Biological Cluster Validity Indices Based on the Gene Ontology

Nora Speer, Christian Spieth, and Andreas Zell

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

Abstract. With the invention of biotechnological high throughput methods like DNA microarrays and the analysis of the resulting huge amounts of biological data, clustering algorithms gain new popularity. In practice the question arises, which clustering algorithm as well as which parameter set generates the most promising results. Little work is addressed to the question of evaluating and comparing the clustering results, especially according to their biological relevance, as well on distinguishing biologically interesting clusters from less interesting ones. This paper presents two cluster validity indices intended to evaluate clusterings of gene expression data in a biological manner.

1 Introduction

In an attempt to understand complex biological regulatory mechanisms of a cell, biologists tend to use large scale techniques to collect huge amounts of gene expression data. Thus, DNA microarrays became a popular tool in the past few years. A problem inherent in the use of DNA arrays is the tremendous amount of data produced, whose analysis itself constitutes a challenge. Data mining techniques like cluster algorithms are utilized to extract gene expression patterns inherent in the data and thus find potentially co-regulated genes [14]. Various methods have been applied, such as Self-Organizing-Maps (SOMs) [22], K-Means [23], Hierarchical Clustering [7] as well as Evolutionary Algorithms [13,20].

Since different cluster algorithms or different runs of the same algorithm generate different solutions given the same data set, in practice, biologists are faced with the problem of choosing an appropriate algorithm with appropriate parameters for the data set. The evaluation of cluster results is a process known as cluster validity and is an important task in cluster analysis.

Several cluster validity indices are known in literature, such as Dunn's Index [6], Rand Index [15], Figure of Merit [25], Silhouette Index [18] or Davies-Bouldin Index [5] and many of them have already been used with gene expression data [1,3,25]. All these indices evaluate the mathematical properties of a clustering, but especially for gene expression data, the biological cluster quality plays an important role, too [17,19]. Some attempts in this direction were based on text mining methods for literature abstracts [16]. Others simply count Gene Ontology annotations per cluster [2,17,19], but in contrast to our approach, none of them

A.F. Famili et al. (Eds.): IDA 2005, LNCS 3646, pp. 429–439, 2005.

relies on biological distances between genes, an advantage that enables the use of established cluster indices.

The paper is organized as follows: a brief introduction to the Gene Ontology is given in section 2. Section 3 explains our method in detail. The performance on real world data sets is shown in section 4. Finally, in section 5, we conclude.

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 [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 is independent from any biological species and is rapidly growing. 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 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*.

By 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.

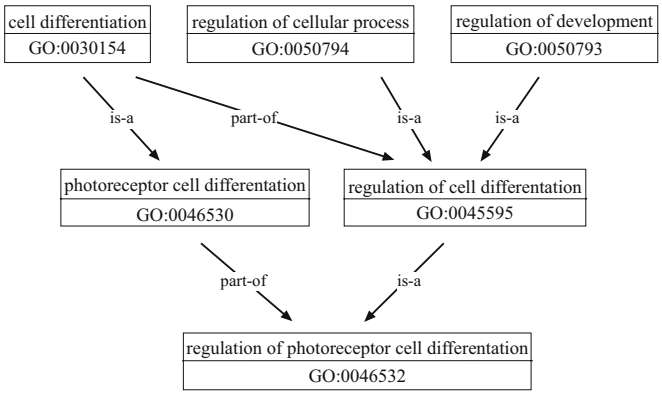

Fig. 1. Relations in the Gene Ontology. Each node is annotated with a unique accession number.

3 Methods

3.1 Mapping Genes to the Gene Ontology

To properly evaluate a clustering result with GO information, a mapping M that relates the clustered genes to the nodes in the GO graph is required. For eucaryotic genes the common biological databases (e.g. TrEMBL or GenBank) provide GO annotation for their entries and also biotech companies like Affymetrix provide GO mappings for their DNA microarrays. Such a mapping is not one-to-one, which means that there are genes annotated with more than one GO term as well as genes without a GO annotation. The first point will be discussed later in this section, the latter reduces the number of genes that can take part in such an analysis.

3.2 Distances Within the Gene Ontology

To calculate biological distances within the GO, we rely on a technique that was originally developed for other taxonomies like WordNet to measure semantic distances between words [11]. The distance measure is based on the information content of a GO term. Following the notation in information theory, the information content (IC) of a term t can be quantified as follows:

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

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

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 compute 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* [12]:

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

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, Jiang and Conrath developed the following distance measure [11]:

$$d(t_i, t_j) = 2 \ln P_{ms}(t_i, t_j) - (\ln P(t_i) + \ln P(t_j)) \tag{3}$$

Since genes can have more than one function and are therefore often annotated with more than one GO term, multiple functional distances can be computed between two genes. Since, we don't know which of these functions play a role in the underlying biological experiment, we assume the best and use the smallest distance between two genes during the calculation of cluster validities.

3.3 Cluster Validities

A good cluster validity index should be independent of the number of clusters, thus allowing to compare two clusterings with different number of clusters. At the same time, it is desirable that genes in one cluster have minimum possible distance to each other and maximum distance to the genes in other clusters, in other words, we seek clusters that are compact and well separated. Two cluster validity measures that fulfill these criteria are the Silhouette and the Davies-Bouldin index [18,5].

Given a set of genes $G = \{g_1, g_2, \ldots, g_n\}$ and a clustering of G in $C = \{C_1, C_2, \ldots, C_k\}$, the Silhouette index is defined as follows [18]: for each gene g_i of cluster C_j, a confidence measure, the Silhouette width $s(g_i)$, is calculated that indicates if gene g_i belongs to cluster C_j. The Silhouette width $s(g_i)$ is defined as follows:

$$s(g_i) = \frac{\min(\bar{d}_B(g_i)) - \bar{d}_W(g_i)}{\max\{\bar{d}_W(g_i), \min(\bar{d}_B(g_i))\}} \tag{4}$$

where $\bar{d}_W(g_i)$ is the average distance from g_i to all other genes of the cluster to which g_i is assigned and $\bar{d}_B(g_i)$ is the average distance between g_i and all other genes assigned to the clusters C_l with $l = 1, \cdots, k \wedge j \neq l$. Observations with a large $s(g_i)$ (almost 1) are very well clustered, a small $s(g_i)$ (around 0) means that the observation lies between two clusters, and observations with a negative $s(g_i)$ are probably placed in the wrong cluster. Thus, for each cluster C_j, a mean Silhouette index

$$S_j(C_j) = \frac{1}{|C_j|} \sum_{i=1}^{|C_j|} s(g_i) \tag{5}$$

can be computed. $|C_j|$ denotes the number of genes included in cluster C_j. The index ranges between 1 (for a perfect cluster/clustering) and -1. Thus, the overall quality of a clustering C can be measured using:

$$S(C) = \frac{1}{n} \sum_{i=1}^{n} s(g_i), \tag{6}$$

Given the same notation as above, the Davies-Bouldin index has been defined in [5] as:

$$DB_j(C_j) = \max_{i \neq j} \left\{ \frac{\Delta(C_i) + \Delta(C_j)}{\delta(C_i, C_j)} \right\} \tag{7}$$

where $\Delta(C_i)$ and $\Delta(C_j)$ represent the inner cluster distance of cluster C_i and C_j and $\delta(C_i, C_j)$ denotes the distance between the clusters C_i and C_j. Usually $\Delta(C_i)$ and $\delta(C_i, C_j)$ are calculated as the sum of distances to the respective cluster center and the distance between the centers of two clusters. Since means are not defined in a DAG, we use the average diameter of a cluster as $\Delta(C_i)$ and the average linkage between two clusters as $\delta(C_i, C_j)$:

$$\Delta(C_i) = \frac{1}{|C_i|(|C_i - 1|)} \sum_{g_i, g_j \in C_i, g_i \neq g_j} d(g_i, g_j) \tag{8}$$

$$\delta(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{g_i \in C_i, g_j \in C_j} d(g_i, g_j) \tag{9}$$

where $d(g_i, g_j)$ defines the distance between the genes g_i and g_j. It is clear from the above definition, that $DB_j(C_j)$ is the average similarity between cluster C_j, and its most similar one. It is desirable for the clusters to have minimum possible similarity to each other. Therefore, we seek clusterings that minimize $DB_j(C_j)$. The index for the whole clustering can be computed as:

$$DB(C) = \frac{1}{k} \sum_{j=1}^{k} DB_j(C_j). \tag{10}$$

4 Results

4.1 Data Sets

The performance of the cluster validity indices are discussed on two real world data sets. For our work, we only use the taxonomy *biological process*, because we are mostly interested in gene function. 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. We selected 14 clusters as indicated in our previous publication [21].

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.* found 388 genes whose expression levels varied significantly [4]. In [9] Hvidsten *et al.* provide a mapping of the data set to the GO. 233 of the 388 genes showed at least one mapping to the *biological process* taxonomy and were thus used for clustering. We selected 10 clusters as indicated in our previous publication [21].

4.2 Computational Experiments

If our proposed cluster indices are able to distinguish biologically meaningful clusterings from less meaningful ones, a functional clustering according to the GO annotations should show better validity index values than a clustering that was produced according to the normalized expression vectors of the genes.

Therefore, in our experiments, we used a clustering algorithm based on an Evolutionary Algorithm from earlier publications [20,21] to produce these two

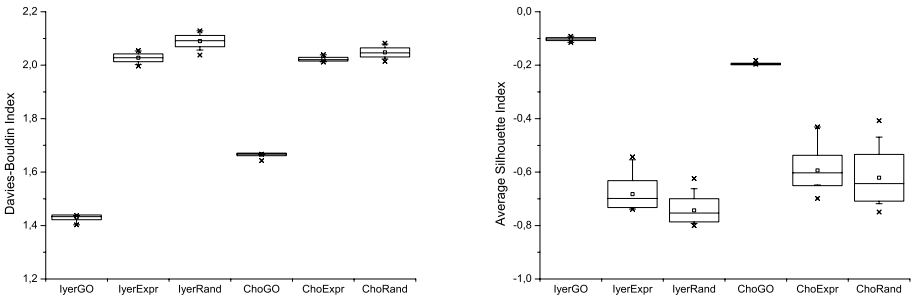

Fig. 2. Davies-Bouldin index (left, small values indicate good clusterings) and Silhouette index (right, large values indicate good clusterings) averaged over 25 runs. Maximum and minimum values are indicated by a cross, the mean by the rectangle, the standard deviation is indicated by the box, the error bars indicate the 5-95 confidence intervals.

different clusterings for each data set: an expression based clustering and a functional clustering. In principle, any cluster algorithm could be used in place that does not rely on mean calculation (this is important for the functional clustering, since we cannot compute means in the GO as mentioned earlier). The only reason why we use this algorithm is that we got good results compared to other non-mean based methods like Average Linkage clustering [20,21]. While producing these two clusterings, all parameters of the algorithm were fixed (200 generations, population size of 40 and 40% mutation and recombination rate), except the distance function used: for the functional clustering, we used the GO distance (Eq. 3) and for the expression based clustering, we used the Euclidean distance of the normalized expression vectors of each gene. The normalization was performed as described in [23]. We also compared the clusterings to a random partition. For the random partition, one result corresponds to the best partition out of 8000 (200 generations $*$ 40 individuals) tries. All results are averaged over 25 runs.

Fig. 2 shows the Davies-Bouldin index (left) and the Silhouette index (right) of the expression based and functional clusterings and for the random partition for both data sets. Maximum and minimum values are indicated by a cross, the mean by the rectangle, the standard deviation is indicated by the box, the error bars indicate the 5-95 confidence intervals. For both indices and both data sets, the GO based clustering obtains significant better values than a gene expression based clustering. These results were of course expected since we used a biological clustering method to produce this clustering. But nevertheless, it indicates that our validity measures are able to detect biological meaningful clusterings. Beside that, it is notable that the expression based clustering is only slightly better than random concerning its biological similarity, which emphasizes the need for methods that can distinguish between biologically interesting and less interesting clusterings.

Table 1. Cluster validity values for the individual clusters for a GO based clustering. A low value of the Davies-Bouldin and a high value for the Silhouette index indicate good clusters. A good and a bad cluster are marked in bold.

Cluster	Davies-Bouldin Index	Silhouette Index
1	1.49	-0.67
2	1.76	-0.55
3	1.32	-0.09
4	1.29	0.24
5	1.55	0.16
6	1.73	-0.20
7	1.39	0.21
8	**1.76**	**-0.40**
9	1.57	-0.22
10	1.29	-0.21
11	1.32	-0.26
12	1.28	-0.16
13	**1.07**	**0.49**
14	1.29	0.05

Furthermore, the presented cluster validity measures can not only be used to distinguish between whole clusterings but also to validate individual clusters and thus find interesting clusters that contain genes that are biologically closely related and already known to be involved in the same pathway. Such a cluster would indicate that a whole biological process might be switched on or off under the given experimental condition, e.g. that cells leave the G_O- phase and enter cell proliferation. Tab. 1 shows the individual cluster validity values for the overall best clustering. As an example, we show two extreme clusters in more detail.

For both cluster validity measures, cluster 13 has a good quality, whereas cluster 8 is much more functionally diverse. The GO annotations of cluster 13 are displayed in Tab. 2 and those of cluster 8 are shown in Tab. 3. The genes of the good cluster are mostly closely related to DNA replication and repair, which is a defined and separated process in biology. So cluster 13 is a small and functionally compact cluster that was also indicated by the validity values. Instead, the other example is larger and much more diverse. Genes in that cluster are related to cell adhesion, cell motility, inter- and intra-cellular signal transduction, metabolism, nervous system development and pregnancy. All theses functions are quite different biological processes, which was already indicated by the validity measures.

We showed that our two biological cluster indices are able to distinguish biologically more homogeneous clusters from less homogeneous ones, a fact that can be used to find those clusters in a clustering that contain genes that are not only co-expressed, but also related to the same biological process. Additionally, we showed that one can use these indices to measure the biological quality of

Table 2. Example of the GO annotation of a good functional cluster (cluster 13)

Probeset Id	GO Term Name
H63374	DNA repair
	pyrimidine-dimer repair, DNA damage excision
N22858	chromosome organization and biogenesis (sensu Eukarya)
	DNA methylation
	DNA recombination
	DNA repair
N68268	DNA replication
	DNA replication, priming
W93122	DNA dependent DNA replication
	DNA replication
N93479	DNA replication
H29274	DNA repair
	DNA replication
	double-strand break repair
	UV protection
AA053076	DNA replication
AA031961	cell cycle
	regulation of cell cycle
	cell proliferation
	DNA repair
	regulation of CDK activity

a whole clustering and therefore find biologically meaningful clusterings out of a bunch of given clusterings. Thus, our two presented biological cluster validity indices can be used to evaluate clusterings and single clusters of genes in a biological manner.

5 Conclusion

In this paper, we presented two biological cluster validity indices that are based on the Gene Ontology. We showed that they can be utilized to detect clusters of genes that share similar functions. This is especially important, because such clusters indicate that a whole regulatory pathway might be affected under the given conditions, which leads to an information gain about the underlying regulatory mechanisms of a cell. The fact that a clustering due to gene expression profiles does not always implicate a biological clustering as shown by our results even emphasizes the need of a tool like the presented biological cluster indices.

The advantage of our method compared to other approaches is, that it is based on biological distances, which enable the usage of established cluster validity measures including the knowledge of their weaknesses and advantages. Beside that, the utilized GO annotation is easy to obtain from biological databases.

One problem of our method is, of course, that for each gene at least one Gene Ontology annotation is needed. In most of the cases the GO annotation is available in public databases. Nevertheless, there are still some genes that do

Table 3. Example of the GO annotation of a bad functional cluster (cluster 8)

Probeset Id	GO Term Name
W89002	peroxidase reaction
H63779	central nervous system development
	epidermal differentiation
	lipid metabolism
	peripheral nervous system development
N79778	cell-matrix adhesion
N67806	respiratory gaseous exchange
R37986	pregnancy
AA029995	pregnancy
W86618	DNA metabolism
	intracellular protein transport
	G2 phase of mitotic cell cycle
	NLS-bearing substrate-nucleus import
	regulation of DNA recombination
	spindle pole body and microtubule cycle (sensu Saccharomyces)
T70079	chemotaxis
	G-protein coupled receptor protein signaling pathway
	inflammatory response
T62835	cell adhesion
N22383	cell adhesion
	cell-matrix adhesion
	cell-substrate junction assembly
	integrin-mediated signaling pathway
AA056401	cellular morphogenesis
	epidermal differentiation
N63308	cell adhesion
	neuronal cell recognition
AA037351	cell adhesion
	neuronal cell recognition
AA045473	cell adhesion
N93476	cell adhesion
	G-protein coupled receptor protein signaling
W49619	cell adhesion
R80217	cell motility
	inflammatory response
	peroxidase reaction
	physiological processes
	prostaglandin metabolism
AA044993	cell adhesion
	cell growth and/or maintenance
	cell motility
	DNA metabolism
	epidermal differentiation

not have that kind of annotation. One way to solve this problem might be to use all genes for clustering, but calculate the validity index only with those that can be annotated. In this case, one might additionally think of giving a score to each cluster, indicating how many genes participate in the validity index. We will address this point in future work.

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. F. Azuaje. A cluster validity framework for genome expression data. *Bioinformatics*, 18(2):319–320, 2001.
2. T. Beißbarth and T. Speed. GOstat: find statistically overexpressed Gene Ontologies within groups of genes. *Bioinformatics*, 20(9):1464–1465, 2004.
3. N. Bolshakova, F. Azuaje, and P. Cunningham. An integrated tool for microarray data clustering and cluster validity assessment. *Bioinformatics*, 21(4):451–455, 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. J.C. Dunn. Well separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4:95–104, 1974.
7. M. Eisen, P. Spellman, D. Botstein, and P. Brown. Cluster analysis and display of genome-wide expression patterns. In *Proceedings of the National Academy of Sciences, USA*, volume 95, pages 14863–14867, 1998.
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. 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. 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.
13. Peter Merz. Clustering gene expression profiles with memetic algorithms. In *Proceedings of the 7th International Conference on Parallel Problem Solving from Nature, PPSN VII*, pages 811–820. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2002.

14. J. Quackenbush. Computational analysis of microarray data. *Nature Reviews Genetics*, 2(6):418–427, 2001.
15. W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66:846–850, 1971.
16. 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.
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. N.H. Shah and N.V. Fedoroff. CLENCH: a program for calculating Cluster ENriCHment using Gene Ontology. *Bioinformatics*, 20(7):1196–1197, 2004.
20. N. Speer, P. Merz, C. Spieth, and A. Zell. Clustering gene expression data with memetic algorithms based on minimum spanning trees. In *Proceedings of the IEEE Congress on Evolutionary Computation (CEC 2003)*, volume 3, pages 1848–1855. IEEE Press, 2003.
21. 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.
22. P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, and T.R. Golub. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. In *Proceedings of the National Academy of Sciences, USA*, volume 96, pages 2907–2912, 1999.
23. S. Tavazoie, J.D. Hughes, M.J. Campbell, R.J. Cho, and G.M. Church. Systematic determination of genetic network architecture. *Nature Genetics*, 22:281–285, 1999.
24. The Gene Ontology Consortium. The gene ontology (GO) database and informatics resource. *Nucleic Acids Research*, 32:D258–D261, 2004.
25. K.Y. Yeung, D.R. Haynor, and W.L. Ruzzo. Validating clustering for gene expression data. *Bioinformatics*, 17:309–318, 2001.