Comparing Functional Visualizations of Genes

Hamid Ghous[1] *, Nicholas Ho[2], Daniel R. Catchpoole[2], and Paul J. Kennedy[1]

[1]Centre for Quantum Computation and Intelligent Systems, University of Technology, Sydney, PO Box 123, Broadway NSW 2007, AUSTRALIA
[2] Biospecimens Research and Tumour Bank, Children's Cancer Research Unit, The Kid's Research Institute, The Children's Hospital at Westmead, Locked Bag 4001, Westmead NSW 2145, AUSTRALIA

Abstract. Biological experiments identify large lists of genes and biologists find functional relationships between them to get a better understanding of data. Gene Ontology (GO) is a database of terms related to genes and gene products that has the potential to assist in visualizing and finding functional relationships between genes. We augment genes with GO terms and compare visualizations using two term–to–term similarity measures for terms associated with genes: a hop-based distance measure and an information-content-based similarity measure (IC). Visualization is with Singular Value Decomposition. Relationships are further explained using Pearson correlation with GO terms. Results show that both methods find the relationships between genes however, difference is observed in visualization of GO terms, where IC method shows tightly-packed clusters in contrast to the loose and scattered clusters found with hop based method.

1 Introduction

The widespread use of microarray-based high–throughput technology has produced masses of gene expression data. Interpretation of these large datasets poses a substantial challenge to biologists. Furthermore, gene names do not provide functional information about genes which makes it harder to make sense of gene lists. Consequently researchers enhance data with ontologies such as the Gene Ontology [1] to add functional information to genes which eventually leads to better understanding of large lists of genes.

The Gene Ontology project includes a database of terms related to genes and gene products in several organisms. A gene is associated with zero or more terms that are divided into three subontologies: molecular function, biological process and cellular component which represent biochemical activity, biological objective and the physical location of the gene products respectively. Each term has a unique identifier (GO:#######), name and type. Gene Ontology terms are represented in a hierarchical structure where relationship between terms is either "is–a" or "part–of" representing class–subclass and containment respectively.

* Corresponding author: hamid.ghous@student.uts.edu.au

Previous use of GO in visualization of genes is widespread. We review literature in two main domains: (i) functional–based clustering and visualization of genes using GO and (ii) different similarity measures used for clustering genes with GO. Richards *et al.* [14] developed a framework to find the functional coherence between genes by constructing a graph-based matrix using GO while Huang *et al.* [8] evaluated different tools available for functional analysis of gene lists. They divided them into three categories: singular enrichment analysis, gene set enrichment analysis and modular enrichment analysis.

Different similarity measures have been used by researchers using GO such as Lee *et al.* [11] who exploit the GO hierarchical structure and use directed acyclic graph similarity measures to find clusters of genes. On the other hand, Fröhlich *et al.* [5] used an information-content-based similarity measure to look for common shared parents terms between GO terms, took the probability of each shared parent term in the dataset and applied k-means clustering to identify gene clusters. In Ghous *et al.* [6] we used a hop-based similarity measure and kernel principal component analysis to find similarities between genes.

In this study we apply singular value decomposition to lists of genes augmented with GO terms using two term-to-term similarity measures: information-content-based similarity measure (IC) and hop-based similarity measure and compare the results with respect to their biological interpretation. Further comparison is done with the trustworthiness measure [15] by comparing the faithfulness of the different low dimensional representations to the high dimensional representation. This approach is applied on two datasets: a simulated list of genes from Kyoto Encyclopedia of Genes and Genomes (KEGG) [10] to validate the approach and a dataset derived from biological experiments in childhood leukemia.Gene Ontology database was downloaded on 17-01-2010. The rest of this paper is organized as follows. Section 2 describes two methods of incorporating GO information into gene lists and our approach to visualization. Section 3 describes our datasets. Section 4 presents results and analysis followed by the conclusion in section 5.

2 Methods

We create two matrices $\mathbf{X}$ and $\mathbf{P}$ with $\mathbf{X}$ showing relationships between genes and terms and $\mathbf{P}$ describing the similarity between terms. Let G be a set of genes and T a set of GO terms related to any of the genes in G. The $\mathbf{X}$ matrix is a binary matrix with dimensions $g \times t$ where g is the number of genes ($|G|$) and t is the number of GO terms ($|T|$). If a term relates to a gene the value of x_{ij} is 1 otherwise 0. The element p_{ij} in matrix $\mathbf{P} \in \mathbb{R}^{t \times t}$ represents the similarity between terms i and j. The augmented data matrix $\mathbf{X}'$ is defined as

$$\mathbf{X}' = \mathbf{XP} \tag{1}$$

2.1 Generating the proximity matrix

Two proximity measures between terms are explored in this paper: a hop-based similarity measure and an information-content-based similarity measure.

Hop-based proximity is based on counting the shortest number of hops between two GO terms over the ontology using only 'is-a' links, which are far more frequent than 'part-of' links. Proximity p_{ij} between terms i and j is defined using d_{ij}, the link distance between terms i and j, as

$$p_{ij} = \frac{1}{d_{ij} + 1} \tag{2}$$

Information-content-based proximity [5], on the other hand, uses information content theory [13] to calculate the semantic similarity between GO terms. It is based on the probability of GO terms in the gene dataset $\mathbf{X}$. The information content measure is defined as

$$IC(t) = -\log_2 P(t) \tag{3}$$

where $P(t)$ is the probability of term t in the data matrix and is calculated as $P(t) = \text{freq}(t)/N$ where N is the total number of GO terms in $\mathbf{X}$ and freq(t) is the number of occurrences of t or any of its the child terms. Similarity between terms i and j is defined as

$$p_{ij} = -\log_2 \min_{\hat{t} \in Q_a(i,j)} P(\hat{t}) = -\log_2 P_{ms}(i,j) \tag{4}$$

where $Q_a(i,j)$ is a function returning the set of common shared parent terms between terms i and j and P_{ms}, the probability of the minimum subsumer [12], is the minimum $P(\hat{t})$ if there is more than one parent.

2.2 Singular Value Decomposition

Singular value decomposition (SVD) [7] is a method that factors a data matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$ into matrices $\mathbf{U} \in \mathbb{R}^{n \times r}$, $\mathbf{V} \in \mathbb{R}^{m \times r}$ and the diagonal $\mathbf{D} \in \mathbb{R}^{r \times r}$

$$\mathbf{A} = \mathbf{UDV}^T \tag{5}$$

where $r \leq m$ is the rank of $\mathbf{A}$, n is the number of data points and m the number of terms for all genes. We visualize the data using the first k rows of $\mathbf{U}$ and $\mathbf{V}$ which are associated with genes and terms respectively. SVD is used to transform data matrix $\mathbf{X}'$ after scaling and centering. Visualization quality is measured using trustworthiness [15]. Trustworthy visualizations are those where neighboring points are the same in both the input and low dimensional spaces. Due to space limitation we are unable to describe trustworthiness in detail.

3 Datasets

Two datasets are used in this study: a "simulated" set of genes selected from known classes used for validation (the "KEGG dataset") and a dataset of "real" genes from the childhood leukemia domain (the "cancer dataset").

The KEGG dataset uses genes with known functionality selected from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [10], a database providing functional classification of genes independent of Gene Ontology. We identified genes based on KEGG Orthology terms from five classes: ribosome (ko03010, class 1, 20 genes), RNA polymerase (ko03020, class 2, 19 genes), transcription (ko01210, class 3, 11 genes), pentose phosphate pathway (ko00030, class 4, 11 genes) and pentose and glucoronate interconversions (ko00040, class 5, 7 genes) with 68 genes in total. Classes 1, 2 and 3 are related to genetic processing while classes 4 and 5 are related to the carbohydrate metabolic process.

The cancer dataset relates to genes identified as important in Acute Lymphoblastic Leukemia (ALL). It was constructed based on results from Flotho *et al.* [4] and Catchpoole *et al.* [3]. Flotho and colleagues identified a fourteen gene signature with expression values able to separate a cohort of ALL patients into two groups that agreed with minimal residual disease (MRD) results. Catchpoole and colleagues [3] examined these genes on a different cohort of ALL patients and also discovered a separation of patients but it did not agree with MRD results nor with clinical presentation. Random forest [2] was used to identify other genes that supported the same separation of patients as achieved by Flotho's gene signature. A random forest run of 50,000 trees was performed on a gene expression dataset generated using Affymetrix Human Genome U133-based chips on diagnostic bone marrow samples from ALL patients. The dataset consists of 127 patients and 22,280 probesets. The 250 probesets with the largest mean decrease in Gini index were selected for the "cancer dataset".

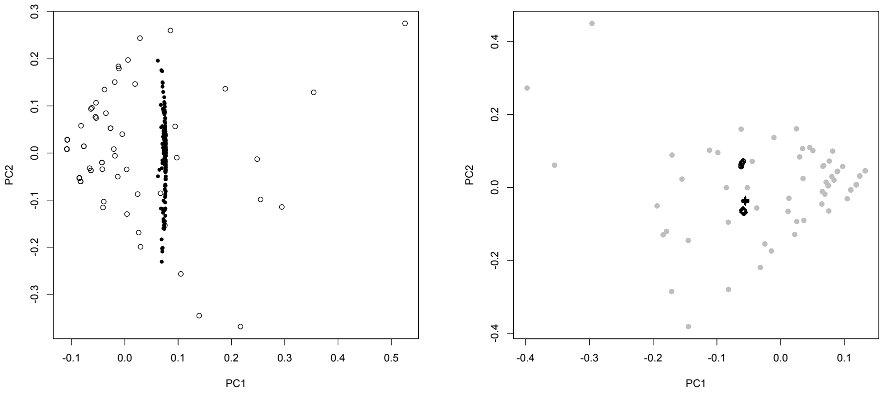
4 Results

4.1 Visualizing KEGG dataset

Results show that principal component 1 (PC1) with both similarity measures is related to the number of GO terms for genes (Fig. 1(a) and Fig. 1(b)) which Jolliffe [9] describes as a "size" component. The high Pearson correlation (0.99) between points projected to PC1 and the number of GO terms associated to genes also confirms this. It is reasonable that the largest amount of variation in dataset is based on the number of the GO terms. Later PCs identify functional relationships between genes. For the hop-based measure, PC2 contrasts genes showing the separation between the genetic information processing genes and the carbohydrate metabolism genes (Fig. 2(a)) while PC2 with the information content (IC) measure separates terms based on the GO subontology (as shown in Fig. 1(b)). The IC method separates genetic information processing and carbohydrate metabolism genes at PC5 (Fig. 2(b)) demonstrating that both methods find the expected functional relationship between genes. Trustworthiness does not show significant differences between the measures (see Fig. 6, top).

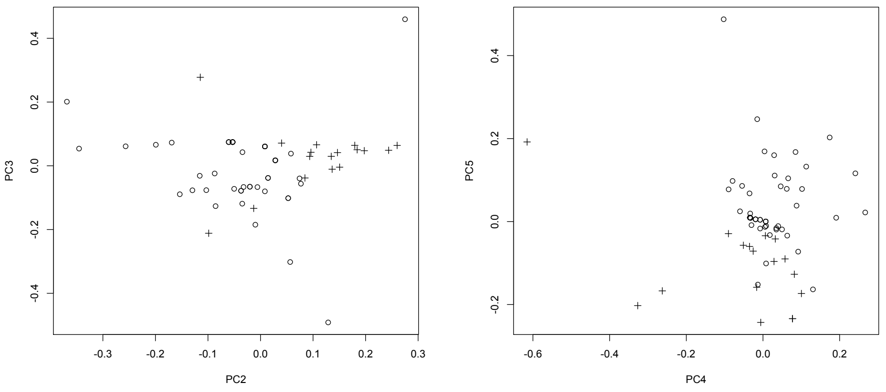
4.2 Visualizing cancer dataset

As with the KEGG dataset, the hop-based and IC approaches both show a strong correlation between the number of GO terms and projected points to



(a) (b)

Fig. 1. Plot for PC1 and PC2 for both methods using KEGG dataset. (a) Hop-based method. Legend: ○ is gene, ● is term. (b) IC similarity measure. Legend: ● is gene, ○ is molecular function GO term, + is biological process GO term and ◇ is cellular component term.



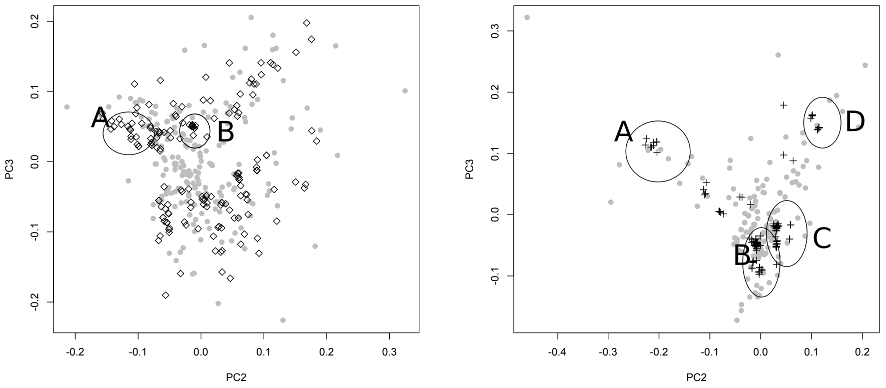
(a) (b)

Fig. 2. (a) Principal component(PC) 2 and 3 for hop-based method and (b) PC4 and 5 for IC method. Legend: (○) is genetic information processing genes and (+) represent carbohydrate metabolism genes.

PC1. Overall, the distributions of GO terms make three clusters, associated with each GO subontology. To untangle the relationships across the subontologies, we applied SVD to the terms from each subontology separately. We examined clusters through correlation and by listing the terms in each cluster.

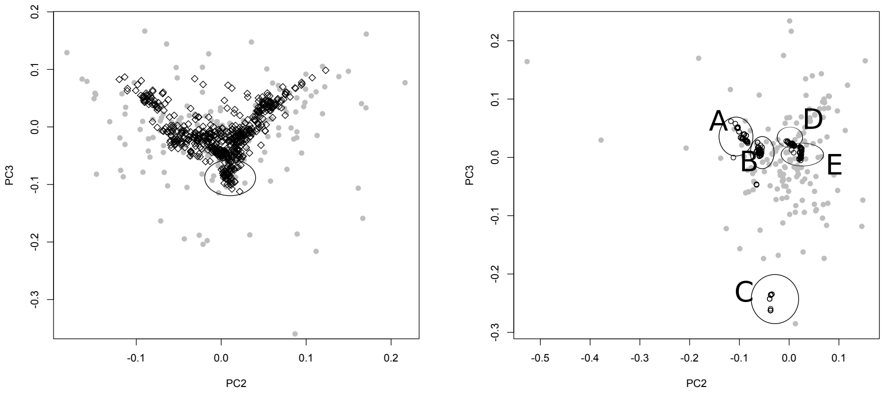
For the Cellular Component GO terms, the hop-based approach revealed a separation between cytoplasmic structure terms and DNA replication terms along PC3 axis as shown in Fig. 3(a). It also highlighted a cluster of terms associated with the membrane on the negative end of PC2 and a cluster of tubulin and kinesin GO terms towards the positive end of PC3 (see Fig. 3(a) clusters A and B respectively). PC2 in the IC approach reveals four small distinct clusters: a cluster of membrane and extracellular-matrix-related terms, a cluster of terms associated to organelles, protein-complex-related terms and a cluster of cell-division-apparatus-related terms as shown in Fig. 3(b) as clusters A, B, C and D respectively. Some of the terms in these clusters are listed in Table 1.

For the Biological Process GO terms PC3 in the hop-based approach reveals a cluster of terms associated with development (e.g. embryonic development, notochord development, forebrain development, embryonic axis specification) as shown in Fig. 4(a). PC2 in the IC approach identifies five tight clusters (shown in Fig. 4(b)): cluster A relates to morphogenesis and early development, cluster B to homeostasis and response to stimulus (within which is a subgroup related to molecular transport in the cell), cluster C relates to gene expression regulation and metabolism, cluster D to differentiation and cluster E to DNA metabolism and function along with a number of small subgroups e.g vesicle transport. As before, some of the terms found in these clusters are listed in Table 1.



(a) (b)

Fig. 3. Plot of principal components 2 and 3 of cancer dataset with cellular component (CC) terms. (a) Hop-based similarity measure. Legend: (•) is genes and (◇) is CC terms.(b) IC similarity measure. Legend: (•) is genes and (+) is CC terms.



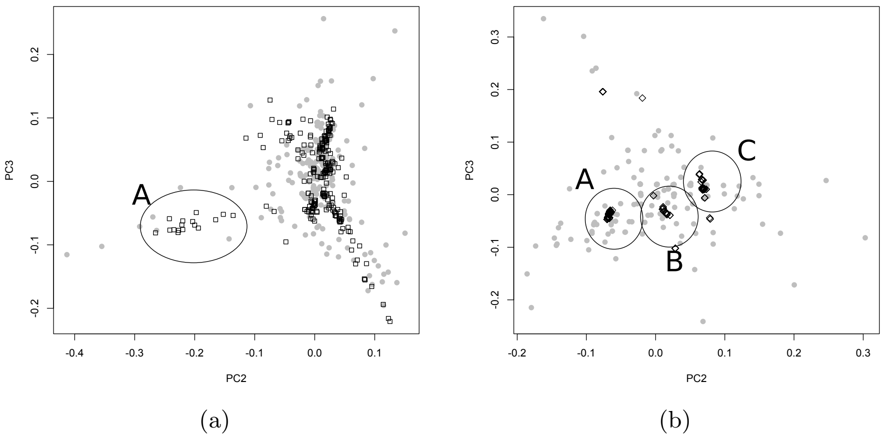
(a) (b)

Fig. 4. plot of principal components 2 and 3 of cancer dataset with biological process(BP) terms. (a) Hop based similarity measure. Legend: (●) is genes and (◇) is BP terms (b) IC similarity measure. Legend: (●) is genes and (○) is BP terms.

For the Molecular Function terms, PC2 in the hop-based approach identifies a cluster of terms associated with DNA helicase activity. In close proximity to this cluster is a loosely packed cluster of six genes that code for mini chromosome maintenance proteins (MCM2, MCM3, MCM4, MCM5, MCM6 and MCM7) as shown in Fig. 5(a). Both MCM proteins and replicative helicase play integral roles in eukaryotic DNA replication. The IC approach also identified this loose cluster of MCM genes and the GO term for DNA helicase activity. Across PC2, the rest of the clusters relate to enzyme activity No. 1, enzyme activity No. 2 and non-enzymatic molecular interactions as shown in Fig.5(b) as A, B and C respectively. Some terms from these clusters are in Table 1. Again, trustworthiness does not show significant differences between either similarity measure on the cancer dataset (see Fig. 6, bottom).

5 Conclusion

This study compares singular value decomposition visualizations of genes and Gene Ontology terms using two inter–term similarity measures: a hop-based measure and an information-content-based measure. Two datasets were investigated: a list of selected genes with known functionality from the KEGG database and a dataset of "real" genes from the childhood leukemia domain. Results show that the first PC for both methods visualize genes based on the number of terms associated with them but that later PCs visualize genes by their functionality. The information-content-based method also clustered GO terms based on the subontology from which they arose. Both methods identified clusters of GO



(a)

(b)

Fig. 5. Plot of principal components 2 and 3 of cancer dataset with molecular function terms. Legend: (●) is genes and (◇) is molecular function terms. (a) Hop based similarity measure (b) IC similarity measure.

terms related to functionality but the IC method resulted in more tightly–packed clusters of terms and genes than the hop–based method. Trustworthiness showed negligible difference in the quality of projections based on nearest neighbors in the high and low dimensional spaces. We recognise that the PCs are challenging for biologists to understand and plan to address this in future work. Biologists will also be involved in validation using biological experiments. In future we also plan to apply clustering methods to better describe the resulting visualizations. We plan also to explore other term–to–term similarity measures such as set–based and vector–based approaches. Finally, we will compare the results of our method with other commonly used gene set enrichment tools.

Acknowledgments. This study was supported by a UTS Challenge Grant and a NSW Cancer Institute Research Innovation Grant.

References

1. Ashburner, M., Ball, C.A., A.Blake, J., et al.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25(1), 25–29 (2000)
2. Breiman, L.: Random forests. Machine learning 45(1), 5–32 (2001)
3. Catchpoole, D., Guo, D., Jiang, H., Biesheuvel, C.: Predicting outcome in childhood acute lymphoblastic leukemia using gene expression profiling: Prognostication or protocol selection? Blood 111(4), 2486–2487 (2008)
4. Flotho, C., Coustan-Smith, E., Pei, D., et al.: A set of genes that regulate cell proliferation predicts treatment outcome in childhood acute lymphoblastic leukemia. Blood 110(4), 1271–1277 (2007)

Table 1. GO term clusters using IC method for Cellular Components(CC), Biological Process(BP) and Molecular Function(MF) of GO based on correlation results.

Clusters	Example Terms	Description
CC Terms		
A	GO:0042175/nuclear envelope-endoplasmic reticulum network, GO:0005887/integral to plasma membrane	Cluster of membrane and extracellular matrix
B	GO:0005635/nuclear envelope, GO:0030117/membrane coat and GO:0000324/fungal-type vacuole	Organelles
C	GO:0042719/mitochondrial inter-membrane space protein transporter complex, GO:0005760/gamma DNA polymerase complex and GO:0031588/AMP-activated protein kinase complex	Protein complexes
D	GO:0044430/cytoskeletal part, GO:0031616/spindle pole centrosome and GO:0000922/spindle pole	Cell division apparatus
BP Terms		
A	GO:0001658/branching involved in ureteric bud morphogenesis, GO:0048754/branching morphogenesis of a tube and GO:0001947 heart looping	Morphogenesis and Early Development (Stem Cells)
B	GO:0006974/response to DNA damage stimulus, GO:0007548/sex differentiation and GO:0007276/gamete generation	Response to Stimulus Transport or Homeostasis
C	GO:0010468/regulation of gene expression GO:0010628/positive regulation of gene expression and GO:0005975/carbohydrate metabolic process	Gene expression regulation and metabolism
D	GO:0048676/axon extension involved in development, GO:0045467/R7 cell development and GO:0007409/axonogenesis	Differentiation
E	GO:0000718/nucleotide-excision repair, DNA damage removal, GO:0000720/pyrimidine dimer repair by nucleotide-excision repair, GO:0000724/double strand break repair via homologous recombination	DNA metabolism and function with a number of small subgroups e.g. vesicle transport
MF Terms		
A	GO:0003678/DNA helicase activity, GO:0004003/ATP-dependent DNA helicase activity and GO:0008026/ATP-dependent helicase activity	Enzyme activity No.1
B	GO:0003777/microtubule motor activity, GO:0003774/motor activity and GO:0003924/GTPase activity	Enzyme activity No.2
C	GO:0016853/isomerase activity, GO:0003689/DNA clamp loader activity and GO:0003916/DNA topoisomerase activity	Molecular interactions non-enzymatic

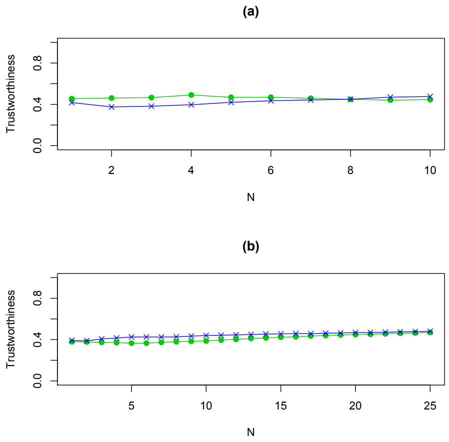


Fig. 6. (a)Trustworthiness plot of KEGG dataset and (b) is trustworthiness plot of cancer dataset. (•) line is for hop based method and (×) line represent IC method where N is the number of neighbors.

5. Fröhlich, H., Speer, N., Spieth, C., Zell, A.: Kernel based functional gene grouping. Int. Joint Conference on Neural Networks, 2006. IJCNN'06 pp. 3580–3585 (2006)
6. Ghous, H., Kennedy, P.J., Catchpoole, D.R., Simoff, S.J.: Kernel-based visualisation of genes with the Gene Ontology. In: Data Mining and Analytics 2008: Proc. of the 7th Australasian Data Mining Conf. Conferences in Research and Practice in IT (CRPIT), vol. 87, pp. 133–140. Australian Computer Society, Sydney (2008)
7. Golub, G., Van Loan, C.: Matrix computations. Johns Hopkins University Press (1996)
8. Huang, D., Sherman, B., Lempicki, R.: Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic acids research 37(1), 1–13 (2009)
9. Jolliffe, I.T.: Principal Component Analysis. Springer, New York, 2nd edn. (2004)
10. Kanehisa, M., Araki, M., Goto, S., et al.: KEGG for linking genomes to life and the environment. Nucleic Acids Research 36, 480–484 (2008)
11. Lee, S., Hur, J., Kim, Y.: A graph-theoretic modeling on GO space for biological interpretation of gene clusters. Bioinformatics 20(3), 381–388 (2004)
12. Lord, P., Stevens, R., Brass, A., Goble, C.: Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics 19(10), 1275–1283 (2003)
13. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: Int. Joint Conference on Artificial Intelligence. pp. 448–453 (1995)
14. Richards, A.J., Muller, B., Shotwell, M., Cowart, L.A., Rohrer, B., Lu, X.: Assessing the functional coherence of gene sets with metrics based on the Gene Ontology graph. Bioinformatics 26(12), i79–i87 (2010)
15. Venna, J., Kaski, S.: Comparison of visualization methods for an atlas of gene expression data sets. Information Visualization 6(2), 139–154 (2007)