Available online at www.sciencedirect.com
ScienceDirect
Journal of Biomedical Informatics 41 (2008) 232–241
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

BioLattice: A framework for the biological interpretation of microarray gene expression data using concept lattice analysis

Jihun Kim [a,c], Hee-Joon Chung [a], Yong Jung [a], Kack-Kyun Kim [c], Ju Han Kim [a,b,*]

[a] *Seoul National University Biomedical Informatics (SNUBI), Seoul National University College of Medicine, 28 Yongon-dong Chongno-gu, Seoul 110-799, Republic of Korea*
[b] *Human Genome Research Institute, Seoul National University College of Medicine, 28 Yongon-dong Chongno-gu, Seoul 110-799, Republic of Korea*
[c] *Department of Oral Microbiology, Seoul National University College of Dentistry, Seoul 110-799, Republic of Korea*

Received 25 December 2006
Available online 1 November 2007

Abstract

Motivation. A challenge in microarray data analysis is to interpret observed changes in terms of biological properties and relationships. One powerful approach is to make associations of gene expression clusters with biomedical ontologies and/or biological pathways. However, this approach evaluates only one cluster at a time, returning long unordered lists of annotations for clusters without considering the overall context of the experiment under investigation.

Results. BioLattice is a mathematical framework based on concept lattice analysis for the biological interpretation of gene expression data. By considering gene expression clusters as objects and associated annotations as attributes and by using set inclusion relationships BioLattice orders them to create a lattice of concepts, providing an 'executive' summary of the experimental context. External knowledge resources such as Gene Ontology trees and pathway graphs can be added incrementally. We propose two quantitative structural analysis methods, 'prominent sub-lattice' and 'core–periphery' analyses, enabling systematic comparison of experimental concepts and contexts. BioLattice is implemented as a web-based utility using Scalable Vector Graphics for interactive visualization. We applied it to real microarray datasets with improved biological interpretations of the experimental contexts.

Keywords: DNA microarray; Gene expression; Clustering; Concept analysis; Concept lattice

1. Introduction

One of the challenges in DNA microarray data analysis is to extract biological meanings from massive amounts of gene expression data. Clustering has been one of the most successful methods for extracting coordinately regulated sets of genes [1,2]. The 'post-analytical challenge' of interpreting clusters using biological knowledge is under active investigation. Many Gene Ontology (GO)-based tools for gene expression analysis have been developed [3–9]. Several groups have proposed interpretation methods using biological pathways [10–13]. Gene Set Enrichment Analysis (GSEA) uses predefined gene sets and ranks of genes to identify significant biological changes in gene expression datasets [14,15].

Despite the undoubted importance of ontology and pathway-based annotation methods, they have limitations. The result, for example, is typically a long unordered list of annotations for tens or hundreds of clusters. The methods evaluate only one cluster at a time in a sequential manner without considering the informative association network of clusters and annotations. It is very time-consuming to read the massive annotation lists for a large number of clusters. Moreover, it is unthinkably hard to manually assemble the 'puzzle pieces' (i.e., the cluster–annotation

* Corresponding author. Address: Seoul National University Biomedical Informatics (SNUBI), Seoul National University College of Medicine, 28 Yongon-dong Chongno-gu, Seoul 110-799, Republic of Korea. Fax: +82 2 742 5947.
E-mail address: juhan@snu.ac.kr (J.H. Kim).
URL: http://www.snubi.org/software/biolattice/ (J.H. Kim).

1532-0464/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2007.10.003

sets) into an 'executive summary' (i.e., the context of the whole experiment). Ideally, the assembly should involve eliminating redundant attributes and organizing the pieces in a well-defined order for better biological understanding and insight into the underlying 'context' of the experiment under investigation.

Here, we propose BioLattice, a mathematical framework based on concept lattice analysis to organize traditional clusters and associated annotations into a lattice of concepts for better biological interpretation of microarray gene expression data. Concept lattice analysis was introduced by Rudolf Wille [16]. The theoretical foundation rests on mathematical lattice theory. It studies how objects can be grouped hierarchically according to their common attributes.

BioLattice considers gene expression clusters as objects and annotations as attributes and provide a graphical summary of the order relations by arranging them on a concept lattice in an order based on set inclusion relation. By thinking in terms of concepts and contexts rather than in terms of individual clusters and annotations, this framework sets out the scope of conceptual clustering. The rest of this paper is organized as follows. In Sections 2.1–2.3, we introduce concept lattice theory in general and describe datasets, annotation methods and techniques for the construction of biological concept lattices. In Section 2.4, we propose two structural analysis methodologies that can be applied to a complex biological lattice to extract central and peripheral concepts and major sub-contexts of differing biological significance from the lattice. Section 3 describes the analysis results and how to read and navigate a biological lattice. Structural robustness of a lattice was evaluated. Finally, conclusions and future works are detailed in the last section.

2. Methods

2.1. Concept lattice

Context is a triplet (G, M, I) consisting of two sets G and M and a relation I between G and M. The elements of G are called the objects and the elements of M are called the attributes. To show that object g has attribute m, we write gIm or $(g,m) \in I$. For a set $A \subseteq G$ of objects, we define $A' := \{m \in M | gIm \text{ for all } g \in A\}$ (i.e., the set of attributes common to the objects in A). Correspondingly, for a set $B \subseteq M$ of attributes, we define $B' := \{g \in G |\ gIm \text{ for all } m \in B\}$ (i.e., the set of objects that have all attributes in B).

The concept analysis models concepts as units of thought, consisting of two parts. A concept of the context (G, M, I) is a pair (A, B) with $A \subseteq G$, $B \subseteq M$, $A' = B$ and $B' = A$. We call A the extent and B the intent of concept (A, B). The extent consists of all objects belonging to the concept while the intent contains all attributes shared by the objects. The set of all concepts of the context (G, M, I) is denoted by $C(G, M, I)$. A concept lattice is drawn by ordering (A, B), which are defined as concepts of the context (G, M, I). The set of all concepts of a context together with the partial order $(A1, B1) \leqslant (A2, B2) :\leftrightarrow A1 \subseteq A2$ (which is equivalent to $B1 \supseteq B2$) is called a concept lattice.

We can regard A as defining gene expression clusters that share common knowledge attributes and B as defining the knowledge terms that are annotated to the clusters. The concepts are arranged in a hierarchical order so that the order of $C1 \leqslant C2 \leftrightarrow A1 \subseteq A2 \leftrightarrow B1 \supseteq B2$ is defined at $C1 = (A1, B1)$, $C2 = (A2, B2)$. Fig. 1 demonstrates a context (or a gene expression dataset) with clusters and annotations. Note that the relation matrix between objects (i.e., rows or clusters) and attributes (i.e., columns or annotations) can be represented by a directed graph (Fig. 1(b)) or a concept lattice with nonreduced (Fig. 1(c)) and reduced labeling (Fig. 1(d)). A concept lattice organizes all clusters and annotations of a relation matrix into a single unified structure with no redundancy and no loss of information. If $E1$ is a set of $\{(K2), (b,d,f,j)\}$ and $E2$ is a set of $\{(K1, K2), (b,f,j)\}$, then $E2$ subsides $E1$ because $\{K2\} \subseteq \{K1, K2\}$ and $\{b,d,f,j\} \supseteq \{b,f,j\}$ (Fig. 1(c)).

The top element of a lattice is a unit concept, representing a concept that contains all objects. The bottom element is a zero concept having no object. Specifically, the direct upper neighbors of the zero concept are called atoms and the direct lower neighbors of the unit concept are called coatoms. Fig. 1(c) and (d) are different visual representations of the same context (i.e., Fig. 1(a)). Fig. 1(d) demonstrates reduced labeling, where objects and attributes that can be omitted without losing information are omitted for easier reading. The extent of a concept is formed by collecting all objects that can be reached by descending line paths from the concept and vice versa to the intents. If a label of attribute A (object O) is attached to a certain concept, the attribute label occurs in all intent (extent) members of the concept, reachable by all descending (ascending) paths in the lattice from this concept to zero (unit) concept of the lattice.

In many applications, background knowledge may be available that can be used to model and analyze the data represented in a context [17]. Fig. 1(d)–(f) illustrates that background knowledge (or the GO trees in (e)) can be added easily to a concept lattice (d), returning an expanded concept lattice (f) (i.e., (d) + (e) = (f)).

2.2. Datasets

Four publicly available datasets were used to evaluate BioLattice. The mouse anti-GBM IgA nephropathy model (AGBM) dataset has 15 hybridizations at five time points with triplicates [18]. We used the 1112 genes showing significant temporal patterns by permutation analysis as described in the original manuscript. The human HeLa cell-division cycle (HCDC) dataset contains 26 hybridizations [19]. We used 2626 probes having pathway information. The yeast cell-division cycle (YCDC) dataset is a large collection of 59 time-course hybridizations, alpha factor, cdc15 and cdc28 [20]. We selected 2446 genes after removing all genes whose maximum minus minimum val-

a

	(a) NF-kB Pathway	(b) TGF-beta Pathway	(c) Response to wounding	(d) Defense response	(e) Response to hypoxia	(f) Anti-inflammatory response	(g) Extracellular space	(h) Intracellular	(i) Membrane	(j) Interstitial matrix
K1	O	O				O				O
K2		O		O		O				O
K3				O					O	
K4			O				O			
K5					O			O		

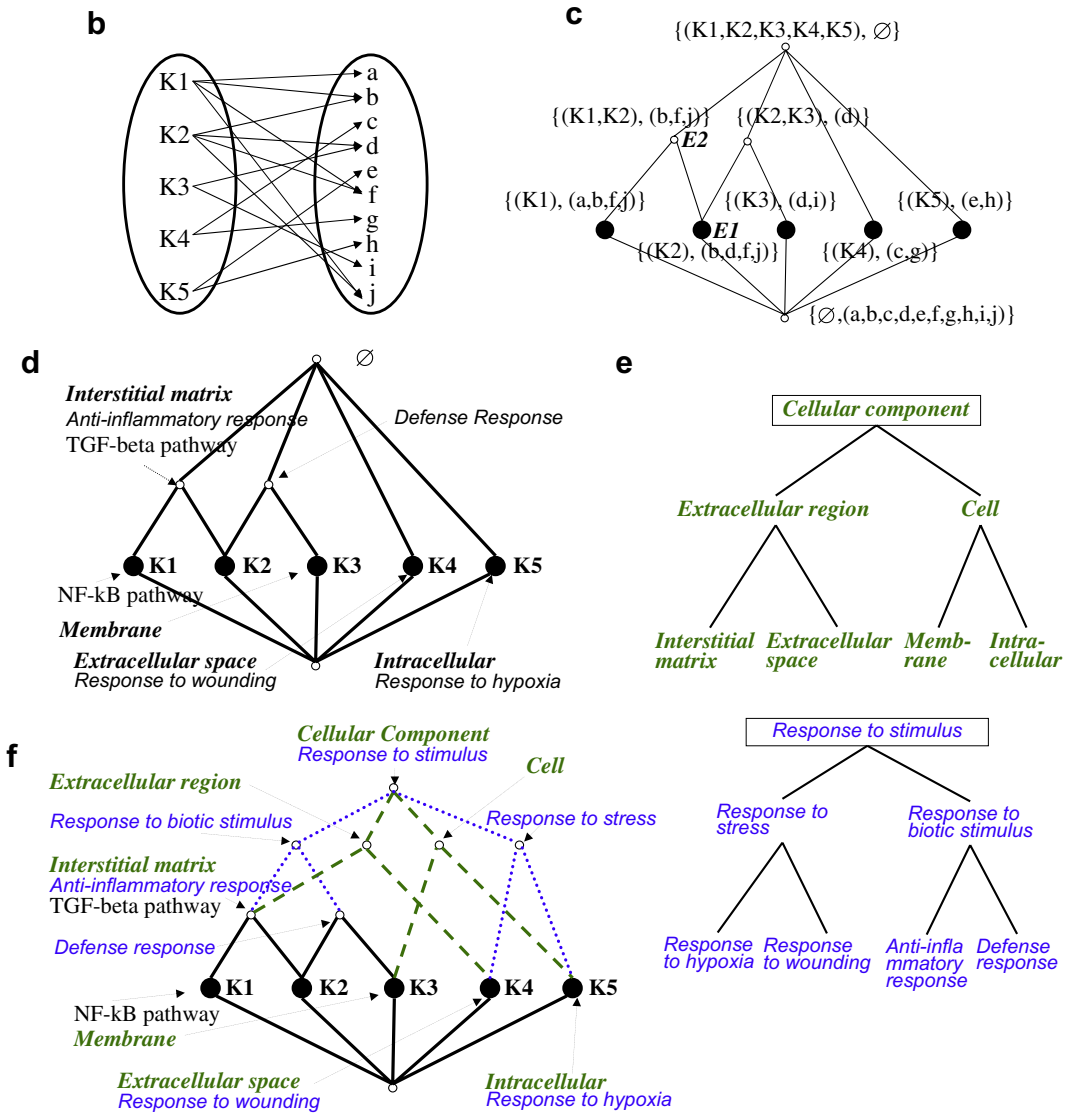

Fig. 1. Concept lattice and concept lattice expansion. The binary relation set $R = \{(K1,a), (K1,b), (K1,f), (K1,j), (K2,b), (K2,d), \ldots, (K5,e), (K5,h)\}$ can be represented by (a) a relation matrix, (b) a directed graph and a concept lattice, (c) with nonreduced and (d) reduced labeling (if R is a partial ordered set). A concept lattice can be expanded by adding background knowledge such as (e) GO trees. GO terms in the cellular component category are written in bold italic and those in the biological process category in italic (f) an expanded concept lattice, (f)=(d)+(e).

ues are less than 2.0. The heat shock 25–37 °C (YHS) dataset consists of six time points [1], and 2467 genes were downloaded from the Eisen laboratory (http://genome-www.stanford.edu/clustering/). We created 100 clusters equally for each dataset. The MITree-K clustering algorithm [21] was used to cluster genes in AGBM, as in the original manuscript [18] and in HCDC. We applied K-means clustering for yeast datasets, YCDC and YHS. Note that the higher concept-level analysis provided by BioLattice allows us an integrative analysis of heterogeneous datasets possessing different experimental designs and data processing steps.

2.3. Knowledge annotation and lattice construction

We annotated genes with three knowledge resources, (1) GO terms, (2) biological pathways and (3) transcription factors (TFs). GO-based annotation was performed according to [11]. Specifically, we used both implicit and explicit GO annotations [6]. We mapped clusters onto biologic pathways using ArrayXPath [10,11]. We matched the probe identifiers from microarrays to the pathway–node identifiers from KEGG, GenMAPP, BioCarta and/or PharmGKB Pathways. Pathway annotation provided by ArrayXPath was available only for the human and mouse datasets (i.e., AGBM and HCDC). YEASTRACT contains information about regulatory TFs in yeast. We used the documented TFs (http://www.yeastract.com/) to annotate clusters for yeast microarray datasets, YCDC and YHS.

To determine the binary relation between cluster and annotation in the relation matrix, we applied Fisher's exact test by constructing 2-by-2 contingency tables containing two cluster memberships (within and without a cluster) as row variables and attribute membership (within and without an annotation) as column variables [10]. As in our previous article, we applied a false discovery rate control to adjust the P-values from multiple-hypotheses testing. We determined the binary relation by testing if an annotation was over-represented significantly for any cluster at a certain threshold.

For each microarray dataset, we constructed a relation matrix. If the experimental context was S, $S := (G, M, I)$, each component in (G, M, I) indicated clusters, annotation terms and relation. We extracted all concepts from each context in an order that was convenient to draw a lattice. The Ganter algorithm [22] was applied. BioLattice was implemented as a web-based tool using Perl, JavaScript and Scalable Vector Graphics (http://www.snubi.org/software/biolattice/).

2.4. Structural analysis

Even an 'executive summary' can be too huge and complex for human eyes to read when there are many concepts. We propose two structural analysis methods: prominent sub-lattice analysis and core–periphery structure analysis. An atom sub-lattice is defined as a sub-lattice whose top element is a unit concept and the bottom element is an atom (i.e., a sub-lattice whose elements are the upper bounds of an atom). A coatom sub-lattice is defined as a sub-lattice whose top element is a coatom and the bottom element is a zero concept (Fig. 2(a)). A score is assigned according to the number of the concepts in a sub-lattice. Prominent (or bigger) sub-lattices are interpreted as more important substructures in the experimental context (Fig. 2).

Core–periphery analysis decomposes BioLattice into four disjoint core–periphery substructures: 'core', 'communicating', 'peripheral' and 'independent'. The 'core' sub-lattice is defined as the maximal atom sub-lattice in terms of size (i.e., the number of the red colored concepts in Fig. 2(b)). The set of all lower bounds to the nodes included in the 'core' sub-lattice (excluding those included in the 'core') is defined as the 'communicating' substructure (in green). When an atom equals a coatom, we call it 'independent' (in yellow). All the rest are defined as 'peripheral' (in gray).

3. Results

3.1. Construction and visualization

As shown in Fig. 1(a), the same GO terms (or pathways) are frequently assigned to multiple clusters. This redundancy problem that complicates data interpretation is completely removed by constructing a BioLattice that displays each object and attribute only once without redundancy (see Fig. 1(d)). Table 1 demonstrates that redundancy is widespread in all our four datasets. In AGBM, for example, 25 (=147–122) among the 147 significantly over-represented GO terms ($P < 0.01$) are redundantly assigned to more than one cluster.

Fig. 3 shows that BioLattice organizes all cluster–annotation relations systematically into a single unified structure with ordering and without redundancy. The upper semicircle of a concept node contains the concept ID and the lower semicircle the list of the cluster IDs belonging to the node. Only 36 among the 100 clusters from the AGBM dataset have at least one statistically significant annotation ($P < 0.01$; Table 1). The top node contains the 64 (=100–36) clusters (omitted) with no significant annotation. All terms below statistical significance are assigned to the bottom node (omitted). The 147 significant annotations consist of 122 unique GO terms (see Table 1) and are depicted as the connected attributes to the nodes.

BioLattice summarizes the complex relations among clusters and annotations. One can easily notice, for example, that there is no GO term shared by all clusters (i.e., the intent of the top element is empty). *Immune response* assigned to cluster 1 (in concept C1) and *immune cell migration*, *immune cell chemotaxis* and *neutrophil chemotaxis* assigned to cluster 2 (in concept C2) are shared by the lower common neighbor concept C31 (having cluster 22), which has 26 more annotations exclusively assigned to it. The term *immune response* is common to the three clusters 1, 22 and 24. All terms annotated to the yellow independent sub-lattices are exclusive to the individual nodes. More comprehensive analysis results for other datasets and other GO categories at various threshold levels are available at the Supplement site (http://www.snubi.org/software/biolattice/).

3.2. Structural analysis

Table 2 shows the five most prominent atom or coatom sub-lattices extracted from the concept lattice in Fig. 3.

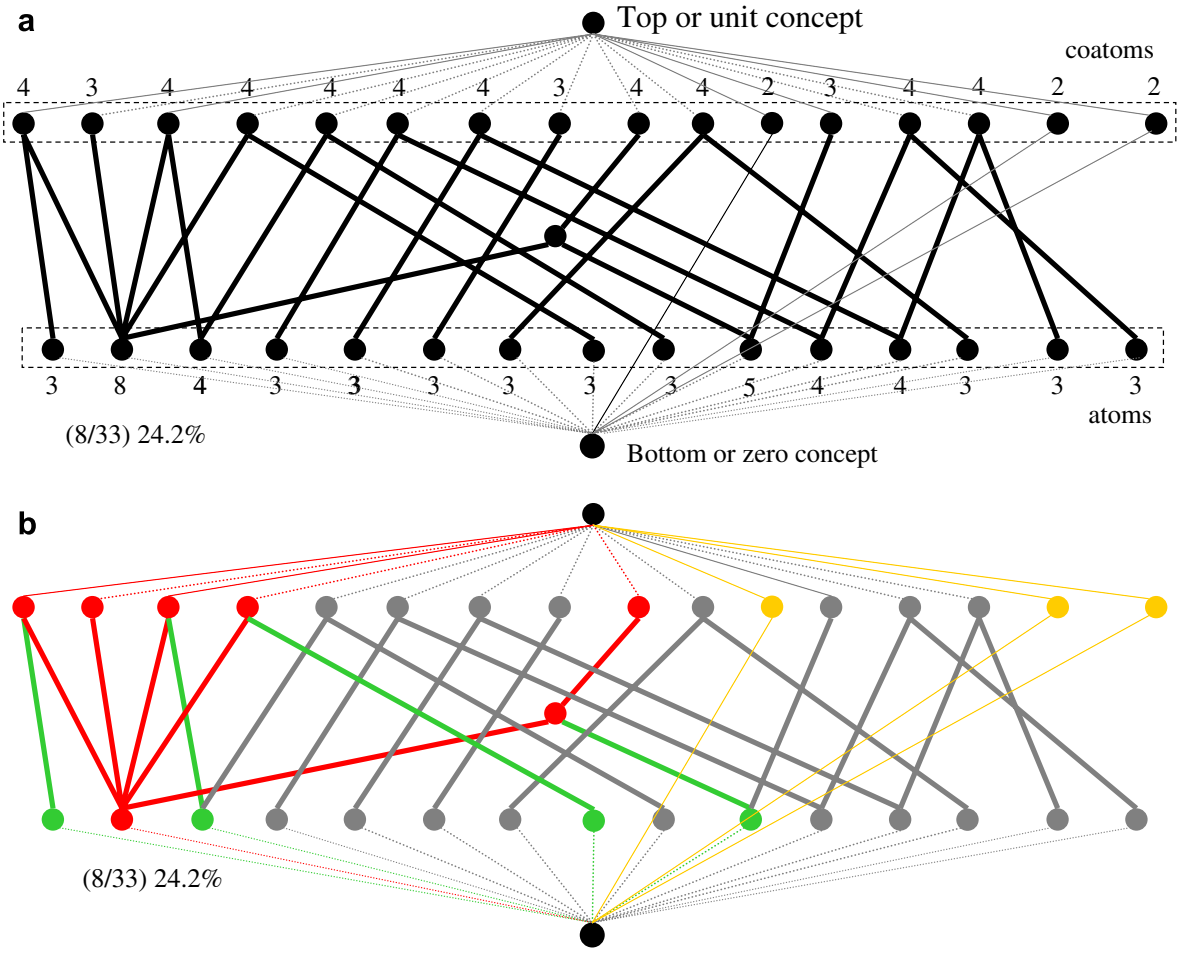

Fig. 2. Structural analysis. (a) Prominent sub-lattices are determined by the number of concepts in atom or coatom sub-lattices. (b) Core–periphery structures are marked in colors. The 'core' (red), communicating (green), 'peripheral' (gray) and 'independent' (yellow) substructures are used in combination for biological interpretation (see methods). The most prominent sub-lattice in (a) coincides with the 'core' sub-lattice in (b), occupying 24.2% (8/33) of all nodes in the context.

Table 1
Distribution of redundant annotations for gene expression clusters

Dataset	GO category	No. of clusters with significant GO annotations	No. of significant GO annotations (a)	No. of unique GO terms in GO annotations (b)	Redundancy = $(a)/(b)$
AGBM	BP	36	147	122	1.20
	MF	30	57	51	1.11
	CC	19	28	23	1.21
HCDC	BP	74	513	359	1.42
	MF	67	317	247	1.28
	CC	31	85	54	1.57
YCDC	BP	66	483	305	1.58
	MF	45	178	129	1.37
	CC	41	238	107	2.24
YHS	BP	70	435	270	1.61
	MF	61	187	133	1.40
	CC	53	248	122	2.03

BP, biological process; MF, molecular function; CC, cellular component; AGBM, anti-glomerular basement membrane; HCDC, human cell-division cycle; YCDC, yeast cell-division cycle; YHS, yeast heat shock.

Prominent sub-lattices appear to represent distinct functional areas. The first and the second sub-lattices are closely related to chemotaxis and cellular and humoral immune responses. The third is in association with apopto-

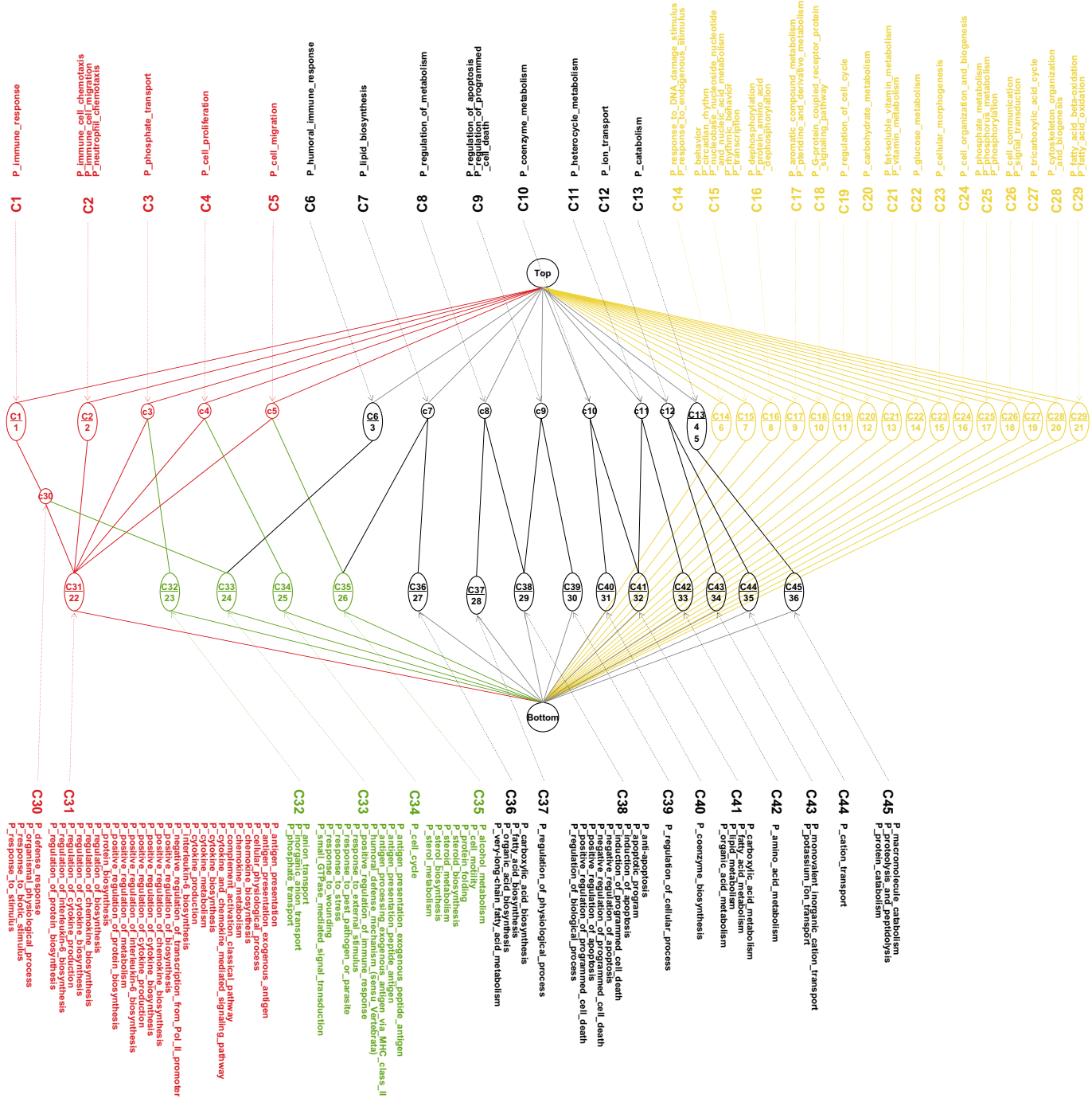

Fig. 3. Concept lattice constructed from the AGBM dataset having 100 clusters annotated by GO terms in the biological process category. Only 36 among 100 clusters demonstrate at least one significant GO term(s) ($P < 0.01$). Overall, the dataset shows 147 significant annotations with 122 unique GO terms (see Table1). The core–periphery substructures marked with colors (i.e., core in red, communicating in green, independent in yellow and peripheral in gray). The prominent substructures in Table 2 provide valuable information for the biological interpretation of the overall experimental context.

sis and programmed cell death. The fourth and the fifth are related to metabolic processes. This summarization with immune response, apoptosis and metabolic processes for a study of an immunoglobulin-mediated disease (i.e., IgA nephropathy) is in very good agreement with the biological interpretation of the original experiment [18]. This type of structure-based interpretation without concept lattice formation is hard to be made directly from the long unordered list of annotations for 100 clusters.

Core–periphery analysis determined the core sub-lattice (marked in red in Fig. 3) having three clusters that are annotated with chemotaxis-related GO terms (i.e., *immune cell chemotaxis*, *immune cell migration* and *neutrophil chemotaxis* for clusters 2 and 22 and *immune*

Table 2
Prominent sub-lattice analysis: the five largest atom or coatom sub-lattices with their objects and attributes are listed for the ABGM dataset with 100 clusters annotated by the GO biological process category ($P < 0.01$)

The largest prominent sub-lattice has 8 nodes having 3 clusters, 1, 2 and 22, with 90 genes and 37 GO terms		
1	*Antigen presentation*	*Organismal physiological process*
	Antigen presentation exogenous antigen	*Phosphate transport*
	Cell migration	*Positive regulation of biosynthesis*
	Cell proliferation	*Positive regulation of chemokine biosynthesis*
	Cellular physiological process	*Positive regulation of cytokine biosynthesis*
	Chemokine biosynthesis	*Positive regulation of cytokine production*
	Chemokine metabolism	*Positive regulation of interleukin-6 biosynthesis,*
	Complement activation classical pathway	*Positive regulation of metabolism*
	Cytokine and chemokine mediated signaling pathway	*Positive regulation of protein biosynthesis*
	Cytokine biosynthesis	*Protein biosynthesis*
	Cytokine metabolism	*Regulation of biosynthesis*
	Cytokine production	*Regulation of chemokine biosynthesis*
	Defense response	*Regulation of cytokine biosynthesis*
	Immune cell chemotaxis	*Regulation of cytokine production*
	Immune cell migration protein biosynthesis	*Regulation of interleukin-6 biosynthesis*
	Immune response	*Regulation of protein biosynthesis*
	Interleukin-6 biosynthesis	*Response to biotic stimulus*
	Negative regulation of transcription from Pol II promoter	*Response to stimulus*
	Neutrophil chemotaxis	
The second largest sub-lattice has 5 nodes having 3 clusters, 1, 24 and 3, with 57 genes and 16 GO terms		
2	*Antigen presentation exogenous peptide antigen*	*Positive regulation of immune response*
	Antigen presentation peptide antigen	*Response to biotic stimulus*
	Antigen processing exogenous antigen via MHC class II	*Response to external stimulus*
	Defense response	*Response to pest pathogen or parasite*
	Humoral defense mechanism (sensu Vertebrata)	*Response to stimulus*
	Humoral immune response	*Response to stress*
	Immune response	*Response to wounding*
	Organismal physiological process	*Small GTPase mediated signal transduction*
The third largest sub-lattice has 4 nodes having 1 cluster, 29, with 26 genes and 12 GO terms		
3	*Anti-apoptosis*	*Regulation of programmed cell death*
	Apoptotic program	*Positive regulation of apoptosis*
	Induction of apoptosis	*Positive regulation of programmed cell death*
	Induction of programmed cell death	*Regulation of apoptosis*
	Negative regulation of apoptosis	*Regulation of biological process*
	Negative regulation of programmed cell death	*Regulation of metabolism*
The forth largest sub-lattice has 4 nodes having 1 cluster, 26, with 7 genes and 9 GO terms		
4	*Alcohol metabolism*	*Steroid biosynthesis*
	Cell migration	*Steroid metabolism*
	Cell motility	*Sterol biosynthesis*
	Lipid biosynthesis	*Sterol metabolism*
	Protein folding	
The fifth largest sub-lattice has 4 nodes having 1 cluster, 32, with 14 genes and 6 GO terms		
5	*Coenzyme metabolism*	*Heterocycle metabolism*
	Carboxylic acid metabolism	*Lipid metabolism*
	Fatty acid metabolism	*Organic acid metabolism*

response for clusters 1 and 22 with the 26 terms specific to cluster 22 including humoral immune response, antigen presentation, chemokine, cytokine, interleukin and complement pathways). The communicating substructure (marked in green) has clusters 23–26 involved both in immune responses and metabolic processes. The peripheral substructure (marked in black) is associated with apoptosis and lipid metabolism. The 16 independent sub-lattices (marked in yellow) are related to various metabolic processes. The core sub-lattice extracted from the lattice with pathway-based annotation for the same dataset also shows high correlation with immune response-related pathways including IL signaling-, MAPK signaling-, growth factor- and T-cell-related pathways (Supplementary Tables S1 and S2). The results for detailed analyses and other datasets are available on the Supplement page.

HCDC shows cell cycle- and DNA replication-related GO terms in prominent sub-lattices (Table S3). The core sub-lattice shows cell cycle-related GO terms (*cell cycle*, *cell proliferation*, *DNA metabolism*, *response to stress*, etc., Table S4) and pathways (Tables S5 and S6).

YCDC, in a similar fashion, shows cell cycle-related GO terms in the prominent sub-lattices (Table S7). The core sub-lattice also contains cell cycle-related terms including *cell cycle*, *response to stimulus*, *DNA metabolism* and *development* (Table S8). We applied TFs to annotate yeast datasets. TFs associated with the prominent sub-lattices are related to cell cycle and cell growth (Table S9). Cell cycle-related TFs are rich in the core sub-lattice, including *gal4*, *gat3*, *hcm1*, *mbp1*, *pdr1*, *phd1*, *sok2*, *ste12*, *swi4*, *swi5*, *tos8* and *yox1*. *Yox1* is known to bind to the promoters of cell cycle-regulated genes. *Swi4* and *swi5* are well-known cell cycle-related TFs that regulate G1 phase transcription. *Hcm1* and *mbp1* are also known to be cell cycle-related (Table S10).

For YHS, the prominent sub-lattices are significantly associated with the stress-related translational processes, *ribosomal processes*, *biosynthesis*, *metabolism* and *translation* (Table S11). The core sub-lattice also contains the biological process terms *protein biosynthesis*, *ribosome biogenesis and assembly*, *translation* and *metabolism* (Table S12). The TFs annotated to the core sub-lattice are *abf1*, *fhl1*, *ifh1*, *rap1*, *rpn4* and *sfp1*. *Abf1* is a TF gene involved in DNA repair (Table S10). Ribosomal protein genes in eukaryotes are coordinated in response to growth stimuli and environmental stress, thereby permitting cells to adjust ribosome numbers and overall protein synthetic capacity to physiological conditions. The transcriptional regulator *rap1* binds to most yeast ribosomal protein promoters, and *fhl1* and *ifh1* associate almost exclusively with ribosomal protein promoters [23]. In yeast, *rpn4* is regulated transcriptionally by various stress responses, whereas *Sfp1* is a stress- and nutrient-sensitive regulator of ribosomal protein gene expression [24].

3.3. Structural robustness

Because applying different significance thresholds may result in different lattices and substructures, it is essential to evaluate the robustness of the analysis results using different thresholds. We evaluated the structural robustness by measuring the consistency for a cluster (or a GO term) to remain in the same substructure across six different threshold cutoffs (i.e., $P <$ 0.01, 0.005, 0.001, 5.0×10^{-4}, 1.0×10^{-5} and 5.0×10^{-5}). Large portions of the clusters in all datasets tend to remain in the same core–periphery substructures at all six threshold cutoffs (64% for ABGM, 51% for HCDC, 47% for YCDC and 45% for YHS) (Table S13). Chi-square testing on the four (substructure)-by-six (cutoff) levels shows that all clusters (or all annotations) show a statistically significant tendency to remain in the same substructure (Table 3). In ABGM, for example, *immune response* remains in the core sub-lattice across all cutoff levels, in agreement with the context of the immunoglobulin-mediated kidney disease, IgA nephropathy.

For HCDC and YCDC, *ATPase activity* and *helicase activity* are consistently annotated in the core and communicating substructures across all cutoff levels. Both coimmunoprecipitation and two-hybrid assays showed that ATPase is essential for cell cycle progression. Yeast mutational analyses demonstrated that mutations affecting ATPase activity also abolished helicase activity [25]. Mutational studies of human DNA helicase B suggest that its activity is critical for the G1/S transition of the mitotic cell cycle [26].

Ribosomal processes remain in the core sub-lattice at all significance levels for YHS: *structural constituent of ribosome* in the molecular function and *ribosome* and *cytosolic ribosome* in the cellular component. Genes whose expressions are repressed by multiple environmental stresses almost entirely consist of forms that encode proteins associated with ribosomal structure, function or biogenesis [27].

Not only thresholds but also different numbers of clusters (or k) may result in different lattices, we also evaluated the robustness of the structural analysis results using different k's. We created 10, 25, 50, 75, 100, 125, 150 and 200 clusters for each dataset and evaluated the consistency of GO terms to remain in the same core–periphery substructure across different k's. Chi-square testing on the four (substructures)-by-eight (k) groups showed statistically significant tendency for all GO terms to remain in the same

Table 3
Statistical significance of gene expression clusters and annotations to remain in the same substructure of a concept lattice across six different threshold levels

Dataset	Concept	BP	MF	CC	All categories
AGBM	Cluster	$4.8 \times 10^{-3*}$	0.242	0.268	0.052
	GO	$2.2 \times 10^{-13*}$	$2.2 \times 10^{-3*}$	$9.2 \times 10^{-3*}$	$2.5 \times 10^{-23*}$
HCDC	Cluster	$7.0 \times 10^{-5*}$	$9.2 \times 10^{-8*}$	$1.2 \times 10^{-4*}$	$8.5 \times 10^{-14*}$
	GO	$8.4 \times 10^{-25*}$	$1.2 \times 10^{-17*}$	$8.7 \times 10^{-15*}$	$3.8 \times 10^{-77*}$
YCDC	Cluster	$1.4 \times 10^{-13*}$	$2.7 \times 10^{-19*}$	$3.7 \times 10^{-21*}$	$2.9 \times 10^{-22*}$
	GO	$1.0 \times 10^{-145*}$	$8.0 \times 10^{-79*}$	$2.4 \times 10^{-157*}$	$\approx 0^{*}$
YHS	Cluster	$1.5 \times 10^{-7*}$	$1.2 \times 10^{-6*}$	$9.6 \times 10^{-22*}$	$8.9 \times 10^{-34*}$
	GO	$6.9 \times 10^{-61*}$	$1.2 \times 10^{-32*}$	$4.4 \times 10^{-174*}$	$1.2 \times 10^{-306*}$

BP, biological process; MF, molecular function; CC, cellular component; AGBM, anti-glomerular basement membrane; HCDC, human cell-division cycle; YCDC, yeast cell-division cycle; YHS, yeast heat shock.
* $P < 0.01$.

core–periphery substructure across different numbers of clusters (Table S14).

4. Discussion and conclusion

Biomedical ontology or pathway-based annotation of gene expression clusters is one of the most powerful approaches for interpreting DNA microarray experiments [28]. Rather than interpreting one cluster at a time, BioLattice integrates all gene expression clusters and annotations into a unified framework: a lattice of concepts. BioLattice replaces a long unordered list of annotations for clusters with a unified structure, a context of concepts (or units of thought) that considers both clusters and annotations simultaneously. Annotation redundancy is completely avoided. Complex relations among clusters and annotations are clarified, ordered and visualized.

There have been systematic efforts to organize clusters to reveal meta-structures. Self-Organizing Maps (SOM), for example, impose topographic ordering on the cluster. The Self-Organizing Tree Algorithm [29] imposes a binary tree structure on the data by combining SOM and hierarchical clustering. Matrix Incision Tree-*K* [21] is a divisive hierarchical clustering algorithm that provides multilevel threshold-graph representation of clusters based on expression profile-similarity measures. However, to the best of our knowledge, there is no algorithm that considers both expression profile and annotation similarities simultaneously. Moreover, while previous ontology and pathway-based analysis methods analyze each cluster–annotation set separately, BioLattice considers all existing relations among clusters and annotations simultaneously, providing comprehensive insight into the overall experimental context.

Genetic and functional co-regulations do not necessarily coincide. Therefore, genes sharing common functions may show completely different expression profiles. BioLattice analysis helps to explore the modular organization of functional processes by extracting experimental concepts or biological modules. External knowledge resources can be added to better explore the underlying structures (Fig. 1(d)–(f)). Knowledge resources may include GO, bio-pathways, transcription factor binding, chromosomal co-location, protein–protein interaction networks and so forth. While GO-based annotation may be regarded as direct functional attributes of genes, different knowledge resources like pathway membership must be carefully interpreted according to the nature of the association between genes and annotations.

We applied the Ganter algorithm for concept lattice generation. Algorithmic research on concept generation is being pursued by other research groups [30]. The concern is when both numbers of clusters and annotations are very big. The performance of Ganter algorithm is mainly dependent upon the number of objects (clusters) and the time complexity ($O(|G|^2*|M|*|L|)$) increases exponentially with the growth of the number of clusters ($|G|$, $|M|$ and $|L|$ are sizes of the clusters, annotations and concept lists, respectively). However, it is not a problem for hundreds of clusters. By increasing the number of clusters for the AGBM dataset, we measured time complexity 10 times and found that less than a minute is required for 3000 clusters (Supplement Fig. S1) by a computer with dual Xeon processors.

The two structural analysis methods for BioLattice enable us to extract biological processes relevant to the experimental context under investigation. Prominent sub-lattices may help us to extract central concepts and related sub-contexts of varying biological importance. While prominent sub-lattices may overlap with each other, core–periphery analysis decomposes a concept lattice into four disjoint subsets with different biological roles. Both prominent sub-lattice and core–periphery analyses facilitate structured biological interpretation of entire microarray experiments.

Graphical representation using a concept lattice may provide formalism for knowledge-based conceptual clustering. BioLattice provides a shared platform for comparing different experiments (or contexts) in a systematic manner at a semantic level. Suggested future work is mathematical processing of the lattices for formal analysis, association rule mining, clustering of complex data, integrating heterogeneous biological knowledge resources, and comparative analysis of different microarray experiments.

Acknowledgments

This study was supported by a grant from Ministry of Health & Welfare, Korea (A040163). J.K.'s educational training was partly supported by a grant from Ministry of Health & Welfare, Korea (A060711).

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;95:14863–8.
[2] Walker MG, Volkmuth W, Sprinzak E, Hodgson D, Klingler T. Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes. Genome Res 1999;9:1198–203.
[3] Al-Shahrour F, Diaz-Uriarte R, Dopazo J. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics 2004;20:578–80.
[4] Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, et al. GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics 2004;20:3710–5.
[5] Dennis Jr G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, et al. DAVID: database for annotation, visualization, and integrated discovery. Genome Biol 2003;4:P3.
[6] Robinson PN, Wollstein A, Bohme U, Beattie B. Ontologizing gene-expression microarray data: characterizing clusters with Gene Ontology. Bioinformatics 2004;20:979–81.
[7] Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, et al. GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol 2003;4:R28.
[8] Zhang B, Schmoyer D, Kirov S, Snoddy J. GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. BMC Bioinformatics 2004;5:16.

[9] Zhong S, Storch KF, Lipan O, Kao MC, Weitz CJ, Wong WH. GoSurfer: a graphical interactive tool for comparative analysis of large gene sets in gene ontologytrade mark space. Appl Bioinformatics 2004;3:261–4.
[10] Chung HJ, Kim M, Park CH, Kim J, Kim JH. ArrayXPath: mapping and visualizing microarray gene-expression data with integrated biological pathway resources using Scalable Vector Graphics. Nucleic Acids Res 2004;32:W460–4.
[11] Chung HJ, Park CH, Han MR, Lee S, Ohn JH, Kim J, et al. ArrayXPath II: mapping and visualizing microarray gene-expression data with biomedical ontologies and integrated biological pathway resources using Scalable Vector Graphics. Nucleic Acids Res 2005;33:W621–6.
[12] Hosack DA, Dennis Jr G, Sherman BT, Lane HC, Lempicki RA. Identifying biological themes within lists of genes with EASE. Genome Biol 2003;4:R70.
[13] Pandey R, Guru RK, Mount DW. Pathway Miner: extracting gene association networks from molecular pathways for predicting the biological significance of gene expression microarray data. Bioinformatics 2004;20:2156–8.
[14] Damian D, Gorfine M. Statistical concerns about the GSEA procedure. Nat Genet 2004;36:663. [author reply 663].
[15] Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 2003;34:267–73.
[16] Wille R. Restructuring Lattice Theory: an approach based on hierarchies of concepts. Dordrecht-Boston: Reidel; 1982.
[17] Carpineto C, Romano GE. Concept data analysis: theory and applications. Hoboken, NJ: Wiley; 2004.
[18] Kim JH, Ha IS, Hwang CI, Lee YJ, Kim J, Yang SH, et al. Gene expression profiling of anti-GBM glomerulonephritis model: the role of NF-kappaB in immune complex kidney disease. Kidney Int 2004;66:1826–37.
[19] Whitfield ML, Sherlock G, Saldanha AJ, Murray JI, Ball CA, Alexander KE, et al. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Mol Biol Cell 2002;13:1977–2000.
[20] Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, et al. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell 1998;9:3273–97.
[21] Kim JH, Kohane IS, Ohno-Machado L. Visualization and evaluation of clustering structures for gene expression data analysis. J Biomed Inform 2002;35:25–36.
[22] Ganter B, Wille R. Formal concept analysis: mathematical foundations. Berlin, New York: Springer; 1999.
[23] Wade JT, Hall DB, Struhl K. The transcription factor Ifh1 is a key regulator of yeast ribosomal protein genes. Nature 2004;432:1054–8.
[24] Marion RM, Regev A, Segal E, Barash Y, Koller D, Friedman N, et al. Sfp1 is a stress- and nutrient-sensitive regulator of ribosomal protein gene expression. Proc Natl Acad Sci USA 2004;101:14315–22.
[25] Pause A, Methot N, Svitkin Y, Merrick WC, Sonenberg N. Dominant negative mutants of mammalian translation initiation factor eIF-4A define a critical role for eIF-4F in cap-dependent and cap-independent initiation of translation. EMBO J 1994;13:1205–15.
[26] Gu J, Xia X, Yan P, Liu H, Podust VN, Reynolds AB, et al. Cell cycle-dependent regulation of a human DNA helicase that localizes in DNA damage foci. Mol Biol Cell 2004;15:3320–32.
[27] Loar JW, Seiser RM, Sundberg AE, Sagerson HJ, Ilias N, Zobel-Thropp P, et al. Genetic and biochemical interactions among Yar1, Ltv1 and Rps3 define novel links between environmental stress and ribosome biogenesis in Saccharomyces cerevisiae. Genetics 2004;168:1877–89.
[28] Yue L, Reisdorf WC. Pathway and ontology analysis: emerging approaches connecting transcriptome data and clinical endpoints. Curr Mol Med 2005;5:11–21.
[29] Herrero J, Valencia A, Dopazo J. A hierarchical unsupervised growing neural network for clustering gene expression patterns. Bioinformatics 2001;17:126–36.
[30] Kuznetsov S. Comparing performance of algorithms for generating concept lattices. J Exp Theor Art Int 2002;2/3(14):189–216.