Incorporating biological domain knowledge into cluster validity assessment

Nadia Bolshakova[1], Francisco Azuaje[2], and Pádraig Cunningham[1]

[1] Department of Computer Science, Trinity College Dublin, Ireland
{nadia.bolshakova, padraig.cunningham}@cs.tcd.ie
[2] School of Computing and Mathematics, University of Ulster, Jordanstown, BT37 0QB, U.K
fj.azuaje@ulster.ac.uk

Abstract. This paper presents an approach for assessing cluster validity based on similarity knowledge extracted from the Gene Ontology (GO) and databases annotated to the GO. A knowledge-driven cluster validity assessment system for microarray data was implemented. Different methods were applied to measure similarity between yeast genes products based on the GO. This research proposes two methods for calculating cluster validity indices using GO-driven similarity. The first approach processes overall similarity values, which are calculated by taking into account the combined annotations originating from the three GO hierarchies. The second approach is based on the calculation of GO hierarchy-independent similarity values, which originate from each of these hierarchies. A traditional node-counting method and an information content technique have been implemented to measure knowledge-based similarity between genes products (biological distances). The results contribute to the evaluation of clustering outcomes and the identification of optimal cluster partitions, which may represent an effective tool to support biomedical knowledge discovery in gene expression data analysis.

1 Introduction

Over the past few years DNA microarrays have become a key tool in functional genomics. They allow monitoring the expression of thousands of genes in parallel over many experimental conditions (e.g. tissue types, growth environments). This technology enables researchers to collect significant amounts of data, which need to be analysed to discover functional relationships between genes or samples. The results from a single experiment are generally presented in the form of a data matrix in which rows represent genes and columns represent conditions. Each entry in the data matrix is a measure of the expression level of a particular gene under a specific condition.

A central step in the analysis of DNA microarray data is the identification of groups of genes and/or conditions that exhibit similar expression patterns. Clustering is a fundamental approach to classifying expression patterns for biological and biomedical applications. The main assumption is that genes that

are contained in a particular functional pathway should be co-regulated and therefore should exhibit similar patterns of expression [1]. A great variety of clustering algorithms have been developed for gene expression data. The next data analysis step is to integrate these numerical analyses of co-expressed genes with biological function information. Many approaches and tools have been proposed to address this problem at different processing levels. Some methods, for example, score whole clustering outcomes or specific clusters according to their biological relevance, other techniques aim to estimate the significance of over-represented functional annotations, such as those encoded in the Gene Ontology (GO), in clusters [2], [3], [4], [5]. Some approaches directly incorporate biological knowledge (e.g. functional, curated annotations) into the clustering process to aid in the detection of relevant clusters of co-expressed genes involved in common processes [6], [7]. Several tools have been developed for ontological analysis of gene expression data (see review by Khatri and Drăghici [8], for instance) and more tools are likely to be proposed in the future.

The prediction of the correct number of clusters in a data set is a fundamental problem in unsupervised learning. Various cluster validity indices have been proposed to measure the quality of clustering results [9], [10]. Recent studies confirm that there is no universal pattern recognition and clustering model to predict molecular profiles across different datasets. Thus, it is useful not to rely on one single clustering or validation method, but to apply a variety of approaches. Therefore, combination of GO-based (knowledge-driven) validation and microarray data (data-driven) validation methods may be used for the estimation of the number of clusters. This estimation approach may represent a useful tool to support biological and biomedical knowledge discovery.

We implemented a knowledge-driven cluster validity assessment system for microarray data clustering. Unlike traditional methods that only use (gene expression) data-derived indices, our method consists of validity indices that incorporate similarity knowledge originating from the GO and a GO-driven annotation database. We used annotations from the *Saccharomyces Genome Database* (SGD) (October 2005 release of the GO database). A traditional node-counting method proposed by Wu and Palmer [11] and an information content technique proposed by Resnik [12] were implemented to measure similarity between genes products. These similarity measurements have not been implemented for clustering evaluation by other research.

The main objective of this research is to assess the application of knowledge-driven cluster validity methods to estimate the number of clusters in a known data set derived from *Saccharomyces cerevisiae*.

2 The GO and cluster validity assessment

The automated integration of background knowledge is fundamental to support the generation and validation of hypotheses about the function of gene products. The GO and GO-based annotation databases represent recent examples of such knowledge resources. The GO is a structured, shared vocabulary that

allows the annotation of gene products across different model organisms. The GO comprises three independent hierarchies: molecular function (MF), biological process (BP) and cellular component (CC). Researchers can represent relationships between gene products and annotation terms encoded in these hierarchies. Previous research has applied GO information to detect over-represented functional annotations in clusters of genes obtained from expression analyses [13]. It has also been suggested to assess gene sequence similarity and expression correlation [14]. For a deeper review of the GO and its applications, the reader is referred to its website (http://www.geneontology.org) and Wang et al. [14].

Topological and statistical information extracted from the GO and databases annotated to the GO may be used to measure similarity between gene products. Different GO-driven similarity assessment methods may be then implemented to perform clustering or to quantify the quality of the resulting clusters. Cluster validity assessment may consist of data- and knowledge-driven methods, which aim to estimate the optimal cluster partition from a collection of candidate partitions [15]. Data-driven methods mainly include statistical tests or validity indices applied to the data clustered. A data-driven, cluster validity assessment platform was previously reported by Bolshakova and Azuaje, [9], [10]. We have previously proposed knowledge-driven methods to enhance the predictive reliability and potential biological relevance of the results [15].

Traditional GO-based cluster description methods have consisted of statistical analyses of the enrichment of GO terms in a cluster. Currently, there is a relatively large number of tools implementing such an approach [8]. At the same time, this approach is severely limited in certain regards (for detailed review on ontological analysis see by Khatri and Drăghici [8]). For instance, overestimation of probability values describing over-representation of terms. This may be due to the lack of more complete knowledge or the incorporation of biased datasets to make statistical adjustments and detect spurious associations. However, the application of GO-based similarity to perform clustering and validate clustering outcomes has not been widely investigated. A recent contribution by Speer et al. [16], [17] presented an algorithm that incorporates GO annotations to cluster genes. They applied data-driven Davies-Bouldin and Silhouette indices to estimate the quality of the clusters.

This research applies two approaches to calculating cluster validity indices. The first approach processes overall similarity values, which are calculated by taking into account the combined annotations originating from the three GO hierarchies. The second approach is based on the calculation of independent similarity values, which originate from each of these hierarchies. The second approach allows one to estimate the effect of each of the GO hierarchies on the validation process.

3 GO-based similarity measurement techniques

For a given pair of gene products, g_1 and g_2, sets of GO terms $T_1 = t_i$ and $T_2 = t_j$ are used to annotate these genes. Before estimating between-gene similarity it

is first necessary to understand how to measure between-term similarity. We implemented GO-based between-term similarity using a traditional approach proposed by Wu and Palmer [11] and an information content technique proposed by Resnik [12].

3.1 Wu and Palmer-based method

Similarity was defined by Wu and Palmer [11] as follows:

$$sim(t_i, t_j) = \frac{2N}{N_i + N_j + 2N} \tag{1}$$

where N_i and N_j are the number of links (edges) from t_i and t_j to their closest common parent in the GO hierarchy, T_{ij}, and N is the number of links from T_{ij} to the GO hierarchy root.

This similarity assessment metric may be transformed into a distance, d, metric:

$$d(t_i, t_j) = 1 - sim(t_i, t_j) \tag{2}$$

then the average inter-set similarity value across each pair of t_i and t_j is computed [13]:

$$d(g_k, g_m) = \underset{i,j}{\text{avg}}(d(t_{ki}, t_{mj})) \tag{3}$$

This between-term distance aggregation may then be used as an estimate of the GO-based similarity between two genes products g_k and g_m, which is defined as:

$$d(g_k, g_m) = \underset{i,j}{\text{avg}}(1 - \frac{2N}{N_{ki} + N_{mj} + 2N}) \tag{4}$$

3.2 Resnik-based similarity measurement

This similarity was defined by Resnik [12] as follows:

$$sim(t_i, t_j) = max(-log(p(T_{ij}))) \tag{5}$$

T_{ij} is defined as above and has the highest information value V defined as $-log(p(T_{ij}))$,where $p(T_{ij})$ is a probability, of finding term T_{ij} (or its descendants) in the dataset of genes under study, i.e. the SGD in this study.

Such similarity assessment metric may be transformed into a distance metric:

$$d(t_i, t_j) = \frac{1}{1 + sim(t_i, t_j)} \tag{6}$$

Based on the average value across each pair of t_i and t_j, as computed by Azuaje and Bodenreider [13], the GO-based similarity between two genes products g_1 and g_2 is defined as:

$$d(g_k, g_m) = \underset{i,j}{\text{avg}}\left(\frac{1}{1 + max(-log(p(T_{kmij})))}\right) \quad (7)$$

In this research we first study an approach based on the aggregation of similarity information originating from all three GO. We also proposed and implemented three hierarchy - specific similarity assessment techniques, each based on information individually extracted from each GO hierarchy (BP, MF and CC).

4 Clustering and cluster validation methods

4.1 Clustering

The data analysed in this paper comprised yeast genes described by their expression values during the cell cycle [18]. Previous research has shown that disjoint clusters of genes are significantly associated with each of the five cell cycle stages: **early G1, late G1, S, G2, M**. Several cluster partitions (with numbers of clusters from two to six clusters), obtained with the k-means algorithm, were analysed to estimate the optimum number of clusters for this dataset. Clustering was performed with the Machaon CVE tool [10].

4.2 Cluster validation methods

Cluster validation was performed using two validity indices: the C-index [19] and the Goodman-Kruskal index [20], whose data-driven versions have been shown to be effective cluster validity estimators for different types of clustering applications. Nevertheless, each of the implemented validation methods has their advantages and limitations. For example, Goodman-Kruskal index is expected to be robust against outliers because quadruples of patterns are used for its computation. However, its drawback is its high computational complexity in comparison, for example, with the C-index.

C-index The *C-index* [19], C, is defined as follows:

$$C = \frac{S - S_{min}}{S_{max} - S_{min}} \quad (8)$$

where S, S_{min}, S_{max} are calculated as follows. Let p be the number of all pairs of samples (conditions) from the same cluster. Then S is the sum of distances between samples in those p pairs. Let P be a number of all possible pairs of samples in the dataset. Ordering those P pairs by distances we can select p pairs with the smallest and p pairs with the largest distances between samples. The sum of the p smallest distances is equal to S_{min}, whilst the sum of the p largest is equal to S_{max}. From this formula it follows that the nominator will be small if pairs of samples with small distances are in the same cluster. Thus, small values of C correspond to good clusters. We calculated distances using the knowledge-driven methods described above. The number of clusters that minimize *C-index* is taken as the optimal number of clusters, c.

Goodman-Kruskal index For a given dataset, $X_j(j = 1, , k$, where k is the total number of samples (gene products in this application), j, in the dataset, this method assigns all possible quadruples [20]. Let d be the distance between any two samples (w and x, or y and z) in X_j. A *quadruple* is called *concordant* if one of the following two conditions is true:

$d(w, x) < d(y, z)$, w and x are in the same cluster and y and z are in different clusters.

$d(w, x) > d(y, z)$, w and x are in different clusters and y and z are in the same cluster.

By contrast, a *quadruple* is called *disconcordant* if one of following two conditions is true:

$d(w, x) < d(y, z)$, w and x are in different clusters and y and z are in the same cluster.

$d(w, x) > d(y, z)$, w and x are in the same cluster and y and z are in different clusters.

We adapted this method by calculating distances using the knowledge-driven methods described above.

A good partition is one with many *concordant* and few *disconcordant quadruples*. Let N_{con} and N_{dis} denote the number of *concordant* and *disconcordant quadruples*, respectively. Then the *Goodman-Kruskal index, GK*, is defined as:

$$GK = \frac{N_{con} - N_{dis}}{N_{con} + N_{dis}} \tag{9}$$

Large values of GK are associated with good partitions. Thus, the number of clusters that maximize the *GK index* is taken as the optimal number of clusters, c.

5 Results

The clustering algorithm was applied to produce different partitions consisting of 2 to 6 clusters each. Then, the validity indices were computed for each of these partitioning results. The two GO-based similarity assessment techniques introduced above were used for all cases to calculate biological distances between the genes.

Tables 1 to 4 show the predictions made by the validity indices at each number of clusters. Bold entries represent the optimal number of clusters, c, predicted by each method. In the tables the first cluster validity index approach processes overall GO-based similarity values, which are calculated by taking into account the combined annotations originating from the three GO hierarchies. The other indices are based on the calculation of independent similarity values, independently obtained from each of the GO hierarchies.

The C-indices based on Resnik similarity measurement and similarity information from the MF, BP and the combined hierarchies indicated that the optimal number of clusters is $c = 5$, which is consistent with the cluster structure expected [18]. The C-indices based on Wu and Palmer similarity measurement

Table 1. C-index predictions based on Wu and Palmer's GO-based similarity metric for expression clusters originating from yeast data

Validity indices based on:	c=2	c=3	c=4	c=5	c=6
Combined hierarchies	0.51	0.472	0.464	**0.453**	0.463
Biological process	0.501	0.321	0.259	**0.235**	0.237
Molecular function	0.501	0.32	0.274	**0.243**	0.272
Cellular component	**0.514**	0.586	0.602	0.614	0.615

Table 2. C-index values predictions based on Resnik's GO-based similarity estimation technique for expression clusters originating from yeast data

Validity indices based on:	c=2	c=3	c=4	c=5	c=6
Combined hierarchies	0.504	0.395	0.373	**0.349**	0.369
Biological process	0.503	0.321	0.261	**0.234**	0.243
Molecular function	0.501	0.32	0.278	**0.25**	0.29
Cellular component	**0.517**	0.645	0.69	0.723	0.759

Table 3. Goodman-Kruskal index values used Wu and Palmer's similarity metric for expression clusters originating from yeast data

Validity indices based on:	c=2	c=3	c=4	c=5	c=6
Combined hierarchies	-0.023	-0.01	-0.018	**0.004**	-0.017
Biological process	-0.013	0.005	-0.005	**0.034**	0.018
Molecular function	-0.02	0.009	0.005	**0.066**	-0.026
Cellular component	-0.025	**-0.022**	-0.032	-0.046	-0.025

Table 4. Goodman-Kruskal index values used Resnik's similarity metric for expression clusters originating from yeast data

Validity indices based on:	c=2	c=3	c=4	c=5	c=6
Combined hierarchies	-0.026	-0.001	-0.02	**0.016**	-0.01
Biological process	-0.018	0.014	-0.012	**0.055**	0.044
Molecular function	-0.02	0.012	0.004	**0.087**	-0.016
Cellular component	-0.025	-0.035	**-0.024**	-0.037	-0.025

and similarity information from the MF and BP indicated that the optimal number of clusters is $c = 5$. In all cases only the method based on the CC hierarchy suggested the partition with two clusters as the optimal partition, which confirms that cellular localization information does not adequately reflect relevant functional relationships in this dataset.

For the Goodman-Kruskal method again only the method based on the CC hierarchy suggested the partition different from $c = 5$ as the optimal partition.

6 Accompanying tool

The approaches described in this paper are available as part of the *Machaon CVE* (Clustering and Validation Environment) [10]. This software platform has been designed to support clustering-based analyses of expression patterns including several data- and knowledge-driven cluster validity indices. The program and additional information may be found at http://www.cs.tcd.ie/ Nadia.Bolshakova/GOtool.html

7 Conclusion

This paper presented an approach to assessing cluster validity based on similarity knowledge extracted from the GO and GO-driven functional databases. A knowledge-driven cluster validity assessment system for microarray data clustering was implemented. Edge-counting and information content approaches were implemented to measure similarity between genes products based on the GO. Edge-counting approach calculates the distance between the nodes associated with these terms in a hierarchy. The shorter the distance, the higher the similarity. The limitation is that it heavily relies on the idea that nodes and links in the GO are uniformly distributed.

The research applies two methods for calculating cluster validity indices. The first approach process overall similarity values, which are calculated by taking into account the combined annotations originating from the three GO hierarchies. The second approach is based on the calculation of independent similarity values, which originate from each of these hierarchies. The advantage of our method compared to other computer-based validity assessment approaches lies in the application of prior biological knowledge to estimate functional distances between genes and the quality of the resulting clusters. This study contributes to the development of techniques for facilitating the statistical and biological validity assessment of data mining results in functional genomics.

It was shown that the applied GO-based cluster validity indices could be used to support the discovery of clusters of genes sharing similar functions. Such clusters may indicate regulatory pathways, which could be significantly relevant to specific phenotypes or physiological conditions.

Previous research has successfully applied C-index using knowledge-driven methods (GO-based Resnik similarity measure) [15] to estimate the quality of the clusters.

Future research will include the comparison and combination of different data- and knowledge-driven cluster validity indices. Further analyses will comprise, for instance, the implementation of permutation tests as well as comprehensive cluster descriptions using significantly over-represented GO terms.

The results contribute to the evaluation of clustering outcomes and the identification of optimal cluster partitions, which may represent an effective tool to support biomedical knowledge discovery in gene expression data analysis.

8 Acknowledgements

This research is partly based upon works supported by the Science Foundation Ireland under Grant No. S.F.I.-02IN.1I111.

References

1. Fitch, J., Sokhansanj, B.: Genomic engineering: Moving beyond DNA. Sequence to function. In: Proceedings of the IEEE. Volume 88. (2000) 1949–1971
2. Gat-Viks, I., Sharan, R., Shamir, R.: Scoring clustering solutions by their biological relevance. Bioinformatics **19**(18) (2003) 2381–2389
3. Lee, S., Hur, J., Kim, Y.: A graph-theoretic modeling on go space for biological interpretation on gene clusters. Bioinformatics **20**(3) (2004) 381–388
4. Goeman, J., van de Geer, S., de Kort, F., van Houwelingen, H.: A global test for groups of genes: testing association with a clinical outcome. Bioinformatics **20**(1) (2004) 93–99
5. Raychaudhuri, S., Altman, R.: A literature-based method for assessing the functional coherence of a gene group. Bioinformatics **19**(3) (2003) 396–401
6. Hanisch, D., Zien, A., Zimmer, R., Lengauer, T.: Co-clustering of biological networks and gene expression data. Bioinformatics **18** (2002) S145–S154
7. Sohler, F., Hanisch, D., Zimmer, R.: New methods for joint analysis of biological networks and expression data. Bioinformatics **20**(10) (2004) 1517–1521
8. Khatri, P., Drăghici, S.: Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics **21**(18) (2005) 3587–3595
9. Bolshakova, N., Azuaje, F.: Cluster validation techniques for genome expression data. Signal Processing **83**(4) (2003) 825–833
10. Bolshakova, N., Azuaje, F.: Machaon CVE: cluster validation for gene expression data. Bioinformatics **19**(18) (2003) 2494–2495
11. Wu, Z., Palmer, M.: Verb semantics and lexical selection. In: 32nd Annual Meeting of the Association for Computational Linguistics, New Mexico State University, Las Cruces, New Mexico (1994) 133–138
12. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI). (1995) 448–453
13. Azuaje, F., Bodenreider, O.: Incorporating ontology-driven similarity knowledge into functional genomics: an exploratory study. In: Proceedings of the fourth IEEE Symposium on Bioinformatics and Bioengineering (BIBE 2004). (2004) 317–324
14. Wang, H., Azuaje, F., Bodenreider, O., Dopazo, J.: Gene expression correlation and gene ontology-based similarity: An assessment of quantitative relationships. In: Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology. (2004) 25–31

15. Bolshakova, N., Azuaje, F., Cunningham, P.: A knowledge-driven approach to cluster validity assessment. Bioinformatics **21**(10) (2005) 2546–2547
16. Speer, N., Spieth, C., Zell, A.: 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), IEEE Press (2004) 252–259
17. Speer, N., Spieth, C., Zell, A.: Functional grouping of genes using spectral clustering and gene ontology. In: Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN 2005), IEEE Press (2005) 298–303
18. Cho, R., Campbell, M., Winzeler, E., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T., Gabrielian, A., Landsman, D., Lockhart, D., Davis, R.: A genomewide transcriptional analysis of the mitotic cell cycle. Molecular Cell **2** (1998) 65–73
19. Hubert, L., Schultz, J.: Quadratic assignment as a general data-analysis strategy. British Journal of Mathematical and Statistical Psychologie (1976) 190–241
20. Goodman, L., Kruskal, W.: Measures of associations for cross-validations. Journal of Ameracan Statistical Association (1954) 732–764