CarGene: Characterisation of sets of genes based on metabolic pathways analysis

Jesus S. Aguilar-Ruiz,
Domingo S. Rodriguez-Baena*,
and Norberto Diaz-Diaz*

School of Engineering,
Pablo de Olavide University,
41013 Seville, Spain,
E-mail: aguilar@upo.es
E-mail: dsrodbae@upo.es
E-mail: ndiaz@upo.es
*Corresponding authors

Isabel A. Nepomuceno-Chamorro

Computer Science Department,
University of Seville,
41012 Seville, Spain
E-mail: inepomuceno@us.es

Abstract: The great amount of biological information provides scientists with an incomparable framework for testing the results of new algorithms. Several tools have been developed for analysing gene-enrichment and most of them are Gene Ontology-based tools. We developed a Kyoto Encyclopedia of Genes and Genomes (Kegg)-based tool that provides a friendly graphical environment for analysing gene-enrichment. The tool integrates two statistical corrections and simultaneously analysing the information about many groups of genes in both visual and textual manner. We tested the usefulness of our approach on a previous analysis (Huttenshower et al.). Furthermore, our tool is freely available (http://www.upo.es/eps/bigs/cargene.html).

Keywords: clustering; biclustering; Kegg; biological validation; gene-enrichment.

Reference to this paper should be made as follows: Aguilar-Ruiz, J.S., Rodrignez-Baena, D.S., Diaz-Diaz, N. and Nepomuceno-Chamorro, I.A. (2011) 'CarGene: Characterisation of sets of genes based on metabolic pathways analysis', *Int. J. Data Mining and Bioinformatics*, Vol. 5, No. 5, pp.558–573.

Biographical notes: Jesus S. Aguilar-Ruiz is an Associate Professor at Pablo de Olavide University, Seville, Spain. He leads a research group on data mining and bioinformatics, and has published over 150 papers in international conferences and journals. Currently, he is the Dean of the School of Engineering and the co-Editor-in-Chief of the *BioData Mining Journal*.

Domingo S. Rodriguez-Baena is an Assistant Professor at Pablo de Olavide University, Seville, Spain. His main interests include data mining techniques, as biclustering and clustering, applied to gene expression datasets.

Norberto Diaz-Diaz is a Lecturer at Pablo de Olavide University, Seville, Spain. His main focus is upon the extraction of interesting knowledge from gene expression data and biological evaluation.

Isabel A. Nepomuceno-Chamorro is a lecturer at the University of Seville, Spain. Her research interests include microarray analysis, inference of gene networks, network-based prognostic approaches and systems-level approach for translational bioinformatics.

1 Background

Clustering techniques have proven to be useful to elucidate gene function, gene regulation and cellular processes and play an important role in the analysis of gene expression data (Dougherty et al., 2002; Eisen et al., 1998). Biclustering techniques search for subsets of genes that show similar activity patterns under a subset of experimental conditions (Cheng and Church, 2000). The development of new clustering and biclustering techniques is very appealing in bioinformatics due to the ability of extracting hidden and potentially useful biological knowledge. However, to prove whether a new technique is valid indeed, it is necessary to evaluate whether or not the results are interesting. The definition of 'interesting' in this realm usually incorporates some analytical criterion, although it is also important the practical aspect, related to the biological relevance. Most evaluation techniques (Azuaje, 2002; Halkidi et al., 2001) quantify the quality of clusters in terms of metrics that do not consider the prior biological relevance of the groups of genes. For instance, Bolshakova and Azuaje (2003) present a cluster validation tool that partitions samples or genes into groups characterised by similar expression patterns, and evaluates the quality of the clusters obtained by means of several measures such as the C-index and the Davis–Bouldin's measure. Instead, other approaches include biological knowledge in the evaluation process (Gat-Viks et al., 2003). The main aim of these approaches is to find whether a group of genes obtained by a clustering or biclustering technique reveals any unknown biological property. In this context, the great amount of biological information stored in public databases provides to scientists an incomparable framework for verifying and testing the results of these algorithms. Hence, the integration of these databases with searching techniques in order to automatically validate results from different sources is a key factor in bioinformatics. Several tools have been developed, such as FuncAssociate (Berriz et al., 2003), GoSurfer (Zhong et al., 2004), GO::TermFinder (Boyle et al., 2004), GOstat (Falcon and Gentleman, 2007) and David (Huang et al., 2007), which are mainly based on Gene Ontology (GO) (The Gene Ontology Consortium, 2000; Khatri and Drăghici, 2005). Other approaches use some functionality from the Kyoto Encyclopedia of Genes and Genomes (Kegg) (Ogata et al., 1999; Kanehisa and Goto, 2000; Kanehisa et al., 2006, 2007), such as the analysis of carbohydrate

sugar chains (KCaM) (Aoki et al., 2004), the pathway mapping of prokaryotic genomes (PathwayVoyager) (Altermann and Klaenhammer, 2005), or dynamic visualisation of pathway diagrams (KGML-ED) (Klukas and Schreiber, 2007).

In this work, we introduce a software tool, called CarGene, that validates sets of genes by means of verified biological knowledge. However, unlike other tools mentioned above, the validation is carried out by using the biological knowledge on metabolic pathways stored in the Kegg database instead of GO. The Kegg database contains pathway maps representing the current knowledge on the molecular interaction and reaction networks for metabolism (e.g., carbohydrate, energy, or lipid), genetic information processing (e.g., transcription, translation, or degradation), environmental information processing (e.g., membrane transport or signal transduction), cellular processes (e.g., cellular motility, cell growth and death, endocrine system, or immune system), and human diseases (e.g., cancers, immune disorders, neurodegenerative or infectious diseases). CarGene offers interesting analytical, visualisation and statistical features, which will be described further. A great effort has been put on creating an user-friendly and well-designed interface that shows results graphically and textually, processes folders at once, and analyses statistically many results generated by several algorithms.

Next, we first describe the features of CarGene, the methods of extracting and representing the biological data, the statistical tests incorporated in the tool, and a brief remark on the software technology. Later, the sets of genes provided by eight clustering and biclustering techniques are statistically compared over one of the six datasets from the study of Huttenhower et al. (2007). Finally, the main conclusions are summarised.

2 Implementation

CarGene is designed to retrieve information from Kegg and process it in order to validate several (bi)clustering results by contrasting them with the biological information. One of the most important features of CarGene is the possibility of simultaneously comparing and statistically analysing the information about many groups of genes in both visual and textual manner. Furthermore, it includes its own browser to explore in detail the information extracted from Kegg.

The main functionalities of CarGene will be described as follows: firstly, how to query the Kegg database; secondly, the statistical measures used in data processing are described; thirdly, how to interpret graphical outputs and use the CarGene browser; and finally, some details on the software technology that supports the project.

2.1 Extracting biological information

The user can extract biological information from Kegg by using two types of database queries. The first type is based on the genes and pathways from a specific organism. In the Kegg database there is genomic information about a large amount of organisms, which is daily updated, increasing in about 10 organisms per month. When an organism is selected, CarGene retrieves all the genes and the metabolic pathways associated with it. Using this information, the user can obtain reports on the set of pathways with which a selected gene or list of genes is involved.

The second type of query is based on the verification of the degree of coherence of a group of genes obtained from (bi)clustering techniques with respect to the involvement of those genes in metabolic pathways stored in Kegg. In order to use this functionality, users can refer to (bi)clusters in two different ways. In the first way, each set of genes is stored in a file with the next format: ORF names in single column (each row stands for an element of the set). In the second one, all the (bi)clusters can be stored in a single file in which each ORF name is associated with the number of the (bi)cluster it belongs to. Thus, CarGene is adapted to the output format of Expander (Shamir et al., 2005), one of the most important bioinformatics resources used to apply (bi)clustering techniques to microarray datasets. In both file formats there are no constraints about the name or the extension, and comments are allowed by using the symbol % at the beginning of the line. Henceforth, we will refer to the results of clustering or biclustering techniques as *clusters*, since we will only deal with the set of genes contained in those results (ignoring the subsets of experimental conditions involved in biclusters).

The user can select a group of files (each one related to a single set of genes) or even a complete folder (containing all the files), and start the execution (with a given name). The groups of genes will be validated by means of Fisher's exact test (the reader may refer to Rivals et al. (2007) for a survey). Finally, the reported p-values are adjusted using Bonferroni or Westfall-Young procedures, both implemented in CarGene.

2.2 Statistical measures

For each run, CarGene uses the information provided by Kegg about metabolic pathways to measure the enrichment level of the input genes included in clusters. In other words, CarGene accepts or refuses the following null hypothesis: "to belong to an input set of genes Q is independent of belonging to a specific metabolic pathway P". CarGene calculates the probability of finding at least or at most m genes in the pathway P in a query of length q (number of genes in Q) if the null hypothesis is true. This probability (p-value) is computed using the Fisher's exact test, which employs exact hypergeometric probabilities (Sheskin, 2004). If the p-value is equal to or less than a level of significance α, then the null hypothesis is refuted, which means that the input set of genes Q is dependent of belonging to a specific pathway P. However, if we test a null hypothesis which is in fact true, using $\alpha = 0.05$, we have a probability of 0.95 of coming to the conclusion that the null hypothesis is true. If we test two independent true null hypotheses, the probability of coming to the conclusion that the two null hypotheses are true is $0.95 \times 0.95 = 0.90$. This probability decreases with the number of null hypotheses. To solve this problem, multiple testing corrections are used to adjust the p-values.

The Bonferroni correction (Bland and Altman, 1995) is a single-step procedure that adjusts the level of significance $\alpha' = \frac{\alpha}{k}$, where k is the number of hypotheses. An individual null hypothesis is rejected when its p-value is smaller than α'. Some tools use Bonferroni or Holm adjustments to the computed p-values to correct multiple hypothesis. However, if there is dependency among hypotheses – in this case, among Kegg pathways – other tests might be more appropriate (Berriz et al., 2003; Gibbons and Roth, 2002). Unlike the Bonferroni correction, in which each p-value is corrected independently, the Westfall and Young (1989) correction takes advantage

of the dependence structure between pathways, by permuting all the pathways at the same time. This procedure estimates an adjusted p-value from the result of 1000 simulated null hypothesis queries by using Monte Carlo simulations.

2.3 Representing the biological information

Once the biological information referred to a certain input set is retrieved from the Kegg database, CarGene processes that information and represents it in a detailed and friendly way. The main goal is to provide researchers with resources that can be used to compare and validate quickly set of genes (or in other words, to be able to reject bad results easily). The user has the possibility of visualising either simultaneously or individually the results related to the input clusters. The former possibility is provided by the *global view mode*, whereas the latter by the *single view mode*. Furthermore, the user has a complete report of all the executions in the *textual summary*, from which the most promising results can be filtered.

2.3.1 Global view mode

The aim of this view is to show in a single window all the executions carried out by different (bi)clustering methods and the results produced by every one of them (or a selected group). Three main parts can be distinguished in Figure 1: the execution three, the top graph and the bottom graph. To the left, the execution tree, in which three algorithms were run (Click, BiMax and Opsm), and the clusters obtained by them are depicted. Each (bi)clustering technique produced 6, 5 and 5 clusters, respectively. The user can choose which sets of genes have potential to be analysed in detail (in this case, the first two results from each technique).

When comparing different methods, a number of clusters are generated, and it is not easy to get an insight of the overall experimentation. The global view mode provides a quick, organised and visual comparison of the set of clustering results. For each method's execution, the results are ordered increasingly by the best adjusted p-value of each cluster (from top to bottom). For instance, for the Click algorithm, cluster3 has an adjusted p-value lower than that of the cluster5.

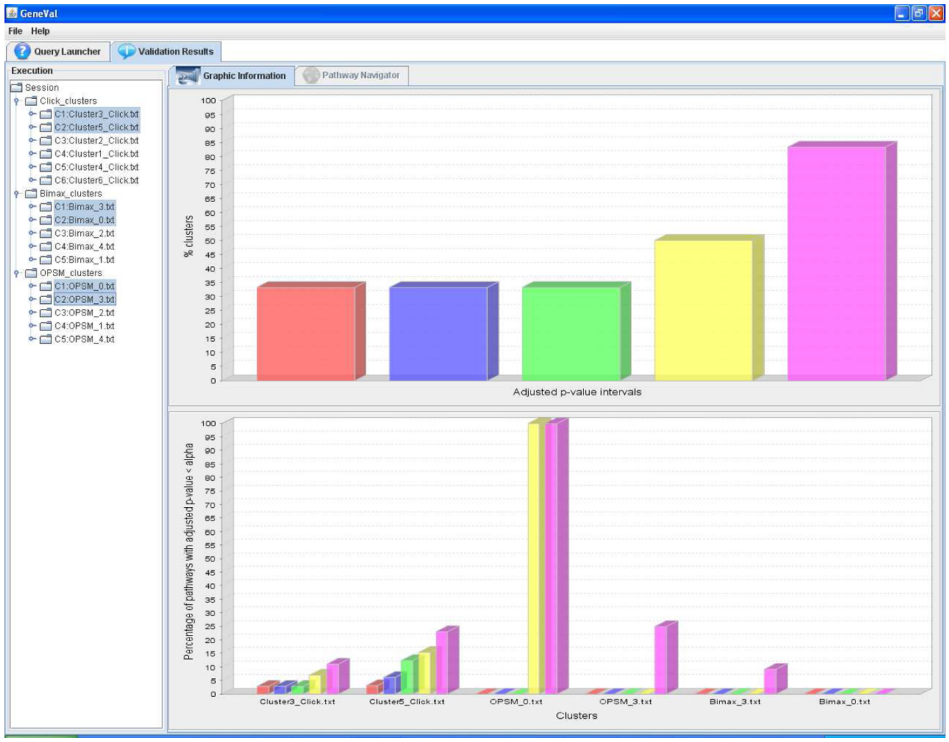
The graph at the top shows the percentage of clusters that contain any pathway with an adjusted p-value in one of the following intervals: [0,0.001), [0, 0.01), [0, 0.1), $[0, 1)$ and $[0, 5)$. The bars are cumulative in order to clearly determine the percentage of clusters whose p-values are lower than a given α. For instance, 33% of the selected clusters (2 out of 6) are significant ($\alpha \leq 0.001$). The first three 3d-bars show the same percentage value, so it means that over 33% of the clusters has at least one pathway with an adjusted p-value lower than 0.001 and none in the next two intervals. The fourth bar includes another cluster, achieving 50%. The last 3d-bar, with a percentage value near to 83%, tells us that about 33% of clusters has some pathway in the last interval. The remaining 16% corresponds to clusters (only one in this case) with at least one pathway having an adjusted p-value greater than 5.

At the bottom, the second graph shows the percentage of pathways with adjusted p-values in some of the aforementioned intervals. A five-bar graph is associated with each cluster, which can be selected directly by the users by clicking on the clusters they are interested in (from the execution tree to the left).

The first two clusters from three executions have been selected from the execution tree in Figure 1. The executions are referred to three different gene grouping

techniques applied to the same yeast microarray dataset. The first cluster, produced by Click (cluster3), presents around 2.74% of the pathways with p-value lower than 0.001 (see in the graph at the bottom the first red bar from the first five-bars graph, from left to right). Taking into account that the clusters are ordered according to the lowest adjusted p-values, all this information can be used to determine what group of genes are more significantly enriched by Kegg metabolic pathways. It can be also observed that the first two clusters from Click present better results than clusters from Opsm (Karp et al., 2002) and BiMax (Prelic et al., 2006) techniques, which only have pathways with an adjusted p-value equal or greater than 0.1 (see that the first three bars for Opsm and the first four bars for BiMax are not present). It is worth to note that the fact that the first bar of cluster3 is shorter than the first bar of cluster5 (both from Click) does not mean that cluster3 is worse than cluster5, as bars are representing the percentage of pathways with adjusted p-value lower than a given α. However, the clusters are ordered by the best adjusted p-value of any pathway in the cluster. Thus, the user has double information at a glance.

Figure 1 This is the global view mode in which the results of three different (bi)clustering techniques are shown. The best two clusters from three different runs have been selected in order to compare their results. The graph on top shows the percentage of the selected clusters that contains any pathway with an adjusted p-value in the following cumulative intervals: [0,0.001), [0, 0.01), [0, 0.1), [0, 1) and [0, 5). At the bottom graph, all the selected clusters appear ordered by the lowest adjusted p-value. For each cluster, the percentage of pathways with adjusted p-values in some of the aforementioned intervals is shown (see online version for colours)



CarGene also provides more information about executions. Associated with every one of them there is a menu with other visualisation options, such as full screen visualisation (from every input gene set: the percentage of genes in pathways, the percentage of genes with p-values in the respective intervals; from every present pathway: the percentage of genes in input sets, the percentage of pathways with adjusted p-values lower than α) and execution summary. In addition, CarGene offers a statistical summary of all the executions, including for each one the number of pathways involved and the adjusted p-values for each aforementioned interval.

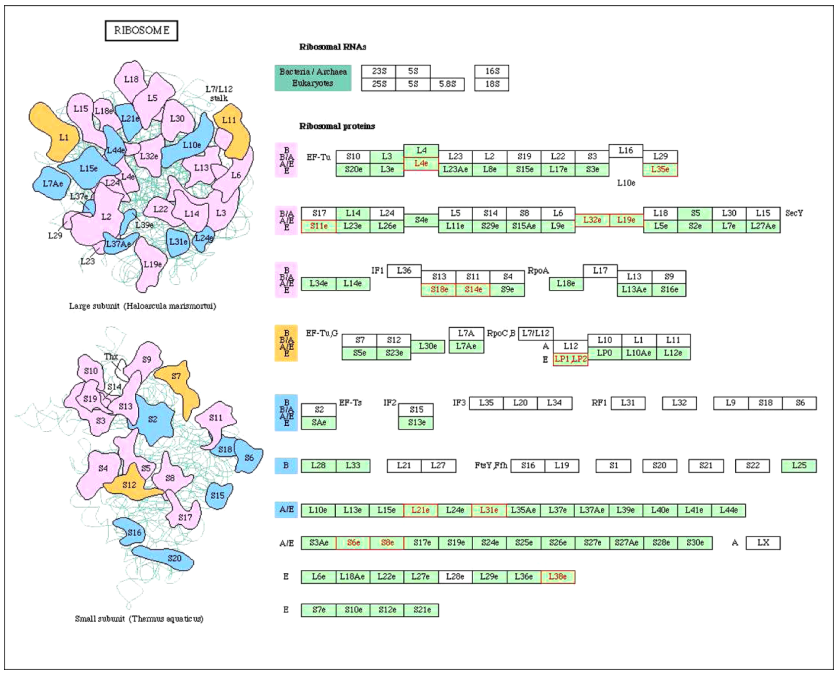
2.3.2 Single view mode

After a global view in which one gets an overall idea of the results, the analysis focus on a small number of interesting results detected previously. For every single cluster selected, the single view mode provides two different types of information.

The first type is related to the adjusted p-value of every biological pathway detected in the cluster. Like in the second type of graphics described in the former subsection, this view mode presents graphically the percentage of pathways in the cluster with an adjusted p-value in every one of the already known intervals. In addition, other type of graph is depicted, in which the real adjusted p-value is represented for every single biological pathway present in the cluster. A complete textual report, including the names of the genes associated with every pathway, the exact p-value and the adjusted p-value of every pathway, the date of execution, among others, is also available. These graphs and tables are provided in full screen format through the menu linked with every cluster, as well as other types of interesting illustrations, such as the percentage of genes from the cluster in every pathway, or the percentage of genes from the pathways that appear in the cluster.

The second type of information is depicted by the CarGene browser. Besides consulting all the information related to the verification of the degree of coherence of a group of genes with respect to the participation of these genes in metabolic pathways, the user can explore through every one of these pathways. Thus, CarGene provides its own browser and the user can view the map representation of every pathway associated with a certain cluster. For example, in Figure 2 the map representation of the ribosomal proteins associated with the Ribosome metabolic pathway is illustrated. This pathway has been detected in the clusters Cl-20 and Cl-37 from the Click clustering technique mentioned in previous subsections (see also Table 2, which indicates that their Bonferroni corrections are 3.16E-12 and 1.24E-1, respectively). For the sake of clarity, CarGene highlights those genes from the input set that are present in the pathways of Kegg (in red). In addition, the user can interact with these maps as with a web page, that is, all the links of this image can be used to navigate to other parts of Kegg web page. Navigation buttons are incorporated in order to make easier the use of this browser. For instance, the user has access to a more detailed description of L21e (in red in Figure 2), including *entry* (YBR191W), *gene name* (RPL21A), *definition* (Protein component of the large (60S) ribosomal subunit, nearly identical to Rp121Bp and has similarity to ret L21 ribosomal protein), *orthology* (K02889), *pathway* (sce03010), and more information like *class*, *motif*, *structure*, *position* or *AA seq.*

Figure 2 The representation map of ribosomal proteins generated by Kegg and provided by CarGene (see online version for colours)



2.3.3 *Statistical summary*

The possibility of comparing different results, concerning certain clusters or whole executions, is a very useful feature of CarGene. Moreover, a user can compare all the executions from CarGene explorer by means of a textual report that contains a statistical summary of them. In this report, for every execution, the following information is provided:

- number of different pathways detected in all the clusters (this information shows the variety of biological processes detected in the executions)
- percentage of clusters that present a given adjusted p-value (this information measures the quality of the clusters in relation to the biological processes).

2.4 *Software technology*

CarGene is written in Java language, with multithread technology, and uses a client Simple Object Access Protocol (SOAP), that is XML-based protocol for exchanging structured information in a decentralised, distributed environment. This means that users can launch several executions at the same time, so they can keep on working with the tool while these tasks are being processed. In addition, in multiprocessors platforms with n central processing units, users can launch n executions in parallel, improving the efficiency. The statistical component is also developed in Java and uses the Java Native Interface (JNI) framework to call native methods written in C.

3 Discussion

In order to illustrate the usefulness of CarGene, a test based on the study carried out by Huttenshower et al. (2007) has been analysed. The paper describes a graph-based algorithm (Nearest Neighbor Networks, NNN) used to generate sets of genes with similar expression profiles. The approach was tested by comparing the results of eight different clustering and biclustering methods using six different *Saccharomyces Cerevisiae* data sets.

To show the CarGene capabilities, the results of several learning methods applied to one out of the six *Saccharomyces Cerevisiae* data sets used in the previous paper are compared. Concretely, we selected the well-known yeast cell cycle dataset published by Spellman et al. (1998). The dataset recorded expression profiles of 6178 genes in different time points during the cell cycle.

With the aim of reproducing the same input data used in Huttenhower et al. (2007), 24 time instants corresponding to temperature–sensitive mutant conditions (CDC15) have been selected. Besides, genes with missing data for 50% or more of the conditions have been removed. Finally, any remaining missing values were replaced using KNNImpute application with $k = 10$ (Troyanskaya et al., 2001). Thus, the final dataset is composed by 5672 genes, each represented by an expression vector of length 24 containing no missing values.

Eight different clustering and biclustering methods, summarised in Table 1, have been applied to the Spellman's dataset. In this table, 'Technique' denotes the clustering or biclustering algorithm applied; 'Software' stands for the software used; finally, 'N^o (Bi)Clusters' represents the number of clusters or biclusters found (references are also included).

Table 1 Clustering and biclustering techniques

Technique (*Reference*)	*Software* (*Reference*)	N^o *(Bi) clusters*
NNN (Huttenhower et al., 2007)	Huttenhower et al. (2007)	36 clusters
Click (Sharan et al., 2003)	Expander Shamir et al. (2005)	49 clusters
Samba (Tanay et al., 2002)	Shamir et al. (2005)	31 biclusters
K–Means (Ball and Hall, 1967)	BiCat Barkow et al. (2006)	10 clusters
BiMax (Prelic et al., 2006)	Barkow et al. (2006)	463 biclusters
Opsm (Karp et al., 2002)	Barkow et al. (2006)	11 biclusters
Isa (Bergmann et al., 2003)	Barkow et al. (2006)	5 biclusters
CC (Cheng and Church, 2000)	Barkow et al. (2006)	20 biclusters

Summary of clustering and biclustering techniques describing the output for the yeast cell cycle dataset.

In addition, each technique has been run using default parameters except for NNN, where the parameter values were $g = 5$ and $n = 25$. Note that the first three techniques were used by Huttenhower et al. (2007), and their parameter values are proposed by these authors. Moreover, as it was selected in that work, Pearson's correlation has been used in those algorithms based on a distance measure.

Once the selected clustering and biclustering techniques have been applied to the yeast dataset, CarGene is used to compare the clusters obtained. With this aim, the Bonferroni and Westfall & Young correction for multiple hypothesis testing are chosen.

An overview of the analysis produced by CarGene can be observed in Figure 3, that shows in tabular format the global results for each algorithm, where 'Execution' denotes the name of the algorithm, 'Pathways' the number of different pathways that are involved in the results of the method, and the rest of columns stand for the percentage of clusters with one or more pathways with an adjusted p-value within each of the five intervals described in Section 2.3.1. Furthermore, the selected algorithms are ordered by the percentage of cluster in the most restrictive interval $[0, 0.001)$.

Figure 3 Comparative summary using the Bonferroni correction (see online version for colours)

Statistical Summary

Execution	Pathways	%cluster <0.001	%cluster <0.01	%cluster <0.1	%cluster <1	%cluster <5
SAMBA	88	19.35	29.03	41.93	77.42	96.77
OPSM	109	18.18	18.18	18.18	45.45	90.91
CLICK	114	12.24	14.29	16.33	40.82	79.59
KMEANS	114	10.0	30.0	60.0	70.0	90.0
BIMAX	76	8.42	8.42	14.04	24.41	72.14
ISA	71	0.0	0.0	60.0	60.0	80.0
CC	114	0.0	0.0	5.0	35.0	65.0
NNN	101	0.0	0.0	2.78	13.89	69.44

Focusing on the 'Pathways' column, we can observe that the NNN results are involved in 101 different pathways with only 36 clusters. The result is expected according to Huttenhower et al. (2007), since the authors stated that NNN is an efficient tool for extracting functionally diverse clusters from co-expression data. However, these results do not seem to be statistically significant, as none or very small percentage of clusters have adjusted p-value lower than 0.1. On the other hand, K-Means is involved in 114 pathways with only 10 clusters, and these clusters are not due to chance, given the percentage of significant clusters shown in columns 3–7. Particularly, the fifth column (%cluster< 0.1) presents an acceptable result, since 60% of the clusters are significant from the point of view of the coherent participation of genes in metabolic pathways.

The most relevant fact is that at first sight it can be noticed that the last three methods (Isa, CC and NNN) failed to provide any relevant result, as the percentages of significant clusters for the columns with thresholds 0.001 and 0.01, respectively, are null. Hence, these methods might be discarded for further analysis of this dataset.

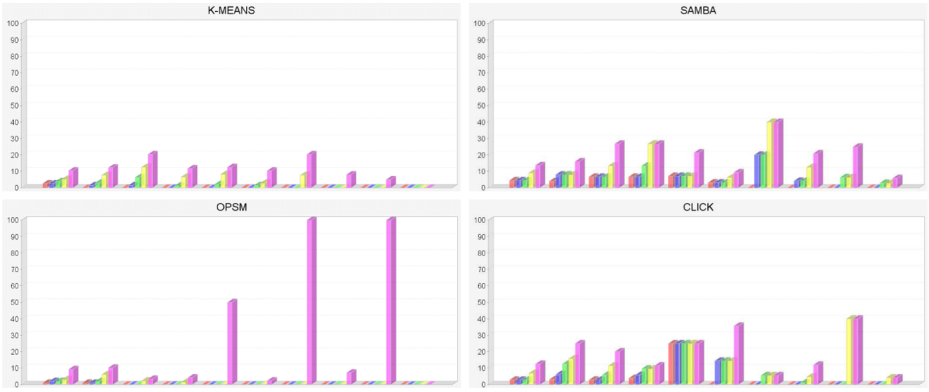
According to Figure 3, Samba has 19.35% clusters with at least one pathway in the first interval, followed by Opsm, Click and K-Means with values 18.18%, 12.24% and 10.0%, respectively. Although Samba appears at the first place, it does not mean that this technique always produces the best results. If we focus on the second interval, there K-Means triples its value and Samba increases up to 29.03.

In general, a good criterion is to focus on those techniques and results that are significant at level 0.001, and to analyse those clusters that are highly represented in metabolic pathways. It is likely that some genes in these clusters and not included in pathways have similar patterns than those belonging to pathways, which might

suggest to emphasise the role of these genes within the biological processes that they seem to be involved in.

In order to deeply analyse the results from the best four techniques, Figure 4 offers a visual comparison of the best 10 clusters of each of them. Each graph is automatically depicted by CarGene when marking the best 10 cluster of the methods. Figure 4 shows that K-Means has only one cluster with p-value lower than 0.01 and the rest of clusters might be due to chance at this statistical level. Opsm only has three clusters in the first two intervals, whereas Samba and Click generate greater number of clusters that are not due to chance. These three clusters are involved each in only one metabolic pathway. The results for Samba and Click are substantially better, as the first six clusters of both techniques are involved in metabolic pathways with p-values lower than 0.001. In fact, it is worth to note that Samba and Click succeed in producing precise clusters spanning a wider variety of biological processes. This is evidenced, for example, in the first cluster of both clustering techniques, which are significantly involved in two metabolic pathways (Proteasome and Ribosome, which are observed in the textual report from the single view mode).

Figure 4 The best 10 (bi)clusters from K-Means, Samba, Opsm and Click (see online version for colours)



Once the best techniques have been identified, the next step is to analyse the clusters generated by them. CarGene also offers detailed information of the best results. In Table 2 is summarised the information about Samba and Click executions for the best 10 results, which have been depicted in Figure 4. The names of the clusters are presented in column 'Technique', whereas 'Best pathway' and 'Correction' columns denote the best pathway name associated with the cluster and its adjusted p-value, respectively. For example, the best Samba bicluster ('Sa-8') presents a Bonferroni correction of 1.14E-08 associated with the Proteasome pathway. The best cluster of Click is 'Cl-3' and it is also associated with the Proteasome pathway with a Bonferroni correction of 1.23E-15. These two sets of genes are statistically significant under the Westfall and Young's correction. It is important to note that only for three cases (Sa-0, Sa-14 and Cl-17) the Westfall and Young correction was significant whereas the Bonferroni correction indicated the opposite.

In addition, CarGene offers more information about each cluster. For example, the cluster 'Cl-3' is involved in 73 metabolic pathways, and 71% of the genes in

the pathway *Proteasome* are present in the cluster, which is shown in Figure 5. This image has been obtained using CarGene *browser* and the 23 genes in red belong to the aforementioned cluster.

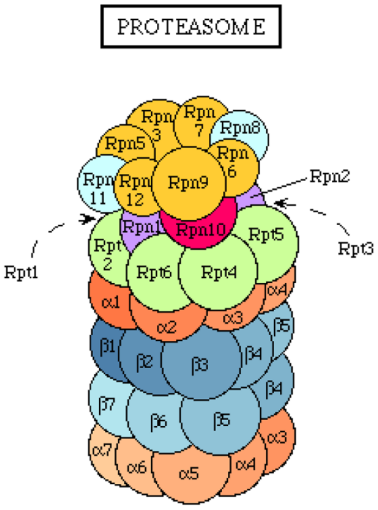
Finally, it is worth to mention that the number of known genes is much greater than the number of genes annotated in metabolic pathways. This means that only about 30% of genes of the best clusters are involved in known pathways. In spite of this ratio, some algorithms find groups of genes statistically significant with regard to the cohesive participation in biological processes. To discover the role of genes in clusters that are not found in pathways is an interesting problem. Recently, the assignment of putative functional annotations to unclassified genes has been addressed by means of biclustering techniques successfully (Bryan and Cunningham, 2008).

Table 2 Best 10 (bi)clusters found by Samba and Click

Tech.	*Bonferroni*		*Westfall and Young*	
Samba	*Best pathway*	*Correction*	*Best pathway*	*Correction*
Sa-8	Proteasome	1.14E-8	Proteasome	<0.001
Sa-7	Glycolysis/ Gluconeogenesis	6.54E-8	Glycolysis/ Gluconeogenesis	<0.001
Sa-1	DNA replication	8.45E-5	Cell cycle – yeast	<0.001
Sa-22	Proteasome	1.38E-4	Proteasome	<0.001
Sa-12	MAPK signalling pathway – yeast	3.06E-4	MAPK signalling pathway – yeast	<0.001
Sa-13	Proteasome	5.66E-4	Proteasome	<0.001
Sa-5	Glycolysis/ Gluconeogenesis	2.51E-3	Glycolysis/ Gluconeogenesis	1.00E-3
Sa-9	Arginine and proline metabolism	5.21E-3	Arginine and proline metabolism	2.00E-3
Sa-0	DNA replication	5.94E-3	DNA replication	<0.001
Sa-14	MAPK signaling pathway – yeast	1.44E-2	MAPK signaling pathway – yeast	<0.001
Cl-3	Proteasome	1.23E-15	Proteasome	<0.001
Cl-5	Glycolysis/ Gluconeogenesis	2.99E-15	Glycolysis/ Gluconeogenesis	<0.001
Cl-20	Ribosome	3.16E-12	Ribosome	<0.001
Cl-2	RNA polymerase	8.31E-5	Pyrimidine metabolism	<0.001
Cl-32	Cell cycle – yeast	6.15E-4	Cell cycle – yeast	<0.001
Cl-8	Pyrimidine metabolism	6.50E-4	Pyrimidine metabolism	<0.001
Cl-17	Cell cycle – yeast	1.26E-3	Cell cycle – yeast	<0.001
Cl-12	Basal T.F.	2.37E-2	Basal T.F.	6.00E-3
Cl-37	Ribosome	1.24E-1	Ribosome	2.40E-2
Cl-1	Pyrimidine metabolism	1.97E-1	Galactose metabolism	2.54E-1

Description of the best 10 (bi)clusters found by Samba and Click techniques, respectively, including Bonferroni, and Westfall and Young corrections.

Figure 5 71% of genes the Proteasome pathway are involved in the cluster Cl-3 (exactly 23 genes out of 32, which are represented in red) (see online version for colours)



Rpn1	Rpn2	Rpn3	Rpn4	Rpn5	Rpn6	
Rpn7	Rpn8	Rpn9	Rpn10	Rpn11	Rpn12	
Rpt1	Rpt2	Rpt3	Rpt4	Rpt5	Rpt6	
α1	α2	α3	α4	α5	α6	α7
β1	β2	β3	β4	β5	β6	β7

4 Conclusions

Assuming that the objective of clustering is to bring genes of similar function together, the best method of clustering a particular dataset will depend on the underlying clustering properties and the intrinsic characteristics of data. A priori, there is no universal clustering paradigm that work always better than others. The question "which is the clustering algorithm that provides better clusters for a particular dataset?" is not solved yet. Therefore, given the great variety of clustering algorithms that have been proposed, it makes sense for further biological analysis to generate as many clusters as possible and select those clusters that are potential high-quality candidates.

CarGene is intended to be an useful tool for rapidly characterising the results of different clustering or biclustering techniques over the same gene expression dataset.

CarGene can be used to draw conclusions from set of genes obtained by any grouping technique, evaluating the relevance of the coherence of groups of genes with respect to the participation of them in metabolic processes.

We have noticed that, in general, different methods might achieve similar results, however they do not contain exactly the same genes. For instance, Sa-8 and Cl-3 share many genes, even though at the level of a specific pathway, such as proteasome. Some genes remain at the intersection even when they do not belong to the pathway. Future work will be addressed in this direction, in order to elucidate the role of these common genes by means of other sources of biological knowledge.

Availability

The software is platform independent and available upon request.

Acknowledgements

This work has been funded by the Ministry of Science and Innovation, projects TIN2007-68084-C02-00, PCI2006-A7-0575 and by Junta de Andalucia, projects P07-TIC-02611 and TIC-200.

References

Altermann, E. and Klaenhammer, T. (2005) 'Pathwayvoyager: pathway mapping using the kyoto encyclopedia of genes and genomes (Kegg) database', *BMC Genomics*, Vol. 6, No. 1, pp.60–67.

Aoki, K., Yamaguchi, A., Ueda, N., Akutsu, T., Mamitsuka, H., Goto, S. and Kanehisa, M. (2004) 'KCAM (kegg carbohydrate matcher): a software tool for analyzing the structures of carbohydrate sugar chains', *Nucleic Acids Res.*, Vol. 32, pp.267–272.

Azuaje, F. (2002) 'A cluster validity framework for genome expression data', *Bioinformatics*, Vol. 18, No. 2, pp.319–320.

Ball, G. and Hall, D. (1967) 'A clustering technique for summarizing multivariate data', *Behavioral Sciences*, Vol. 12, No. 2, pp.153–155.

Barkow, S., Bleuler, S., Prelic, A., Zimmermann, P. and Zitzler, E. (2006) 'Bicat: a biclustering analysis toolbox', *Bioinformatics*, Vol. 22, No. 10, pp.1282–1283.

Bergmann, S., Ihmels, J. and Barkai, N. (2003) 'Iterative signature algorithm for the analysis of large-scale gene expression data', *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, Vol. 67, No. 2, Pt 1, p.03190201–18.

Berriz, G., King, O., Bryant, B., Sander, C. and Roth, F. (2003) 'Characterizing gene sets with FuncAssociate', *Bioinformatics*, Vol. 19, No. 18, pp.2502–2504.

Bland, J. and Altman, D. (1995) 'Multiple significance tests – the bonferroni method, 1.00,0.00,0.00', *British Medical Journal*, Vol. 310, pp.1–170.

Bolshakova, N. and Azuaje, F. (2003) 'Machaon CVE: cluster validation for gene expression data', *Bioinformatics*, Vol. 19, No. 18, pp.2494–2495.

Boyle, E., Weng, S., Gollub, J., Jing, H., Botstein, D., Cherry, J. and Sherlock, G. (2004) 'GO::TermFinder – open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes', *Bioinformatics*, Vol. 20, No. 18, pp.3710–3715.

Bryan, K. and Cunningham, P. (2008) 'Extending bicluster analysis to annotate unclassified orfs and predict novel functional modules via expression data', *BMC Genomics*, Vol. 9, Suppl. 2, p.s20.

Cheng, Y. and Church, G. (2000) 'Biclustering of expression data', *Proc. ISMB'00.*, San Diego, USA, pp.93–103.

Dougherty, E., Barrera, J., Brun, M., Kim, S., Cesar, R., Chen, Y., Bittner, M. and Trent, J. (2002) 'Inference from clustering with application to gene-expression microarrays', *Journal of Computational Biology*, Vol. 9, No. 1, pp.105–126.

Eisen, M., Spellman, P., Brown, P. and Botstein, D. (1998) 'Cluster analysis and display of genome-wide expression patterns', *Proceedings of the National Academy of Sciences*, Vol. 95, No. 25, p.14863.

Falcon, S. and Gentleman, R. (2007) 'Using GOstats to test gene lists for GO term association', *Bioinformatics*, Vol. 23, No. 2, p.257.

Gat-Viks, I., Sharan, R. and Shamir, R. (2003) 'Scoring clustering solutions by their biological relevance', *Bioinformatics*, Vol. 19, No. 18, pp.2381–2389.

Gibbons, F. and Roth, F. (2002) 'Judging the quality of gene expression-based clustering methods using gene annotation', *Genome Res.*, Vol. 12, pp.1574–1581.

Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2001) 'On Clustering Validation Techniques', *Journal of Intelligent Information Systems*, Vol. 17, No. 2, pp.107–145.

Huang, D.W.a.W., Sherman, B.T., Tan, Q., Kir, J., Liu, D., Bryant, D., Guo, Y., Stephens, R., Baseler, M.W., Lane, H.C. and Lempicki, R.A. (2007) 'DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists', *Nucleic Acids Research* (Web Server issue), Vol. 35, p.W169.

Huttenhower, C., Flamholz, A., Landis, J., Sahi, S., Myers, C., Olszewski, K., Hibbs, M., Siemers, N., Troyanskaya, O. and Coller, H. (2007) 'Nearest neighbor networks: clustering expression data based on gene neighborhoods', *BMC Bioinformatics*, Vol. 8, No. 1, pp.250–263.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T. and Yamanishi, Y. (2007) 'Kegg for linking genomes to life and the environment', *Nucl. Acids Res.*, Vol. 36, pp.gkm882+.

Kanehisa, M. and Goto, S. (2000) 'KEGG: Kyoto Encyclopedia of Genes and Genomes', *Nucleic Acids Res.*, Vol. 28, pp.27–30.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K., Itoh, M., Kawashima, S., Katayama, T., Araki, M. and Hirakawa, M. (2006) 'From genomics to chemical genomics: new developments in KEGG', *Nucleic Acids Res.*, Vol. 34, pp.D354–D357.

Karp, R., Ben-Dor, A., Chor, B. and Yakhini, Z. (2002) 'Discovering local structure in gene expression data: the order-preserving submatrix problem', *RECOMB '02: Proceedings of the Sixth Annual International Conference on Computational Biology*, ACM, New York, NY, USA, pp.49–57.

Khatri, P. and Drăghici, S. (2005) 'Ontological analysis of gene expression data: current tools', limitations, and open problems', *Bioinformatics*, Vol. 21, No. 18, pp.3587–3595.

Klukas, C. and Schreiber, F. (2007) 'Dynamic exploration and editing of Kegg pathway diagrams', *Bioinformatics*, Vol. 23, No. 3, pp.344–350.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. and Kanehisa, M. (1999) 'KEGG: Kyoto encyclopedia of genes and genomes', *Nucleic Acids Res.*, Vol. 27, pp.29–34.

Prelic, A., Bleuler, S., Zimmerman, P., Wille, A., Buhlmann, P., Gruissem, W., Hennig, L., Thiele, L. and Zitzler, E. (2006) 'A systematic comparison and evaluation of biclustering methods for gene expression data', *Bioinformatics*, Vol. 22, No. 9, pp.1122–1129.

Rivals, I., Personnaz, L., Taing, L. and Potier, M. (2007) 'Enrichment or depletion of a go category within a class of genes: Which test?', *Bioinformatics*, Vol. 23, No. 4, pp.401–407.

Shamir, R., Maron-Katz, A., Tanay, A., Linhart, C., Steinfield, I., Sharan, R., Shiloh, Y. and Elkon, R. (2005) 'EXPANDER, an integrative program suite for microarray data analysis', *BMC Bioinformatics*, Vol. 6, No. 1, pp.232–234.

Sharan, R., Maron-Katz, A. and Shamir, R. (2003) 'CLICK and EXPANDER: a system for clustering and visualizing gene expression data', *Bioinformatics*, Vol. 19, No. 14, pp.1787–1799.

Sheskin, D. (2004) *Handbook of Parametric and Nonparametric Statistical Procedures*, CRC Press., Chapman and Hall/CRC, USA.

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D. and Futcher, B. (1998) 'Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization', *Mol. Biol. Cell*, Vol. 9, No. 12, pp.3273–3297.

Tanay, A., Sharan, R. and Shamir, R. (2002) 'Discovering statistically significant biclusters in gene expression data', *Bioinformatics*, Vol. 18, Suppl 1, pp.136–144.

The Gene Ontology Consortium (2000) 'Gene ontology: tool for the unification of biology', *Nature Genet*, Vol. 25, pp.25–29.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R.B. (2001) 'Missing value estimation methods for dna microarrays', *Bioinformatics*, Vol. 17, No. 6, pp.520–525.

Westfall, P. and Young, S. (1989) 'p-value adjustments for multiple tests in multivariate binomial models', *Journal of the American Statistical Association*, Vol. 84, No. 407, pp.780–786.

Zhong, S., Storch, F., Lipan, O., Kao, M., Weitz, C. and Wong, W. (2004) 'GoSurfer: a graphical interactive tool for comparative analysis of large gene sets in Gene Ontology space', *Applied Bioinformatics*, Vol. 3, No. 4, pp.1–5.