columns. PCA is used to construct linear combinations of gene expressions (i.e., PCs) that can effectively represent effects of the original measurements. In this way, correlated genes can be projected in the same direction because the data dimension reduction is based solely on gene expression.

Because PCs are constructed with the goal of explaining variation, it makes them difficult to interpret. Thus, we use the common rules for choosing how many PCs to retain. That is, we keep enough PCs so that the cumulative variance explained by the PCs is larger than a pre-specified threshold (e.g., 70% in this work, based on a preliminary test). With these PCs retained, the original expression data can be reduced to a lower dimensional subspace. Thus, each data record for gene $g_i$ original represented as $[s_{i1}, s_{i2}, ..., s_{im}]$ in which $m$ is the sample size of the expression data, can now be represented as a new feature vector $[g_{i1}, g_{i2}, ..., g_{in}]$ with a smaller data dimension $n$ (the number of PCs retained). The Euclidean distance measure is then applied for gene-gene similarity calculations. That is, for any two genes $g_1, g_2$
represented as the above PC vectors, the similarity between them (based on the expression data) is calculated by

\[ \text{Sim}_{\text{exp}}(g_1, g_2) = \sqrt{\sum_{d=1}^{D} (g_{1d} - g_{2d})^2} \]  

(1)

The other important part of measuring gene-gene distance is to take domain knowledge into consideration so that genes can be further analyzed from the perspective of molecular function. Here, we use the most popular and prominent biological knowledge source, gene ontology, for gene analysis. The GO resource contains three independent ontologies: molecular functions, biological processes, and cellular components. Each ontology serves as an organizing principle for describing gene products and is a hierarchical classification scheme structured as a directed acyclic graph. We choose to use the biological process ontology to annotate each gene (i.e., the corresponding proteins) of a network, because two proteins interacting physically are more likely to be involved in similar biological processes. In this study, two popular GO-based similarity measure methods are used to determine the data distance in the feature space, they are described as below.

The first method is an information content-based method adopted from Resnik’s algorithm ([15,16]). In this algorithm, while measuring the similarity of two concepts in a specific taxonomy (such as GO) the main focus is on investigating how they share information. Following the information theory, the information content of a concept \( c \) can be quantified as \( -\log p(c) \), where \( p(c) \) is the probability of encountering an instance of concept \( c \) in taxonomy. This means while the probability increases (the more abstract a concept), the informativeness decreases (the lower its information content). The similarity measure of two GO terms relies on the information content of the minimum subsumer, which is their lowest common ancestor in the GO hierarchy. In this way, the similarity of two GO terms \( g_1 \) and \( g_2 \) can be defined as in the following Equation (2)

\[ \text{Sim}_{\text{GO}}(g_1, g_2) = \max_{g = S(g_1, g_2)} [-\log p(g)] \]  

(2)

In the above equation, \( S(g_1, g_2) \) is the set of all common ancestor nodes that subsume both \( g_1 \) and \( g_2 \). A node with the maximum value is termed the most informative subsume. Here, taking the maximum with respect to information content is analogous to taking the first intersection in semantic network marker-passing or the shortest path with respect to edge distance. More calculation details are referred to in [15,16].

In the above equation, \( S(g_1, g_2) \) is the set of all common ancestor nodes that subsume both \( g_1 \) and \( g_2 \). That is, \( S = \{ s | s \in \text{path}(g_1, \text{root}) \land s \in \text{path}(g_2, \text{root}) \} \) in which \( \text{path} \) denotes a path from its first argument node to its second argument node, and \( \text{root} \) is the root node of the ontology. A node with the maximum value is termed the most informative subsume. Here, taking the maximum with respect to information content is analogous to taking the first intersection in semantic network marker-passing or the shortest path with respect to edge distance. Several assessments have shown that this method can provide consistently high correlation with sequence similarity and gene co-expression.

Differing from the above content-based method, the topology-based measures use the intrinsic topology of the GO direct acyclic graph. Taking into account that the
specificity of a GO term is usually determined by its location in the GO graph, Wang et al. proposed a graph-based strategy to compute semantic similarity using the topology of the GO graph structure [17]. In Wang’s method, the semantics of GO terms are encoded into a numeric format, and the different semantic contributions of the distinct relations are taken into consideration. In this approach, the semantic value of a GO term \( A \) (i.e., \( SV(A) \)) is given by

\[
SV(A) = \sum_{i \in T_A} S_i(t) \quad (3)
\]

In the above equation, \( T_A \) denotes the set of ancestors of term \( A \); \( S_i(t) \) is 1 if \( t = A \), otherwise \( S_i(t) \) is defined as \( \max \{w_r \times S_i(t') \mid t' \in C(A)\} \), in which \( C(A) \) is the set of children of term \( A \); and \( w_r \) represents the semantic contribution factor for “is_a” and “part_of” relations in the hierarchy. These relations are set to the often used values 0.8 and 0.6, respectively. Taking the case shown in [17] as an example (illustrated in Fig. 2), to determine the semantic value for a GO term 0043231, we have to find the graph including all nodes (terms) on the paths from term 0043231 to the root. Then, the semantic value for each node is calculated according to Equation (3). For example, starting from the node 0043221, the value for node 0043229 is 0.8 (i.e., 1 × 0.8), and the value for node 0005622 is 0.48 (i.e., 1 × 0.8 × 0.6). The semantic value for node 0043231 is thus the sum of all values of the nodes in the graph.

![Diagram of GO terms](image)

**Fig.2.** Example of the graph formed by the GO term 0043231

With the above calculation, the term-to-term similarity can be derived. For two genes \( g_1 \) and \( g_2 \) with sets of GO terms \( T_1 \), \( T_2 \) that annotate \( g_1 \) and \( g_2 \), respectively, the similarity of two GO terms \( t_1 \in T_1 \) and \( t_2 \in T_2 \) is defined as

\[
Sim_{GO}(t_1, t_2) = \frac{\sum_{i \in T_1 \cap T_2} (S_i(t_1) + S_i(t_2))}{SV(t_1) + SV(t_2)} \quad (4)
\]

Because none of the existing measures account for all aspects of GO, it is hypothesized that integrating multiple measures can improve the performance. Therefore, in this work, the above two measures are both taken and then averaged to
obtain the semantic similarity between genes. Also, it is notable that the main use of GO semantic similarity measures is the computation of protein semantic similarity (or functional similarity) between proteins based on their GO annotations. Because each gene may be annotated by multiple GO terms, different similarity measures for a pair of genes can thus be obtained. In this work, we take the average value as the similarity between genes.

2.2 Gene Clustering

After defining the two types of similarity measures (i.e., data-based and knowledge-based) between genes, this section describes how they are used in the clustering algorithm to group genes. Because each single gene in the network may be involved in different biological functions and interact with many other genes, a flexible strategy for gene discrimination in the clustering process provides a better choice. Therefore, we adopt the fuzzy \( c \)-means algorithm ([18,19]) to cluster genes. This algorithm attempts to divide a set of gene elements into a set of fuzzy clusters based on a given criterion. Each gene element may belong to more than one cluster simultaneously with different degrees during the clustering process. More calculation details on fuzzy clustering are referred to [18,19].

As mentioned previously, using gene expression or gene ontology alone is not enough to capture the interaction relationships among genes. Therefore, in our approach the similarity between any two gene elements \( g_1 \) and \( g_2 \) is calculated by a hybrid of data-based (\( \text{Sim}_{\text{exp}} \)) and knowledge-based (\( \text{Sim}_{\text{GO}} \)) measures with two weighting factors \( w_1 \) and \( w_2 \), as below:

\[
\text{Sim}(g_1, g_2) = w_1 \times \text{Sim}_{\text{exp}}(g_1, g_2) + w_2 \times \text{Sim}_{\text{GO}}(g_1, g_2) \quad (5)
\]

The algorithm repeatedly generates the membership levels for each gene element, which are then used to calculate the new centers of the clusters. It terminates when no improvement is observed (i.e., the change in the maximal membership is less than a pre-specified threshold \( \varepsilon \)).

The performance of gene clustering is evaluated by the protein-protein interaction (PPI) rate. It is defined as the fraction of interacting pairs found among all gene pairs that end up in the same cluster. As a validation information resource, we use a genomic database (i.e., BioGRID, that attempts to catalogue all known PPI ([20,21]) as a straightforward extension of the idea) to validate expression derived gene clusters using GO. More precisely, the PPI of a cluster \( c_i \) is defined as

\[
PPI_i = \frac{\text{num of (interacting gene pairs } \in \text{ } c_i)}{\text{num of (all gene pairs } \in \text{ } c_i)} \times 100\% \quad (6)
\]

In the above equation, \( i \) is the index of a certain cluster \( c_i \) and where \( k \) is the total number of clusters (i.e., \( 1 \leq i \leq k \)). Taking the average across all \( k \) clusters, one may also define the global measure assessing the quality (biological significance) of the whole partition.

2.3 Network Reconstruction

As indicated previously, Boolean networks rely on the simple logical theory of a binary system. It is a relatively simple model that is applicable to large-scale data
analysis. Because our goal is to verify the effectiveness of the proposed hybrid similarity measurement in gene clustering rather than the inference method, we adopt the Boolean network model and use the popular software BoolNet ([14]) to infer networks from clustered expression data.

BoolNet is an R package that provides tools for assembling, analyzing, and visualizing Boolean networks. We choose the synchronized mode for network reconstruction. In the synchronous model, the assumption is that all genes are updated at the same time. This simplification facilitates the analysis of the networks. In this tool, a novel binarization method is developed to address the gene profiles of continuous values. This method is based on a threshold determination using scan statistics, and can provide a measure of threshold validity. The main idea is to search for at least one cluster in the measurements whose probability $p$ is lower than a specified significance level. If there is no cluster with the required significance, then no reliable binarization can be determined because the data are distributed uniformly. Based on the detected clusters with high significance (i.e., a low value of $p$), the binarization is performed in such a way that the points within a cluster are assigned to the same binary value.

To evaluate the results of network reconstruction (or, more precisely, link prediction), different criteria often used in binary classification are taken. The first is precision, defined as the fraction of retrieved instances that are relevant. The second is recall, defined as the fraction of relevant instances that are retrieved. Although often in conflict in nature, the measures of precision and recall are both important in evaluating the performance of a prediction approach. Therefore, these two measures can be combined with equal weights to obtain a single metric, the $F$ metric. The three performance evaluation metrics are defined as follows:

$$\text{precision} = \frac{TP}{TP + FP} \quad (7)$$

$$\text{recall} = \frac{TP}{TP + FN} \quad (8)$$

$$F - \text{measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (9)$$

In the above measures, $TP$, $FP$, $TN$, and $FN$ are the true positive, false positive, true negative, and false negative, respectively. It is notable that in a real gene network, the rate of links (i.e., regulatory relationships) between nodes (genes) is quite low, which leads to a high true negative in network inference. Therefore, as in other relevant studies, we do not measure accuracy due to the large bias caused by the true negative.

3 Results and Discussion

To evaluate our hybrid approach for gene clustering and network reconstruction, this section describes the experiments conducted and presents the results. Several datasets were collected and pre-examined. Two datasets (called datasets A and B hereafter) were chosen, because they are real datasets suitable for performing both PCA and GO
semantic mapping procedures, and the constructed networks can be verified (most of the network links are known). Dataset A was collected from [21] and it included 800 genes. Dataset B was collected from [20]; it described the yeast *S. cerevisiae* regulatory network and contained more than 6000 genes. Considering the constraints for dealing with large datasets by the available packages, we sampled randomly two subsets of 400 genes (i.e., B₁ and B₂) for experiments. It is notable that the original studies that contained these datasets were not for the purposes of gene clustering and network reconstruction, the results obtained from our experiments cannot be compared to those studies. We thus focus on developing a hybrid approach for constructing large gene networks and verifying the proposed approach from different perspectives.

### 3.1 Results of Gene Clustering

The first set of experiments investigated the effect of using different distance measurement strategies on gene clustering. Both of the datasets described above were used for evaluation. The PCA procedure was first performed on the time series expression data for feature extraction, and a set of features was selected that contained at least 70% of the original data’s information content. As mentioned in Section 2.1, two types of similarity measures were applied to the data using the data-based method (i.e., PCA) and the knowledge-based method. Next, the two types of similarities were combined and used for distance calculation in the clustering process (i.e., Equation 5).

In this set of experiments, the fuzzy \( c \)-means clustering method was used. To investigate the effects of expression data and semantic similarity on gene clustering, different combinations of \( w_1 \) and \( w_2 \) defined in Equation 5 were arranged. As described, the PPI measurement was taken to evaluate the results. For each weight combination, 100 experimental trials were conducted for gene clustering, and the PPI values of the trials were averaged. The results are presented in Figs. 3 (for dataset A) and 4 (for datasets B₁ and B₂).

In these figures, the \( x \)-axis indicates the weights of the two components (i.e., expression data-based distance and semantic similarity between genes). The indices on the left side of the \( x \)-axis (from the leftmost to the middle) represent the weights of the GO-based distance (i.e., values for \( w_2 \)), and they range from 0 to 1 (stepped by 0.1). In this interval, the weight of the expression data-based distance (i.e., \( w_1 \)) remains 1. Meanwhile, the indices on the right side of the \( x \)-axis are the weights of the expression data-based distance (i.e., values for \( w_1 \)); the weights vary from 1 (the middle) to 0 (the rightmost), and the weight of GO-based distance (i.e., \( w_2 \)) is 1 in this interval. The two extreme combinations of weights (the leftmost and the rightmost) thus mean the cases that consider only expression data or GO similarity. In Figs. 3 and 4, the \( y \)-axis indicates the PPI values for different combinations of the above two components. Furthermore, the curves with different colors represent the results produced in the experiments by different numbers of clusters (indicated by the integers on the right-hand side of the figure).

The results presented here demonstrate the effectiveness of using domain knowledge (i.e., semantic similarity) in gene clustering. They also suggest that certain combinations with large weights on data-based distance and relatively small weights on knowledge-based distance can have the best performance. It is notable that the PPI values vary in different datasets (depending on the gene-gene interaction in the
network and the number of clusters), and they are thus not directly comparable. In general, the PPI values obtained by our approach have been relatively high, in contrast to those reported in relevant studies [20]. Because our study mainly focuses on investigating the effectiveness of incorporating expression data and domain knowledge for gene clustering first, and then for network reconstruction (rather than on developing a novel clustering algorithm), we thus did not systematically perform further performance comparisons.

3.2 Evaluation of Network Reconstruction

To investigate the effect of gene clustering on network reconstruction, in the second series of experiments, we reconstructed the gene networks based on the clusters obtained from previous experiments. As mentioned in Section 2.3, we adopted the online software BoolNet to model the genes into Boolean networks. In the experiments with BoolNet, we chose the “synchronous Boolean network” as the target model and the algorithm “Best-Fit Extension” for constructing and assembling networks. This algorithm retrieves a set of functions with minimal error on the input and is thus suitable for processing noisy data. This software also includes a binarization procedure with scan statistics to transfer the real-value time series data into the binary form before network construction.
We first conducted a set of experimental trials for each dataset to verify whether the clustering results with high PPI values are beneficial to network reconstruction. In the experiments, the cluster sets with the five highest PPI values and those with the five lowest PPI values were used for network reconstruction. The results for the three datasets (A, B₁, B₂) are presented in Fig. 5, in which the precision (p), recall (r), and the F-measure (f) are shown for each set of clusters. In the figures, the values were averaged from five trials (“best-5” represents results averaged from five trials with the highest PPI values, and “worst-5”, for the lowest PPI values). The PPIs corresponding to each set of clusters are also shown. As can be seen, for each dataset, in network reconstruction the cluster sets with high PPIs have consistently better results (i.e., relatively higher precision, recall, and F-measure) than those with low PPIs. From these figures, we can also observe that the precision of the trials with higher PPIs is significantly better than that of the trials with lower PPIs. This is because that the false positive rate of the former is much lower than that of the latter.

It is worth noting that the false positive rate in Boolean network reconstruction is often high, particularly in the reconstruction of large gene networks. This is mainly due to the disagreement between the experimental front and the computational front for link detection. On the experimental front, the PPI link detection techniques, such as yeast two-hybrid screens and large-scale affinity purification with mass
spectrometry, attempt to discover direct physical interactions between proteins. However, on the computational front, the approaches consider protein-protein interactions in the most general context and often refer to “functionally interacting proteins”. This implies that the proteins cooperate to perform a given task without necessarily involving any physical contact. The different considerations make the overlap between sets of interacting proteins identified by functional and physical relationships limited (particularly when the binarization procedure is used to address the real-value data) [22]. Under such circumstances, precision and recall often have low values.

After showing that high PPI clusters can lead to better performance in network reconstruction, we conducted a second set of experimental runs to further examine whether random clustering could produce the same results. To confirm the effect of high PPI clusters, we randomly partitioned the genes in each dataset into the same numbers of clusters with the highest PPI values, and then performed the same procedure to construct networks. The results of network reconstruction are also shown in Fig. 5. As can be observed, in contrast to random clustering, the proposed gene clustering method could obtain better PPI values for all datasets and consequently resulted in better network reconstruction performance. It indicates that the results obtained from the proposed gene clustering method are better than those without clustering. These results demonstrate the effectiveness of the proposed approach.

3.3 Semantic Verification
In addition to the quantitative experiment results presented above, we take one of the sub-clusters for dataset B as an example to further examine the correctness of the results. The selected sub-cluster (named yeast-14 herein) consists of 14 genes (they are RDH54, DUN1, MSH6, RAD51, RAD54, RAD27, RAD5, RHC18, UNG1, OGG1, PMS1, MSH2, DHS1, and RAD53) and the inferred topology is depicted in Fig. 6. The hub of this network is RAD5, which has the most linkages compared to its related genes, followed by RAD27, DHS1, and RAD54. Meanwhile, these genes control over half of the interactions.

To investigate the mechanism of yeast-14, we accessed the Saccharomyces Genome database (SGD, http://www.yeastgenome.org/) for evaluating the biological meaning (i.e., the correctness of the clustering and reconstruction) of yeast-14. Table 1 shows the validation result that includes the recognized 30 GO terms among genes. False discovery rate (FDR) for each GO term with corresponding classified genes are also provided. The FDR was calculated by running 50 simulations with random genes, and counting the average number of times a p-value as good as or better than a p-value generated from the real data is seen. The closer the p-value (or the corresponding FDR) is to zero, the more significant the particular GO term associated with the group of genes is. As seen in this table, all the FDR values are less than 10%. This illustrates that a promising cluster group was built.

All 14 genes, in detail, belong to DNA metabolic process. On the one hand, for a widely definition in this cluster, 13 and 12 (see the third column in Table 1) out of the 14 genes were identified as cellular response to DNA damage stimulus and DNA repair, respectively. These genes are in response to a change in state of a cell due to a stimulus indicating damage to its DNA. On the other hand, for knowing a specific definition of genes, we can capture gene interactions more adequately when
reconstructing a network. For instance, from the SGD results we find that MSH2 and MSH6 are related to the GO term *interstrand cross-link repair*, which explains the relationship between the two nodes in Fig. 6. In detail, MSH2 forms heterodimers with MSH6 to active mispair recognition complex.

The analyses and interpretations on this sub-network show that *yeast-14* can be regarded as functioning the DNA damage stimulus and DNA repair processes. By looking into the genetic interactions (Fig. 6) and each gene’s corresponding GO terms (not listed here due to the space limitation), we not only can observe the genetic relationships in a sub-network, but also can refer interactions to several genetic functions as well. The approach of integrating data-driven measurement (i.e. time series data) and knowledge-driven measurement (i.e. gene ontology) to infer biological networks is thus confirmed.

### Table 1. The yeast-14 network was evaluated by the *Saccharomyces* Genome database.

<table>
<thead>
<tr>
<th>GO ID</th>
<th>GO term</th>
<th>(# of genes, %)</th>
<th>FDR(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0006259</td>
<td>DNA metabolic process</td>
<td>(14, 100%)</td>
<td>6.40%</td>
</tr>
<tr>
<td>0033554</td>
<td>cellular response to stress</td>
<td>(13, 92.9%)</td>
<td>8.30%</td>
</tr>
<tr>
<td>0006974</td>
<td>cellular response to DNA damage stimulus</td>
<td>(13, 92.9%)</td>
<td>4.20%</td>
</tr>
<tr>
<td>0006950</td>
<td>response to stress</td>
<td>(13, 92.9%)</td>
<td>9.10%</td>
</tr>
<tr>
<td>0006281</td>
<td>DNA repair</td>
<td>(12, 85.7%)</td>
<td>3.40%</td>
</tr>
<tr>
<td>0022402</td>
<td>cell cycle process</td>
<td>(9, 64.3%)</td>
<td>8.20%</td>
</tr>
<tr>
<td>0007049</td>
<td>cell cycle</td>
<td>(9, 64.3%)</td>
<td>8.80%</td>
</tr>
<tr>
<td>0006310</td>
<td>DNA recombination</td>
<td>(7, 50.0%)</td>
<td>2.50%</td>
</tr>
<tr>
<td>0000003</td>
<td>reproduction</td>
<td>(7, 50.0%)</td>
<td>6.50%</td>
</tr>
<tr>
<td>1903046</td>
<td>meiotic cell cycle process</td>
<td>(6, 42.9%)</td>
<td>3.30%</td>
</tr>
<tr>
<td>0051321</td>
<td>meiotic cell cycle</td>
<td>(6, 42.9%)</td>
<td>3.90%</td>
</tr>
<tr>
<td>0006312</td>
<td>mitotic recombination</td>
<td>(5, 35.7%)</td>
<td>0.80%</td>
</tr>
<tr>
<td>0007127</td>
<td>meiosis I</td>
<td>(5, 35.7%)</td>
<td>1.40%</td>
</tr>
<tr>
<td>0006302</td>
<td>double-strand break repair</td>
<td>(5, 35.7%)</td>
<td>1.50%</td>
</tr>
<tr>
<td>0007126</td>
<td>meiotic nuclear division</td>
<td>(5, 35.7%)</td>
<td>2.40%</td>
</tr>
<tr>
<td>0006298</td>
<td>mismatch repair</td>
<td>(4, 28.6%)</td>
<td>0.30%</td>
</tr>
<tr>
<td>0060249</td>
<td>anatomical structure homeostasis</td>
<td>(4, 28.6%)</td>
<td>1.10%</td>
</tr>
<tr>
<td>0000723</td>
<td>telomere maintenance</td>
<td>(4, 28.6%)</td>
<td>1.10%</td>
</tr>
<tr>
<td>0032200</td>
<td>telomere organization</td>
<td>(4, 28.6%)</td>
<td>1.10%</td>
</tr>
<tr>
<td>0030491</td>
<td>heteroduplex formation</td>
<td>(3, 21.4%)</td>
<td>0.10%</td>
</tr>
<tr>
<td>0000710</td>
<td>meiotic mismatch repair</td>
<td>(3, 21.4%)</td>
<td>0.10%</td>
</tr>
<tr>
<td>0007534</td>
<td>gene conversion at mating-type locus</td>
<td>(3, 21.4%)</td>
<td>0.20%</td>
</tr>
<tr>
<td>0035822</td>
<td>gene conversion</td>
<td>(3, 21.4%)</td>
<td>0.20%</td>
</tr>
<tr>
<td>0071897</td>
<td>DNA biosynthetic process</td>
<td>(3, 21.4%)</td>
<td>0.40%</td>
</tr>
<tr>
<td>0007533</td>
<td>mating type switching</td>
<td>(3, 21.4%)</td>
<td>0.40%</td>
</tr>
<tr>
<td>0045165</td>
<td>cell fate commitment</td>
<td>(3, 21.4%)</td>
<td>0.50%</td>
</tr>
<tr>
<td>0007530</td>
<td>sex determination</td>
<td>(3, 21.4%)</td>
<td>0.50%</td>
</tr>
<tr>
<td>0007531</td>
<td>mating type determination</td>
<td>(3, 21.4%)</td>
<td>0.50%</td>
</tr>
<tr>
<td>0036297</td>
<td>interstrand cross-link repair</td>
<td>(2, 14.3%)</td>
<td>0.10%</td>
</tr>
</tbody>
</table>
4. Conclusion

To tackle the problem of scalability in gene network inference, we have developed an efficient gene clustering method to perform network decomposition that is beneficial to network reconstruction. Because simply using gene expression or gene ontology alone is not enough to capture the interaction relationships among genes, we thus proposed a hybrid approach to calculate gene distance. Our method incorporates both data and knowledge-based techniques to extract specific features from the original gene profiles. PCA method is employed to derive data features that are used to calculate data similarity, and knowledge ontology is used to measure the semantic similarity between genes. The above two types of similarities are then used to calculate the gene distance in a fuzzy gene clustering procedure.

A series of experiments have been conducted to demonstrate how the presented approach can be used to infer large networks. Because biological knowledge about the general properties of genetic networks can alleviate some of the data requirements, the first set of results shows that the proposed approach with a hybrid similarity measure can bring about better performance (i.e., clusters of high PPIs). These results also suggest that certain combinations with large weights on data-based distance and relatively small weights on knowledge-based distance can have the best performance. Moreover, the second set of experiments demonstrates that the cluster sets with high PPIs have consistently better results than those with low PPIs in network reconstruction. Finally, a small network obtained from our two-stage procedure has been analyzed and interpreted for further verification. These experimental evaluations verify the effectiveness of our approach.
References