ISSN 1672-9145 **Acta Biochimica et Biophysica Sinica** 2006, **38**(6): 379–384 **CN** 31-1940/Q

Comparisons of Graph-structure Clustering Methods for Gene Expression Data

Zhuo FANG[1], Lei LIU[2,4], Jiong YANG[3], Qing-Ming LUO[1]*, and Yi-Xue LI[2]*

[1] *Hubei Bioinformatics and Molecular Imaging Key Laboratory, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China;*
[2] *Shanghai Center for Bioinformatics Technology, Shanghai 200235, China;*
[3] *Department of EECS, Case Western Reserve University, Cleveland 44106, USA;*
[4] *W. M. Keck Center for Comparative and Functional Genomics, University of Illinois at Urbana-Champaign, Urbana 61801, USA*

Abstract Although many numerical clustering algorithms have been applied to gene expression data analysis, the essential step is still biological interpretation by manual inspection. The correlation between genetic co-regulation and affiliation to a common biological process is what biologists expect. Here, we introduce some clustering algorithms that are based on graph structure constituted by biological knowledge. After applying a widely used dataset, we compared the result clusters of two of these algorithms in terms of the homogeneity of clusters and coherence of annotation and matching ratio. The results show that the clusters of knowledge-guided analysis are the kernel parts of the clusters of Gene Ontology (GO)-Cluster software, which contains the genes that are most expression correlative and most consistent with biological functions. Moreover, knowledge-guided analysis seems much more applicable than GO-Cluster in a larger dataset.

Key words clustering; expression pattern; biological function

Expression profiling and large-scale proteomics have revolutionized biology by generating a vast amount of data. Many mathematical clustering algorithms have been adapted or directly applied to gene expression data analysis [1–12]. However, all of these algorithms only pay attention to the mathematical similarity of genes and conditions. The essential step in the analysis of those experiments is biological interpretation by manual inspection [13].

The main challenge to the biologist is contained in the next step of the analysis, which consists of identifying the biologically relevant expression changes. The universal grounds behind expression clustering are that genes with similar expression patterns are possibly involved in similar biological process. However, it has been observed that genes with similar expression profiles sometimes do not share similar biological functions [14,15], and genes involved in the same biological process are not always perfectly correlated [16]. Thus, incorporating prior knowledge in the clustering process became necessary as it helps to generate more refined and biologically relevant clusters.

Generally, the prior biological knowledge is the evidence, which provides connections among differently expressed genes. This evidence includes the function correlation of genes, such as a shared annotation, joint participation in some physiological process and physical interaction at the protein level [17]. All the evidence can be represented as a graph, for example, a Gene Ontology (GO) network graph, a metabolic and signaling pathway graph or a protein interaction map. After mapping relevant genes on the graph, the clustering procedure can be processed based on the graph structure. The results of these clustering methods can identify most biological processes and regulatory mechanisms [17,18] and show more advantages than conventional clustering methods [19].

In this paper, we introduce some current clustering

Received: February 7, 2006 Accepted: March 27, 2006

This work was supported by the grants from the National Basic Research Program of China (No. 2004CB518606), the Fundamental Research Program of Shanghai Municipal Commission of Science and Technology (No. 04DZ14003) and the National Key Technologies R&D Program of China (No. 2005BA711A04)

*Corresponding authors:

Qing-Ming LUO: Tel, 86-27-87792033; Fax, 86-27-87792034; E-mail, qluo@mail.hust.edu.cn

Yi-Xue LI: Tel, 86-21-61313680; Fax, 86-21-61313670, E-mail, yxli@sibs.ac.cn

DOI: 10.1111/j.1745-7270.2006.00175.x

algorithms, based on graph structure constituted by biological knowledge, then compare the results in terms of homogeneity of clusters, coherence of annotation and matching ratios.

Materials and Methods

Biological annotation evidence

The biological knowledge can be obtained from scientific literature or public databases, such as GO [20], metabolic networks [21] and Medline [22].

GO is a cross-species, controlled vocabulary describing three domains of molecular biology [20]: molecular function, cellular component and biological process. It is currently the most popular source for biological terminology and used often in interpretion or validation of microarray results [15]. GO has a hierarchical classification scheme structured as a direct acyclic graph, with each node designating a biological term and each edge representing the relationship of "is a" or "part of", meaning that a child term is either a part of the parent or a more specific example of the parent term. To facilitate calculation, the original digraph of GO will be transformed into an ordered tree, that is to say, a directed tree with an order defined for the children of every node of the tree. Because the same GO term might occur in different levels of the ontology, each appearance of a GO term is considered distinct when an ordered tree is built. This may be justified from a biological viewpoint that in the gene ontology, what counts is not a GO term itself but which path the GO term takes from the root. Each appearance of a GO term is considered distinct if a distinct path leads to it from the root.

The complete metabolic network in cells can be represented as a bipartite undirected graph [23], called a metabolic graph. In the metabolic graph, metabolites as well as enzymes are represented as nodes, and interactions between them are represented as edges. Thus, a metabolic node is connected to all of the enzyme nodes that catalyze reactions involving the particular metabolite, and an enzyme node is connected to all of the metabolites that take part in the corresponding reaction.

Graph-structure clustering methods

Here we introduce three graph-structure clustering methods, which depend on GO or the metabolic network.

GO-Cluster is an executable program for Microsoft Windows 98/2000/NT/XP [13]. This software uses the tree structure of the GO database as a framework for numerical clustering, thus allowing a simple visualization of gene expression data at various levels of the ontology tree [13]. Compared with other known visualization tools, such as MAPPFinder [24] or GoMiner [25], GO-Cluster does not judge statistically the regulation of a GO term, but carries out hierarchical average-distance clustering by applying Pearson's correlation coefficient to the genes that are allocated to the corresponding term. The advantage of this clustering is that no "rules" have to be predefined and all of the available datasets are informative. In the GO-Cluster, every node of the GO can be selected for cluster analysis, the corresponding tree can be calculated in real time and simultaneously displayed as a left-to-right tree structure.

In the knowledge-guided analysis of microarray data [19], GO information is introduced to guide the clustering process. Both expression pattern and biological function similarities are considered. The steps of the algorithm are as follows [19]. Subsequently, in the construction of the GO tree, genes in the expression dataset are mapped to this tree through a species-related database, and unmapped nodes (terms) in the GO tree are excluded. Thereafter, every node in this GO tree is checked from top to bottom. Genes mapped to a node as well as its descendant nodes form an initial cluster, and the expression similarity of this cluster is calculated. If high expression similarity is obtained, the cluster is output. The node and its descendant nodes are excluded from the GO tree. Otherwise, no action is taken. Once the whole GO tree is checked, the output clusters are refined by average trend constraint filtering to obtain clusters with both high expression similarities and high function similarities (**Fig. 1**).

The recently published work by Breitling *et al*. [17] provided a statistical method, known as graph-based iterative group analysis (GIGA), to identify active subgraphs in the biological knowledge graph, which might contain high connection genes with expression correlation. Genes sharing an annotation are connected to build the graph. In addition to the graph, a complete list of genes sorted by differential expression is provided and each node is ranked according to the allocated gene. Then local minima are identified in the graph, in which the nodes have a lower rank than all their direct neighbors. Next, subgraphs are iteratively extended from each of those local minima by including the neighboring node with the next highest rank (m) and, if present, all adjacent nodes of ranks equal or smaller than m. For each extension, a *P*-value will be calculated. The extension process continues until all nodes reachable from the local minimum are included or the sub-

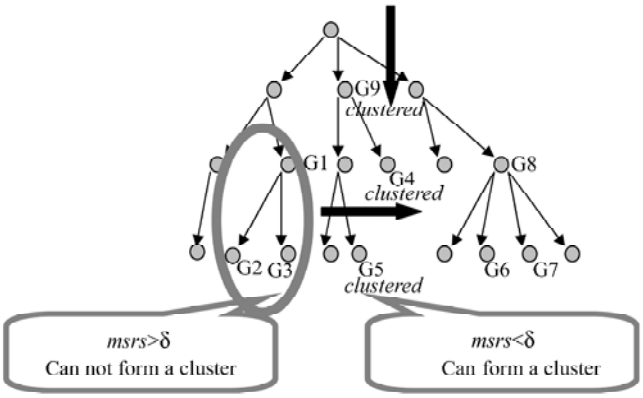

Fig. 1 Flow chart of knowledge guided analysis

The clustering starts at the first level of the GO tree. For each level, the algorithm first picks up the leftmost node of this level (G_9 node in level 2, since only nodes mapped with genes will be considered in our process). All the descendant nodes of this target gene will be found (G_4 and G_5). Thereafter, we calculate the expression similarity (*msrs*) of the initial cluster, which is formed by the genes mapped to the target node and its descendant nodes (G_9, G_4 and G_5). If the *msrs* value is below a predefined threshold δ, this cluster will be output. Simultaneously, the genes in the cluster will be marked with 'clustered', which means these genes will be excluded from next analysis. Otherwise, no action will be taken and the algorithm goes to the next node of current GO tree level. In this example, the node corresponding to G_9 is the only marked node of level 2. Thus we go to the nodes in level 3. This process repeats for all nodes in the level and then goes to the next level. The iteration terminates when every level of the GO tree has been visited. The arrows represent the move direction of clustering, that are, from up to down as well as from left to right.

graph reaches an arbitrary maximum size. After sorting by increasing *P*-value, the subgraphs at the top positions will be considered as ideal relevant regions of the biological graph.

Additionally, Ideker *et al.* [18] introduced an approach based on simulated annealing to searching active subnetworks in a molecular interaction network. Segal *et al.* [26] also described an approach for identifying "pathways" from gene expression and protein interaction data by the Expectation Maximization algorithm. As with GIGA, the guideline for finding gene clusters is the degree of expression activity, but not expression similarity. Therefore, we only compared the clustering results of GO-Cluster and knowledge-guided analysis, outlined below.

Data preparation

The well-known dataset by Eisen *et al.* [1] was applied to these algorithms. Gene expression in the budding yeast *Saccharomyces cerevisiae* was studied during the diauxic shift, the mitotic cell division cycle, sporulation, and temperature and reducing shocks using microarrays. Each cell in the expression matrix represented the measured Cy5/Cy3 fluorescence ratio at the corresponding target element. All ratio values were log transformed (base 2 for simplicity). Using the hierarchical clustering methods, Eisen *et al.* successfully clustered the gene expression profiles [1]. We investigated the genes reported previously [1], which can be accessed from http://genome-www.stanford.edu/clustering/. There are 2467 genes and 79 conditions, which relate with certain experiments. As the maximal number of experiments which can be dealt with by GO-Cluster is 16, we divided the whole expression matrix into several parts according to different types of conditions. The expression data of Elutriation (14 conditions), Sporulation (11 conditions) and Diauxic Shift (7 conditions) were studied. The *Saccharomyces* Genome Database downloaded from http://www.geneontology.org/ was used to extract GO terms and generate annotation reference space [27].

Cluster validation

To validate and compare the clustering results, some criteria were used for various aspects.

The homogeneity of a cluster is used to measure the similarity degree of the gene expressions within one cluster [28], which is a basal guideline for assessing the quality and reliability of the cluster. For example, if the gene set of a cluster is *C*, *G* represents the gene in *C*. The definition of the homogeneity of the cluster is defined in **Equation 1** as:

$$H(C)=\frac{\sum_{G_i,G_j\in C,G_i\neq G_j} dist(G_i,G_j)}{\|C\|\cdot(\|C\|-1)} \quad 1$$

where $dist(G_i,G_j)$ means the Euclidean distance between two expression vectors in *C* and $\|C\|$ means the norm of expression matrix corresponding to *C*. Here the Frobenius norm is used. This definition represents the homogeneity of cluster *C* by the average pairwise object similarity within *C*, and a small *H* value represents a good cluster.

To assess the reliability of the clusters, a function *WR* is defined to evaluate the coherence of annotation [19]. For a cluster *C* whose annotation space is *A*, we suppose that a is the most frequently occurring annotation in *A* (**Equation 2**):

$$WR=1-\frac{cor\_num}{whl\_num} \quad 2$$

where *cor_num* is the number of genes annotated by *a* and *whl_num* is the total number of genes in *C*. Obviously, *WR* can measure the inconsistency of annotations in a cluster. Smaller *WR* value implies stronger coherence.

We defined matching ratio to examine the consistency

of the clustering results of different methods. For a certain GO term, C_1 and C_2 are the clusters annotated by this term of two different algorithms. Suppose C_1*_num* is the gene number of C_1 and C_2*_num* is the gene number of C_2, and *ins_num* is the gene number of the intersection of C_1 and C_2. The matching ratio from C_1 to C_2 is shown in **Equation 3**:

$$MR(C_1, C_2) = \frac{ins\_num}{C_1\_num} \quad 3$$

Analogously, the matching ratio from C_2 to C_1 is as follows (**Equation 4**):

$$MR(C_2, C_1) = \frac{ins\_num}{C_2\_num} \quad 4$$

Results

We compared the clustering results from GO-Cluster and knowledge-guided analysis from the three aspects outlined in previous sections, using the expression dataset of Elutriation, Sporulation and Diauxic Shift. For each level from 4 to 8 of the GO tree, we selected three clusters randomly as the indication of this level and calculated the average values of the three criteria. The *Saccharomyces* Genome Database was used as the annotation space for *WR* calculation.

Elutriation

There are 14 conditions for Elutriation. Using knowledge-guided analysis, we found 94 significative clusters when the threshold was set as 0.15. Three clusters were selected for each level and corresponding clusters with the same GO terms were found from the results of GO-Cluster. After that, all the parameters were calculated for the clusters of the two methods. For each level, we took the average of the three clusters. The results are shown in **Fig. 2**. From **Fig. 2(A,B)**, we can see that the result clusters of knowledge-guided analysis have advantages both in the similarity of expression pattern and convergence of annotation reference. From **Fig. 2(C)**, we can see that the value of *MR*(knowledge, GO-Cluster) is close to 1 whereas *MR*(GO-Cluster,knowledge) is obviously low, which means that the clusters from knowledge-guided analysis are almost subsets of clusters from GO-Cluster.

Sporulation

There are 11 conditions for Sporulation. When the threshold was set as 0.15, 73 clusters were found. As with Elutriation, three clusters were selected from the results of both the knowledge-guided analysis and GO-Cluster for each GO tree level. All the parameters for validation were calculated and averaged for every level. **Fig. 3** shows the results. From **Fig. 3(A,B)**, we can see that the result clusters of knowledge-guided analysis have advantages both in the similarity of expression pattern and convergence of annotation reference. From **Fig. 3(C)**, we can see that *MR*(knowledge,GO-Cluster) is close to 1 whereas *MR*(GO-Cluster,knowledge) is obviously low. It means that the clusters from knowledge-guided analysis are almost subsets of clusters from GO-Cluster.

Diauxic Shift

There are seven conditions for Diauxic Shift. We obtained 82 significative clusters using knowledge-guided analysis, when the threshold was set as 0.15. Three clusters were selected for each level from the results of the two methods and all the parameters were calculated for each cluster. The comparisons of each level, after averaging, are shown in **Fig. 4**. As with Elutriation and Sporulation, the result clusters of knowledge-guided analy-

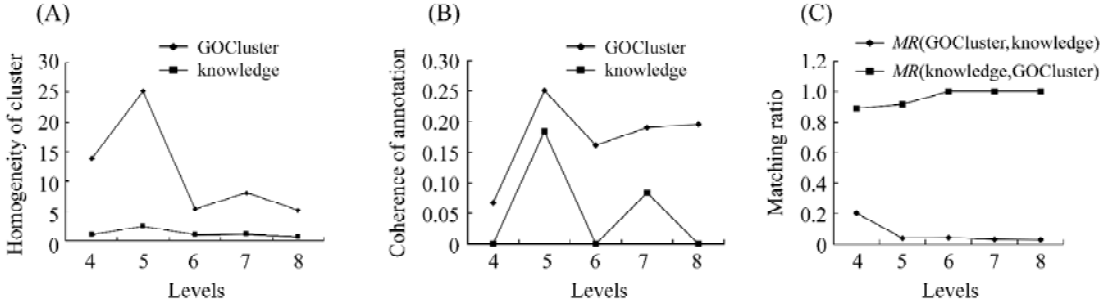

Fig. 2 Comparisons for Elutriation

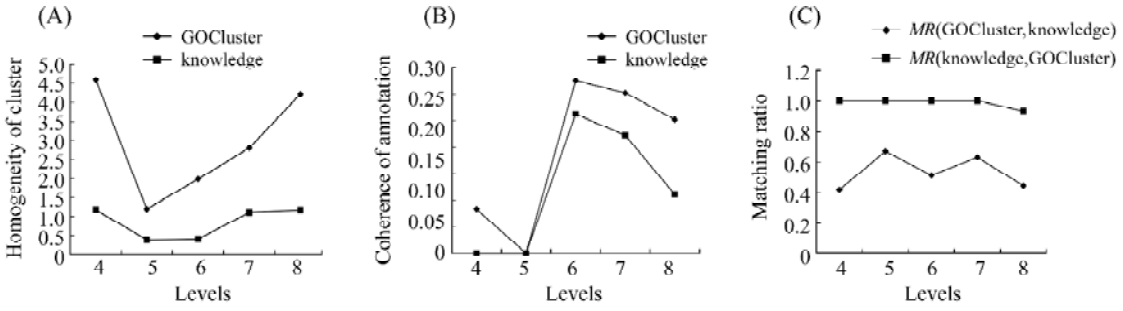

Fig. 3 Comparisons for Sporulation

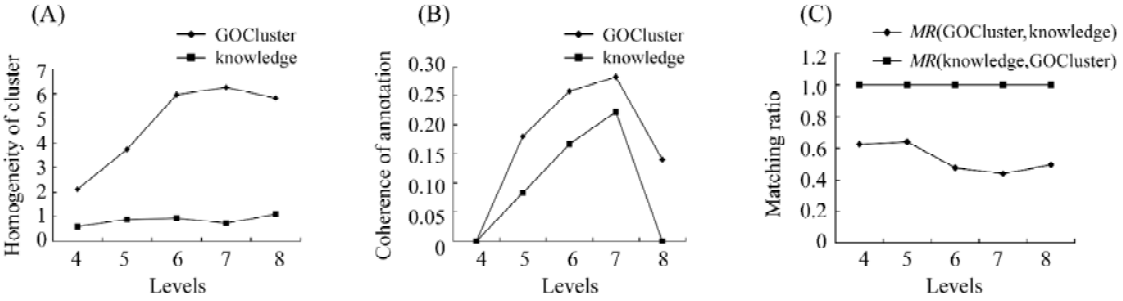

Fig. 4 Comparisons for Diauxic Shift

sis have advantages both in the similarity of expression pattern and convergence of annotation reference. From **Fig. 4(C)**, we can also draw a conclusion that the clusters from knowledge-guided analysis are almost subsets of clusters from GO-Cluster.

Discussion

So far we have discussed some graph-structure clustering methods for microarray data analysis. These algorithms can find gene clusters with expression pattern and biological function similarities through the modified structure of biological annotation evidence. The results showed not only that the genes were clustered according to their expression profiles, but also the clusters were annotated automatically. This makes it more convenient to combine the external profiles with essential functions.

These methods have their innovations as well as limitations. GO-Cluster is convenient for rapid visualization and can evaluate the gene expression profiles in every node of a GO tree. However, it applies hierarchical average distance clustering to the genes that are allocated to only a certain GO term, which will lead to the incompleteness of clustering results because the genes corresponding to the children term of a GO term also belong to the category of this GO term. Furthermore, the Pearson's correlation coefficient used in GO-Cluster has been proven not to be robust with respect to outliers [3]. The knowledge-guided analysis is based on a similarity measure that depends on both expression profiles and biological functions, which are equally essential for gene clusters. Nevertheless, the analysis depends strongly on the accurateness and completeness of the GO hierarchy, for example, it can not deal with the genes corresponding to a very high level of a GO tree because the expression similarity of these genes can not satisfy the threshold. The simulated annealing approach combines a rigorous statistical measure for scoring with a search algorithm for identifying subnetworks, but it requires relatively complex parameter estimation and can not guarantee to find the optimally scoring subnetworks. The GIGA algorithm brings a simple and fast method to screening significant subgraphs, but it can only address one condition at a time, making it inadequate for expression data analysis.

We have also compared the results of two of the above methods from three different aspects: expression homogeneity, annotation coherence and matching ratio. For the expression homogeneity and annotation coherence, the results of knowledge-guided analysis were much bet-

ter than the results of GO-Cluster. It was shown in **Figs. 2–4** that the differences of the two criteria between the results of knowledge-guided analysis and GO-Cluster were much smaller in Sporulation and Diauxic Shift than in Elutriation. This might be attributed to the data magnitude. In other words, knowledge-guided analysis is much more applicable than GO-Cluster in a larger dataset.

The comparison of matching ratios shows that the clusters of knowledge-guided analysis are almost subsets of the clusters of GO-Cluster. Hence, we can see that the results of knowledge-guided analysis are the kernel parts of the results of GO-Cluster, containing the genes that are most expression correlative and most consistent to a certain biological function. These results are possibly more acceptable to biologists because they involve more precise and detailed information.

References

1 Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 1998, 95: 14863–14868

2 Herwig R, Poustka AJ, Muller C, Bull C, Lehrach H, O'Brien J. Large-scale clustering of cDNA-fingerprinting data. Genome Res 1999, 9: 1093–1105

3 Heyer LJ, Kruglyak S, Yooseph S. Exploring expression data: Identification and analysis of coexpression genes. Genome Res 1999, 9: 1106–1115

4 De Smet F, Mathys J, Marchal K, Thijs G, De Moor B, Moreau Y. Adaptive quality-based clustering of gene expression profiles. Bioinformatics 2002, 18: 735–746

5 Herrero J, Valencia A, Dopazo J. A hierarchical unsupervised growing neural network for clustering gene expression patterns. Bioinformatics 2001, 17: 126–136

6 Shamir R, Sharan R. CLICK: A clustering algorithm for gene expression analysis. Proc Int Conf Intell Syst Mol Biol 2000, 8: 307–316

7 Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL. Model-based clustering and data transformations for gene expression data. Bioinformatics 2001, 17: 977–987

8 Jiang D, Pei J, Zhang A. DHC: A density-based hierarchical clustering method for time series gene expression data. In: Proceedings of BIBE2003: 3rd IEEE International Symposium on Bioinformatics and Bioengineering; August, 2003, Washington USA

9 Getz G, Levine E, Domany E. Coupled two-way clustering analysis of gene microarray data. Proc Natl Acad Sci USA 2000, 97: 12079–12084

10 Lazzeroni L, Owen A. Plaid models for gene expression data. Statistica Sinica 2002, 12: 61–86

11 Cheng Y, Church GM. Biclustering of expression data. Proc Int Conf Intell Syst Mol Biol 2000, 8: 93–103

12 Yang J, Wang W, Wang H, Yu P. delta-Clusters: Capturing subspace correlation in a large dataset. In: Proceedings of 18th International Conference on Data Engineering (ICDE) 2002, 517–528

13 Adryan B, Schuh R. Gene-Ontology-based clustering of gene expression data. Bioinformatics 2004, 20: 2851–2852

14 Clare A, King RD. How well do we understand the clusters found in microarray data? In Silico Biol 2002, 2: 511–522

15 Gibbons FD, Roth FP. Judging the quality of gene expression-based clustering using gene annotation. Genome Res 2002, 12: 1574–1581

16 Cheng J, Cline M, Martin J, Finkelstein D, Awad T, Kulp D, Siani-Rose MA. A knowledge-based clustering algorithm driven by Gene Ontology. J Biopharm Stat 2004, 14: 687–700

17 Breitling R, Amtmann A, Herzyk P. Graph-based iterative group analysis enhances microarray interpretation. BMC Bioinformatics 2004, 5: 100

18 Ideker T, Ozier O, Schwikowski B, Siegel AF. Discovering regulatory and signaling circuits in molecular interaction networks. Bioinformatics 2002, 18: S233–S240

19 Fang Z, Yang J, Li Y, Luo Q, Liu L. Knowledge guided analysis of microarray data. J Biomed Inform 2005 (accepted)

20 Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP *et al*. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000, 25: 25–29

21 Hanisch D, Zien A, Zimmer R, Lengauer T. Co-clustering of biological networks and gene expression data. Bioinformatics 2002, 18: S145–S154

22 Masys DR, Welsh JB, Fink JL, Gribskov M, Klacansky I, Corbeil J. Use of keyword hierarchies to interpret gene expression patterns. Bioinformatics 2001, 17: 319–326

23 Patil KR, Nielsen J. Uncovering transcriptional regulation of metabolism by using metabolic network topology. Proc Natl Acad Sci USA 2005, 102: 2685–2689

24 Doniger SW, Salomonis N, Dahlquist KD, Vranizan K, Lawlor S, Conklin BR. MAPPFinder: Using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. Genome Biol 2003, 4: R7

25 Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, Narasimhan S *et al*. GoMiner: A resource for biological interpretation of genomic and proteomic data. Genome Biol 2003, 4: R28

26 Segal E, Wang H, Koller D. Discovering molecular pathways from protein interaction and gene expression data. Bioinformatics 2003, 19: 264–271

27 Dwight SS, Harris MA, Dolinski K, Ball CA, Binkley G, Christie KR, Fisk DG *et al*. *Saccharomyces* Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). Nucleic Acids Res 2002, 30: 69–72

28 Jiang D, Tang C, Zhang A. Cluster analysis for gene expression data: A survey. IEEE Trans Knowl Data Eng 2004, 16: 1370–1386

Edited by
Shigeto SENO