Validating Gene Clusterings by Selecting Informative Gene Ontology Terms with Mutual Information

Ivan G. Costa[1], Marcilio C. P. de Souto[2], Alexander Schliep[1]

[1] Max-Planck-Institut für Molekulare Genetik,
Ihnestr., 63 D-14195 Berlin, Germany
{ivan.filho,alexander.schliep}@molgen.mpg.de
[2] Dep. de Informática e Matemática Aplicada - UFRN
Campus Universitário, 59072-970 Natal (RN), Brazil
marcilio@dimap.ufrn.br

Abstract. We propose a method for global validation of gene clusterings. The method selects a set of informative and non-redundant GO terms through an exploration of the Gene Ontology structure guided by mutual information. Our approach yields a global assessment of the clustering quality, and a higher level interpretation for the clusters, as it relates GO terms with specific clusters. We show that in two gene expression data sets our method offers an improvement over previous approaches.

Key words: cluster validation, external index, gene ontology, mutual information

1 Introduction

With the advent of DNA microarrays there has been a great deal of work on clustering methods for the analysis of data from large-scale gene expression experiments. The main idea behind these approaches is to find clusters of co-expressed genes, providing biologists with genes regulated in a similar manner [9]. While most of these approaches yielded useful analysis of gene expression data, the evaluation of the biological relevance of the clusters is still a difficult task. There is little guidance available for choosing a clustering method [8]. There is also no established framework for evaluation of gene clusterings resulting from these methods exists.

The biological interpretation of clusters has been addressed, for instance, by comparing the results with available functional genomics data, such as provided by the Gene Ontology (GO) project [2] (see Section 2.1 for more details). One common approach is to search for GO terms (functional annotations) that are significantly enriched within a cluster of genes [3, 4]. Although this allows a biological interpretation of individual clusters of genes, it gives no global assessment on the "quality" of a gene clustering (or a set of clusters) returned by a clustering method.

Recently, there have been proposals of global indices for the validation of gene clusterings [6, 10]. However, in contrast to the approaches in [3, 4], these validation methods provide no "biological" interpretation of their assessments. Furthermore, they do not take several important features of GO into account. For example, the GO structure (direct acyclic graph or DAG) presents a parent-child relation, which implies that a term inherits all annotations of its immediate descendent [1, 11]. This makes the annotations of a GO term highly redundant with respect to terms "near" in the GO DAG. The use of redundant terms possibly introduces a bias in the global index, since contributions of GO terms that have many siblings will have a higher weight [10].

Motivated by the limitations presented above, we present a method that provides a global validation measure of gene clusterings. The method works by selecting a set of informative and non-redundant GO terms through an exploration of the Gene Ontology structure with the mutual information measure [7]. By informative, we mean terms that help to discriminate between clusters in a clustering. Additionally, by taking the parent-child relationship into account, our method detects a list of non-redundant GO terms within the informative ones. With this set of terms, we can calculate, as in [6, 10], a global fitness measure of the clustering. Furthermore, our method relates a set of informative GO terms to a particular cluster, which provides a biological interpretation of the results.

1.1 Related Work

One of the first applications that used GO for evaluating groups of genes was the so called GO Term Enrichment (`TE`) analysis. By means of a statistical test, such as the Fisher exact test, one can estimate a p-value indicating whether a significant fraction of genes in a cluster is annotated with a specific GO term [3, 4]. This approach has some limitations as it assumes independence between GO terms, and it suffers from the multiple testing problem [**?**]. More recent methods [1, 11] take the dependencies of GO terms caused by the parent-child relations into account. In particular, the Parent-Term Enrichment method (`PE`) [11] assumes that whenever a particular term is enriched, so are its parents. Thus, it yields a more refined selection of GO terms.

All these methods have been shown useful and have found widespread use in the interpretation of individual clusters of genes. However, as previously mentioned, they do not produce a global assessment of how "biologically relevant" a given gene clustering is.

A global index for evaluating gene clusterings with GO was presented in [10]. This index, based on an approximation of mutual information, is able to discriminate between results of clustering methods from random cluster assignments. In [6], an external index was proposed for a similar task. However, neither of these two approaches account for any biological interpretation of the results. A further extension of [10] was presented in [14], where an informative set of GO terms is collected and the exact mutual information is computed. This method, however, has a high computational cost. It is exponential in the number of selected GO terms. Thus, in practice, only a small set of GO terms can be chosen.

Putting our approach into perspective, it combines the characteristics of "global indices", such as [10], with the interpretability of the "local" approaches such as [1, 3, 11]. Also, in contrast to [14], we constrain the selection of terms within the GO structure, yielding a more efficient computational procedure. This also makes the identification of redundant GO terms possible, which decreases bias of the global index towards GO terms having many siblings.

2 Method

2.1 Gene Ontology

The Gene Ontology (GO) project is a collaborative effort to address the need for consistent descriptions of gene products in different databases [2]. Three structured controlled vocabularies (ontologies) describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner—cellular component describes components in which genes are active (e.g., *rough endoplasmic reticulum*); molecular function contains concepts related to gene function (e.g., *catalytic activity*); and biological process describes the processes that a gene can take part of (e.g., *cellular physiological process*).

More formally, a given Gene Ontology (GO) is represented by a directed acyclic graph (DAG), in which each node t_i in a set $T = \{t_1, ..., t_N\}$ represents a biological term (controlled vocabulary or GO term) and the edges stand for a set of relationships $\mathcal{R}$ among these terms. A relationship $R(t_i, t_j) \in \mathcal{R}$ means that term t_i is a parent of term t_j. Such a relation is interpreted as t_j is a subclass of t_i—i.e., t_i is a more general concept than t_j. For instance, the biological term "*cell cycle*" is related to the more specific terms "*mitotic cell cycle*" and "*meiotic cell cycle*".

A set of genes $G = \{g_1, ..., g_M\}$ is related to a given GO by an annotation set $\mathcal{A}$, where $A(t_i, g_m) \in \mathcal{A}$ indicates that gene g_m is annotated with term t_i. Genes often have multiple biological roles, so they are usually annotated with several GO terms. Furthermore, the parent-child relation of GO implies that genes annotated to a term are also annotated to all parents of this term. That is, for all $R(t_i, t_j) \in \mathcal{R}$, given a gene g_m, $A(t_j, g_m) \rightarrow A(t_i, g_m)$.

2.2 Selecting Informative GO Terms by Mutual Information Gain

In order to select a set of non-redundant and informative GO terms, we explore the DAG structure of GO and the parent-child relation. By informative terms we refer to terms that help to discriminate a cluster from others in a clustering. This can be measured with the mutual information, which is a general measure of dependence between two random variables [7]. In our case, the mutual information provides a systematic quantitative measure of the relationship between cluster membership and GO term membership of a set of genes. We call redundant terms the ones that annotate a similar set of genes. Recall the parent-child

relation $R(t_i, t_j)$, as t_i also annotates all terms t_j annotates, we expect that t_i is informative whenever t_j is.

Our selection procedure—called `MutSel`—works bottom-up as follows. For a given GO, a set of genes G and its respective annotation set $\mathcal{A}$, we start with a candidate collection of terms S with unitary sets, each one containing a leaf node (a node of the DAG without descendants). Such a collection corresponds to the most specific annotations present in GO for genes in G. From these we calculate the gain in mutual information, with respect to the cluster membership, when joining each set $\mathbf{s}_i \in S$ either with other adjacent (or neighboring) set or with parent terms not included in the candidate sets S.

The set of adjacency relations, $\mathcal{D}$, is defined by the parent-child relation, where sets $\mathbf{s}_p$ and $\mathbf{s}_q$ are adjacent, $D(\mathbf{s}_p, \mathbf{s}_q)$, if and only if there exists terms $t_i \in \mathbf{s}_p$ and $t_j \in \mathbf{s}_q$, such that $R(t_i, t_j) \in \mathcal{R}$ or $R(t_j, t_i) \in \mathcal{R}$. At each step, we select the pair of adjacent sets that yields the higher non-negative mutual information gain, joining them in a new set of terms. This step is equivalent to looking for more general terms in the GO DAG, which are more informative to the clustering results. We repeat this step until no mutual information gain is possible.

More formally, let X^p be a discrete random variable with alphabet $\mathcal{X} = \{0, 1\}$ representing the annotation of $\mathbf{s}_p$, where an observation x takes the value 1 if a term in $\mathbf{s}_p$ annotates it, or zero otherwise. Respectively, the random variable Y with alphabet $\mathcal{Y} = \{1, ..., K\}$ represents the cluster assignment, where a observation y takes value k if it belongs to cluster k. The mutual information gain, $\mathrm{MIG}(X^p, X^q|Y)$, of joining two adjacent sets $\mathbf{s}_p$ and $\mathbf{s}_q$ in the context of cluster membership Y is defined as

$$\mathrm{MIG}(X^p, X^q|Y) = \mathrm{MI}(X^p \vee X^q, Y) - \mathrm{MI}(X^p, Y) - \mathrm{MI}(X^q, Y), \tag{1}$$

where MI denotes the mutual information, and $X^p \vee X^q$ the variable resulting in the union of sets s_p and s_q. The mutual information, MI, is defined as,

$$\mathrm{MI}(X^i, Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \mathbf{P}[X^i = x, Y = y] \log \left(\frac{\mathbf{P}[X^i = x, Y = y]}{\mathbf{P}[X^i = x]\mathbf{P}[Y = y]} \right), \tag{2}$$

$\mathrm{MI}(X^i, Y) \geq 0$, with equality only if both variables X^i and Y are independent.

For a given set of genes G, we have a set of observations $\{x_1^i, ..., x_M^i\}$, where $x_m^i = 1$ if t^i annotates gene m, 0 otherwise. Respectively, we have a set of observations $\{y_1, ..., y_M\}$, where $y_m = k$ denotes that gene m belongs to cluster k. From these observations, we can obtain the following estimates for computing $\mathrm{MI}(X^i, Y)$,

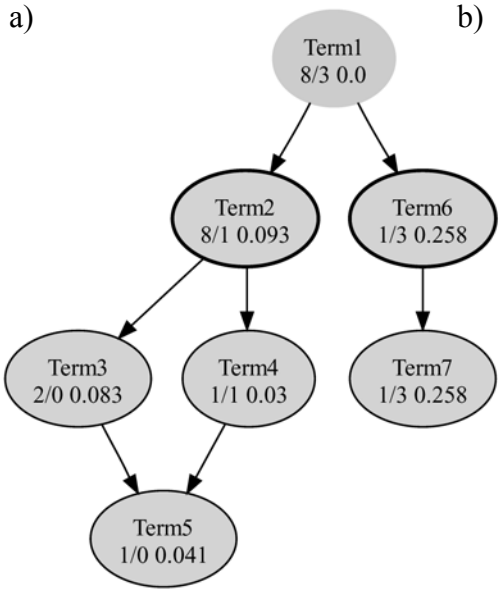
$$\mathbf{P}[X^i = j, Y = k|G] = \frac{1}{M} \sum_{m=1}^{M} 1\{x_m^i = j\} 1\{y_m = k\}, \tag{3}$$

$$\mathbf{P}[Y = k|G] = \frac{1}{M} \sum_{m=1}^{M} 1\{y_m = k\} \tag{4}$$

where 1 is a indicator function, $j \in \mathcal{X}$ and $k \in \mathcal{Y}$.

Figure 1 illustrates our method. On the left (Figure 1 (a)), we depict a simple example of a DAG with 7 terms. In Figure 1(b), we display a table, where the rows corresponds to the random variables X^i and the columns the genes from set G. An one in position (i, j) indicates that gene j is annotated with term i. The last line, Y, indicates the assignment of genes to one of the two the clusters considered. At each node of the DAG in Figure 1(a), we display the cluster counts and the mutual information of the respective term. For example, in Term7, "1/3" means that this term annotates one gene from cluster 1 and three genes from cluster 2. The value 0.258 corresponds to the mutual information. Terms with good discriminative power in relation to Y display a higher MI (e.g., Term2 and Term7) than non-discriminative terms (e.g Term1 and Term4).

a)



b)

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}	g_{11}
X^1	1	1	1	1	1	1	1	1	1	1	1
X^2	1	1	1	1	1	1	1	1	1	0	0
X^3	1	1	0	0	0	0	0	0	0	0	0
X^4	1	0	0	0	0	0	0	0	1	0	0
X^5	1	0	0	0	0	0	0	1	0	0	0
X^6	0	0	0	0	0	0	0	1	1	1	1
X^7	0	0	0	0	0	0	0	1	1	1	1
Y	1	1	1	1	1	1	1	1	2	2	2

c)

Terms Groups	MIG
$X^3 \vee X^5$	0.04
$X^4 \vee X^5$	-0.01
$X^3 \vee X^4 \vee X^5$	0.08
$X^2 \vee X^3 \vee X^5$	0.01
$X^1 \vee X^2 \vee X^3 \vee X^5$	-0.09
$X^6 \vee X^7$	0.00
$X^1 \vee X^6 \vee X^7$	-0.25

Fig. 1. We depict on the left (a) an example of a simple DAG, on the top right (b) a table describing the terms annotations to a set of 11 genes and on the bottom right (c) a list of candidate join operations and the respective MIG.

Starting with a collection $S = \{\mathbf{s}_1, ..., \mathbf{s}_P\}$ such that $\mathbf{s}_p = \{t_l\}$ where t_l is a leaf from GO DAG, and $\mathcal{D}$ is the adjacency list, the algorithm works as follows:

1. while $\max_{D(\mathbf{s}_i, \mathbf{s}_j) \in \mathcal{D}} MIG(X^i, X^j|Y) \geq 0$ do
2. $\quad D(\mathbf{s}_p, \mathbf{s}_q) = \arg\max_{D(\mathbf{s}_i, \mathbf{s}_j) \in \mathcal{D}} MIG(X^i, X^j|Y)$
3. $\quad \text{join}(\mathbf{s}_p, \mathbf{s}_q)$
4. $\quad \text{update}(\mathcal{D})$

The algorithm returns a collection S of groups of GO terms. Given that we join only parent terms, each of these groups constitutes a sub-DAG from GO. From these, we can select the most general terms, or the terms without any parent node within a group $\mathbf{s}_p$ as the representative term(s) of $\mathbf{s}_p$. All other terms in the group can be considered as redundant, since they will carry the same or less information than the representative terms. Furthermore, we can also relate a given group of terms $\mathbf{s}_p$ with a cluster in k'

$$k' = \arg\max_{k \in \mathcal{Y}} \mathrm{MI}(X^p, Y = k). \tag{5}$$

Figure 1(c) illustrates a simple example of the method. There, we display the MIG from joining candidate sets of terms. The selection method starts with the leaf nodes Term5 and Term7. It then looks for neighboring terms, whose unions with the leaves has non-negative MIG. For example, Term5 has Term3 and Term4 as parents. While joining Term4 and Term 5 ($X^4 \vee X^5$) yields a negative MIG, merging Term3 and Term5 ($X^3 \vee X^5$) produces a positive MIG. Thus, the latter are chosen. In the end, the method returns two groups of terms {Term2,Term3,Term4,Term5} and {Term6,Term7}: the former is related to cluster 1 and the latter to cluster 2. From these groups, the method selects Term2 and Term6 as representative terms, since they constitute the most general terms within these groups; and the other terms in the sets {Term3,Term4,Term5} and {Term7} are regarded as uninformative, since their annotations are also present in the informative terms Term2 and Term6.

2.3 Validation Index

We use the index proposed in [10] to obtain a global measure of fitness by comparing a clustering (partition) with the set of terms selected with `MutSel`. Again, we have a random variable Y defining the clustering results, and the random variables $\{X^1, ..., X^p, ..., X^P\}$ corresponding to the annotation vectors of group of terms selected above. The measure in based on the approximation of the joint mutual information $\mathrm{MI}^{app}(X, Y)$ as proposed in [10],

$$\mathrm{MI}^{app}(X, Y) = \sum_{p=1}^{P} \mathrm{MI}(X^p, Y). \tag{6}$$

As discussed in [14], this approximation assumes independence between variables from X, which does not hold for most selections of GO terms, given the high dependency between GO term annotations. An alternative to improve the approximation of Eq. 6 is to select a set of terms with low annotation redundancy. To tackle this problem, [10] introduces a parameter U, also based in the mutual information, which excludes redundant terms from the computation. The smaller the value of U is, the less redundancy will be allowed in the set of terms X used for computing Eq. 6. Note that `MutSel` joins terms displaying dependencies caused by the parent-child property of GO annotations in a principled fashion, automatically excluding redundant terms and requiring no extra parameter.

To quantify deviation from randomness, we compute a z-score by repeating the `MutSel` procedure with random cluster assignments as performed in [10]. The random clusterings are draw with the same cluster size distribution as the evaluated clustering. More formally, from a given real clustering Y, its selection of GO terms X, a random clustering Y^r, its selection of GO terms X^r, then we have,

$$z\mathrm{MI}^{app} = \frac{\mathrm{MI}^{app}(X,Y) - \mu^r}{\sigma^r}. \tag{7}$$

where $\mu^r = \mathrm{Mean}(\mathrm{MI}^{app}(X^r, Y^r))$ is the mutual information mean for L random clusterings and $\sigma^r = \mathrm{Var}(\mathrm{MI}^{app}(X^r, Y^r))^{1/2}$ is the standard deviation of the mutual information from L random clusterings. Hereafter, we refer to z^{MI} as the index proposed [10], and z^{MutSel} as the index from Eq. 7 after selection of GO terms by `MutSel`.

3 Experiments

We evaluate our method on two typical scenarios of gene expression data analysis. First, we inspect the selection of GO terms in a differential gene expression analysis, where a group of induced and a group of repressed genes after treatment of yeast were identified [12]. This data, where two clusters of genes are given beforehand and no clustering analysis is needed, allow us to evaluate the "biological relevance" of the selection of GO terms, since the biological processes behind these two clusters are well characterized. In the second experiment, we perform a small scale comparison of clustering methods on a yeast cell cycle data set. This data set has been manually labeled [5], allowing us to compare our index and the prior approach [10] to the expert manual annotation.

3.1 Yeast Treatment (`YT`)

Gene expression of yeast was measured at particular time points after the treatment with sulfometuron methyl (SM) [12]. We use a group of 241 induced genes and a group of 121 repressed genes 4h after treatment with 5μg/ml of SM. This clustering gives a simple scenario to evaluate our method, since the biological processes behind theses two clusters are well characterized [12].

3.2 Yeast Cell Cycle (`YCC5`)

This dataset represents the expression levels of over 6,000 genes during two cell cycles from Yeast measured in 17 time points [5]. We used a subset `YCC5`, of 384 genes visually identified to peak at five distinct time points [5], each representing a distinct phase of cell cycle (Early G1, Late G1, S, G2 and M). Hereafter, this subset will be referred to as `YCC5`. The expression values of each gene were standardized, which can enhance the performance of model-based clustering methods, when the original data consists of intensity levels.

In relation to the clustering methods, we performed analysis with hierarchical clustering (`Hier`) [9], k-means [**?**], mixture of multivariate Gaussians with diagonal covariance matrix (`MixGaus`) [**?**] and mixtures of Hidden Markov models (`MixHMM`) [13]. We set the number of clusters to be equal to 5 in all methods (as this is the number of classes in the manual annotation). For k-means, `MixGauss` and `MixHMM`, we initialize models randomly, perform clustering 15 times, and selected the solution with minimal error criteria (see [13] for details). For k-means and hierarchical clustering, we used Pearson correlation as a similarity measure.

4 Results

4.1 GO Term Selection

In order to evaluate our method with respect to the selection of "biologically relevant" GO terms, we use the set of repressed and induced genes from the study on response of yeast to a inhibitor of amino acid synthesis [12] introduced in Section 3.1. Table 1 depicts the top five informative GO terms, from the Biological Process GO, for the induced genes (first five rows), as well as for the repressed ones (last five rows). The columns represent the GO term id, the GO term name, the counts of induced genes, the counts of repressed genes, and the mutual information.

As highlighted in [12], induced genes were mainly related to molecule transport, amino acid biosynthesis and nitrogen metabolism. Indeed, all terms from Table 1, with exception of "*vitamin biosynthetic process*", are directly related to these processes. Among the repressed genes, the study detected genes related to carbohydrate and lipid biosynthesis, translation, cell cycle and ribosome. All terms listed in Table 1 bottom are either directly related or more general terms describing these processes.

Table 1. Top five informative GO terms, from the Biological Process GO, for induced (top) and repressed (bottom) genes

Term ID	Term Name	#I	#R	MI
GO:0006807	nitrogen compound metabolic process	68	14	0.022
GO:0009110	vitamin biosynthetic process	14	0	0.022
GO:0006519	amino acid and derivative metabolic process	60	13	0.018
GO:0016769	transferase activity, transferring nitrogenous groups	11	0	0.017
GO:0009059	macromolecule biosynthetic process	16	36	0.051
GO:0051301	cell division	3	10	0.019
GO:0008610	lipid biosynthetic process	4	11	0.018
GO:0022613	ribonucleoprotein complex biogenesis and assembly	2	7	0.014
GO:0044265	cellular macromolecule catabolic process	9	14	0.013

We also compare, in the context of `YT`, the GO terms selected with `MutSel` with the ones obtained with well-known methods, such as the Term Enrichment

(TE) [3] and the Parent-Term enrichment (PE) [11]. Table 2 summarizes this comparison. Its rows correspond to, respectively, the set of induced and repressed genes in the dataset YT. The columns of the first part correspond to, respectively, the number of terms selected with PE (p-value lower then 0.05), the number of terms selected with MutSel, and the intersection of both sets. Likewise, in following columns, we present the number of terms selected with TE (p-value lower then 0.05), the number of all terms (informative and redundant) selected with MutSel (we refer to this set as MutSelAll), and the intersection of both sets.

Analyzing the results presented in Table 2, the informative terms selected by our method are mainly a smaller subset of genes enriched in PE; 85% of terms related to the cluster of induced genes and 81% of terms related to the cluster of repressed genes detected by MutSel are also selected in PE. Likewise, the result obtained with TE, which does not filter redundant terms, is comparable to the set of all terms (informative and redundant) selected by MutSel. Again, the terms indicated by the MutSelAll was a small subset of PE; 84% for the cluster of induced genes and 100% for the cluster of repressed genes.

To further investigate the distinction between these methods, we measure the redundancy of annotation of GO terms. For two GO terms, redundancy can be measure by computing their mutual information (MI): redundant terms have higher mutual information values. More precisely, we compute the mutual information between all pairs of GO terms from a given set, select the maximum MI for each term and average the values. For the cluster of induced genes, terms obtained with MutSel, MutSelAll, PE, and TE had a MI mean of, respectively, 0.154, 0.219, 0.271, and 0.275. For the set of repressed genes, these values were, respectively, 0.097, 0.168, 0.198, and 0.185. In both cases, the methodologies taking the parent-child property into account displayed lower MI than their counterparts. In general, MutSel presented lower MI values, which demonstrates its ability to select a set of non-redundant terms.

Table 2. Comparison of the number of GO terms selected with MutSel, MutSelAll, TE and PE in the analysis of dataset YT.

	PE	MutSel	∩	TE	MutSelAll	∩
Induced	41	13	11	79	39	33
Repressed	79	22	18	159	80	80

4.2 Comparison of Clustering Methods

We display in Table 3, for dataset YCC5, the rankings of the results from the four clustering methods, according to the different indices. More precisely, we list the rank of the methods according to z^{MI} for five choices of U and z^{MutSel}. After each method name we display the mean values for 10 replications of the

z score. The last line corresponds to the corrected Rand (CR) [**?**] of comparing the clustering assignment with the manual labeling. We used the original implementation to obtain values from z^{MI} (available at `http://llama.med.harvard.edu/cgi/ClusterJudge/cluster_judge.pl`).

Table 3. We list the rank of the methods given by the indices z^{MI} for several choices of U, z^{MutSel}, and CR comparing the clustering assignment with the manual labeling.

Indices	Rank 1	Rank 2	Rank 3	Rank 4
z^{MI} $U = 0.8$	k-means (3.26)	`MixHMM` (2.87)	`Hier.` (2.86)	`MixGaus` (2.32)
z^{MI} $U = 0.4$	k-means (3.74)	`Hier.` (2.91)	`MixHMM` (2.87)	`MixGaus` (2.26)
z^{MI} $U = 0.2$	k-means (1.41)	`MixHMM` (0.27)	`MixGaus` (0.06)	`Hier.` (-0.17)
z^{MI} $U = 0.1$	k-means (0.86)	`MixGaus` (0.37)	`MixHMM` (0.36)	`Hier.` (-0.1)
z^{MI} $U = 0.01$	k-means (1.4)	`MixGaus` (0.83)	`Hier.` (0.64)	`MixHMM` (-0.1)
z^{MutSel}	k-means (1115.3)	`MixGaus` (1034.0)	`MixHMM` (791.9)	`Hier.` (616.3)
CR	k-means (0.5)	`Hier.` (0.46)	`MixGaus` (0.43)	`MixHMM` (0.39)

In general, k-means was ranked as the first one by all indices. In contrast, all others ranking positions differed from index to index. One important result that can be observed in this table is the impact of parameter U, the uncertainty index, in the values obtained by z^{MI} [10] and on the resulting rankings. For instance, for higher U values, where some redundancy in annotation is allowed, hierarchical clustering was ranked second; for more stringent values of U (i.e., 0.1 and 0.2), the result of this algorithm presented a negative z^{MI} score, which indicates results obtained by chance. These results contradict the claims in [10], where the authors state that the parameter U had small influence on the rankings of methods. In comparison to z^{MI}, z^{MutSel} yielded higher z-scores. This is explained by the fact that for random clusterings, `MutSel` makes very few merging operations. In this situation, the resulting selection of terms is mainly composed of leaf terms with few annotated genes. These terms have also very low information regarding Y. In other words, `MutSel` can easily discriminate clusterings from random generated ones.

No index was able to recover the ranking given by CR. Although we cannot take the annotation used to calculate the CR as the actual and only "ground truth" for dataset `YCC5`, since it was made via visualization of profiles, such an annotation still provides a basis for comparing the clusterings. With regard to z^{MutSel}, the difference was mainly in the ranking of the hierarchical clustering. An inspection of the contingency table, cluster against annotation labels, shows that hierarchical clustering placed genes that correspond to two different classes of the manual annotation (phases S and G2) into a single cluster, and had a small cluster with 10 genes from all distinct classes. On the other hand, the other clustering solutions had no such small cluster. This indicates that z^{MutSel} penalize this merge of groups S and G2 more strongly than CR. On the other

hand, z^{MI} did not yield a definitive solution, while its rankings vary from values of U. Furthermore, for lower Us, the value of the index for the hierarchical clustering are negative, which indicates that its results are comparable with a random solution. This strongly contradicts the CR values derived from the manual annotation. As manual annotation is usually not provided in the majority of gene expression data sets, z^{MutSel} represents a better alternative to z^{MI}, since it requires no extra parameters, while it selects the set of most informative and non-redundant terms.

5 Conclusion

In this paper, we present the `MutSel` method for computing a global validity measure of a clustering of genes. The main advantage of this method is a selection of relevant and non-redundant terms in relation to the evaluated clustering. In order to do so, we use of a characteristic intrinsic to Gene Ontology (GO), the parent-child relation, which makes annotations of GO terms highly redundant. The set of informative and non-redundant GO terms resulted from the application of `MutSel` yields not only a global index of "biological validity" of the clustering, but it also relates GO terms to clusters yielding a "biological interpretation" of individual clusters .

A comparison of `MutSel` to established methods for providing interpretation of a cluster of genes, such as Term Enrichment analysis and Parent-Term Enrichment analysis, showed that `MutSel` mainly selects a set of GO terms also found to be relevant by these methods. Furthermore, the set of selected terms has a lower degree of annotation redundancy.

In relation to a global evaluation index for clusterings, we show that the selection of terms from `MutSel` improves the mutual information-based measure proposed in [10]. Our experimental results show that the choice of parameters of the original index [10] has a great impact on the resulting rankings of clustering methods. Thus, `MutSel` represents an improvement to the original proposal, as it requires no parameter settings, while its results are consistent with manual annotation of genes in a benchmark data set. As an extension of this work, we plan to accomplish a large scale evaluation, including more clustering methods and gene expression data sets.

Acknowledgments. The first author would like to acknowledge funding from the DAAD/CNPq (Brazil).

References

1. Adrian Alexa, Jorg Rahnenfuhrer, and Thomas Lengauer. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, 22(13):1600–1607, 2006.
2. Michael Ashburner. Gene ontology: tool for the unification of biology. *Nat Genet*, 25(1):25–29, 2000.

3. Tim Beissbarth and Terence P. Speed. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, 20(9):1464–1465, 2004.
4. Elizabeth I. Boyle, Shuai Weng, Jeremy Gollub, Heng Jin, David Botstein, J. Michael Cherry, and Gavin Sherlock. GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, 20(18):3710–3715, 2004.
5. R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart, and R. W. Davis. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell*, 2(1):65–73, Jul 1998.
6. I. Costa and A. Schliep. On external indices for mixtures: validating mixtures of genes. In M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nurnberger, and W. Gaul, editors, *From Data and Information Analysis to Knowledge Engineering*, pages 662–669. Springer 2005, 2005.
7. T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley - Interscience, 1991.
8. P. D'haeseleer. How does gene expression clustering work? *Nat Biothech*, 24(12):1499–1501, 2005.
9. Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, and David Botstein. Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95(25):14863–14868, 1998.
10. Francis D. Gibbons and Frederick P. Roth. Judging the Quality of Gene Expression-Based Clustering Methods Using Gene Annotation. *Genome Res.*, 12(10):1574–1581, 2002.
11. Steffen Grossmann, Sebastian Bauer, Peter N. Robinson, and Martin Vingron. *An Improved Statistic for Detecting Over-Represented Gene Ontology Annotations in Gene Sets*. 2006.
12. Melissa H. Jia, Robert A. LaRossa, Jian-Ming Lee, Antoni Rafalski, Ellen DeRose, Greg Gonye, and Zhixiong Xue. Global expression profiling of yeast treated with an inhibitor of amino acid biosynthesis, sulfometuron methyl. *Physiol. Genomics*, 3(2):83–92, 2000.
13. Alexander Schliep, Ivan G. Costa, Christine Steinhoff, and Alexander Schonhuth. Analyzing gene expression time-courses. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 2(3):179–193, 2005.
14. Ralf Steuer, Peter Humburg, and Joachim Selbig. Validation and functional annotation of expression-based clusters based on gene ontology. *BMC Bioinformatics*, 7(1):380, 2006.