Spectral Clustering Gene Ontology Terms to Group Genes by Function

Nora Speer, Christian Spieth, and Andreas Zell

University of Tübingen, Centre for Bioinformatics Tübingen (ZBIT),
Sand 1, D-72076 Tübingen, Germany
nspeer@informatik.uni-tuebingen.de

Abstract. With the invention of biotechnological high throughput methods like DNA microarrays, biologists are capable of producing huge amounts of data. During the analysis of such data the need for a grouping of the genes according to their biological function arises. In this paper, we propose a method that provides such a grouping. As functional information, we use Gene Ontology terms. Our method clusters all GO terms present in a data set using a Spectral Clustering method. Then, mapping the genes back to their annotation, genes can be associated to one or more clusters of defined biological processes. We show that our Spectral Clustering method is capable of finding clusters with high inner cluster similarity.

1 Introduction

In the past few years, high-throughput techniques like microarrays have become major tools in the field of genomics. In contrast to traditional methods, these technologies enable researchers to collect tremendous amounts of data, whose analysis itself constitutes a challenge. Since these techniques provide a global view on the cellular processes as well as on their underlying regulatory mechanisms, they are quite popular among biologists. After the analysis of such data, using filtering methods, clustering techniques or statistical approaches, researchers often end up with long lists of interesting candidate genes that need further examination. Then, in a second step, they categorize these genes by known biological functions.

In this paper, we address the problem of finding functional clusters of genes by clustering Gene Ontology terms. Based on methods originally developed for semantic similarity, we are able to compute a functional similarity between GO terms [13]. This information is fed into a spectral clustering algorithm [15]. This has the advantage, that after mapping the genes back to the GO terms, a gene with more than one associated term (function) can be present in more than one cluster which seems biologically plausible.

The organization of this paper is as follows: a brief introduction to the Gene Ontology is given in section 2. Related Work is discussed in section 3. Section 4 explains our method in detail. The experimental setup and the results on real world data sets are shown in section 5. Finally, in section 6, we conclude.

R. Casadio and G. Myers (Eds.): WABI 2005, LNBI 3692, pp. 1–12, 2005.
© Springer-Verlag Berlin Heidelberg 2005

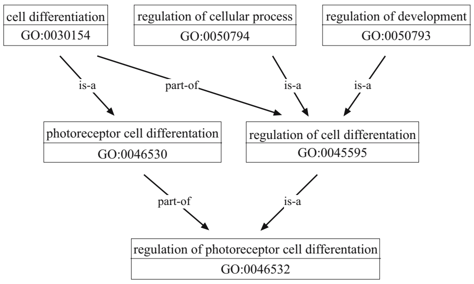
cell differentiation
GO:0030154
regulation of cellular process
GO:0050794
regulation of development
GO:0050793
is-a
part-of
is-a
is-a
photoreceptor cell differentiation
GO:0046530
regulation of cell differentiation
GO:0045595
part-of
is-a
regulation of photoreceptor cell differentiation
GO:0046532

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

2 The Gene Ontology

The Gene Ontology (GO) is one of the most important ontologies within the bioinformatics community and is developed by the Gene Ontology Consortium [21]. 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 is independent from any biological species. It represents terms in a Directed Acyclic Graph (DAG), covering three orthogonal taxonomies or "aspects": *molecular function*, *biological process* and *cellular component.* The GO-graph consists of over 18.000 terms, represented as nodes within the DAG, 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 (*photoreceptor cell differentiation* is, for example, a child of *cell differentiation*) and the "part-of" relationship that describes, for instance, that *regulation of cell differentiation* is part of *cell differentiation.*

Providing a standard vocabulary across any biological resources, the GO enables researchers to use this information for automatic data analysis done by computers and not by humans.

3 Related Work

While GO analysis is an increasingly important field, existing techniques suffer from some weaknesses: Many methods consider the GO simply as a list of terms, ignoring any structural relationships [2,7,17,23]. Others regard the GO primarily as a tree and convert the GO graph into a tree structure for determining distances between nodes [11]. Again others use a pseudo-distance that does not fulfill all metric conditions and relies on counting path lengths [3]. This is a delicate approach in unbalanced graphs like the GO those subgraphs have different degrees of detail.

Besides, the aim of some methods is primary either to use the GO as preprocessing [1] or as visualization tool [6]. Only few approaches utilize its structure for computation. Many methods are scoring techniques describing a list of genes annotated with GO terms [2,6,7,11,17,23]. But to our knowledge and apart from our earlier publications [20,19], there exists no automatic functional GO-based clustering method. One method is related to clustering and can be used to indicate which clusters are present in the data [3]. However, it suffers from the weaknesses that come along with using pseudo-distances as mentioned earlier.

4 Methodology

Our method consists of different steps that will be explained separately in this section: the mapping of the genes to the Gene Ontology, the calculation of functional similarities on GO terms, the spectral clustering algorithm and finally how the appropriate number of clusters is determined.

4.1 Mapping the Genes to the Gene Ontology

The functional similarity measure operates on pairs of GO nodes in a DAG, whereas in general, researchers are dealing with database ids of genes or probes. Therefore, a mapping M that relates the genes of a microarray experiment to nodes in the GO graph is required. Many databases (e.g. TrEMBL (GOA-project)) provide GO annotation for their entries and companies like Affymetrix provide GO mappings to their probe set ids as well. We used GeneLynx [8] to map the genes of dataset I. Hvidsten *et al.* [9] provide a mapping for dataset II.

4.2 Similarities Within the Gene Ontology

To calculate functional similarities between GO nodes, we rely on a technique that was originally developed for other taxonomies like WordNet to measure semantic similarities between words [12].

Following the notation in information theory, the information content (IC) of a term t can be quantified as follows [13]:

$$IC(t) = -\ln P(t) \tag{1}$$

where $P(t)$ is the probability of encountering an instance of term t in the data.

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 $P(t)$ 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" and take, for example, "biological process" as our root node instead.

To compute a similarity between two terms, one can use the IC of their common ancestor. As the GO allows multiple parents for each term, two terms can share ancestors by multiple paths. We take the minimum $P(t)$, if there is more than one ancestor. This is called P_{ms}, for *probability of the minimum subsumer* [13]. Thereby, it is guaranteed, that the most specific parent term is selected:

$$P_{\mathrm{ms}}(t_i, t_j) = \min_{t \in S(t_i, t_j)} P(t) \tag{2}$$

where $S(t_i, t_j)$ is the set of parental terms shared by both t_i and t_j. Based on Eq. 1 and 2, Lin extended the similarity measure, so that the IC of each single node was also taken into account [12,13]:

$$s(t_i, t_j) = \frac{2 \ln P_{ms}(t_i, t_j)}{\ln P(t_i) + \ln P(t_j)} \ . \tag{3}$$

Since $P_{ms}(t_i, t_j) \geq P(t_i)$ and $P_{ms}(t_i, t_j) \geq P(t_j)$, its value varies between 1 (for similar terms) and 0.

One should note, that the probability of a term as well as the resulting similarity between two terms differs from data set to data set, depending on the distribution of terms. Therefore, our clustering differs from a general clustering of the GO and a subsequent mapping of the genes to such a general clustering. Due to our approach, we are able to arrange the resulting cluster boundaries depending on the distribution of the GO terms either more specific (if the terms concentrate on a specific part of the GO) or more general (if the terms are widely spread).

4.3 Spectral Clustering

We decided to cluster GO terms, not genes, because of two reasons: first, we do not face the problem of combining different similarities per gene like in earlier publications [19,20] and second, after mapping the genes back to the GO, they can be present in more than one functional cluster which is biologically plausible, since they can also fulfill more than one biological function.

Recently, Spectral Clustering methods haven been growing in popularity. Several new algorithms have been published [22,18,14,15]. A set of objects (in our case GO terms) to be clustered will be denoted by T, with $|T| = n$. Given an affinity measure $A_{ij} = A_{ji} \geq 0$ for two objects i, j, the affinities A_{ij} can be seen as weights on the undirected edges ij of a graph G over T. Then, the matrix $A = [A_{ij}]$ is the real-valued adjacency matrix for G. Let $d_i = \sum_{j \in T} A_{ij}$ be called the degree of node i, and D be the diagonal matrix with d_i as its diagonal. A clustering $C = \{C_1, C_2, \ldots, C_K\}$ is a partitioning of T into the nonempty mutually disjoint subsets $C_1, C_2, \ldots, C_K$. In the graph theoretical paradigm a clustering represents a multiway cut in the graph G.

In general, all Spectral Clustering algorithms use Eigenvectors of a matrix (derived from the affinity matrix A) to map the original data to the K-dimensional vectors $\{\gamma_1, \gamma_2, \ldots, \gamma_n\}$ of the spectral domain $\Re^K$. Then, in a second step, these vectors are clustered with standard clustering algorithms. Here, we use K-means. We chose the newest Spectral Clustering algorithm by Ng *et al.* [15] and we will now review it briefly:

1. From the affinity matrix A and its derived diagonal matrix D, compute the Laplacian matrix $L = D^{-1/2}AD^{-1/2}$.
2. Find $v^1, v^2, \ldots, v^K$, the Eigenvectors of L, corresponding to the K largest Eigenvalues.
3. Form the matrix $V_{n\times k} = \left[v^1, v^2, \ldots, v^K\right]$ with these Eigenvectors as columns.
4. Form the matrix Y from V by renormalizing each of X's rows to have unit norm.
5. Cluster the rows of $Y = [\gamma_1, \gamma_2, \ldots, \gamma_n]$ as points in a K-dimensional space.
6. Finally assign the original object i to cluster j if and only if row γ_i of the matrix Y was assigned to j.

Since Spectral Clustering relies on the affinity matrix A, it is easy to apply it to any kind of data, where affinities can be computed. For numerical data, affinities are usually computed with a kernel function, e.g. $A_{ij} = \exp(\frac{-d(i,j)^2}{2\sigma^2})$, with $d(i,j)$ denoting the Euclidean distance between point i and j and σ denoting the kernel width. For non-numerical data, like GO terms, affinity can either be defined in the same way, given a distance measure d. This approach has the advantage of non-linearity, controlled by the kernel width σ, which allows for sharper separation between clusters. But it has also disadvantages: the question of how to deduce σ in a meaningful way arises and additionally, for many data types, especially the GO, similarity is much easier to define since it does not need to fulfill any metric conditions. As noted in [16], there is nothing magical about the definition of affinity. Therefore, we directly apply our similarity matrix as affinity matrix.

4.4 Cluster Validity

We selected the number of clusters K in our data according to the Davies-Bouldin index [5]. Given a clustering $C = \{C_1, C_2, \ldots, C_K\}$, it is defined as:

$$DB(C) = \frac{1}{K}\sum_{i=1}^{K} \max\left\{\frac{\Delta(C_i) + \Delta(C_j)}{\delta(C_i, C_j)}\right\} \tag{4}$$

where $\Delta(C_i)$ represents the inner cluster distance of cluster C_i and $\delta(C_i, C_j)$ denotes the inter cluster distance between cluster C_i and C_j. K is the number of clusters. Small values of $DB(C)$ indicate a good clustering.

$\Delta(C_i)$ and $\delta(C_i, C_j)$ are calculated as the sum of distances to the respective cluster mean and the distance between the centers of two clusters, respectively. Since we use similarities, not distances, and cannot compute means in the GO, we apply the DB-Index in the spectral domain $\Re^K$ (after the Eigenvector decomposition) where we are dealing with simple numerical data.

5 Computational Experiments

5.1 Data Sets

One possible scenario where researchers would like to group a list of genes according to their function is when they received lists of up- or down-regulated genes from the analysis of an DNA microarray experiment. Thus, we chose two publicly available microarray data sets, annotated the genes with the GO and used them for functional clustering. We only use the taxonomy *biological process*, because we are mainly interested in gene function in a more general sense. However, our method can be applied in the same way for the other two taxonomies.

The authors of the first data set 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 using GeneLynx [8]. After mapping to the GO, 238 genes showed one or more mappings to *biological process* or a child term of *biological process*. These 238 genes were used for the clustering.

In order to study gene regulation during eukaryotic mitosis, the authors of the second data set 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.* [4] found 388 genes whose expression levels varied significantly. Hvidsten *et al.* [9] provide a mapping of the data set to GO. 233 of the 388 genes showed at least one mapping to the GO *biological process* taxonomy and were thus used for clustering.

5.2 Experimental Design

In the experiments, we had the problem of how to compare our method to other known clustering algorithms, because to our best knowledge, there is no clustering method that does a clustering only due to a similarity matrix. Instead, most algorithms need distances. Beside that, most clustering techniques were originally developed for numerical data and therefore utilize means during the clustering process which we cannot compute in the GO. Only linkage methods work on a proximity matrix, although this is also usually a distance matrix. Average Linkage clustering is known to be its most robust, non-means based representative. Therefore, we compare our approach to a modified version of an Average Linkage algorithm that joins the most similar clusters, instead of joining those with the smallest distance. Inner cluster similarity of cluster C_i is computed as follows:

$$s(C_i) = \frac{1}{|C_i|(|C_i - 1|)} \sum_{t_i, t_j \in C_i, t_i \neq t_j} s(t_i, t_j) \tag{5}$$

with $s(t_i, t_j)$ denoting the similarity between term t_i and t_j and $|C_i|$ denoting the number of terms in cluster C_i.

For Spectral Clustering, K-means was carried out 25 times and the solution with the minimum distortion was taken as proposed in [15]. For both algorithms, we performed runs for different values of K, ranging from $K = 5, 6, \ldots, 25$.

5.3 Results

Fig. 2 shows the average inner cluster similarity for Average Linkage and Spectral Clustering for both data sets and different numbers of K. It is clearly visible that except for one exception ($K = 5$, data set I), Spectral clustering always shows a much higher inner cluster similarity than Average Linkage clustering.

Additionally, we wanted to evaluate the best solutions generated by Spectral Clustering in more detail. Since inner cluster similarity is not independent from the number of clusters K, we chose the best solution according to the Davies-Bouldin index (Eq. 4) that was calculated after the Eigenvalue decomposition in the spectral domain $\Re^K$. Fig. 3 shows the Davies-Bouldin index for the cluster numbers $K = 5, ..., 25$ for data set I and II, respectively. For data set I, the best clustering was achieved with 10 clusters and for data set II with 9 clusters. These two solutions (indicated by an arrow in Fig. 3) were then used for further examination.

Figure 4 shows the Euclidean distance matrix calculated after the Eigenvector decomposition for data set I (left) and II (right). Higher values are indicated

Table 1. Cluster 5 of dataset I. This cluster contains mainly GO terms associated with mitosis

Term Acc.	GO Term Name
GO:0007050	cell cycle arrest
GO:0000074	regulation of cell cycle
GO:0008151	cell growth and/or maintenance
GO:0007049	cell cycle
GO:0007095	mitotic G2 checkpoint
GO:0000079	regulation of CDK activity
GO:0008284	positive regulation of cell proliferation
GO:0008283	cell proliferation
GO:0006878	copper ion homeostasis
GO:0008285	negative regulation of cell proliferation
GO:0006260	DNA replication
GO:0006874	calcium ion homeostasis
GO:0008156	negative regulation of DNA replication
GO:0006269	DNA replication, priming
GO:0007093	mitotic checkpoint
GO:0007096	regulation of exit from mitosis
GO:0006298	mismatch repair
GO:0000080	G1 phase of mitotic cell cycle
GO:0007088	regulation of mitosis
GO:0000067	DNA replication and chromosome cycle
GO:0007089	start control point of mitotic cell cycle
GO:0000085	G2 phase of mitotic cell cycle
GO:0007079	mitotic chromosome movement
GO:0000089	mitotic metaphase
GO:0007080	mitotic metaphase plate congression
GO:0006261	DNA dependent DNA replication

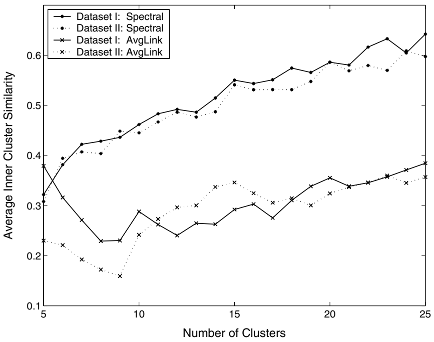

Fig. 2. Average Inner Cluster Similarity for Average Linkage and Spectral clustering for data set I and II

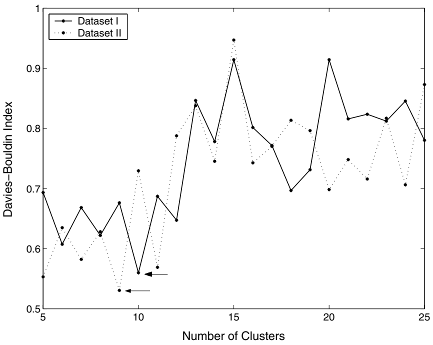

Fig. 3. Davies-Bouldin index in the spectral domain $\Re^K$ for Spectral Clustering of data set I and II

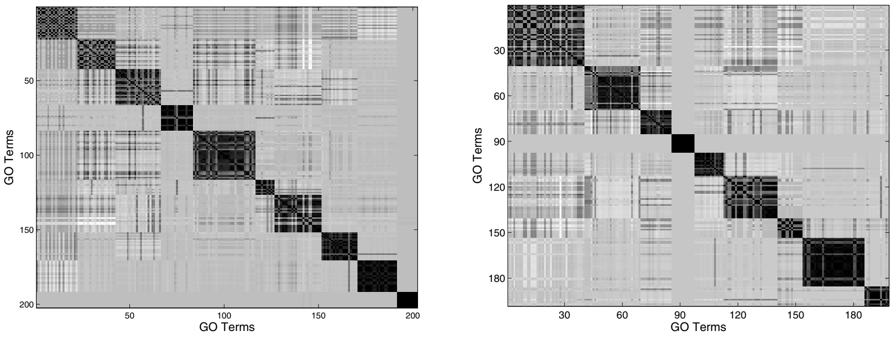

Fig. 4. Distance matrices in the spectral domain $\Re^K$ (after the Eigenvector decomposition) of data set I (left) and II (right): The 10 (left) and 9 (right) clusters are clearly visible.

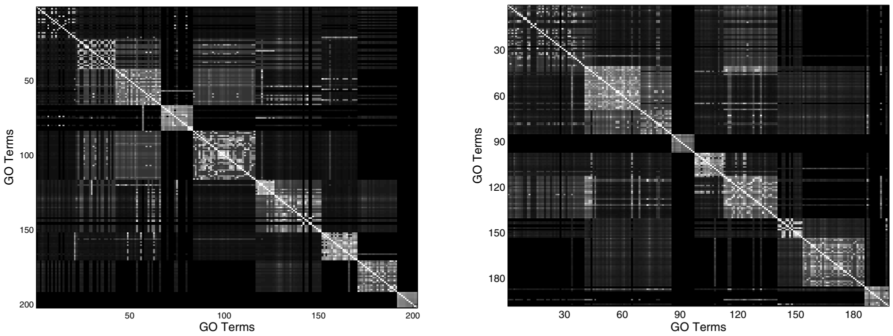

Fig. 5. The original similarity matrix of data set I (left) and II (right): Again the 10 (left) and the 9 (right) clusters are clearly visible.

Table 2. Cluster 8 of dataset I: this cluster contains mainly GO terms associated with signal transduction

Term Acc.	GO Term Name
GO:0000188	inactivation of MAPK
GO:0008277	regulation of G-protein coupled receptor protein signaling pathway
GO:0007165	signal transduction
GO:0007267	cell-cell signaling
GO:0007166	cell surface receptor linked signal transduction
GO:0007200	G-protein signaling, coupled to IP3 second messenger (phospholipase C activating)
GO:0007186	G-protein coupled receptor protein signaling pathway
GO:0007181	TGFbeta receptor complex assembly
GO:0007155	cell adhesion
GO:0008038	neuronal cell recognition
GO:0007179	TGFbeta receptor signaling pathway
GO:0007156	homophilic cell adhesion
GO:0007229	integrin-mediated signaling pathway
GO:0007178	transmembrane receptor protein serine/threonine kinase signaling pathway
GO:0007160	cell-matrix adhesion
GO:0007268	synaptic transmission
GO:0007173	EGF receptor signaling pathway
GO:0000165	MAPKKK cascade
GO:0000187	activation of MAPK
GO:0007169	transmembrane receptor protein tyrosine kinase signaling pathway
GO:0007243	protein kinase cascade

by a light color and lower values by a dark color. Thus, the 10 squares (left) and the 9 squares (right) indicate regions of small distances corresponding to the 10 and clusters, respectively. Figure 4 demonstrates that the clusters in the spectral domain $\Re^K$ have small inner cluster distances and high distances between them. The original affinity (or similarity) matrices for both data sets are visualized in Fig. 5. Again, light colors indicate higher values, thus, in this case a higher similarity. The 10 (left) and 9 (right) clusters are still clearly visible as regions of high inner cluster similarity compared to the similarity between the clusters.

Additionally, we examined clusters of a solution in more detail, but due to space limitations, we cannot show all clusters of both data sets. Therefore, we confine ourselves to show three selected clusters of data set I: cluster 5, 8 and 9. Tab. 1 - Tab. 3 show the GO terms of each of these clusters, respectively. A closer study of the GO term names reveals that our method produces from each other distinct functional clusters each containing GO terms that belong to a defined biological process. The GO terms of cluster 5 (Tab. 1) are mainly related to mitosis like cell cycle regulation or CDK activity regulation and DNA replication. In Tab. 2, the GO terms of cluster 8 are listed. They are mostly related to processes associated with signal transduction pathways like the TGF-

Table 3. Cluster 9 of dataset I: this cluster contains mainly GO terms associated with metabolism

Term Acc.	GO Term Name
GO:0006101	citrate metabolism
GO:0015936	coenzyme A metabolism
GO:0006629	lipid metabolism
GO:0006768	biotin metabolism
GO:0006633	fatty acid biosynthesis
GO:0006564	L-serine biosynthesis
GO:0006729	tetrahydrobiopterin biosynthesis
GO:0006048	UDP-N-acetylglucosamine biosynthesis
GO:0006631	fatty acid metabolism
GO:0016042	lipid catabolism
GO:0005989	lactose biosynthesis
GO:0006096	glycolysis
GO:0006700	C21-steroid hormone biosynthesis
GO:0008203	cholesterol metabolism
GO:0008202	steroid metabolism
GO:0006695	cholesterol biosynthesis
GO:0008299	isoprenoid biosynthesis
GO:0006694	steroid biosynthesis
GO:0006529	asparagine biosynthesis
GO:0006541	glutamine metabolism
GO:0006635	fatty acid beta-oxidation
GO:0006809	nitric oxide biosynthesis
GO:0006559	phenylalanine catabolism
GO:0006520	amino acid metabolism
GO:0006563	L-serine metabolism
GO:0006636	fatty acid desaturation
GO:0006004	fucose metabolism
GO:0006099	tricarboxylic acid cycle
GO:0006693	prostaglandin metabolism
GO:0006207	'de novo' pyrimidine base biosynthesis
GO:0006780	uroporphyrinogen III biosynthesis

β pathway or G-protein coupled signaling and these GO terms form cluster 8. Finally, cluster 9 (Tab. 3) contains GO terms associated with metabolic processes like amino acid synthesis, lipid metabolism or fatty acid biosynthesis, just to name a few.

6 Discussion

In this paper, we presented a clustering method for GO terms that can be used to cluster genes or any other gene products that can be annotated with the Gene Ontology. We showed that the clusters produced by our method have a higher average inner cluster similarity than those produced by a similarity-based variant of Average Linkage Clustering. Beside that, we showed for the best two solutions in detail that their GO terms have a much higher similarity to each other than to those in the other clusters. This is not only true for the data in the spectral domain $\Re^K$, but also for the original affinity matrix. Furthermore, we evaluated three clusters in more detail and could show that the GO terms in each cluster belong to a defined and separated biological process.

The Spectral Clustering technique enables us to cluster those objects, like GO terms, where it is easy to calculate similarities but more difficult to calculate distances or even means, that are needed by many popular clustering methods. In contrast to these methods, Spectral Clustering is able to produce a clustering only due to an affinity matrix. To be suitable for clustering, the affinity matrix only needs to reflect the natural relationships of the data.

Additionally, the fact that we are using GO terms for clustering and not genes like in our previous publications has the advantage that now, one gene can belong to more than one cluster. This makes also biologically sense, since one gene can also have more than one function. Thus, our method facilitates the functional 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. We thank Holger Fröhlich for fruitful discussions.

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. Flmer C.A. Joslyn, S.M. Mniszewski and G. Heaton. The gene ontology categorizer. *Bioinformatics*, 20(Suppl. 1):i169–i177, 2004.
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. J.L. Davies and D.W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1:224–227, 1979.
6. S.W. Doniger, N.Salomonis, K.D. Dahlqusi, K. Vranizan, S.C. Lawlor, and B.R. Conklin. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biology*, 4(1):R7, 2003.
7. I. Gat-Viks, R. Sharan, and R. Shamir. Scoring clustering solutions by their biological relevance. *Bioinformatics*, 19(18):2381–2389, 2003.
8. Gene Lynx. http://www.genelynx.org, 2004.
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. S.G. Lee, J.U. Hur., and Y.S. Kim. A graph-theoretic modeling on go space for biological interpretation on gene clusters. *Bioinformatics*, 20(3):381–388, 2004.
12. D. Lin. An information-theoretic definition of similarity. In Morgan Kaufmann, editor, *Proceedings of the 15th International Conference on Machine Learning*, volume 1, pages 296–304, San Francisco, CA, 1998.
13. P.W. Lord, R.D. Stevens, A. Brass, and C.A. Goble. Semantic similarity measures as tools for exploring the gene ontology. In *Proceedings of the Pacific Symposium on Biocomputing*, pages 601–612, 2003.
14. M. Meila and J. Shi. Learning segmantation by random walks. In T.K. Leen, T.G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13, pages 873–879, 2001.
15. A.Y. Ng, M.I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14, pages 849–856, Cambridge, MA, 2002. MIT Press.
16. P.Perona and W. Freeman. A factorization approach to grouping. In *Lecture Notes in Computer Sience*, 1406, pages 655–670. Springer, 1998.
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. J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.
19. N. Speer, H. Fröhlich, C. Spieth, and A. Zell. Functional grouping of genes using spectral clustering and gene ontology. In *To appear in Proceedings of the IEEE International Joint Conference on Neural Networks*, 2005.
20. 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 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, pages 252–259, 2004.
21. The Gene Ontology Consortium. The gene ontology (GO) database and informatics resource. *Nucleic Acids Research*, 32:D258–D261, 2004.
22. Y. Weiss. Segmentation using eigenvectors: a unifying view. In *Proceedings of IEEE International Conference on Computer Vision*, volume 2, pages 975–982, 1999.
23. B.R. Zeeberg, W. Feng, G. Wang, and A.T. Fojo *et al.* GOminer: a resource for biological interpretation of genomic and proteomic data. *Genome Biology*, 4(R28), 2003.