Functional Grouping of Genes Using Spectral Clustering and Gene Ontology

Nora Speer, Holger Fröhlich, Christian Spieth and Andreas Zell
University of Tübingen
Centre for Bioinformatics Tübingen (ZBIT)
Sand 1, D-72076 Tübingen, Germany
Email: {nspeer,froehlic,spieth,zell}@informatik.uni-tuebingen.de

***Abstract*— With the invention of high throughput methods, researchers are capable of producing large amounts of biological data. During the analysis of such data the need for a functional grouping of genes arises. In this paper, we propose a new method based on spectral clustering for the partitioning of genes according to their biological function. The functional information is based on Gene Ontology annotation, a mechanism to capture functional knowledge in a shareable and computer processable form. Our functional cluster method promises to automatize, speed up and therefore improve biological data analysis.**

I. Introduction

In the past few years, DNA microarrays have become major tools in the field of functional genomics. In contrast to traditional methods, these technologies enable researchers to collect tremendous amounts of data, whose analysis itself constitutes a challenge. On the other side, these high-throughput methods provide a global view on the cellular processes as well as on their underlying regulatory mechanisms and are therefore quite popular among biologists.

During the analysis of such data, researchers use different approaches in order to deal with the huge amounts of data they gathered. Some use statistics to find significantly regulated genes that may be involved in the underlying process due to their change in expression. Others apply pattern recognition methods to cluster the genes according to their expression profiles. The hypothesis is, that genes with expression pattern similar to those of known genes involved in the examined biological process, may play a role in the process, too. In both cases, researchers often end up with long lists of interesting candidate genes that need further examination. At this point, a second step is almost always applied: biologists categorize these genes by known biological functions and thus try to combine a pure numerical analysis with biological information.

So far, many approaches are known that address the problem of combining new experimental data with existing biological knowledge. Some methods score whole clusterings or each single cluster due to their biological relevance [5], [12], [7], [15]. Others evaluate all annotations in a group of genes and score each single annotation using sophisticated methods [2], [17], [20]. In order to receive more meaningful clustering results, some methods use the Gene Ontology as a filter to find genes that belong to a special functional category. These genes are then clustered according to their expression pattern [1]. Approaches intending to find clusters of co-expressed genes that share a common function directly incorporate the biological knowledge into the clustering process [8], [23], [21].

In this paper we address the problem of finding functional gene clusters only based on Gene Ontology terms. The advantage of such a method is that no *a priori* knowledge about relevant pathways is necessary except a mapping from genes to their ontological information. The latter is often available in public databases. Given the GO terms we are able to compute a functional similarity between genes [13]. This information is fed into a clustering algorithm. To our best knowledge, so far there exists no automatic method that produces a biologically plausible functional clustering of genes just based on the GO apart from our earlier publication [22]. In contrast to this earlier publication, in this paper we represent each gene by its functional similarity to all other genes. This encoding allows us to construct a valid mathematical distance measure between genes. There is also a deeper connection to "Kernel Methods" [19], which will be discussed later on in this paper. The final grouping of the genes is performed by a spectral clustering method [14].

The organization of this paper is as follows: a brief introduction to the Gene Ontology is given in section II. Section III explains our method in detail. The performance of our functional clustering algorithm on real world datasets is shown in section IV. Finally, in section V, we conclude.

II. The Gene Ontology

The Gene Ontology (GO) is one of the most important ontologies within the bioinformatics community and is developed by the GO Consortium [24]. It is specifically intended for annotating gene products with a consistent, controlled and structured vocabulary. Gene products are for instance sequences in databases as well as measured expression profiles. The GO independent from any biological species and additionally new ontologies covering other biological or medical aspects are being developed.

The GO represents terms in a directed acyclic graph (DAG) covering three orthogonal taxonomies or "aspects": *molecular*

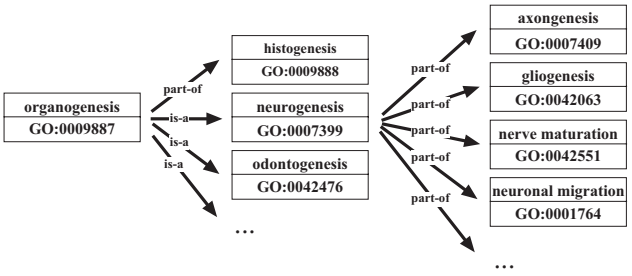

Fig. 1. Relations in the Gene Ontology. Each node is annotated with a unique accession number.

function, *biological process*, and *cellular component*. The GO graph consists of over 18.000 terms represented as nodes, which are connected by relationships represented as edges. Terms are allowed to have multiple parents as well as multiple children. Two different kinds of relationship exist: the "is-a" relationship (*neurogenesis* and *odontogenesis* are for example children of *organogenesis*) and the "part-of" relationship, which describes e.g. that *histogenesis* is part of *organogenesis* or *axongenesis* is part of *neurogenesis*. Providing a standard vocabulary across any biological resources, the GO enables researchers to use this information for automated data analysis.

III. Methods

A. Distances within the Gene Ontology

There are a couple of semantic similarity and distance measures of different complexity [3], most of them were originally developed for taxonomies like WordNet. In this paper we use a distance measure based on the information content of each GO term developed by Jiang and Conrath in [11]. The information content of a term is defined as the probability of occurrence of this term or any child term in a dataset [16]. Following the notation in information theory, the information content (IC) of a term c can be quantified as follows:

$$IC(c) = -\ln P(c)$$

where $P(c)$ is the probability of encountering an instance of term c.

In the case of a hierarchical structure, such as the GO, where a term in the hierarchy subsumes those lower in the hierarchy, this implies that $P(c)$ is monotonic as one moves towards the root node. As the node's probability increases, its information content or its informativeness decreases. The root node has a probability of 1, hence its information content is 0. As the three aspects of the GO are disconnected subgraphs, this is still true if we ignore the root node (*Gene Ontology*, GO:0003673) and take, e.g., *cellular component* (GO:0005575) as our root node instead. $P(c)$ is simply computed using maximum likelihood estimation: $P(c) = \text{freq}(c)/N$, where N is the total number of terms occurring in the dataset and $\text{freq}(c)$ is the number of times term c or any child term of c occurs in the dataset.

The similarity of two terms c_i, c_j can then be defined as followed:

$$\text{sim}(c_i, c_j) = -\ln \min_{c \in S(c_i,c_j)} P(c) = -\ln P_{ms}(c_i, c_j) \quad (1)$$

where $S(c_i, c_j)$ is the set of parental terms shared by both c_i and c_j. As the GO allows multiple parents for each term, two terms can share parents by multiple paths. We take the minimum $P(c)$, if there is more than one parent. This is called P_{ms}, for *probability of the minimum subsumer* [13]:

$$P_{ms}(c_i, c_j) = \min_{c \in S(c_i,c_j)} P(c)$$

Given the similarity score (Eqn. 1), Jiang and Conrath [11] developed a distance measure, which is the inverse of similarity. They defined the semantic distance of two classes c_i, c_j as follows:

$$d(c_i, c_j) = 2 \ln P_{ms}(c_i, c_j) - (\ln P(c_i) + \ln P(c_j)) \quad (2)$$

Since genes are often annotated with more than one GO term, multiple functional distances can be computed between two genes. Therefore, we need to combine all or choose one of the calculated distances. We decided to use the smallest distance found. Obviously, this causes a loss of information (from multiple known gene functions, only one is used). Additionally, the problem with using the smallest GO-distance (Eqn. 2) between two genes x and y is that it can be 0, even if two genes are not identical, because they belong to the same functional class. This prevents us from using (Eqn. 2) directly as a metric for clustering. We solve both problems, by using a feature vector representation for each gene.

B. Distances between Genes Using Feature Vectors

For each gene x we construct a feature vector $\phi_p(x)$ relative to prototypes $\mathbf{p} = (p_1, ..., p_N)^T$

$$\phi_p(x) = (d(x, p_1), ..., d(x, p_N))^T$$

This construction is known as an *empirical feature map* [19]. In our case prototypes are just all genes from our data set. That means each gene x is represented by its smallest functional distance to all other genes. Now, the distance between two genes x and y is simply given by $\hat{d}(x, y) = \|\phi(x) - \phi(y)\|$.

There exists a deep connection to the construction of so called "kernel functions", which can be viewed as a general similarity measure $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with the property of being symmetric and positive definite: More specifically, we have the equality (c.f. [19])

$$\begin{aligned} \hat{d}^2(x, y) &= \|\phi(x) - \phi(y)\|^2 \\ &= \langle \phi(x), \phi(x) \rangle - 2\langle \phi(x), \phi(y) \rangle + \langle \phi(y), \phi(y) \rangle \\ &= k(x, x) - 2k(x, y) + k(y, y) \end{aligned}$$

That means by defining $\phi : \mathcal{X} \to \mathcal{H}$ we map our data into some Hilbert space $\mathcal{H}$. The scalar product in this space defines a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and hence a similarity measure between two genes x and y in our original input space $\mathcal{X}$. If

we take the normalization $\phi_{norm}(x) = \frac{\phi(x)}{\|\phi(x)\|}$, we recover the normalized kernel [19]

$$k_{norm}(x,y) = \langle \phi_{norm}(x), \phi_{norm}(y) \rangle = \frac{k(x,y)}{\sqrt{k(x,x)k(y,y)}}$$

C. Spectral Clustering using Feature Vector Representation

Given our representation of each gene as a feature vector, we can choose any clustering algorithm to group our data. In this paper we took the spectral clustering algorithm by Ng *et al.* [14]: given the distance measure $\hat{d}$ on data $x_1, ..., x_n$ one computes the k largest eigenvalues and corresponding Eigenvectors of the graph Laplacian $L = D^{-1/2}KD^{-1/2}$ where $K = (\exp(\hat{d}^2(x_i, x_j)/2\sigma^2))_{ij}$ and D is a diagonal matrix with $D_{jj} = \sum_i K_{ij}$. After renormalization to unit length the Eigenvectors are then clustered e.g. by k-means. Here we choose the k-means algorithm by Zha *et al.* [25], which leads to a unique and global optimal solution. This has the advantage that no restarts are necessary. The parameter σ can be tuned automatically such that the average distortion of the points in eigenvector space becomes minimal (c.f. [14]).

D. Cluster Validity

We selected the number of clusters k in our data according to the maximal mean Silhouette index [18]. The Silhouette value for each point is a measure of how similar that point is to points in its own cluster vs. points in other clusters, and ranges from -1 to +1. It is defined as:

$$S(i) = \frac{\min(\bar{d_B}(i,j)) - \bar{d}_W(i)}{\max(\bar{d}_W(i), \min(\bar{d_B}(i,j))} \quad (3)$$

where $\bar{d}_W(i)$ is the average distance from the j-th point to the other points in its own cluster, and $\bar{d_B}(i,j)$ is the average distance from the i-th point to points in another cluster j.

IV. Experiments

A. Datasets

One possible scenario where researchers would like to group a list of genes according to their function is when they examine gene expression with DNA microarray technology, afterwards apply some filtering or statistical analysis and end up with a list of genes that show a significant change in their expression according to a control experiment. Thus, we chose two publicly available microarray datasets, annotated the genes with GO information and used them for functional clustering.

The authors of the first dataset examined the response of human fibroblasts to serum on cDNA microarrays in order to study growth control and cell cycle progression. They found 517 genes whose expression levels varied significantly, for details see [10]. We used these 517 genes for which the authors provide NCBI accession numbers. The GO mapping was done via GeneLynx Ids [6]. Since we are interested in gene function, we only use the taxonomy *biological process* of the GO. Out of the 517 genes, 238 genes showed one or more GO mappings to *biological process* or a child term of *biological process*. These 238 genes were used for the functional clustering.

In order to study gene regulation during eukaryotic mitosis, the authors of the second dataset examined the transcriptional profiling of human fibroblasts during cell cycle using microarrays [4]. Duplicate experiments were carried out at 13 different time points ranging from 0 to 24 hours. Cho *et al.* found 388 genes whose expression levels varied significantly. Hvidsten *et al.* [9] provide a mapping of the dataset to GO. 233 of the 388 genes showed at least one mapping to the GO *biological process* taxonomy and were thus used for clustering.

B. Results

In the experiments, we compared our method to k-means and Single Linkage clustering which are also based on the proposed feature vector representation, and evaluated them by means of the Silhouette clustering index (Eqn. 3). Beside that, we show the actual GO annotations of some selected clusters. Due to space limitations, we cannot show all clusters.

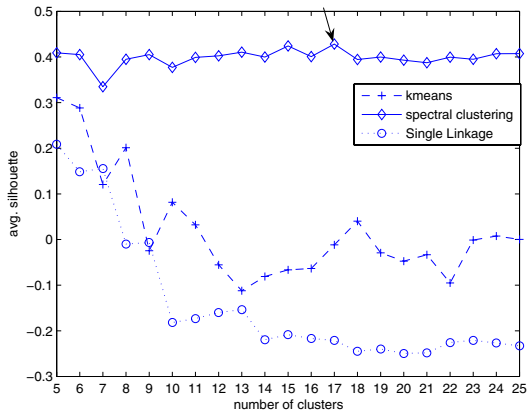

Fig. 2. Average Silhouette index of dataset I. The arrow indicates the solution with the best Silhouette index that was examined in more detail.

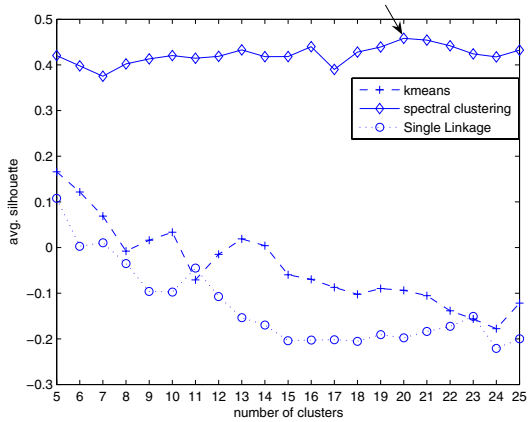

Fig. 3. Average Silhouette index of dataset II. The arrow indicates the solution with the best Silhouette index that was examined in more detail.

Figures 2 and 3 show the average Silhouette index for

cluster numbers $k = 5, ..., 25$ for all three clusterings (spectral, k-means and Single Linkage). Both figures show that the spectral clustering method gives significant better results than the other two approaches.

TABLE I

CLUSTER 7 FROM DATASET I: APOPTOSIS RELATED GENES

Acc. number	Gene Ontology terms
AA029909	**apoptosis** RNA splicing response to stress
AA012996	**anti-apoptosis**
N67978	**anti-apoptosis**
N79013	**apoptosis** **induction of apoptosis**
R62600	**apoptosis** axon guidance embryogenesis and morphogenesis neurogenesis proteolysis and peptidolysis
AA025275	**apoptosis** **induction of apoptosis by extracellular signals** protein amino acid phosphorylation
R51770	**apoptosis**
AA053239	**apoptosis**
AA037369	electron transport **induction of apoptosis**

TABLE II

CLUSTER 12 FROM DATASET I: PROTEIN METABOLISM AND MODIFICATION RELATED GENES

Acc. number	Gene Ontology terms
AA044619	**proteolysis and peptidolysis**
AA027277	**protein biosynthesis**
AA043103	**protein modification**
AA044425	**amino acid activation** **protein biosynthesis**
W73157	**protein amino acid dephosphorylation**
AA045480	**protein biosynthesis**
AA039663	response to oxidative stress **protein amino acid phosphorylation**
AA004517	**protein modification**
H94471	**protein complex assembly**
AA024572	**protein biosynthesis**
AA057638	**protein biosynthesis**
AA056621	**protein folding**
AA043969	**proteolysis and peptidolysis** vision
N49296	**protein folding**
AA045437	**protein modification**
N98463	**protein modification**
AA057826	**protein biosynthesis**
AA057359	**protein amino acid phosphorylation** sodium ion transport response to stress

According to these plots, we picked 17 clusters for dataset I and 20 clusters for dataset II. These solutions were then used for further examination. For dataset I, we show three selected clusters: cluster 7, 12, and 13. Each gene in cluster 7 is beside other functions related to apoptosis (Tab. I).

All genes of cluster 12 have at least one, but in most of the cases more than one GO annotation that is related to

TABLE III

CLUSTER 13 FROM DATASET I: REGULATION OF TRANSCRIPTION RELATED GENES

Acc. number	Gene Ontology terms
W90080	cellular morphogenesis embryogenesis and morphogenesis microtubule cytoskeleton organization and biogenesis pattern specification **regulation of transcription, DNA-dependent**
AA034054	cellular defense response **regulation of transcription, DNA-dependent** **regulation of transcription from Pol II promoter**
AA029205	**transcription from Pol II promoter**
W70150	**regulation of transcription, DNA-dependent**
N39221	response to heat **transcription from Pol II promoter** **regulation of transcription, DNA-dependent**
H14569	**regulation of transcription, DNA-dependent** **regulation of transcription from Pol II promoter**
AA040156	**transcription from Pol II promoter** **regulation of transcription, DNA-dependent**
AA035360	**regulation of transcription, DNA-dependent**
N98485	**transcription from Pol II promoter** **regulation of transcription, DNA-dependent**
T91871	anterior compartment specification oncogenesis posterior compartment specification **regulation of transcription, DNA-dependent**
H27557	**regulation of transcription, DNA-dependent**
T50056	**regulation of transcription, DNA-dependent**
R39209	**regulation of transcription, DNA-dependent**
W44416	drug resistance glutamine metabolism nucleobase, nucleoside, nucleotide and nucleic acid metabolism 'de novo' pyrimidine base biosynthesis
R49309	**regulation of transcription, DNA-dependent**
AA026120	protein modification **regulation of transcription, DNA-dependent**
N99070	**regulation of transcription, DNA-dependent** **regulation of transcription from Pol II promoter**
AA055585	**regulation of transcription, DNA-dependent**

TABLE IV

CLUSTER 12 FROM DATASET II: DNA REPLICATION, REPAIR AND RECOMBINATION RELATED GENES

Acc. number	Gene Ontology terms
D26018_at	**DNA dependent DNA replication**
D38073_at	**DNA replication initiation**
D38551_at	**double-strand break repair** **DNA recombination** **meiotic recombination**
D50370_at	**nucleosome assembly**
J04611_at	**DNA ligation** **double-strand break repair** **double-strand break repair via nonhomologous end-joining** **DNA recombination**
L07541_at	**DNA strand elongation**
M87339_at	**DNA strand elongation**
U27516_at	**double-strand break repair** **mitotic recombination** **meiotic recombination**
X62153_at	**DNA replication initiation**
X74331_at	**DNA replication, priming**

TABLE VI

CLUSTER 19 FROM DATASET II: CELL CYCLE, CELL PROLIFERATION RELATED GENES.

Acc. number	Gene Ontology terms
L11353_at	**negative regulation of cell proliferation**
L22005_at	**cell cycle checkpoint** DNA replication checkpoint **G1/S transition of mitotic cell cycle**
M60974_at	**regulation of cell cycle** **regulation of CDK activity** DNA repair apoptosis response to stress **cell cycle arrest**
M81933_at	**regulation of cell cycle** **regulation of CDK activity**
M90657_at	N-linked glycosylation **cell proliferation** pathogenesis
S81914_at	apoptosis anti-apoptosis embryogenesis and morphogenesis **cell growth and/or maintenance**
U05340_at	**regulation of cell cycle** ubiquitin-dependent protein catabolism **cell cycle**
U33286_at	nucleocytoplasmic transport apoptosis **cell proliferation**
U37426_at	**mitotic spindle assembly** **mitosis**
U40343_at	**regulation of CDK activity** **cell cycle arrest** **negative regulation of cell proliferation**
U47414_at	**cell cycle checkpoint**
U53446_at	**cell proliferation**
U56816_at	**regulation of CDK activity** **mitosis** **regulation of mitosis**
Z36714_at	**regulation of cell cycle**
U63743_at	**centromere binding** **mitosis** **cell proliferation**
X00588_at	cellular morphogenesis EGF receptor signaling pathway **cell proliferation**
X05360_at	**regulation of cell cycle** **start control point of mitotic cell cycle**
X54941_at	**regulation of cell cycle** **regulation of CDK activity** **cell proliferation**
X54942_at	**regulation of CDK activity** **cell proliferation**
X58377_at	**cell-cell signaling cell proliferation** **positive regulation of cell proliferation**
X62048_at	**regulation of cell cycle**
X65550_at	**regulation of cell cycle** **cell proliferation**
X66364_at	**cell proliferation**
X80230_at	**regulation of cell cycle** transcription initiation from Pol II promoter RNA elongation from Pol II promoter **cell proliferation**
X81851_at	chemotaxis immune response cellular defense response **cell proliferation**
X85137_at	**mitotic spindle assembly** **mitosis**
Z24725_at	**regulation of cell cycle** **cell proliferation**
Z29066_at	**regulation of cell cycle** **mitosis** **regulation of mitosis**
Z29067_at	**cell cycle**

protein modification, either by (de-)phosphorylation, protein folding, protein complex assembly or protein biosynthesis in general (Tab. II). The genes of cluster 13 are mainly involved in transcription and regulation of transcription (Tab. III). Other clusters (the data is not shown due to space limitations) contain genes that share the three functions cell growth, cell-cell-signalling and transcription regulation (cluster 6). Others are related to development (cluster 8), DNA repair and replication (cluster 9), cell adhesion in combination with cell-cell-signalling (cluster 10), immune and stress response (cluster 11), electron transport, gycolysis and small molecule transport (cluster 14), signal transduction (cluster 15), fatty acid, amino acid and cholesterol biosynthesis and metabolism (cluster 16) and cell cycle (cluster 17).

For dataset II we show 3 clusters: cluster 10 (Tab. V), 12 (Tab. IV) and 19 (Tab. VI). The genes of cluster 10 are completely annotated with GO terms related to DNA replication, repair and recombination whereas those of cluster 12 are related to cell cycle (mitosis) in combination with oncogenesis. Oncogenes are cancer inducing genes and cancer is often known to occur due to defects in cell cycle regulation. Cluster 19 genes are also related to cell cycle (mitosis), but are not related to oncogenesis. Beside these three, similar clusters are found in dataset II as in dataset I (data not shown), e.g. protein modification and catabolism (cluster 13), energy pathways and metabolism (cluster 16), signal transduction (cluster 17), cell-cell signalling (cluster 18) and transcription and RNA processing (cluster 20). Beside that, four smaller clusters are present containing genes with identical GO annotations of two or three completely independent biological functions.

V. CONCLUSION

In this paper we presented a new functional clustering method for genes based on the GO that is available in most public databases. The fact that we use the smallest distance to combine different GO term distances to one functional distance between on the gene level previously caused the two problems: first too much information is discarded and second one does not operate in a proper metric space. With the feature vector representation of each gene used in this method, we are now able to overcome this problem. We showed that our method is able to detect functional clusters of genes. Additionally, we are able to distinguish between clusters of genes that share one, but differ in a second function, e.g. cell cycle genes related to

TABLE V

CLUSTER 10 FROM DATASET II: CELL CYCLE, CELL PROLIFERATION AND ONCOGENESIS RELATED GENES

Acc. number	Gene Ontology terms
M13150_at	**oncogenesis** G-protein coupled receptor protein signaling pathway embryogenesis and morphogenesis **cell proliferation**
M31423_at	**oncogenesis** **cell growth and/or maintenance**
M86699_at	**regulation of cell cycle** **oncogenesis** **spindle assembly** **mitotic spindle assembly** **mitotic spindle checkpoint** **positive regulation of cell proliferation**
U01038_at	**regulation of cell cycle** **oncogenesis** **mitosis** **cell proliferation**
U09579_at	**regulation of cell cycle** **regulation of CDK activity** **oncogenesis** **cell cycle arrest** **negative regulation of cell proliferation** induction of apoptosis by intracellular signals
U33203_at	**oncogenesis** **negative regulation of cell proliferation**
U33761_at	**regulation of cell cycle** **G1/S transition of mitotic cell cycle** **oncogenesis** **cell proliferation**
U43916_at	**oncogenesis** development cell death **cell proliferation** epidermal differentiation
U58090_at	**G1/S transition of mitotic cell cycle** **oncogenesis** **cell cycle arrest** **negative regulation of cell proliferation** induction of apoptosis by intracellular signals
X51688_at	**regulation of CDK activity** **oncogenesis** **mitotic G2 checkpoint**

oncogenesis and cell cycle genes not related to oncogenesis. Our experiments revealed that the spectral clustering algorithm using our feature vector representation lead to significantly better results than k-means and Single Linkage clustering. The clusters found by our method contain genes annotated with the same or very similar functions. Thus, our method enormously facilitates the analysis of high throughput data.

ACKNOWLEDGMENT

This work was supported by the National Genome Research Network (NGFN) of the Federal Ministry of Education and Research in Germany under contract number 0313323.

REFERENCES

[1] B. Adryan and R. Schuh. Gene Ontology-based clustering of gene expression data. *Bioinformatics*, 20(16):2851–2852, 2004.

[2] T. Beißbarth and T. Speed. GOstat: find statistically overexpressed Gene Ontologies within groups of genes. *Bioinformatics*, 20(9):1464–1465, 2004.

[3] A. Budanitsky and G. Hirst. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and other Lexical Resources, Second meeting of the Nord American Chapter of the Association for Computational Linguistics*. Pittsburgh, 2001.

[4] R.J. Cho, M. Huang, M.J. Campbell, H. Dong, L. Steinmetz, L. Sapinoso, G. Hampton, S.J. Elledge, R.W. Davis, and D.J. Lockhart. Transcriptional regulation and function during the human cell cycle. *Nature Genetics*, 27(1):48–54, 2001.

[5] I. Gat-Viks, R. Sharan, and R. Shamir. Scoring clustering solutions by their biological relevance. *Bioinformatics*, 19(18):2381–2389, 2003.

[6] Gene Lynx. http://www.genelynx.org, 2004.

[7] J.J. Goeman, S.A. van de Geer, F. de Kort, and H.C. van Houwelingen. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99, 2004.

[8] D. Hanisch, A. Zien, R. Zimmer, and T. Lengauer. Co-clustering of biological networks and gene expression data. *Bioinformatics*, 18 (Supplement):S145–S154, 2002.

[9] T.R. Hvidsten, A. Laegreid, and J. Komorowski. Learning rule-based models of biological process from gene expression time profiles using Gene Ontology. *Bioinformatics*, 19(9):1116–1123, 2003.

[10] V.R. Iyer, M.B. Eisen, D.T. Ross, G. Schuler, T. Moore, J.C.F. Lee, J.M. Trent, L.M. Staudt, J. Hudson Jr, M.S. Boguski, D. Lashkari, D. Shalon, D. Botstein, and P.O. Brown. The transcriptional program in response of human fibroblasts to serum. *Science*, 283:83–87, 1999.

[11] J.J. Jiang and D.W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the International Conference on Research in Computational Linguistics*, Taiwan, 1998. ROCLING X.

[12] S.G. Lee, J.U. Hur, and Kim Y.S. A graph-theoretic modeling on go space for biological interpretation on gene clusters. *Bioinformatics*, 20(3):381–388, 2004.

[13] P.W. Lord, R.D. Stevens, A. Brass, and C.A. Goble. Semantic similarity measures across the gene ontology: the relationship between sequence and annotation. *Bioinformatics*, 19:1275–1283, 2002.

[14] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Adv. in Neural Inf. Proc. Syst. 14*, 2002.

[15] S. Raychaudhuri and R.B. Altman. A literature-based method for assessing the functional coherence of a gene group. *Bioinformatics*, 19(3):396–401, 2003.

[16] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, volume 1, pages 448–453, Montreal, 1995.

[17] P.N Robinson, A. Wollstein, U. Böhme, and B. Beattie. Ontologizing gene-expression microarray data: characterizing clusters with gene ontology. *Bioinformatics*, 20(6):979–981, 2003.

[18] P.J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational Applications in Math*, 20:53–65, 1987.

[19] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[20] N.H. Shah and N.V. Fedoroff. CLENCH: a program for calculating Cluster ENriCHment using Gene Ontology. *Bioinformatics*, 20(7):1196–1197, 2004.

[21] F. Sohler, D. Hanisch, and R. Zimmer. New methods for joint analysis of biological networks and expression data. *Bioinformatics*, 20(10):1517–1521, 2004.

[22] N. Speer, C. Spieth, and A. Zell. A memetic clustering algorithm for the functional partition of genes based on the Gene Ontology. In *Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2004)*, pages 252–259. IEEE Press, 2004.

[23] N. Speer, C. Spieth, and A. Zell. A memetic co-clustering algorithm for gene exression profiles and biological annotation. In *Proceedings of the IEEE 2004 Congress on Evolutionary Computation, CEC 2004*, volume 2, pages 1631–1638. IEEE Press, 2004.

[24] The Gene Ontology Consortium. The gene ontology (GO) database and informatics resource. *Nucleic Acids Research*, 32:D258–D261, 2004.

[25] H. Zha, C. Ding, X. He M. Gu, and H. Simon. Spectral relaxation for k-means clustering. In *Proc. Neural Inf. Proc. Syst. 14*, pages 1057 – 1064, 2001.