Multi-Aspect Gene Relation Analysis

H. Yamakawa, K. Maruhashi, and Y.Nakao

Pacific Symposium on Biocomputing 10:233-244(2005)

MULTI-ASPECT GENE RELATION ANALYSIS

H. YAMAKAWA, K. MARUHASHI, Y.NAKAO
Fujitsu Laboratories Ltd., 1-1, Kamikodanaka 4-chome, Nakahara-ku, Kawasaki, Kanagawa,211-8588, Japan

M. YAMAGUCHI
Fujitsu Limited, 9-3, Nakase 1-chome, Mihama-ku, Chiba, Chiba, 261-8588, Japan

Recent progress in high-throughput screening technologies has led to the production of massive amounts data that we can use to understand biological systems. To interpret this data, biologists often need to analyze the characteristics of a set of genes by using Gene Ontology (GO) annotation. We are proposing a novel method for assisting such an analysis. Given a set of genes, the method automatically extracts several analyzing aspects in terms of a subset of genes that are attached to some related GO terms. It then creates a gene-attribute bipartite graph that highlights the aspect selected by the user according to his/her interests. We describe this method in detail and report on an experiment where the proposed method is applied to the analysis of rat kidney expression data.

1 Introduction

The DNA microarray is an effective tool for monitoring and profiling gene expression patterns. It can measure the expression levels of thousands of genes simultaneously and provide a set of expression patterns for a given list of genes. Biologists analyze gene expression patterns to determine their biological meanings (e.g., interactions between specific genes, dependencies between changes in gene expressions, and patient 's responses to treatment). In the bioinformatics field, many methods, including a kind of data mining, have been applied to assisting such analysis[1]. For example, Kennedy et al. used a clustering method to assist in microarray dataset analysis. They extracted the gene list by preprocessing the gene expression data. They then applied a clustering method to the gene list and then presented the resulting gene clusters together with meaningful descriptions using the functional information obtained from the Gene Ontology (GO) [2].

The GO is a vocabulary that describes the attributes of genes (for example, their biological functions). Each term in the vocabulary, called a GO term, represents a possible attribute value that is possessed by a gene. The GO has a hierarchical structure, i.e., GO terms are connected by *is-a* relations and construct a directed acyclic graph. The GO Consortium is currently creating three standard gene ontologies that will describe the associated biological

processes, cellular components, and molecular functions for genes and their products (RNA or protein products encoded by genes). Many biological resources, including LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink), use the GO terms to annotate gene properties.

Because there are many kinds of GO terms, attaching GO annotations to a gene list produces high-dimensional data. In fact, computer-aided analysis methods are required, such as clustering, principal component analysis (PCA), and self-organizing maps (SOM). Many of these conventional methods are helpful in understanding the overall characteristics of a gene list (e.g., the major gene relationships found within that gene list). For example, as shown by Kennedy et al.[3], a dendrogram display of hierarchical clustering with GO terms can be used to illustrate the overall structure of the major characteristics of a gene list.

There is a case, however, in which the overall structure of the gene list characteristics is not appropriate. Because high-dimensional data can contain many aspects, an overall structure is incapable of illustrating all of these aspects, such that the user may miss some important aspects relating to his or her interests. For example, consider a case in which collagen activity is observed in a biological phenomenon related to osteoblasts and osteoclasts. One possible analysis aspect is the collagen metabolism. A biologist whose main interest resides in the metabolism process will want to understand the collagen biosynthetic pathway from the viewpoint of how the collagen biosynthesis pathway interacts with other metabolic pathways, or the requirements for collagen biosynthesis. In this case, it is important to be able to distinguish between the intracellular and the extracellular phenomena, as well as distinguish the metabolism phenomena from other phenomena (see " The extracellular matrix of animals ", pp 971–995, in[4]). Another possible aspect is animal development. A biologist whose main interest is in the developmental program of animals will want to identify the types of collagen-related developmental processes that occur. In this case, it is appropriate to classify the active genes based on their relationship to the development processes (e.g. the formation of an extracellular matrix, arrangement of the cytoskeleton) (see " Fibroblasts and Their Transformations: The Connective-Tissue Cell Family ", pp.1179–1187, in[4]).

In response to the above demands, we have developed a tool for supporting gene relationship analysis, called Genesphere Connection Miner (Cminer)[5], which creates a gene-term bipartite graph that can be focused on the user's interests. The contents of the display can be changed according to the user ' s settings, as specified by a list of GO terms. For example, Wagatsuma et al. reported that they had obtained results related their interest by using this tool to analyze the temporal expression data for a hepatitis model rat

(about 500 genes)[6]. There is a problem with this technique, however, in that it takes a long time to create an appropriate list of GO terms by hand. To solve this problem, this paper proposes a method for automatically extracting the analysis aspects. Given a gene list, the method automatically extracts several aspects in terms of a subset of genes attached with related GO terms. The user can easily specify his/her interest by selecting one appropriate aspect from those that are available. In Chapter 2, we describe the aspect extraction method in detail. In Chapter 3, we report on an experiment in which the proposed method is applied to the analysis of rat kidney expression data.

2 Multi-aspect gene relation analysis system

We are proposing a multi-aspect gene relation analysis system, which outputs multiple aspects about a given gene list. For an aspect selected by the user, this system displays a bipartite graph consisting of gene symbols and GO terms. This system uses the Gene Ontology (GO) that is a vocabulary used to describe the attributes of genes, LocusLink that is a gene database in which genes are annotated with GO terms, and HomoloGene that is used to provide orthologous information.

A unique feature of this system is that it automatically extracts multiple aspects for analyzing a gene list by using a conceptual clustering technique called ETMIC situation decomposition (E-SD) [7,8]. The E-SD method simultaneously selects a gene subset and a GO term subset as an aspect. To date, however, the E-SD method has faced two issues related to this function. One involves the comparison of GO terms of different abstraction levels. The second relates to the fact that only a few combinations of genes can be compared because many genes have no GO term annotations [a]. Our new system overcomes the first problem by: [a] a GO term is inferred by using a transitive relationship. Then, to overcome the latter problem: [b] a GO term is extended by using orthologous information, [c] a GO term is summarized by using weighted singular value decomposition (SVD) [b]

To realize these countermeasures, our new system is composed of the seven subprocesses described below (see Figure 1).

(1) Inference of GO terms using transitive relationships [a]
(2) Logarithmic probability weighted SVD [c]
(3) Additions of terms of orthologous genes for each gene[b]

[a] Actually, over 20% of genes of human are annotated with GO term, but only 4% of genes of rat are annotated.

[b] Although these countermeasures may introduce noisy information, the advantage is expected to outweigh the disadvantages in such a case of very poor gene annotations.

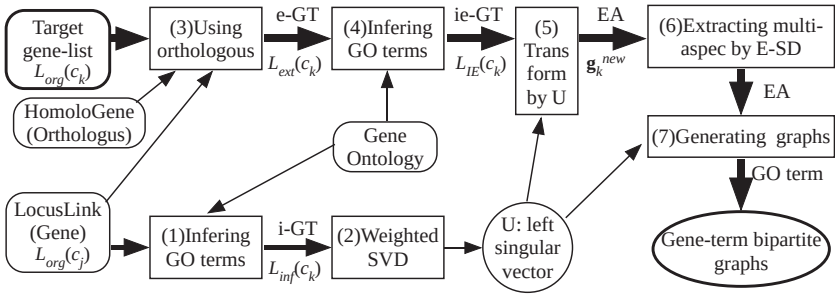

Figure 1: Multi-aspect gene relationship analysis system.
Rounded-rectangles indicate input data, rectangles are processing, and circles are output. Bold arrows indicate the gene list flow for which attributes are added. i-GT: inferred GO term, e-GT: extended GO term, ie-GT: inferred extended GO term, EA: eigen-attributes.

(4) Inference of GO terms using transitive relationships [a]
(5) Transformation of GO terms to eigen-attributes.[c]
(6) Extraction of multi-aspects by E-SD method
(7) Generation of gene-term bipartite graph for each aspect [c]

The main process path for the target gene list runs from steps (3) to (7), with the multi-aspect being extracted in (6). The process path from (1) to (2) generates a left singular vector using all the available genes. To collect terms that have a similar appearance distribution, the vector transforms the target gene list from GO terms to eigen-attributes in (5).

2.1 Inferring GO terms using transitive relations

LocusLink contains a GO term list $L_{org}(c_j)$ for each gene c_j. Each term list $L_{org}(c_j)$ in LocusLink is converted to an inferred GO term list $L_{inf}(c_j)$. In practice, all of the ancestor terms for each GO term are added to the new term list $L_{inf}(c_j)$, using the transitive relationship of the GO.

2.2 Logarithmic probability weighted SVD

To gather up those GO terms having a similar distribution, a gene representation is transformed into eigen-attributes. This transformation subprocess (5) uses a left singular vector U, calculated by the SVD method in subprocess (2).

Logarithmic probability weighting :

When we give all the GO terms an equal weighting, the SVD process tends to select those terms in the higher layers of the GO hierarchy, which has very little meaning. This is because the SVD process tends to select those GO terms associated with many genes. In addition, those terms in a higher layer tend to associate with many genes (e.g., "binding" in the molecular function ontology).

To overcome this problem, each GO term i is weighted with a logarithmic probability weight W_i which highlights a moderate abstraction level (depth) in the GO hierarchy. W_i is calculated using the following formula.

$$W_i = -\log(\frac{m_i}{m}) \tag{1}$$

Where m is the total number of genes, and m_i is the number of genes with term i. For instance, the topmost concept in a hierarchy is annotated by all the genes, its weighting becomes zero $W_i = 0$, and so is ignored. On the other hand, a special narrower term is emphasized.

The idea of logarithmic probability weighting relates to the formula of similarities in the layered structure as proposed by Resnik, Lin, and Jian [9].

Singular value decomposition (SVD):

In preparation for the SVD process, the inferred terms $L_{inf}(c_j)$ for all the genes are converted to a matrix. Firstly, each inferred term list $L_{inf}(c_j)$ $(j \in [1, m])$ is converted into a GO term vector $\mathbf{g}_j$ of length n. Here, n is the number of GO terms used in all the genes. An element registered in the inferred term list $L_{inf}(c_j)$ is set to W_i while others are set to 0. Secondly, all the GO term vectors $\mathbf{g}_j$ $(\forall j \in [1, m])$ are collected into a matrix G with n columns (GO terms) and m rows (genes). SVD decomposes matrix G, as follows.

$$G = USD^T \tag{2}$$

U and D are a unitary matrix that satisfies $U^TU = I_n$ and $D^TD = I_m$ respectively. The column vector of U is called the left singular vector. The column vector of D is called the right singular vector.

2.3 Adding terms of the orthologous genes to the target gene list

Each gene c_k in the target gene list ζ is annotated by the GO term list $L_{org}(c_k)$. To compensate for insufficient GO term annotation in the target gene list $L_{org}(c_k)$, GO terms are added using orthologous information. Each target

gene c_k of rat has corresponding orthologous genes (human and mouse). GO terms that belong to orthologous genes and rat genes c_k are added to a new GO term list $L_{ext}(c_k)$. This is called the extended GO term list for gene c_k. Orthologous information for three species (human, mouse, and rat) is obtained from HomoloGene. This is in the form of a 13952-row (entry) by 3-column (race) LocusID matrix.

2.4 Infering GO terms using transitive relations

In the same way as in subprocess (1), we obtain an extended inferred term list $L_{IE}(c_k)$ from term list $L_{ext}(c_k)$ for each gene c_k in target gene list ζ.

2.5 Transforming GO terms to eigen-attributes

As part of the preparations for the E-SD process, the genes' representations are transformed into eigen-attribute set α from inferred extended term lists $L_{IE}(c_k)$. The set α consists of the top 20 eigen-attributes acquired by the SVD process. Each term list $L_{IE}(c_k)$ for one gene is converted into a GO term vector $\mathbf{g}_k^{IE}$ of length n. Here, n is the number of GO terms used in all the genes. An element registered in a term list $L_{IE}(c_k)$ is set to W_i and others are set to 0. Here W_i is the weighting for the i-th GO term.

Each gene vector $\mathbf{g}_k^{new}$ which is described by eigen-attribute set α is transformed from $\mathbf{g}_k^{IE}$ using left singular vector U.

$$\mathbf{g}_k^{new} = U^T \mathbf{g}_k^{IE} \tag{3}$$

2.6 Extracting multi-aspects with the E-SD method

The ETMIC situation decomposition (E-SD) method [7,8] is used to extract multiple aspects from the target gene list. The target gene list ζ is a collection of vectors $\mathbf{g}_k^{new}$ described by eigen-attribute set α. This list is described as the large square in the left-hand part of Figure 2. Each extracted aspect $J = \{A, C\}$ is a combination of subset A of eigen-attribute set α (horizontal axis in Fig. 2), and subset C of the gene sets ζ (vertical axis in Fig.2).

The E-SD algorithm selects some useful multi-aspects from the enormous combination of subsets A and C. This process is based on searching for the local maximum point of an ETMIC criterion to change the gene selection C in every partial space A [8]. The ETMIC criterion that evaluates each aspect $J = \{A, C\}$ is as follows.

$$E(A, C) = n_C \left(\min_i \left(I_{X_A^{-i}; X_i}(C) \right) - \max_j \left(I_{X_A; X_j}(C) \right) \right) \tag{4}$$

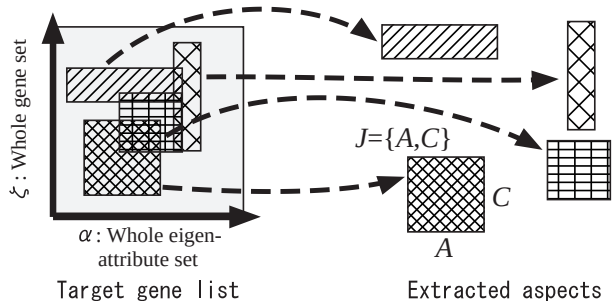

Figure 2: Composition of aspect decomposition process
Because of limitations of the figure, the selections of gene subsets and eigen-attribute subsets are drawn as a continuous area.

Here, n_C is the selected number of genes. $I_{X_A^{-i};X_i}(C)$and $I_{X_A;X_j}(C)$are mutual information on each of two partial spaces for gene subset C. X_A^{-i} is a partial space for eigen-attribute subset i to eigen-attribute subset $A - a_i$. X_i is a partial space for eigen-attribute a_i. X_A is a partial space for eigen-attribute subset A. X_j is a partial space for eigen-attribute j.

2.7 Generation of gene-term bipartite graph for each aspect

For each aspect, genes and GO terms should be selected to draw the gene-term bipartite graph, which is helpful in understanding the gene relations. The members of gene subset C are displayed on this graph. Unfortunately, it is difficult to select the GO terms to be displayed on this graph, because the eigen-attribute subset A corresponds to a weighted selection of GO terms.

The norm of the vector U for i-th GO term $|U_i| = \sqrt{\sum_{a \in A} {U_{ai}}^2}$ is used for selecting the significant GO terms, where U_{ai} is the a-th eigen-attribute and the i-th GO term element of matrix U. GO terms are selected so that every gene connects to at least one GO term. This process selects one GO term from the inferred extended term list $L_{IE}(c_k)$ of each gene. This selection for each gene c_k is based on the equation using the maximum norm of U, as follows.

$$i_{max}(c_j) = \arg \max_{i \in L_{IE}(c_j)} |U_i| \tag{5}$$

Finally, each aspect is drawn by a gene-term bipartite graph, includes the gene subset C and the GO term subset indicated by $i_{max}(c_j)$.

3 Experiment: Analysis of gene-list of rat embryonic kidney

In this section, we describe an experiment in which the proposed method is applied to the analysis of a gene list obtained through microarray expression profiling.

3.1 Experimental setup: Analyzed gene list

Stuart et al.[10] analyzed gene expression patterns during kidney organogenesis using DNA array technology and, as a result, classified 8,740 genes into five discrete clusters based on the temporal patterns of their expression levels. We used the second cluster (group 2), whose expression pattern peaks with mid-nephrogenesis, as an example gene list for our experiment. The following lists the symbols (bold font) and the names of the genes used for our experiment.

> **AGR**: Agrin, **Calm1**: Calmodulin 1 (phosphorylase kinase, delta), **Col1a1**: collagen, type 1, alpha 1, **Dcn**: decorin, **ENP1MR**: Epithelial membrane protein 1, **Erbb2**: Avian erythroblastosis viral (v-erb-B2) oncogene homologue 2 **Erp29**: endoplasmic retuclum protein 29, **Galr3**: galanin receptor 3, **ID125A**: Inhibitor of DNA binding 1, helix-loop-helix protein, **ILGFBPA**: Insulin-like growth factor binding protein 2, **Lamc1**: laminin, gamma 1, **Lbp**: lipopolysaccharide binding protein, **Mmp2**: matrix metalloproteinase 2, **Mtap6**: microtubule-associated protein 6, **Nkaa1b**: ATPase, Na+K+ transporting, alpha 1, **Npr1**: natriuretic peptide receptor 1, **Phb**: Prohibitin, **Pkcb**: protein kinase C, beta 1, **Pmp22**: peripheral myelin protein 22, **SOMATO**: somatostatin receptor 5, **Serpina1**: serine (or cysteine) proteinase inhibitor, clade A, member 1, **Sm22**: Transgelin (Smooth muscle 22 protein), **Sparc**: Secreted acidic cystein-rich glycoprotein (osteonectin), **Ucp2**: Uncoupling protein 2, mitochondrial.

These 24 genes were derived from the group 2 cluster (containing 168 accession numbers of GenBank genes)[c] as follows. Every gene in the cluster was converted into the LocusID or removed if a LocusID could not be found. For each of the 66 genes successfully converted as a result of that process, GO terms were retrieved from the LocusLink database. Finally, we constructed a GO term list $L_{org}(c_k)$ for each gene c_k that has one or more associated GO terms.

[c]The list of genes was acquired from the Kidney Development Gene Expression Database (*http://organogenesis.ucsd.edu/*).

Table 1: Extracted aspects list of top five.

rank	ETMIC score	eigen-attributes	# genes		# different genes				
					$\mathcal{A}$	$\mathcal{B}$	$\mathcal{C}$	$\mathcal{D}$	$\mathcal{E}$
1	8.21	4 6	18	$\mathcal{A}$	0	7	11	8	8
2	8.02	9 13	17	$\mathcal{B}$	7	0	12	11	9
3	7.45	15 18	13	$\mathcal{C}$	11	12	0	15	11
4	7.25	10 13	14	$\mathcal{D}$	8	11	15	0	14
5	6.93	10 11	14	$\mathcal{E}$	8	9	11	14	0

3.2 Experimental setup: Compressing GO term annotation

For the E-SD subprocess (6), we constructed twenty attributes, that is, eigen-attributes, which express the major features of all the GO term annotations in the LocusLink database (see Fig.1). These twenty eigen-attributes A were calculated as follows. Firstly, SVD compressed all the GO term annotations obtained from the LocusLink database into a twenty-dimensional feature space by means of subprocesses (1) and (2). In this process, we used 13,557 ($=n$) effective entries (i.e., entries with GO annotations) among all the 24,489 LocusLink entries (human, mouse and rat) and 5,325 ($=m$) effective GO terms (i.e., GO terms found in the GO annotations of LocusLink entries). Then, to input the E-SD method, each eigen-attribute was digitized to a nominal variable with five domains.

3.3 Experimental results: List of extracted aspects

Table 3.3 summarizes the top five aspects ($\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}, \mathcal{E}$) that the proposed method extracted from the gene list mentioned above.

This table shows that these five aspects have different features. For example, the first aspect $\mathcal{A}$ has 18 genes in a two-dimensional feature space that is spanned by the 4th and 6th axes of the compressed GO annotation, while the second aspect $\mathcal{B}$ shows 17 genes in another feature space spanned by the 9th and 13th axes. They are different not only in the axes spanning the feature space but also in the focused genes, i.e., 7 genes out of 18 and 19 are different, as shown in the right-hand part of this table. This fact indicates a feature of the E-SD process in that it selects genes in such a manner that the selected genes exhibit a relatively simple scattering pattern in the selected feature space. In other words, the selected genes are expected to construct clusters having a relatively simple structure. As a result, it should be easy to interpret each aspect by analyzing the GO annotations associated with some of the clusters of the selected genes.

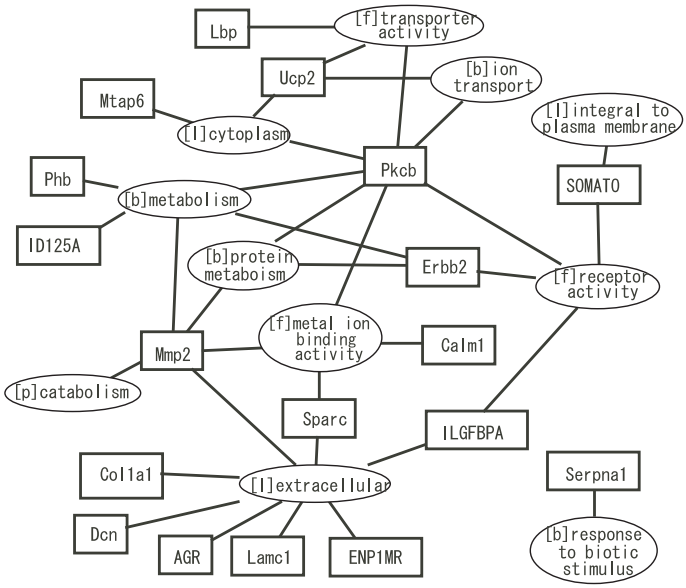

Figure 3: Gene-term bipartite graph for first aspect ($\mathcal{A}$)
Rectangles represent genes. Ovals represent GO terms. [p]=biological_process, [f]=moleculer_function,[l]cellular_componet.eigen-attribute=(4 6) # genes=18.

3.4 Experimental results: Investigation of each aspect

Gene-term bipartite graph for first aspect ($\mathcal{A}$) : As shown in Figure 3, the largest gene cluster in the first aspect $\mathcal{A}$ is *extracellular*. In comparison with the analysis by Stuart et al.[10], the genes of *extracellular* exhibit such a good concordance that they contain all the five representative genes (AGR, Col1a1, Dcn, Sparc, and Mmp2) listed by them Stuart et al. [d] This aspect also presents a contrast between the *extracellular* and *cytoplasm* that are linked through genes related to *metabolism process* and/or *metal ion binding activity* (Pkcb, Mmp2 and Sparc). This could be a good indication for identifying some midnephrogenesis-specific metabolic processes. In addition, it should be noted that gene selection by the E-SD process makes the graph simpler and thus makes the aforementioned focus of the graph clearer. For instance, the graph would be more complicated if the Nkaa1b gene, which connects five GO

[d] Stuart et al.[10] listed ten representative genes of the group 2 cluster. Among them, five genes were used here because no GO terms could be retrieved for the other genes.

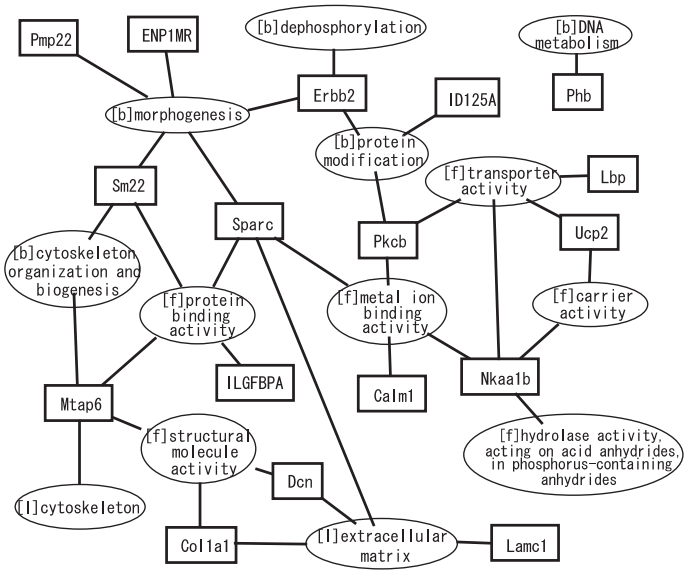

Figure 4: Gene-term bipartite graph for second aspect ($\mathcal{B}$)
Rectangles represent genes. Ovals represent GO terms. [p]=biological_process, [f]=moleculer_function,[l]cellular_componet.eigen-attribute=(9 13) # genes=17.

terms[e], had not been dropped by the E-SD process.

Gene-term bipartite graph for second aspect ($\mathcal{B}$): As shown in Figure 4, the largest gene cluster in the second aspect $\mathcal{B}$ is *morphogenesis*. This aspect also focuses on two cellular components, namely, *cytoskeleton* and *extracellular matrix*. These suggestions for this aspect are similar to the statements made by Stuart et al.[10] in that group 2 was most notable for genes of the extracellular matrix as well as morphogenetic genes.

4 Conclusion

In this paper, we have addressed the need for a method for the multi-aspect analysis of biological data and proposed a novel method for assisting in this kind of analysis. The unique feature of our method is that it automatically

[e] In Fig. 3, five terms (*metabolism*, *catabolism*, *transporter activity*, *metal ion binding activity*, and *integral to plasma membrane*) are associated to Nkaa1b.

extracts multiple aspects for analyzing a gene list by using E-SD (a type of conceptual clustering). We conducted an experiment in which the method was applied to the analysis of rat kidney expression data. In this, our method successfully extracted different analyzing aspects, each of which consisted of relatively few genes and GO terms in a fairly simple structure. This suggests that the analyzing aspects identified by our method can be helpful in examining biological data from a range of viewpoints. For example, the user might mine some interesting viewpoints that he/she had not been aware of previously.

We are currently trying to apply our method to a larger set of genes. A preliminary experiment conducted for about thousand or more genes suggested that our method can extract different analyzing aspects but that extracted aspects tend to consist of too many genes and GO terms (e.g., 500 genes and 30 GO terms) for researchers to quickly find the major characteristics of each aspect. With regard to this point, one of the future issues is to develop a function for summarizing each aspect as well as a function for helping detailed analysis of each aspect. These functions could be realized by using some conventional methods for producing the overall characteristics of a gene list.

References

1. J. Han. How can data mining help bio-data analysis? *Proc. 2nd Workshop on Data Mining in Bioinformatics (BIOKDD02)*, 1–2, 2002.
2. M. Ashburner *et al.* Gene ontology: tool for the unification of biology. *Nature Genet.*, 25:25–29, 2000.
3. P. J. Kennedy *et al.* Extracting and explaining biological knowledge in microarray data. *Proc. PAKDD 2004*, 699–703, 2004.
4. M. Robertson, ed. *Molecular biology of the cell.* Garland, 3 ed., 1994.
5. FUJITSU LIMITED. Biological information mining tool GeneSphere. http://www.fqspl.com.pl/life_science/xminer/GeneSphere.pdf.
6. H. Wagatsuma *et al.* A method of gene expression profiling using Xminer software (in japanese). *Proc. 26th Annual Meetings of the Molecular Biology Society of Japan*, 1PC–146 (154), 2003.
7. H. Yamakawa. Proposing matchability criterion for situation decomposition. *Proc. Int. 1998 Conf. on Neural Information Processing*, 3, 514–517, 1998.
8. H. Yamakawa *et al.* Concept acquisition and reasoning process in a card classification task by situation decomposition using ETMIC criterion. *Cognitive Studies*, 11(2):143–154, 2004.
9. P.W. Lord *et al.* Semantic similarity measures as tools for exploring the gene ontology. *Proc. Pacific Symposium on Biocomputing*, vol. 8, 601–612, 2003.
10. R. O. Stuart *et al.* Changes in global gene expression patterns during development and maturation of the rat kidney. *Proc. Natl, Acad. Sci.*, 98(10):5649–5654, 2001.