Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2016, Article ID 5174503, 10 pages
http://dx.doi.org/10.1155/2016/5174503

Research Article

Semantic Signature: Comparative Interpretation of Gene Expression on a Semantic Space

Jihun Kim,[1,2] Keewon Kim,[1,3,4] and Ju Han Kim[1,5]

[1]*Seoul National University Biomedical Informatics (SNUBI), Seoul 110-799, Republic of Korea*
[2]*LabGenomics Clinical Research Institute, LabGenomics, Seongnam 463-400, Republic of Korea*
[3]*Department of Rehabilitation Medicine, Seoul National University College of Medicine, Seoul 110-799, Republic of Korea*
[4]*Departments of Biomedical Engineering, Seoul National University College of Medicine, Seoul 110-799, Republic of Korea*
[5]*Systems Biomedical Informatics Research Center, Division of Biomedical Informatics, Seoul National University College of Medicine, Seoul 110-799, Republic of Korea*

Correspondence should be addressed to Ju Han Kim; juhan@snu.ac.kr

Received 17 December 2015; Accepted 23 March 2016

Academic Editor: Seiya Imoto

Copyright © 2016 Jihun Kim et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Background. Interpretation of microarray data remains challenging because biological meaning should be extracted from enormous numeric matrices and be presented explicitly. Moreover, huge public repositories of microarray dataset are ready to be exploited for comparative analysis. This study aimed to provide a platform where essential implication of a microarray experiment could be visually expressed and various microarray datasets could be intuitively compared. *Results*. On the semantic space, gene sets from Molecular Signature Database (MSigDB) were plotted as landmarks and their relative distances were calculated by Lin's semantic similarity measure. By formal concept analysis, a microarray dataset was transformed into a concept lattice with gene clusters as objects and Gene Ontology terms as attributes. Concepts of a lattice were located on the semantic space reflecting semantic distance from landmarks and edges between concepts were drawn; consequently, a specific geographic pattern could be observed from a microarray dataset. We termed a distinctive geography shared by microarray datasets of the same category as "semantic signature." *Conclusions*. "Semantic space," a map of biological entities, could serve as a universal platform for comparative microarray analysis. When microarray data were displayed on the semantic space as concept lattices, "semantic signature," characteristic geography for a microarray experiment, could be discovered.

1. Background

Microarray experiments provide high-throughput gene expression profiles to address biological questions for a specific condition. It has been challenging so far to extract biological implication from huge matrices of numeric data and to represent it in a concise and intuitive manner. Also, as microarray datasets accumulate in public repositories, it has become another important issue in microarray analysis how to compare multiple datasets and integrate them [1].

As for the first issue (analysis of an individual microarray experiment), semantic approach was suggested as one of main strategies. For example, clusters of genes with similar expression profiles could be assigned with Gene Ontology (GO) terms or related pathways [2]. However, most of clustering analyses with ontological annotation merely provided long list of biological terms for clusters; they failed to represent functional relationship between clusters and the whole picture of microarray experiments could hardly be grasped. Concept lattice analysis, or formal concept analysis, was proposed as a way to summarize biological information from clusters without annotation redundancy [3]. Concept lattice analysis is a mathematical technique that recognizes hierarchical structure from a relation matrix of objects (clusters) and attributes (annotations) and represents it as a graph (lattice). In this way, clusters are depicted as nodes and set-inclusion relationships of their annotations are drawn as edges in a concept lattice, which can be viewed as an executive summary of the microarray data.

The second issue (comparative microarray analysis) is more elusive. External datasets of various conditions can be compared to interpret one microarray experiment of interest [4–9]. Or multiple experiments can be gathered and analyzed to deduce causality or association between genes [10–20]. Such comparative analyses can reduce noises that individual experiments might contain and can help with arranging vast datasets to reveal novel relationship between phenotypes or gene expressions. However, comparisons are inevitably dependent on reference data being compared and platforms on which comparisons are made. In particular, scope of reference datasets can be restricted if the platform of comparative analysis is not compatible with diverse microarray experiments.

The current study was motivated to provide a universal, not being influenced by formats of data and platform where microarray data could be visually presented and compared. Furthermore, we hoped it could convey biological implication of experiments. For that purpose, we constructed "semantic space," a map of biological entities. Just as a map encompasses the whole territory of interest, the semantic space employed gene sets of Molecular Signature Database (MSigDB) as landmarks. MSigDB was a collection of gene sets from various sources (positional, curated, motifs, computational, and GO gene sets) and could be regarded as a representation of the biological world at this time point [6]. The coordinates of those landmarks were determined by semantic distances between them based on a predefined semantic similarity measure [21].

On the semantic space, microarray data were mapped as a concept lattice. A concept lattices was produced by formal concept analysis, with clusters of similar expression profiles as objects and GO annotations as attributes. Each cluster was located on the semantic space considering its semantic distance from nearest landmarks and edges of the lattice between clusters were also drawn.

A concept lattice described on the semantic space makes a distinctive topography on the semantic space, termed as a "semantic signature." Semantic signature would give information of how gene clusters of a microarray dataset were related to biological landmarks at a glance. Furthermore, we compared various microarray datasets by simply overlapping their semantic signatures on the semantic space. And we tested whether data of similar experiment paradigm resulted in similar semantic signatures and whether those of different paradigm did different semantic signatures. Figure 1 illustrates the brief process for construction of semantic space and plotting of semantic signature.

2. Methods

2.1. Construction of Semantic Space. Gene sets of MSigDB (http://www.broad.mit.edu/gsea/msigdb) imported to generated landmarks of semantic space. MSigDB collected gene sets from various sources and organized them into 5 categories: positional, curated, motifs, computational, and GO gene sets [6]. In this study, we adopted all the gene sets from mouse species, registered in GO biological process: 192 gene sets in total.

Semantic distances among the gene sets were determined as follows. First, we annotated the gene sets significant GO terms; statistical significance was determined by hypergeometric probability with Bonferroni correction [3]. Then, distance (sim) between two gene sets (P_i, P_j) was calculated by best-match-average (bma) of term-to-term distances from each gene set [34]:

$$\underset{\mathrm{bma}(P_i,P_j)}{\mathrm{sim}} = \frac{\mathrm{avg}_{t1\in P_i}\left(\max_{t2\in P_j}\left(\mathrm{sim}_{\mathrm{lin}(t1,t2)}\right)\right) + \mathrm{avg}_{t2\in P_j}\left(\max_{t1\in P_i}\left(\mathrm{sim}_{\mathrm{lin}(t1,t2)}\right)\right)}{2}. \tag{1}$$

Distance between two terms $(t1, t2)$ was computed using Lin's semantic similarity measure [35]. Lin's measure quantified "information content" of a term, by enumerating frequency of the genes that were annotated with the term or its descendant terms:

$$\mathrm{freq}\,(t) = \mathrm{genes}\,(t) + \sum_{x\in \mathrm{descendant}(t)} \mathrm{freq}\,(x)\,. \tag{2}$$

Lin's semantic similarity measure (sim_lin) was then determined as follows:

$$\underset{\mathrm{lin}(t1,t2)}{\mathrm{sim}} = \max_{t\in S(t1,t2)}\left(\frac{2\cdot \log p\,(t)}{\log p\,(t1) + \log p\,(t2)}\right), \quad \text{where } p\,(t) = \frac{\mathrm{freq}\,(t)}{\mathrm{freq}\,(\mathrm{root})}. \tag{3}$$

All pair-wise distances of the gene sets were computed and summarized in a matrix. Afterwards, multidimensional scaling was applied to generate 2D coordinates of the gene sets as landmarks in semantic space (cmdscale package, R, https://www.r-project.org/). Finally, the landmarks were plotted in a plane using scalable vector graphics and finally the semantic space was constructed.

2.2. Concept Lattices from Microarray Datasets. From one microarray dataset, one concept lattice was generated as follows. First, k-means clustering analysis produced 100 clusters for a dataset. Then, clusters were annotated with significant GO terms based on hypergeometric probability with Bonferroni correction. Third, a relation matrix of clusters and annotations was built. Fourth, a formal concept analysis transformed the relation matrix into a concept lattice with clusters as objects (extent) and GO terms as attributes (intent). For graphic representation, this study adopted Ganter's algorithm [36]. Accordingly, a concept was a set of clusters sharing GO terms. Edges between concepts

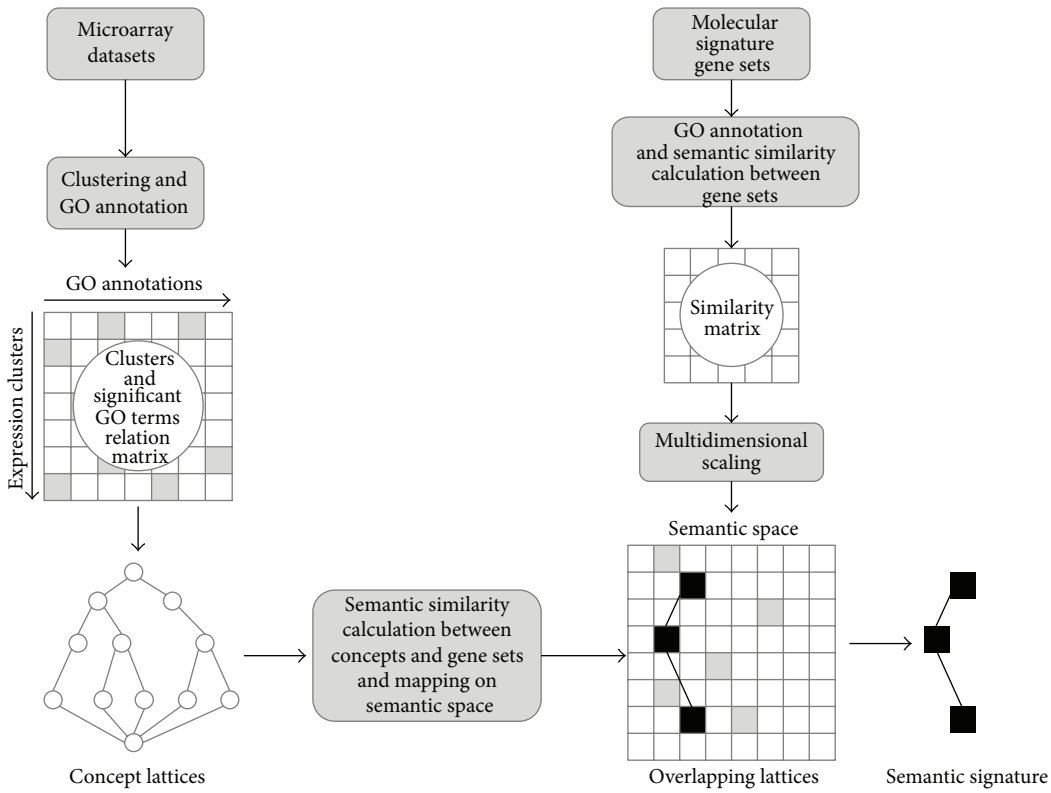


Figure 1: Schematic diagram of the study. Microarray experiments are transformed into concept lattices as illustrated on the left panel of the figure. Semantic space is constructed based on semantic similarities between gene sets as illustrated on the right panel. Concept lattices are then superimposed on the semantic space to reveal a semantic signature and shared geography by homogeneous experiments.

in the lattice indicated set-inclusion relationship. Therefore, the graph structure of a concept lattice could convey the whole information of a relation matrix of clusters and GO annotations. (http://www.snubi.org/software/biolattice/.)

2.3. Discovery of Semantic Signature. Once semantic space was constructed and a concept lattice of microarray experiments was generated, the concept lattice could be mapped on the semantic space. Because extents of a concept lattice were clusters, or sets of genes, semantic distance from the concept to any landmark on the semantic space could be calculated just as distances between landmark gene sets were calculated (see Figure 2). To appropriately place the concept on the semantic space, 3 nearest landmarks were found for each concept. The coordinate of the concept was then determined within the triangle of the 3 nearest landmarks according to relative distances to the 3 vertices. In this way, all the concepts of a concept lattice could be located on the semantic space. Edges between concepts were drawn as well. Finally, the whole concept lattice of microarray experiments was depicted on the semantic space and unique geographic feature could be observed (Figure 2).

In order to compare multiple microarray datasets, concept lattices of them were mapped on the semantic space simultaneously. A common geographic pattern of homogeneous microarray datasets on the semantic space, termed as "semantic signature," was derived by investigating overlapping or closely neighbouring concepts and edges. To visualize the semantic signature, overlapped edges were emphasized by increasing colour intensity according to the overlapping frequency.

2.4. Microarray Datasets. Microarray data of hepatotoxic agent experiments was obtained from a toxicogenomics study by Toxicogenomics Research Center (TGRC). The study applied 12 toxic agents to mice orally or intraperitoneally and observed gene expression profiles from liver specimen according to time course and dosage. Twelve toxic agents were D-galactosamine, ethanol, tetracycline, valproic acid, methotrexate, ANIT, methylenedianiline, phenytoin, thiabendazole, 6-mercaptopurine, phenylbutazone, and diclofenac.

Twenty microarray datasets were downloaded from GEO. Datasets were selected among mouse (*Mus musculus*) datasets if they included sufficient number of conditions (8 or more) for clustering analysis and experimental condition and tissues were explicitly described. The datasets were categorized per condition and tissue: toxin-related (GDS322, GDS2043), development-related (half of GDS2577, GDS2227, GDS2398, GDS2521, GDS1695, GDS568, GDS2202, GDS2203, GDS2098, and GDS2743), and cancer-related (half of GDS2577, GDS1110, GDS604, GDS2640, and GDS2554) conditions; neural (GDS2227, GDS1110, GDS604, GDS887, GDS2850, and GDS2159), hematopoietic (GDS322, GDS2398, GDS2521, GDS1695, GDS568, GDS2640, and

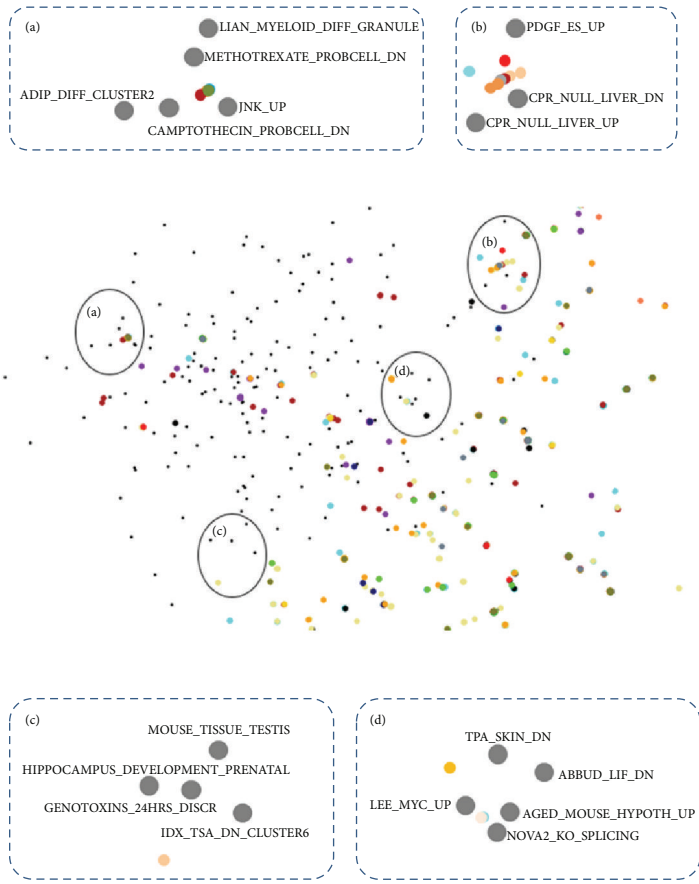


Figure 2: The semantic space. (a) The whole semantic space. (b) Magnified view of specific regions of the semantic space. On magnified views, title of each landmark gene set is presented near it. (a) *LIAN_MYELOID_DIFF_GRANULE*: granule constituents expressed during mouse promyelocytic cell line cell differentiation; *METHOTREXATE_PROBCELL_DN*: downregulated in pro B cells FL5 12 following treatment with methotrexate; *JNK_UP*: upregulated by expression of constitutively active JNK in 3T3 cells; *CAMPTOTHECIN_PROBCELL_DN*: downregulated in pro B cells FL5 12 following treatment with camptothecin; *ADIP_DIFF_CLUSTER2*: strongly upregulated at 2 hours during differentiation of 3T3 L1 fibroblasts into adipocytes cluster 2. (b) *PDGF_ES_UP*: upregulated by PDGF in mouse embryonic stem cells via microarray coupled gene trap mutagenesis; *CPR_NULL_LIVER_DN*: downregulated in mouse liver tissue from mice in which NADPH cytochrome P450 reductase CPR was specifically deleted in the liver by cre lox recombination versus lox only controls; *CPR_NULL_LIVER_UP*: upregulated in mouse liver tissue from mice in which NADPH cytochrome P450 reductase CPR was specifically deleted in the liver by cre lox recombination versus lox only controls. (c) *MOUSE_TISSUE_TESTIS*: genes expressed specifically in mouse testis tissue; *HIPPOCAMPUS_DEVELOPMENT_PRENATAL*: highly expressed in prenatal mouse hippocampus cluster 1; *GENOTOXINS_24HRS_DISCR*: group of genes whose regulation pattern significantly discriminates between direct cisplatin methyl methanesulfonate mitomycin C and indirect taxol hydroxyurea etoposide genotoxins 24 hours following treatment of mouse lymphocytes TK 3 7 2C; *IDX_TSA_DN_CLUSTER6*: strongly downregulated at 2 hours during differentiation of 3T3 L1 fibroblasts into adipocytes with IDX insulin dexamethasone and isobutyl xanthine versus fibroblasts treated with IDX TSA to prevent differentiation cluster 6. (d) *TPA_SKIN_DN*: downregulated in murine dorsal skin cells 6 hours after treatment with the phorbol ester carcinogen TPA; *ABBUD_LIF_DN*: genes downregulated by LIF treatment 10 ng mL overnight in AtT20 cells; *LEE_MYC_UP*: genes upregulated in hepatoma tissue of Myc transgenic mice; *AGED_MOUSE_HYPOTH_UP*: upregulated in the hypothalamus of BALB c mice aged 22 months compared to young 2-month controls; *NOVA2_KO_SPLICING*: genes that are alternatively spliced in the neocortex of mice deficient in the neuron specific splicing factor Nova2 compared to wild type controls.

GDS2554), and germinal (GDS2043, GDS2202, GDS2203, and GDS2098) tissues.

3. Results

3.1. Semantic Space. "Semantic space" is shown in Figure 2. 192 gene sets from MSigDB are marked on it as its landmarks. Figure 2 shows the whole picture of the semantic space. Each vertex indicates one gene set. In this study, the semantic space was constructed using scalable vector graphics (SVG) so that the map could be magnified without being blurred. Magnified views of specific regions are illustrated in Figure 2, (a)~(d). Given that locations of those gene sets were determined according to semantic distances between them, gene sets sharing similar GO annotations congregated in close vicinity. For example, most of gene sets in the upper right circle of Figure 2(b) were related mostly to adipose or secretary cell development.

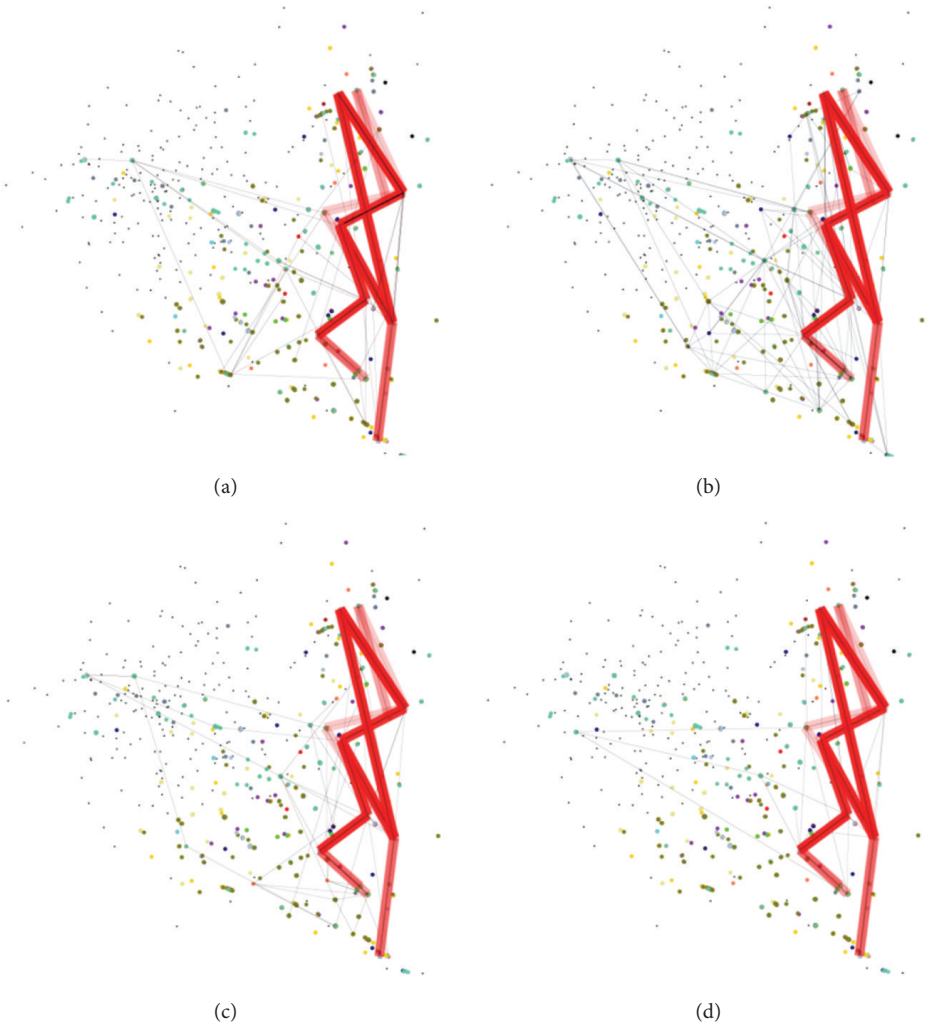


Figure 3: Semantic signature of hepatotoxic agent experiments. Concept lattice of 4 exemplary microarray experiments are represented on the semantic space: (a) tetracycline, (b) ethanol, (c) methylenedianiline, and (d) phenytoin. Shared edges are expressed in bold red lines with colour intensity proportional to overlapping frequency. The common features denote the semantic signature of hepatotoxic agent experiments.

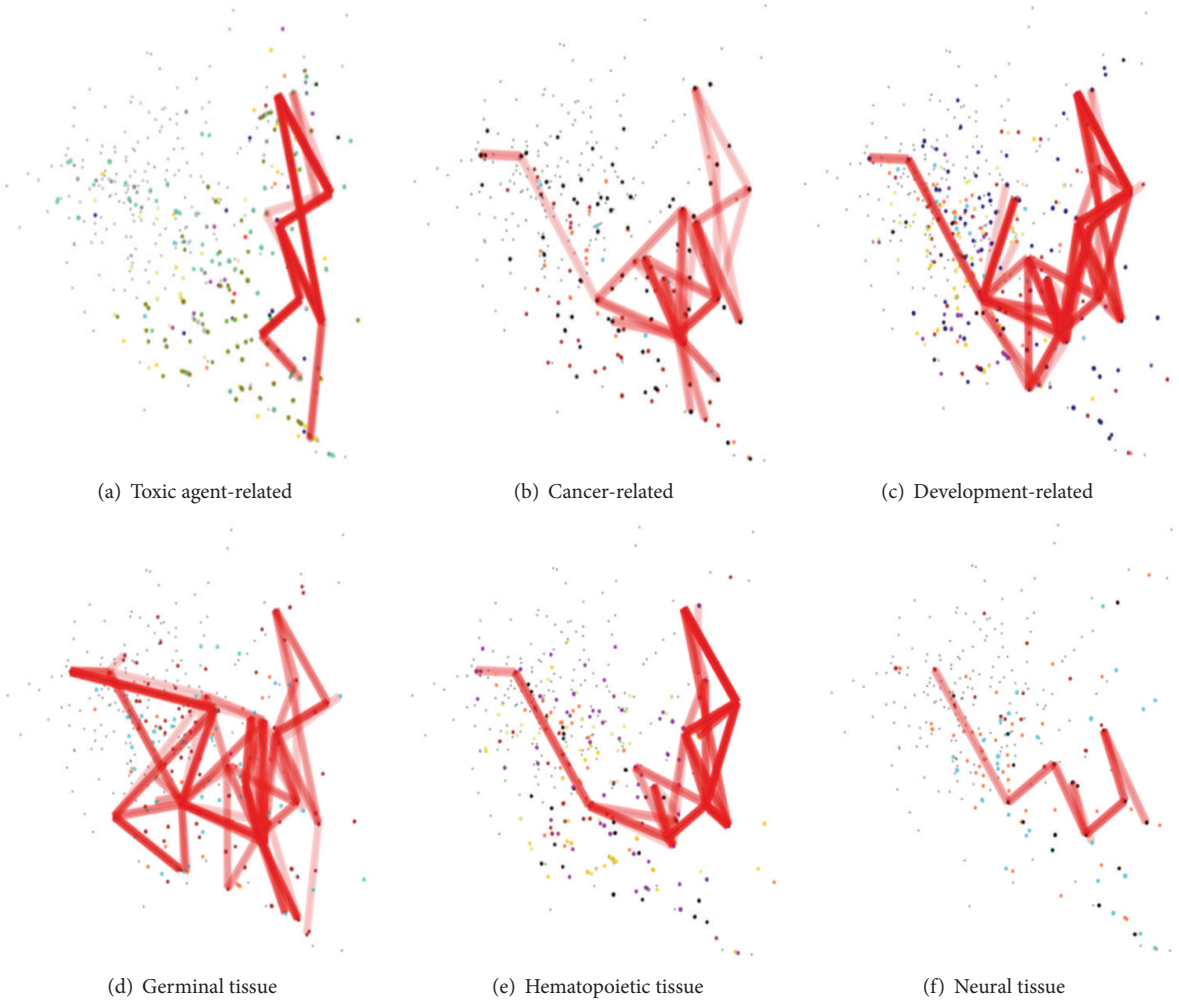
3.2. Semantic Signature: An Example of Hepatotoxic Agent Experiment. As an exemplary case, we assessed the semantic signature of a microarray dataset from hepatotoxic agent experiment (see Section 2). The experiment observed acute hepatotoxic effect of 12 toxicants in mouse after oral or intraperitoneal injection and produced 12 microarray data. Expression profile from each agent was converted into a concept lattice and then mapped on the semantic space. Obtained figures of 12 concept lattices were not identical but their concepts and edges were overlapping or closely neighbouring on the semantic space. Those common edges were visually emphasized by grading the colour intensity in proportion to overlapping frequency (Figure 2). The common concepts and edges were clustered along the right border of the semantic space making figure of a saw tooth, which was then determined as "semantic signature" of the hepatotoxicity experiment.

3.3. Semantic Signatures: Comparison of Heterogeneous Experiments. Other various datasets from different experimental conditions or from different tissues were represented on the semantic space. Twenty heterogeneous microarray datasets were obtained from GEO (Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/geo/) and they were categorized into 3 conditions (toxin-related, cancer-related, and development-related experiments) and 3 tissues (germinal, hematopoietic, and neural tissues) based on descriptions provided by GEO. To discover semantic signatures per condition or tissue, common concepts and edges were depicted in the same way described above (Figure 3). The obtained semantic signatures exhibited distinctive patterns for each experiment category. Each SVG is available in

http://www.snubi.org/software/biolattice/bioclass/canceroverlap.htm;

http://www.snubi.org/software/biolattice/bioclass/devoverlap.htm;

http://www.snubi.org/software/biolattice/bioclass/germoverlap.htm;



(a) Toxic agent-related (b) Cancer-related (c) Development-related

(d) Germinal tissue (e) Hematopoietic tissue (f) Neural tissue

FIGURE 4: Semantic signatures of heterogeneous experiments. Various microarray datasets are categorized per condition or tissue and their semantic signatures are presented in red lines. The conditions are (a) toxic agent-related, (b) cancer-related, and (c) development-related experiments. The specimens are from (d) germinal, (e) hematopoietic, and (f) neural tissues.

http://www.snubi.org/software/biolattice/bioclass/hemaoverlap.htm;

http://www.snubi.org/software/biolattice/bioclass/neuroverlap.htm;

http://www.snubi.org/software/biolattice/bioclass/toxoverlap.htm.

3.4. Validation of Semantic Signature: Semantic Distance among Concept Lattices. As shown in Figures 4 and 5, it might be asserted that geographic patterns from homogeneous experiments looked alike and those from heterogeneous experiments looked different in the semantic space. However, such visual judgement of similarity remained somewhat subjective and arbitrary without quantitative validation. To circumvent the problem, we tested whether concept lattices of the same experiment category were closer than those of different category, based on semantic similarity among concept lattices. The calculation of the semantic distances between two lattices followed the same strategy that was used for construction of the semantic space (Figure 4). Concept lattices of the same experiment conditions were closely located (Figure 4(a)). And the lattices of development-related experiments were closer to the lattices of cancer-related experiments than to those of toxic agent-related ones as was indicated by topographic similarity between their semantic signatures. Also, concept lattices of the same tissue were located in closer vicinity (Figure 4(b)).

4. Discussion

4.1. Biological Interpretation of Semantic Signatures I. For biological interpretation of the semantic signature from the hepatotoxic agent experiments, 8 most frequently neighboured landmark gene sets in the signature were listed (Table 1). One of the landmarks was a gene set that was upregulated by transcription factor Hxc-8, which was known to interact with hematopoietic activities in the liver tissue, suggesting that reactive hematopoiesis is induced by the hepatotoxic agents (LEI_HOXC8_UP gene set) [22]. Another

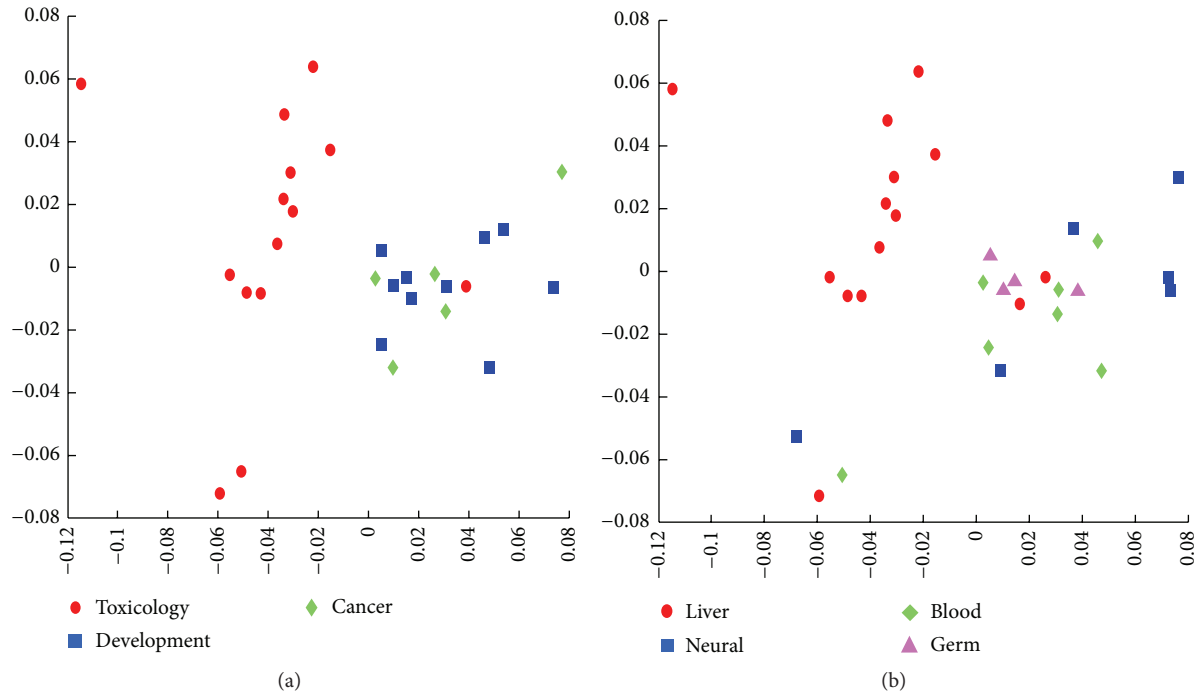


Figure 5: Semantic distance among concept lattices. Concept lattices are plotted on 2D space according to semantic distance between them. Concept lattices of the same (a) experimental condition or (b) tissues are located closely.

Table 1: Landmarks within semantic signature of hepatotoxic agent experiment. Eight most frequently overlapped gene sets among concepts lattices of hepatotoxic agent experiment are listed.

Title of gene set	Description
YEN MYC MUT′	Genes upregulated in mutant MYC mouse model
IGFR IR UP′	Upregulated in common following stimulation of chimeric TrkC IR or TrkC IGFR in NIH3T3 cells
3AB GAMMA DN′	Downregulated synergistically by gamma irradiation and 3-aminobenzamide PARP inhibitor
LEI HOXC8 UP	Upregulated target genes of murine transcription factor Hoxc-8
CHESLER D6MIT150 CIS GLOCUS′	Cis regulatory quantitative trait loci QTLs found at the D6Mit150 region QTLs detected in brain tissue
ROS MOUSE AORTA UP′	Upregulated in mouse aorta by chronic treatment with PPAR-gamma agonist rosiglitazone
YEN MYC WT′	Genes upregulated in wild type MYC mouse model
TNFALPHA ADIP UP′	Upregulated in mature differentiated adipocytes following treatment with TNFalpha

landmark gene set was YEN MYC MUT which was generated from myc mutation mouse; phenobarbital (a hepatotoxic agent included in this experiment) was known to induce apoptosis and carcinogenesis of hepatocytes via c-myc expression [23]. Other landmarks included a gene set which was associated with inflammatory change induced by TNF (TNFALPHA_ADIP_UP) [24] and gene sets related to apoptotic of inflammatory signalling by insulin receptor (ROS_MOUSE_AORTA_UP and IGFR_IR_UP) [25]. While the semantic signatures of hepatotoxic agent experiments shared common landmarks with each other as described above (thick red line in Figure 3), each concept lattice from different hepatotoxic agent also exhibited different patterns (thin red line in Figure 3). For example, concept lattice from ethanol treatment showed more nodes and edges than that from phenytoin, which could be interpreted as ethanol resulting in perturbation in more gene clusters than phenytoin. The finding was consistent with general biological knowledge that ethanol is more toxic than phenytoin to the liver.

4.2. Biological Interpretation of Semantic Signatures II. When geography of the semantic signatures of various microarray datasets was compared, the semantic signature from cancer-related experiments was more similar to that from development-related experiments than to that from toxic agent-related ones. The finding was consistent with an established biological knowledge that one of main oncogenic mechanisms was related to uninhibited activation of

developmental pathways. Specifically, the semantic signature of development-related dataset was closely neighboured by landmarks that included differentially expressed genes during cell differentiation (LIAN_MYELOID_DIFF_GRANULE, HIPPOCAMPUS_DEVELOPMENT_NEONATAL, and PARK_HSC_VS_MPP_UP) and landmarks that were down-regulated genes by immune-modifying or antineoplastic drugs (METHOTREXATE_PROBCELL_DN and CANCERDRUGS_PROBCELL_DN). The semantic signature of cancer-related experiments contained landmark gene sets associated with oncogene activation (LEI_HOXC8_UP and YU_CMYC_UP), oncogene mutation (YEN_MYC_MUT), antineoplastic agents (CANCERDRUGS_PROBCELL_DN and GENOTOXINS_ALL_4HRS_REG), and cell differentiation (LIAN_MYELOID_DIFF_GRANULE and LIAN_MYELOID_DIFF_RECEPTORS).

4.3. Representation of Semantic Space on 2D Space. This study was not the first that attempted to graphically represent biological entities. Several studies have provided "map" of diverse biological objects, such as genes or proteins, based on ontological annotations, structural similarity, or sequence similarities [26–28]. However, those studies employed their map only to illustrate specific study results. To the authors' knowledge, semantic space was the first that developed a comprehensive map of biological entities as a reference frame where heterogeneous experiments could be represented and compared.

In this study, we constructed semantic space on 2D space for convenience purpose. But positional relationship of components, especially if they were numerous, should be distorted when represented on lower dimension. A previous work of "yeast functional map" evaluated deforming stress from multidimensional scaling as the dimensionality was changed [26]. It reported 5D space as optimal and that increasing the dimensionality from 2D to 3D did not yield satisfactory decrease of stress, suggesting that 2D space could be a reasonable choice.

4.4. Measures for Semantic Distance. The current study adopted Lin's semantic similarity measure for the calculation of term-to-term semantic distance and best-match-average of it for the calculation of set-to-set distances [29]. Other various measures for semantic distance could be considered, such as Resnik's [30], Jiang and Conrath's [31], or Wang et al.'s measure [32] instead of Lin's measure. Similarly, best-match-average method could be replaced by simple average, maximum, or maximum weighted by information content for set-to-set calculation. Jaccard distance could also be employed, which did not require calculation of term-to-term distances [33].

To determine the optimal measure for semantic distance, we evaluated distribution of semantic distances between gene sets by different combination of measures. As a result, best-match-average of Lin's measure showed the largest variation of the distance values. Thus, it was chosen for the current study because widely distributed distances reduced distortion stress during dimension reduction. In addition, we also assessed how close were the 3 nearest landmarks that determined a coordinate of each concept in the semantic space, by calculating area of the triangle composed of the 3 landmarks. It was assumed that if 3 nearest landmarks were widely dispersed, placing the concept within the triangle would give less information for biological interpretation. The combination of Lin's measure and best-match-average method resulted in tolerable size of triangles (data not shown).

4.5. Limitations and Future Study. A few limitations should be noted in this study. As already mentioned, 2D representation of numerous landmarks could not avoid inaccuracy in their location. Another limitation is the fact that it could not be claimed that 192 gene sets from MSigDB represented all necessary spots of biological world and the semantic space spanned sufficiently wide territory. Lastly, determination of semantic signature included arbitrary component. Although visual representation of numeric data could help intuitive interpretation, similarity versus difference of geographic pattern should include subjective judgement inevitably.

Future research should consider expression of semantic space in higher dimension and could expand semantic space by merging other sources of biological data. In addition, we should think over to develop a quantitative and objective way by which similarity/difference of semantic signatures could be assessed.

5. Conclusions

Semantic space was constructed as a map of biological entities based on their relative semantic distances. When concept lattices were projected on the semantic space, "semantic signature," which was defined as specific geographic patterns observed on it, allowed intuitive interpretation of microarray experiments. Comparison of semantic signatures of various microarray datasets revealed distinctive features according to experiment conditions or tissues. In conclusion, "semantic space" could serve as a universal platform for comparative microarray analysis and "semantic signature" could be discovered.

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

Jihun Kim collected data, developed program, contributed to analysis, and participated in writing the paper. Keewon Kim participated in study design, contributed to program development, analyzed data, and participated in writing the paper. Ju Han Kim designed the study, acquired funding, supervised the study, and approved the final paper. Jihun Kim and Keewon Kim contributed equally to this work.

Acknowledgments

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea

(NRF) funded by the Ministry of Education, Science and Technology (2012-0000994) and by a grant of the Korean Health Technology R&D Project, Ministry of Health and Welfare (HI13C2164).

References

[1] P. Cahan, F. Rovegno, D. Mooney, J. C. Newman, G. St Laurent III, and T. A. McCaffrey, "Meta-analysis of microarray results: challenges, opportunities, and recommendations for standardization," *Gene*, vol. 401, no. 1-2, pp. 12–18, 2007.

[2] H.-J. Chung, C. H. Park, M. R. Han 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 Research*, vol. 33, no. 2, pp. W621–W626, 2005.

[3] J. Kim, H.-J. Chung, Y. Jung, K.-K. Kim, and J. H. Kim, "BioLattice: a framework for the biological interpretation of microarray gene expression data using concept lattice analysis," *Journal of Biomedical Informatics*, vol. 41, no. 2, pp. 232–241, 2008.

[4] N. O. Fortunel, H. H. Otu, H. H. Ng et al., "Comment on "'stemness': transcriptional profiling of embryonic and adult stem cells" and "a stem cell molecular signature"," *Science*, vol. 302, no. 5644, article 393, 2003.

[5] L. P. Lim, N. C. Lau, P. Garrett-Engele et al., "Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs," *Nature*, vol. 433, no. 7027, pp. 769–773, 2005.

[6] A. Subramanian, H. Kuehn, J. Gould, P. Tamayo, and J. P. Mesirov, "GSEA-P: a desktop application for gene set enrichment analysis," *Bioinformatics*, vol. 23, no. 23, pp. 3251–3253, 2007.

[7] P. Cahan, A. M. Ahmad, H. Burke et al., "List of lists-annotated (LOLA): a database for annotation and comparison of published microarray gene lists," *Gene*, vol. 360, no. 1, pp. 78–82, 2005.

[8] J. C. Newman and A. M. Weiner, "L2L: a simple tool for discovering the hidden significance in microarray expression data," *Genome Biology*, vol. 6, article R81, 2005.

[9] L. Abatangelo, R. Maglietta, A. Distaso et al., "Comparative study of gene set enrichment methods," *BMC Bioinformatics*, vol. 10, article 275, 2009.

[10] A. Aggarwal, D. L. Guo, Y. Hoshida et al., "Topological and functional discovery in a gene coexpression meta-network of gastric cancer," *Cancer Research*, vol. 66, no. 1, pp. 232–241, 2006.

[11] J. K. Choi, U. Yu, S. Kim, and O. J. Yoo, "Combining multiple microarray studies and modeling interstudy variation," *Bioinformatics*, vol. 19, supplement 1, pp. i84–i90, 2003.

[12] L. L. Elo, L. Lahti, H. Skottman, M. Kyläniemi, R. Lahesmaa, and T. Aittokallio, "Integrating probe-level expression changes across generations of affymetrix arrays," *Nucleic Acids Research*, vol. 33, article e193, 2005.

[13] A. T. Kho, Q. Zhao, Z. Cai et al., "Conserved mechanisms across development and tumorigenesis revealed by a mouse development perspective of human cancers," *Genes and Development*, vol. 18, no. 6, pp. 629–640, 2004.

[14] S. A. McCarroll, C. T. Murphy, S. Zou et al., "Comparing genomic expression patterns across species identifies shared transcriptional profile in aging," *Nature Genetics*, vol. 36, no. 2, pp. 197–204, 2004.

[15] G. Parmigiani, E. S. Garrett-Mayer, R. Anbazhagan, and E. Gabrielson, "A cross-study comparison of gene expression studies for the molecular classification of lung cancer," *Clinical Cancer Research*, vol. 10, no. 9, pp. 2922–2927, 2004.

[16] D. R. Rhodes, J. Yu, K. Shanker et al., "Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 25, pp. 9309–9314, 2004.

[17] R. Sandberg and I. Ernberg, "The molecular portrait of *in vitro* growth by meta-analysis of gene-expression profiles," *Genome Biology*, vol. 6, article R65, 2005.

[18] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim, "A gene-coexpression network for global discovery of conserved genetic modules," *Science*, vol. 302, no. 5643, pp. 249–255, 2003.

[19] H. Zhang, K.-H. Pan, and S. N. Cohen, "Senescence-specific gene expression fingerprints reveal cell-type-dependent physical clustering of up-regulated chromosomal loci," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 6, pp. 3251–3256, 2003.

[20] X. J. Zhou, M.-C. J. Kao, H. Huang et al., "Functional annotation and network reconstruction through cross-platform integration of microarray data," *Nature Biotechnology*, vol. 23, no. 2, pp. 238–243, 2005.

[21] Y. Guo and A. K. Manatunga, "Nonparametric estimation of the concordance correlation coefficient under univariate censoring," *Biometrics*, vol. 63, no. 1, pp. 164–172, 2007.

[22] T. Shimamoto, Y. Tang, Y. Naot et al., "Hematopoietic progenitor cell abnormalities in Hoxc-8 null mutant mice," *Journal of Experimental Zoology*, vol. 283, no. 2, pp. 186–193, 1999.

[23] M. Osanai, K. Ogawa, and G.-H. Lee, "Phenobarbital causes apoptosis in conditionally immortalized mouse hepatocytes depending on deregulated c-myc expression: characterization of an unexpected effect," *Cancer Research*, vol. 57, no. 14, pp. 2896–2903, 1997.

[24] H. Tilg, A. Kaser, and A. R. Moschen, "How to modulate inflammatory cytokines in liver diseases," *Liver International*, vol. 26, no. 9, pp. 1029–1039, 2006.

[25] B. Ursø, C. U. Niesler, S. O'Rahilly, and K. Siddle, "Comparison of anti-apoptotic signalling by the insulin receptor and IGF-I receptor in preadipocytes and adipocytes," *Cellular Signalling*, vol. 13, no. 4, pp. 279–285, 2001.

[26] A. Schlicker, F. S. Domingues, J. Rahnenführer, and T. Lengauer, "A new measure for functional similarity of gene products based on gene ontology," *BMC Bioinformatics*, vol. 7, article 302, 2006.

[27] I.-G. Choi, J. Kwon, and S.-H. Kim, "Local feature frequency profile: a method to measure structural similarity in proteins," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 11, pp. 3797–3802, 2004.

[28] J. Hou, G. E. Sims, C. Zhang, and S.-H. Kim, "A global representation of the protein fold space," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 5, pp. 2386–2390, 2003.

[29] D. Lin, "An information-theoretic definition of similarity," in *Proceedings of the 15th International Conference on Machine Learning (ICML '98)*, Madison, Wis, USA, 1998.

[30] P. Resnik, "Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language," *Journal of Artificial Intelligence Research*, vol. 11, pp. 95–130, 1999.

[31] J. Jiang and D. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," in *Proceedings the 10th Inernational Conference on Research on Computational Linguistics (ROCLING '97)*, Taipei, Taiwan, 1997.

[32] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C.-F. Chen, "A new method to measure the semantic similarity of GO terms," *Bioinformatics*, vol. 23, no. 10, pp. 1274–1281, 2007.

[33] P. H. A. Sneath and R. R. Sokal, *Numerical Taxonomy*, Freeman, San Francisco, Calif, USA, 1973.

[34] F. M. Couto, M. J. Silva, and P. M. Coutinho, "Measuring semantic similarity between Gene Ontology terms," *Data and Knowledge Engineering*, vol. 61, no. 1, pp. 137–152, 2007.

[35] D. Lin, "An information-theoretic definition of similarity," in *Proceedings of the 15th International Conference on Machine Learning (ICML '98)*, pp. 296–304, Madison, Wis, USA, 1998.

[36] B. Ganter, "Two basic algorithms in concept analysis," Tech. Rep. FB4-831, TH Darmstadt, 1984.