Using gene ontology annotations in exploratory microarray clustering to understand cancer etiology

Geoff Macintyre[a,b], James Bailey[a,b], Daniel Gustafsson[d], Izhak Haviv[c,e,f], Adam Kowalczyk[b]

[a]*Department of Computer Science and Software Engineering, University of Melbourne, Victoria, Australia*
[b]*NICTA, Victorian Research Lab, Australia*
[c]*Ian Potter Centre for Cancer Genomics and Predictive Medicine, Peter MacCallum Cancer Centre, St. Andrew's Place, East Melbourne, Victoria, Australia*
[d]*Department of Computer Science and Computer Engineering, La Trobe University, Victoria, Australia*
[e]*The Alfred Medical Research and Education Precinct, Baker IDI, Epigenetics Group, Melbourne, Australia*
[f]*Department of Biochemistry and Molecular Biology, University of Melbourne, Victoria, Australia*

Abstract

Gene expression profiling provides insight into the functions of genes at a molecular level. Clustering of gene expression profiles can facilitate the identification of the underlying driving biological program causing genes' co-expression. Standard clustering methods, grouping genes based on similar expression values, fail to capture weak expression correlations potentially causing genes in the same biological process to be grouped separately. We have developed a novel clustering algorithm, which incorporates functional gene information from the Gene Ontology into the clustering process, resulting in more biologically meaningful clusters. We have validated our method using two multi-cancer microarray datasets. In addition, we show the potential of such methods for the exploration of cancer etiology.

Key words: Microarray, Gene Ontology, Clustering, Cancer

1. Introduction

Gene expression profiling using microarrays has become a key tool in the analysis of biological systems at a molecular level. While still producing relatively noisy data, much improvement has been made in noise correcting normalisation procedures and feature selection, providing rich datasets for further biological analysis. Microarray analysis pipelines generally come in two flavours: differential expression analysis and exploratory clustering. The purpose of differential expression analysis is to find a small subset of genes which are differentially expressed between two or more experimental conditions or samples. Having a small gene set makes for more manageable biological interpretation than using the 20,000 genes that are typically profiled on an array. In contrast, exploratory clustering attempts to utilise all genes on an array for biological interpretation by considering sets of genes with similar expression patterns, rather than a per gene analysis. This is useful under the assumption that genes with shared expression patterns have similar function or are involved in similar biological processes. Each of the clusters of genes identified provide a starting point for further biological analysis. Once clusters have been determined, usually a resource such as the Gene Ontology (Ashburner et al., 2000) is used to assist in determining the biological process represented by a set of genes.

The Gene Ontology (GO) (Ashburner et al., 2000) is a curated, structured vocabulary that describes genes and gene products. It is modeled as a directed acyclic graph, with terms as nodes and relationships between terms as arcs. A node (term) can have one or more parents, representing a more general description of the term. A node may also have children that are more specific definitions of the term. The graph is hierarchical with three top parent nodes: molecular function, biological process and cellular component. In the GO, two genes may be annotated to the same term, or they may be related through a shared term higher in the GO hierarchy. Given a set of genes, tools which calculate the terms that are statistically overrepresented in the set are commonly used to describe the biological process represented by the set of genes. For example, GeneMerge (Castillo-Davis and Hartl, 2003), FatiGO (Al-Shahrour et al., 2004) and others (Martin et al., 2004; Lee et al., 2004; Alexa et al., 2006; Zhong et al., 2004).

While a useful procedure, exploratory clustering analysis pipelines commonly face a difficult problem. Clusters can be dominated by strong or noisy expression patterns, forcing genes of similar function or those belonging to the same biological process with less correlated expression, to join another cluster. Therefore the resulting clusters may not represent a biological process in its entirety or majority, making it hard to determine which molecular processes a particular cluster of genes represents. Therefore, to improve the clustering process, additional information can be introduced to ensure that genes with similar function or shared pathways can be clustered together. Sequence similarity, protein structure similarity, shared pathways and functions, are all ways in which genes can be shown to be related. In this paper, we focus on using the GO in the clustering process. While we use only the GO as our additional information source, it is possible that another source of information might be used to further improve the clustering output.

Previous attempts have been made that utilize functional information in the clustering of gene expression profiles, however these have focused mainly on the task of predicting the function of genes with unknown function. The task in this case, is to cluster all genes with known function, and attempt to assign genes with unknown functions to one of these clusters. The unknown function is then inferred from the genes with known function. Huang and Pan (2006) and Pan (2006) used functional annotations shared between genes to modify standard distance and model based clustering algorithms. Boratyn et al. (2007) proposed a general method for modifying the distance measure based on prior shared functional information between genes. However both of these methods only use small numbers of distinct functional categories, which does not apply well when using the GO. The multiple shared functions between genes and large structure would require significant pruning of the GO graph to work in these frameworks. Cheng et al. (2004) attempted to address this by developing a clique-finding algorithm for the GO and used the cliques to perform co-clustering analysis with gene expression profiles. Another attempt developed by Liu et al. (2004) is a biclustering approach that prunes possible cluster assignments based on the GO structure.

There are however two fundamental drawbacks with these approaches. Firstly, the GO is constructed as a directed acyclic graph, with terms lower in the hierarchy being specialisations, or parts of, terms higher in the hierachy. Genes are then annotated to one or more terms in the graph, at the lowest (most specific) level possible. Drawing a path from one gene to another through this graph to determine similarity of the genes does not necessarily imply shared biology. The abstraction of terms across each level of the ontology can be such that two genes with a single shared parent term, may be extremely diverse in terms of their specific function. For example, the two terms *negative regulation of steroid metabolic process* and *positive regulation of steroid metabolic process* share the parent *steroid metabolic process*. Genes annotated to each of these terms have the opposite effect on steroid metabolism. Therefore it would not be correct to state they had similar function based on their shared parent, especially in the context of their co-expression. Secondly, having genes annotated to the same term does not necessarily imply they have similar function or share a biological pathway, in the context of their expression patterns. A single gene can act differently in various biological contexts and thus have context specific roles. It is therefore crucial to consider the expression context of a gene when deciding whether to use the knowledge of shared function to alter the clustering procedure. We define a gene's expression context to be the expression of a gene when considering the expression of all genes that are in the same biological process. It is only within this context, that one can make an informed decision on whether a certain gene should be considered to have a certain function.

Our goal is slightly different from previous approaches, in the sense that we are not attempting to predict genes with unknown function, but generate clusters of genes which are suitable for biological interpretation and encapsulate a particular biological process better than that of a standard clustering approach.

A method is needed that uses shared functional information between genes from the GO, that does not rely on GO structure, and uses GO annotations only when they are relevant to the gene set of interest (the gene's expression context). We previously developed GOMAC: **G**ene **O**ntology assisted **M**icro**a**rray **C**lustering, a modified *k*-means clustering algorithm which incorporates GO information only when it is relevant to the gene's expression context, thus avoiding problems with irrelevant gene similarities (Macintyre et al., 2008). This paper is an extension of the original manuscript, validating the method on two microarray datasets (Tothill et al., 2005; Ramaswamy et al., 2001) spanning 12 and 10 cancer types respectively, demonstrating that our method results in an alternative to *k*-means clustering, providing clusters which are more informative in terms of biological interpretation. We also discuss the biological implications of our results with respect to future research in cancer etiology.

2. Methods

The key biological assumption of the algorithm presented in this paper is that genes that share a particular annotation in the GO, will share a detectable similarity in their microarray expression pattern. There are three key differences between our approach and the previous attempts at clustering using the GO outlined above:

- Only GO terms that are statistically over-represented within a cluster are used to calculate the similarity between genes. This ensures that only GO terms within the gene's expression context are used.

- Calculations of similarities between genes using the GO is done at each iteration of the clustering algorithm, rather than fixing the GO similarities before clustering begins. This means that for each iteration of the algorithm, the terms which are used to describe a set of genes, and consequently alter the distance between a gene and a cluster, are updated to be the set of overrepresented terms for that particular group of genes, at that iteration.

- We do not rely on the structure of the GO, but rather consider every possible term as contributing to our similarity calculation.

In order to construct a model capable of the key points outlined above, each potential cluster of genes to be determined, requires both an expression profile to model the genes' expression measurements and an annotation profile to model the genes' GO similarities. As we are using a *k*-means based clustering algorithm, the number of clusters C is a parameter set by the user.

2.1. Algorithm Overview

1. Initialise using *k*-means clustering, grouping genes based on expression values using the method in Eisen et al. (1998) with C clusters (the value for C is selected by the user).
2. Determine the *expression profile* for each cluster.
3. Determine the *annotation profile* for each cluster.
4. Re-cluster genes based on *both* expression and GO annotations.
5. Re-estimate the expression and annotation profiles.
6. Repeat steps 4 and 5 until convergence.

2.2. Expression Profile

Let κ be the number of samples; let G_c be the set of genes in a cluster c. Each gene g can be viewed as a vector $\mathbf{x}_g = (x_{gi})_{1\leq i\leq\kappa} \in R^\kappa$ of its expression values across all samples. The centroid of cluster c is defined as the vector $\mathbf{X}_c = (x_1^c ... x_\kappa^c) \in R^\kappa$ with entries defined as:

$$x_i^c = \frac{\sum_{g\in G_c} x_{gi}}{|G_c|}. \tag{1}$$

where x_{gi} is the expression measurement for a particular gene g and sample i.

2.3. Gene Ontology Annotation Profile

To generate a Gene Ontology annotation profile for a cluster, all GO terms annotated to the genes in a cluster which are statistically over-represented need to be found. This means that rather than reporting all terms that are annotated to the genes in a cluster, report only those that have sufficiently low probability of being present if we sampled a random selection of genes. For this purpose we are using the hypergeometric distribution to calculate the statistical over-representation of terms, similar to the approach used in Castillo-Davis and Hartl (2003). This procedure uses the hypergeometric distribution with Bonferroni correction to generate a p-value for each term which is annotated to genes in a cluster. We use a threshold of $b \leq 0.05$ of the Bonferroni corrected score, as it provides a biologically meaningful number of terms that describe a cluster. A lower threshold yields clusters based mainly on expression distances with little or no GO terms and a higher threshold results in many GO terms which are less descriptive. All terms reported above the threshold are ignored. Let $\tau(c)$ be the number of terms *below the threshold b* for a given cluster c. From this, a weight d^c is assigned proportional to the number of genes in the cluster that are annotated to that term, normalised over all of a cluster's GO terms. The weight d_t^c shows the degree in which a term t is associated with a particular cluster. Then we can denote a cluster c's annotation profile to be the vector $\mathbf{T}^c = (d_t)^c_{1\leq t\leq\tau(c)}$ with entries defined as

$$d_t^c = \frac{n_t}{\sum_{j=1}^{\tau(c)} n_j}. \tag{2}$$

where n_t is number of genes in the cluster that are annotated to GO term t (below the threshold b).

2.4. Algorithm

The GOMAC algorithm can be summarised in the following steps:

Input.

- Gene list G
- For each gene g, expression measurements $\mathrm{E}_{g1}...\mathrm{E}_{g\kappa}$ for κ samples
- For all GO terms $\mathcal{T}_1...\mathcal{T}_f$, given a particular gene g and the t^{th} term, A_{gt} takes the value 1 if the gene g is directly annotated to the term t, and 0 otherwise. (Obtained by querying October 2008 release of the GO via a relational database interface to a locally stored copy).

Initialisation.

- Form initial groupings of genes using k-means clustering on the expression values using algorithm of Eisen et al. (1998).
- Calculate the cluster centroid (*expression profile*) $\mathbf{X}_c$ for each cluster.
- Calculate the *annotation profile* $\mathbf{T}^c$ for each cluster.

Optimisation.

1. Gene assignment: In the gene assignment step, we re-assign a gene to a cluster based on the current values for the expression and annotation profiles for that cluster. We use a gene's match to a cluster annotation profile to scale the Euclidean expression distance of the gene from that cluster.

   For each gene g let the known expression values be $E_{g\beta}$ where $\beta \in N_g \subset \{1....\kappa\}$ are all indices of samples with known values for gene g. This is due to imperfections in the microarray experimental procedure which may generate data with missing or unknown expression values for a gene. This also means we need to normalise our distance by the number of known gene values used $(1/N_g)$. This may skew the distance measurements for datasets with large number of missing values however in practice, datasets are large enough that this is not a problem. Given this, we define the Euclidean distance of each gene g from cluster c's centroid as:

   $$DE_g^c = \frac{1}{|N_g|} \cdot \sqrt{\sum_{\beta \in N_g} (x_\beta^c - E_{g\beta})^2}. \tag{3}$$

   Then, given a gene g and its GO annotations, we also determine a scaling factor S_g^c (where $0 \leq S_g^c \leq 1$). This is based on how many of the $\tau(c)$ terms in the cluster's annotation profile match the terms annotated to a gene g:

   $$S_g^c = 1 - \sum_{t \in \tau(c)} d_t^c \times A_{gt}. \tag{4}$$

   Next, the expression distance DE of gene g from cluster c is scaled by the degree in which it's annotated terms correlates with that of cluster c:

   $$DES_g^c = DE_g^c \times S_g^c. \tag{5}$$

   Finally, we use the minimum of this modified distance to assign a gene to a particular cluster:

   $$c_g = \arg\min_c (DES_g^c). \tag{6}$$

2. Re-estimation of cluster profiles: With the new assignment of genes, we re-calculate the centroids of each cluster and determine the new GO terms which are over-represented and their associated weights.
3. Repeat steps 1 and 2 until convergence (genes stop changing clusters)

Output.

- A series of gene clusters with associated GO annotations, which can be used as a starting point for further biological analysis.

3. Cancer Microarray Test Data

For testing, microarray datasets with various sample classes were required to demonstrate the potential of GOMAC to uncover biological similarities across classes. We used two published datasets, profiling cancers of unknown primary (CUP): Tothill et al. (2005) which has cDNA microarrays across 12 cancer types and their subtypes, and 10 cancer types profiled using the Affymetrix Hu6800 platform, Ramaswamy et al. (2001). These datasets were useful for our purposes as they have samples in a range of different tissues, allowing us to compare and contrast the clusters of genes output by GOMAC across varying cancers. The Tothill et al. (2005) dataset was filtered retaining only genes with greater than 400 signal intensity in the test channel (Cy5) and greater than 4 fold change (using per gene median normalised data) in at least 5 samples. This left 2185 genes (row of expression matrix) and 165 samples (columns in expression matrix): Breast(23), Colorectal(12), Gastric(7), Lung(Adenocarcinoma 10, Large Cell Carcinoma 8, Squamous Cell Carcinoma

9), Melanoma(11), Mesothelioma(5), Ovarian(21, Mucinous 11), Pancreatic(8), Prostate(5), Renal(12), Squamous Cell Carcinoma(11), Testicular(3), Uterine(9). For the Ramaswamy et al. (2001) dataset, we extracted 5697 unique gene measurements (rows) across 101 samples (columns), with the cancers grouped into the following categories: Bladder(10), Breast(10), Colorectal (9), Glioblastoma (10), Lung(8), Medulloblastoma(10), Ovarian(9), Pancreatic (10), Prostate(7), Renal(8), Uterine(10). When multiple probes mapped to a single gene identifier, we took the probe with the highest mean intensity across all samples.

4. Implementation

The previous version of the GOMAC algorithm was implemented in Perl and used the software GeneMerge (Castillo-Davis and Hartl, 2003) for calculating the over-representation of terms. In order to handle larger datasets and to interface with the latest version of the Gene Ontology, we have re-implemented the experimental set-up in C using a memory resident database as an internal data structure. Using a newer version of the GO (October, 2008) compared to the previous publication (Macintyre et al., 2008) (September 2007), meant that an extra 2631 terms were considered, altering the output of our algorithm by increasing the biological interpretability of the results on the same dataset (Tothill et al., 2005). The main goal of the re-implementation was to facilitate easy integration of any version of the GO, rather than the precompiled GO data files used by GeneMerge. In addition, the memory resident database provides a structured interface to the data used for calculations and the results. By extracting the relevant tuples for our algorithm from the Gene Ontology into our memory resident database we are able to easily interface with the required subset of the Gene Ontology without the overhead of the full Gene Ontology database. The SQL interface to the dataset thus allows for trivial integration with the Gene Ontology, while providing a rich computational platform.

5. Clustering Performance Assessment

External clustering assessment typically uses a ‘gold standard’ clustering determined by external means to compare clusterings. However, in the case of exploratory clustering, there is no ‘gold standard’. Instead, when clustering microarrays, the standard measure to determine whether a new algorithm provides biologically better clusters than a previous algorithm, is to look for statistically over-represented GO terms in each of the clusters and show that the new algorithm has clusters of superior biological relevancy. However because we have used the GO in the clustering process, this measure is not suitable. Therefore we have devised alternative means of validation.

5.1. Cluster partition criterion

When attempting to interpret the biology represented by a cluster of genes, a typical task is to perform additional clustering across the samples, to identify shared or distinct trends across sample classes. In our case this would mean for a given cluster of genes, assessing whether the samples could be grouped in such a way where a distinct expression difference could be seen among sample classes, for that set of genes. If this distinction can be made, then the cluster would be deemed a biologically interesting cluster. We therefore developed a method which assigns a *p*-value to each cluster, which represents the biological interpretability of the cluster. We call this the partition *p*-value.

Specifically, for each cluster of genes, we partition the samples into two groups using the hierarchical clustering algorithm of Eisen et al. (1998). Generally, the two resulting groups consist of one group containing all samples that have higher expressed genes in the cluster than the other group with lower expressed genes. Given a good clustering, a partition of a cluster should contain all of a particular sample class. In our setting of analysing cancer samples, we would expect cancers sharing some (perhaps unknown) biology to be grouped together. To quantify this, the hypergeometric distribution was used to determine the probability of observing a particular enrichment, or saturation, of cancer type(s) in that partition. In order to use the hypergeometric distribution for this purpose, it is necessary to group the cancer types into two classes; positive and negative. To decide which cancer types should belong to the positive class (enriched cancer types) and which should belong to the negative class, we used a majority membership approach: if the majority of one cancer type was contained in the highly expressed sample partition, that cancer type was considered in the positive class, otherwise negative. The enrichment calculation was made as follows:

$$E_c = \frac{\binom{S}{s}\binom{N-S}{n-s}}{\binom{N}{n}}$$

E_c is the probability of observing the partition for cluster c by chance given there are n samples in a partition, of which s are from the positive class, and there are S

total positive class elements and the total number of elements partitioned is N. From this, a Bonferroni corrected p-value is generated which is used to determine the quality of a particular cluster in reference to its biological usefulness. Corrected p-values which pass a threshold of less than 0.05 are considered partitionable clusters.

5.2. Assessing the biological utility of clusters

Interpreting the biology behind a cluster of genes is made easier by having GO terms that describe the cluster. We have therefore designed two ways to quantify the association of GO terms with clusters:

- *GO term abundance:* As mentioned previously, if a cluster can be easily partitioned on its samples, it is likely to be useful for biological exploration. However, without any terms associated with a partitioned cluster, it is difficult to generate any further insight. Therefore it is desirable to have a large number of terms associated with partitioned clusters. Given a collection of partitioned clusters, the number of associated terms can be measured, known as the GO term abundance. Good GO term abundance is achieved when large numbers of clusters have large numbers of terms.

- *GO term significance:* In the case where two clusters have the same terms describing them it is important to know how many of the genes are related to these terms. This is generally reflected in the p-value associating each term with a cluster. The lower the p-value the more of the genes in the cluster this term is likely to be annotated to. Therefore, a good measure of GO term significance is when term p-values are consistently small.

6. Results

Before assessment of the performance of our algorithm is carried out, a value of C, the number of clusters, is required. Only after this can we use the performance criteria outlined above to compare GOMAC with other clustering methods.

6.1. Finding a default value for C: the number of clusters

As our algorithm takes the number of clusters C as a parameter input by the user, we attempted to find a 'default' value for C which produced the most biologically useful results. Standard approaches for determining the number of clusters, such as the average silhouette, have been shown to be ineffective in when using GO clustering algorithms (Huang and Pan, 2006). Therefore, an informative measure for assessing a good value for C might be average number of genes in a cluster. As C increases, a large number of singleton clusters arise, generally with no associated GO terms. Therefore we expect a value of C where an increase beyond it will not yield more biologically informative clusters, but rather breakdown core informative clusters into singleton clusters. Thus by measuring the average number of genes in a cluster while increasing C we can observe the point where we begin to break down core clusters (a plateau in the graph). We performed our analysis of the behavior of C on the Tothill et al. (2005) dataset only, reserving the Ramaswamy et al. (2001) dataset for testing and discussion. Figure 1 shows that a plateau occurs at approximately $C = 60$. Therefore the remainder of our analysis is performed using the value $C = 60$. However, the best value for C will determine largely on how many genes there are in the dataset and how many sample subclasses are in the dataset. We therefore recommend that a number of different values for C are used in practice.

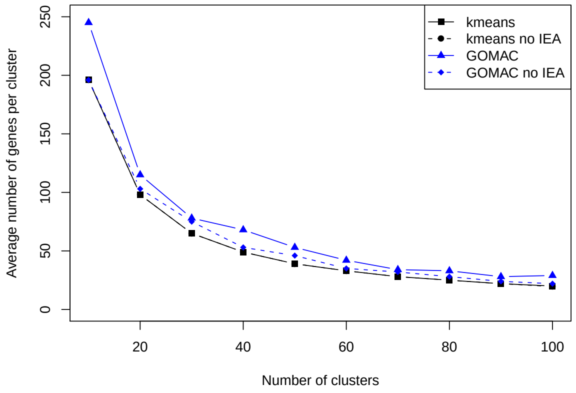


Figure 1: This graph shows the average number of genes in a cluster output by GOMAC and k-means (y-axis) for varying values of the number of clusters (x-axis) using the Tothill dataset.

6.2. Assessing the impact of GO evidence codes on GOMAC

When a gene is annotated to a term in the GO, an evidence code is recorded. This keeps track of the type of evidence used to form the annotation. For example, EXP means the annotation is 'Inferred from Experiment' or IDA is an annotation 'Inferred from a Direct Assay'. While most annotations are curated and likely to be free of noise, one type of annotation may impact on the performance of our method- IEA:'Inferred from Electronic Annotation'. These annotations are automatically generated and are not curated and considering the GO with these annotations drastically increases the number of annotations in the GO. Given the automatic and uncurated nature of these annotations, it is possible that using IEA annotations may add noise. Therefore we have performed our analysis on two versions of the GO, with and without IEA annotations.

6.3. Comparing GOMAC, regular k-means, and k-means with a fixed GO annotation profile

Using the performance assessment criteria outlined above, we analysed the Tothill et al. (2005) dataset to observe key differences between our approach and regular k-means clustering. We also sought to compare GOMAC to previous methods that use the GO in clustering (Huang and Pan, 2006; Pan, 2006; Cheng et al., 2004; Liu et al., 2004). In these methods, the decision on which GO terms to use is fixed before clustering, whereas in our method we decide whether to shrink the distance between a gene and a cluster centroid at each iteration of the algorithm, using only overrepresented GO terms specific to that cluster (the annotation profile). Therefore to make a fair comparison (removing effects from using different underlying clustering algorithms. e.g. k-means vs k-mediods), we modified a version of the k-means algorithm that used a 'fixed' GO representation. That is, after cluster initialisation, we determined the overrepresented terms for each cluster, and fixed these terms for the remaining iterations of the clustering algorithm. All comparisons are made with a cluster number $C = 60$, and the following sections discuss performance under different criteria.

Comparing cluster size

One of the goals of the GOMAC algorithm is to output clusters which encapsulate a biological process in its entirety. In other words, if a small cluster of genes is labeled with a significant term(s) representing a biological process, GOMAC should consolidate this cluster by allowing more genes associated with that process to join the cluster, providing their expression profiles are closely enough related. This results in more biologically interpretable clusters. A consolidation of clusters also results in a shift from evenly distributed cluster size, to clusters of larger size (consolidated clusters) and singleton clusters (genes with no related function). In figure 2 we plotted the distribution of cluster size for GOMAC, k-means and k-means with a fixed GO annotation profile.

k-means with and without IEA show a preference for clusters of size 2-10 genes, with few clusters greater than 50 genes. This is in contrast to the sizes observed for GOMAC, which has a three times increase in singleton clusters and and roughly two times increase in clusters greater than 50 genes. This change in cluster size validates our expectations of GOMAC to favour clusters of larger size with more terms explaining them, and are thus more biologically interpretable. k-means with a fixed GO appears to have behavior half-way between regular k-means and GOMAC, suggesting that methods using GO information before clustering can improve biological interpretability, but not as well as GOMAC. It is the ability of GOMAC to recalculate which terms are useful for calculating gene similarity that sets it apart from other GO clustering algorithms.

Assessing GO term abundance

We looked at GO term abundance by ranking the clusters output from all algorithms on partition p-value, then plotted the range of the number of terms associated with each cluster for the top 20 clusters (Figure 3a). Only the top 20 clusters were considered as many clusters are unparitionable and/or singleton clusters and therefore uninformative. Figure 3a demonstrates a clear increase in GO term abundance of GOMAC over the other algorithms. It is interesting to note that GOMAC with no IEA has more terms than GOMAC with IEA. Intuitively one would expect with more annotations, more terms would be found, however this is not the case. In fact, fewer overall annotations of terms to genes results in more terms passing the significance cut-off when using the hypergeometric distribution for over-representation calculation. This suggests that using a GO version with no IEA version is preferable. In addition, while in figure 2, k-means with a fixed GO looked like an intermediary between GOMAC and regular k-means, it appears to have the worst GO term abundance, suggesting that using the GO to calculate similarities between genes before clustering is done can deteriorate biological GO term abundance. In this case, the use of the GO does not promote a balance between clusters with coherent expression and shared biological function. Instead, a degree of functional similarity is introduced that does not take into account the context of the genes' expression. This introduction of functional similarity that is irrelevant to gene expression is therefore an introduction of noise.

Assessing GO term significance

GO term significance was assessed by ranking the clusters on their number of associated terms, then calcu-

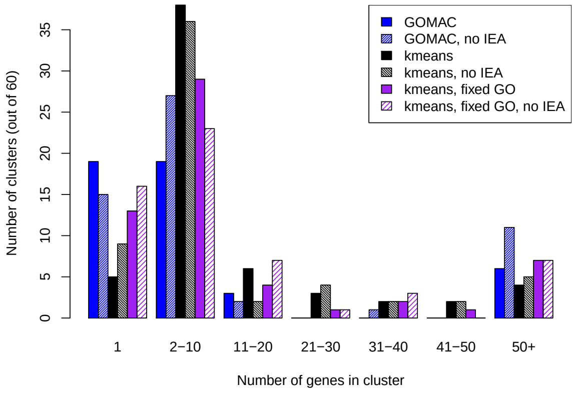
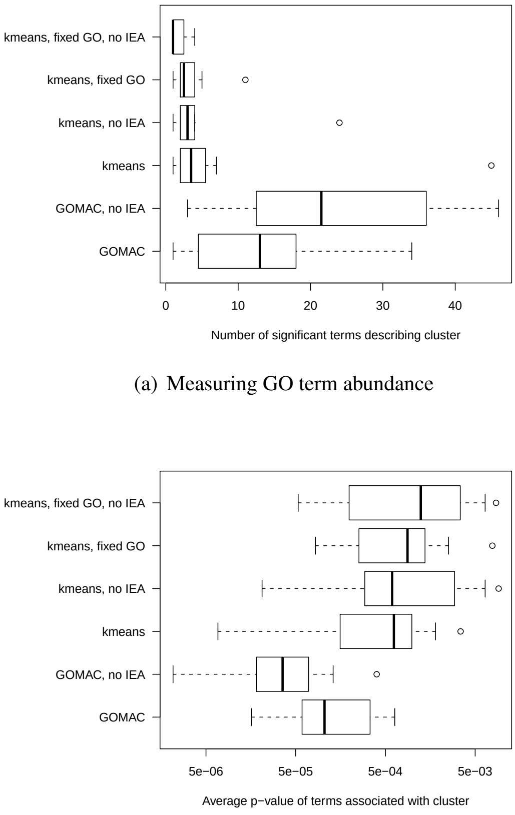


Figure 2: This figure displays the number of clusters (y-axis) that have a particular number of genes (x-axis) for the Tothill dataset. Three algorithms are compared: GOMAC, regular *k*-means and a modified *k*-means that uses a fixed version of the GO. In addition, results on two different versions of the GO are shown, with and without IEA annotations. These results are for $C = 60$.

lating the average *p*-value of associated terms for each cluster. Figure 3b shows the ranges of average term *p*-values for the top 20 clusters of each algorithm. It can be seen that GOMAC provides, on average, more significant terms than the other algorithms and thus has better GO term significance. There also seems to be no advantage to using a fixed GO, versus regular *k*-means for trying to improve GO term significance.

6.4. *Interpretation of clustering output*

Given there are 60 clusters to consider for both datasets used in our study, we have opted to show the output for a subset of the clusters for the Tothill et al. (2005) dataset and use a visualisation tool to look at the results output using the Ramaswamy et al. (2001) dataset. Table 1 contains examples of three clusters output from clustering of the Tothill et al. (2005) dataset. These clusters were chosen as they are significantly partitionable and have significantly overrepresented terms associated with them. For the Ramaswamy et al. (2001) dataset, there were 59/60 clusters that were partitionable with 26 of those having significant terms describing them. Due to these large numbers, we employed a visualisation technique to help interpret the relationship between cancers and certain biological processes represented by each cluster of genes. To do this, we took the most significantly overrepresented term for each cluster, and determined a mean expression value for the two partitions associated with that term (cluster of genes). We then used a heatmap to represent the differences



(a) Measuring GO term abundance

(b) Measuring GO term significance

Figure 3: This figure demonstrates the range of values observed under two criteria, GO term abundance (a) and GO term significance (b), for three algorithms applied to the Tothill dataset: GOMAC, regular *k*-means and a modified *k*-means using a fixed GO. In addition, results on two different versions of the GO are shown, with and without IEA annotations. In (a) the number of significant terms describing a cluster (x-axis) is shown for the top 20 clusters ranked by partition *p*-value, where $C = 60$. In this graph, large values are best. In (b) the average *p*-value of terms annotated to a cluster is shown for the top 20 clusters ranked by total number of terms associated with that cluster ($C = 60$). In this case, smaller values are better. The black line represents the mean value, the box edges represent the first and third quartile and the whiskers extend to 1.5 times the interquartile range of the box. Dots are considered outliers.

C	**Cancer**	***p*-value**	
22	Squamous cell carcinoma	3.388 e-05	**GO Terms**
			anatomical structure development, anatomical structure morphogenesis, axon guidance, biological adhesion, cell adhesion, cell differentiation, cell-cell adhesion, cellular developmental process, collagen fibril organization, developmental process, extracellular structure organization, homophilic cell adhesion, multicellular organismal development, multicellular organismal process, muscle development, nervous system development, organ development, sensory perception of light stimulus, system development, visual perception
			Genes
			ABLIM1, ANKH, ANKRD38, ANXA13, AQP4, ARMCX2, BMP5, C11orf41, C13orf1, CACNA1H, CASKIN2, CASQ2, CD34, CDH11, CDH11, CDH12, CDH16, CDH2, CDH22, CDH3, CDH4, CDH5, CDH6, CDSN, CHRDL2, CLDN1, CLDN10, CLPTM1, COL11A1, COL11A1, COL5A2, COL5A3, COL6A1, COL6A1, COL6A2, COL6A3, COL7A1, COL8A1, COL8A2, COMP, CSRP3, CST6, CSTB, CTNNA2, CYR61, DAB2, DAB2, DCN, DPT, DSC1, DSC2, DSG3, ECM2, EFEMP1, EFHD1, EFNB2, EGFL6, EMCN, FAM81A, FAT2, FBLN2, FBLN5, FEZ1, FHL1, FKTN, FLRT2, GAS1, GAS7, GJA4, GJB1, GJB2, GJB5, GJC1, GPC1, GPC3, GPC4, GPM6B, GPNMB, HMGB3, IFRD1, IMPG1, INA, ISLR, JUP, JUP, KAL1, KIF5C, KRT10, KRT13, KRT14, KRT19, KRT6A, L1CAM, LAD1, LDB2, LMO2, LOC729231, LUM, LY6D, MATN3, MB, MEST, MLLT11, MOSPD3, MPZL2, MTL5, MYH10, MYH3, MYL1, MYO1A, MYOC, MYOM2, NELL1, NPTN, NPTX1, NRCAM, NTNG1, PCDH17, PCM1, PCOLCE, PCP4, PDZD2, PKP2, POSTN, PPL, PRELP, S100A3, SCRG1, SEMA5A, SGCG, SIRPA, SLIT3, SMTN, SNTA1, SPOCK2, SPON1, SPON2, SPP1, SPRR2A, SPRR2C, SRPX, SSPN, TAGLN, TCL1A, THBS2, TNFAIP2, TNFSF11, UPK1A, VCAM1, VCAN, VSNL1, WIPF3, ZFR, ZMYM6, ZMYM6
			Expression
			Genes are more highly expressed in squamous cell carcinoma versus all other cancers.
26	Lung Adenocarcinoma	6.46e-07	**GO Terms**
			homeostatic process, regulation of biological quality, regulation of liquid surface tension, respiratory gaseous exchange
			Genes
			SFTPB, SFTPC, SFTPD
			Expression
			Genes are more highly expressed in lung adenocarcinoma versus all other cancers.
45	Mesothelioma	1.360e-05	**GO Terms**
			none
			Genes
			CALB2, CLDN15, TM4FS1
			Expression
			Genes are more highly expressed in mesothelioma versus all other cancers.

Table 1: This table represents a summary of a number of interesting clusters output by runing GOMAC on the Tothill et al. (2005) dataset. The *p*-values are Bonferroni corrected at 0.05 alpha level.

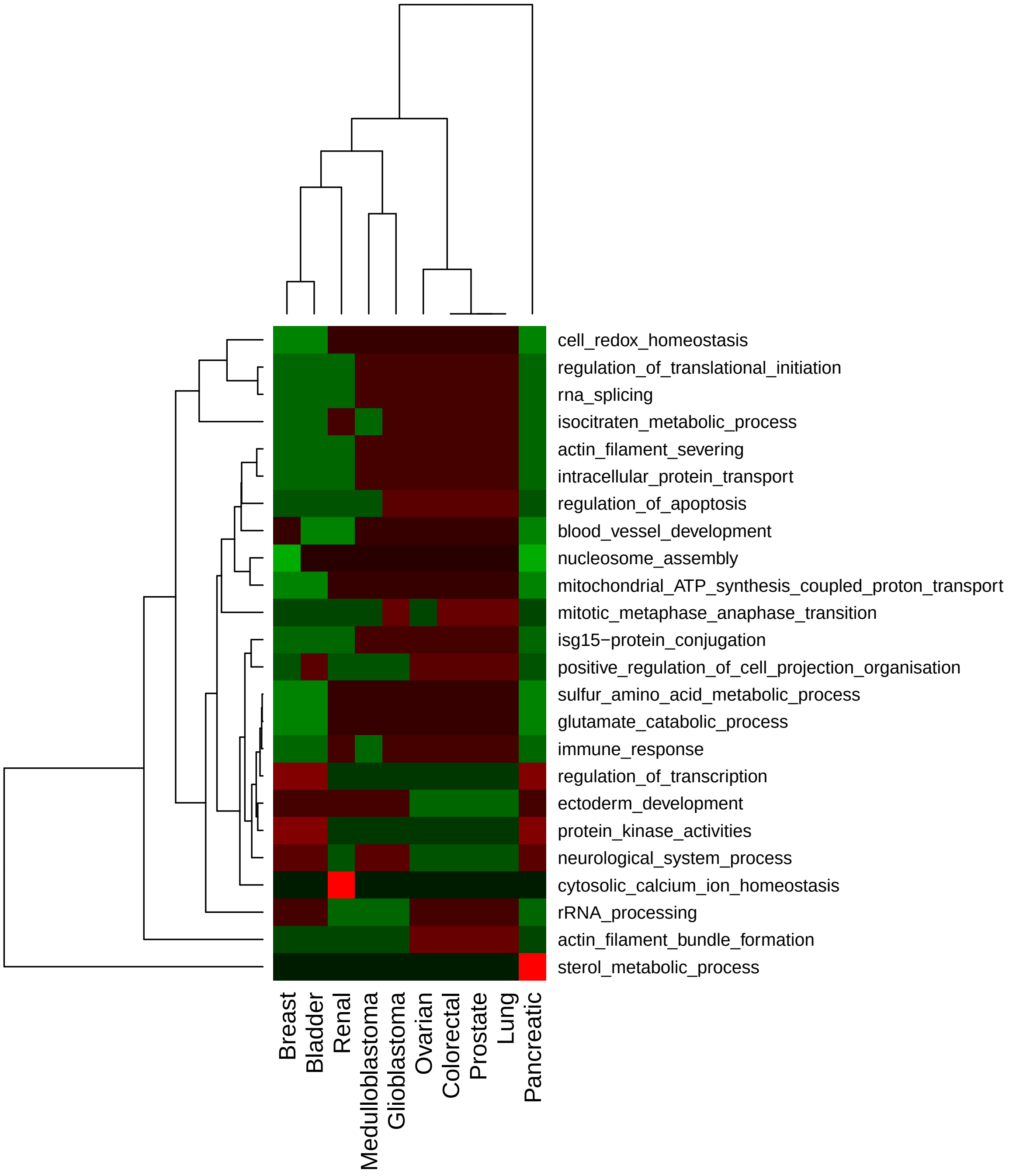


Figure 4: This figure represents the output of GOMAC applied to the Ramaswamy dataset. The heatmap relates clusters of genes described by the GO terms down the right hand side, with the cancer types across the bottom of the diagram. For each gene set, the samples have been partitioned into two groups, groups highlighted in red have higher expression than those in green. The intensity of the colours reflects the difference in the mean expression between two groups. A bright red and green suggest the mean expression of the two groups is very different, whereas a subtle change in red and green mean that although the samples could be partitioned, their mean expression was not significantly different.The colours in the heatmap have been scaled for contrast however the maximum difference between means is 2.78 with means ranged between 5.54 and 12.04. Overall, this graph allows differences and similarities to be observed across cancer types and certain biological processes.

in means between two partitions for each cluster (Figure 4). This figure allows a comparison of cancer types and the differences in their expression across the GOMAC generated gene clusters.

7. Discussion and Future Work

The Gene Ontology is usually used only after the clustering of genes and samples has been done. Here we reasoned that since multiple genes are coordinately expressed by means of biological programs, such as cell types and organs, the use of the GO in the process of clustering would focus the analysis on the driving program rather than individual genes.

We have shown through our analysis, that incorporation of additional biological information into the microarray clustering process in a biologically justified manner, can enhance the interpretability of microarray data. An obvious advantage of benchmarking the ontology-assisted clustering on the carcinoma of unknown primary datasets, is the robust nature of tissue specific gene expression, which is related to the original functions of the cancer organ of origin. For example, cluster 26 of the Tothill et al. (2005) dataset is specific to lung adenocarcinoma, and is composed of genes that code for lung specific surfactants. Other gene clusters also successfully distinguish cancers, whose cell of origin is not a typical epithelial cell, such as mesothelial cells giving rise to ovarian cancer, or squamous cells. For example, Tothill et al. (2005) dataset cluster 22 or cluster 45 are specific to SCC or mesothelial cell type cancers, respectively.

Figure 4 demonstrates the utility of the GOMAC procedure to highlight novel aspects of cancer etiology. The proximity of related cancer types, such as Medulloblastoma and Glioblastoma confirms that GOMAC gives biologically sensible results. Indeed the first remarkable feature is GO terms that are unique to one type of cancer, such as 'cytosolic calcium ion homeostasis' for renal cancer, and 'sterol metabolic process' for pancreatic cancer. Somewhat related to this is the association of Medulloblastoma, Glioblastoma, pancreatic, etc. cancers according to their germ layer of origin, by the 'ectoderm development' and 'neurological system process' GO term. Both these associations are likely driven by the nature of cancer cell precursor and the biological function of the corresponding normal tissue. i.e. osmotic regulation, lipid digestion, and neuroectoderm development, respectively. The GO terms also present some translational potential, as they highlight cancer types more likely to be driven, and thus susceptible to targeting of 'protein kinase activities'. Another feature in the GO-cluster is the association of otherwise unrelated processes, and the reflection of the regulatory interdependence. For example, all GO terms appearing on the top of the cluster, starting from 'cell redox homeostasis' appear to be co-expressed and reflect a co-regulation of energy generation, i.e. 'mitochondrial ATP synthesis coupled proton transport' and 'isocitraten metabolic process' (TCA cycle) and 'mitotic metaphase anaphase transition' implies that cell division in the context of cancer, is limited by the overall ATP/energy load in the cell (TCA being the main source thereof). Furthermore, these processes are likely controlled by epithelial cell polarity, as reflected by co-expression of the 'actin filament severing' and 'intracellular protein transport' GO terms, or epithelial to mesenchymal transition (EMT). The latter also agrees with the metastatic propensity of these cancer types and with their tendency to undergo cell death ('regulation of apoptosis'). Association with 'glutamate catabolic process' representing a primitive form of programmed cell death (King and Gottlieb, 2009). The overall biosynthesis of the cell ('regulation of translational initiation' and 'rna splicing') is linked with 'regulation of apoptosis' (or rather sensitisation to apoptosis via elevated expression of its mediators), is likely a negative feedback control mechanism that confirms cell numbers are held in balance through either lack of expansion, or expansion coupled with death. Many of our observations are in concordance with those observed previously (Segal et al., 2004). Obviously, some associations identified by the empirical results run beyond our understanding of biology, and hopefully will be explained in future research. For example, why 'immune response' is associated with 'glutamate catabolic process' and 'sulfur amino acid metabolic process' in tumour tissue, is difficult to rationalize. Perhaps the latter two are the most critical determinants of cell surface exposed antigenic epitopes. It is these observations that drive the overall conclusion that figure 4 shows that GOMAC unravels novel aspects of cancer expression profiles that deserve further analysis.

We have shown the potential of incorporating additional information into expression profile clustering to unravel the complex nature of the biological processes involved in cancer. Ideally, our method would be repeated multiple times, while alternating the source of the ontology, the cancer types, and genes. Followed by ranking of the segregating lists according to significance, then formation of an integrated summary list, that records all possible drivers of the biological systematic variations among cancers in different organs. A key benefit of such an exercise would be hypothesis generation, in the field of cancer etiology with an organ specific focus.

Acknowledgements

This work was supported by the following grants: ID 400107 NHMRC, W81XWH04-1-0336 DoD OCRP, Komen for the Cure BCTR0707358, and NBCF 509292. This work was partially supported by NICTA. NICTA is funded by the Australian Government's Department of Communications, Information Technology and the Arts, the Australian Research Council through Backing Australia's Ability, and the ICT Centre of Excellence programs. GM is supported by an Australian Postgraduate Award and a NICTA Research Project Award.

References

Al-Shahrour, F., Diaz-Uriarte, R., Dopazo, J., Mar. 2004. Fatigo: a web tool for finding significant associations of gene ontology terms with groups of genes. Bioinformatics 20, 578–580.

Alexa, A., Rahnenfuhrer, J., Lengauer, T., Jul. 2006. Improved scoring of functional groups from gene expression data by decorrelating go graph structure. Bioinformatics 22, 1600–1607.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., Sherlock, G., May 2000. Gene ontology: tool for the unification of biology. Nat Genet 25 (1), 25–29.

Boratyn, G. M., Datta, S., Datta, S., 2007. Incorporation of biological knowledge into distance for clustering genes. Bioinformation 1.

Castillo-Davis, C. I., Hartl, D. L., May 2003. Genemerge–post-genomic analysis, data mining, and hypothesis testing. Bioinformatics 19, 891–892.

Cheng, J., Cline, M., Martin, J., Finkelstein, D., Awad, T., Kulp, D., Siani-Rose, M. A., Aug. 2004. A knowledge-based clustering algorithm driven by gene ontology. Journal of biopharmaceutical statistics 14, 687–700.

Eisen, M. B., Spellman, P. T., Brown, P. O., Botstein, D., Dec. 1998. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences 95, 14863–14868.

Huang, D., Pan, W., May 2006. Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data. Bioinformatics (Oxford, England) 22, 1259–68.

King, A., Gottlieb, E., Oct. 2009. Glucose metabolism and programmed cell death: an evolutionary and mechanistic perspective. Current Opinion in Cell Biology.

Lee, S. G., Hur, J. U., Kim, Y. S., Feb. 2004. A graph-theoretic modeling on go space for biological interpretation of gene clusters. Bioinformatics 20, 381–388.

Liu, J., Wang, W., Yang, J., 2004. Gene ontology friendly biclustering of expression profiles. In: CSB 2004. Proceedings. 2004 IEEE. pp. 436–447.

Macintyre, G., Bailey, J., Gustafsson, D., Boussioutas, A., Haviv, I., Kowalczyk, A., 2008. Gene ontology assisted exploratory microarray clustering and its application to cancer. In: PRIB '08: Proceedings of the Third IAPR International Conference on Pattern Recognition in Bioinformatics. Springer-Verlag, Berlin, Heidelberg, pp. 400–411.

Martin, D., Brun, C., Remy, E., Mouren, P., Thieffry, D., Jacq, B., 2004. Gotoolbox: functional analysis of gene datasets based on gene ontology. Genome biology 5, R101.

Pan, W., Apr. 2006. Incorporating gene functions as priors in model-based clustering of microarray gene expression data. Bioinformatics (Oxford, England) 22, 795–801.

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J. P., Poggio, T., Gerald, W., Loda, M., Lander, E. S., Golub, T. R., Dec. 2001. Multiclass cancer diagnosis using tumor gene expression signatures. Proceedings of the National Academy of Sciences 98 (26), 15149–15154.

Segal, E., Friedman, N., Koller, D., Regev, A., Oct. 2004. A module map showing conditional activity of expression modules in cancer. Nat Genet 36, 1090–1098.

Tothill, R. W., Kowalczyk, A., Rischin, D., Bousioutas, A., Haviv, I., van Laar, R. K., Waring, P. M., Zalcberg, J., Ward, R., Biankin, A. V., Sutherland, R. L., Henshall, S. M., Fong, K., Pollack, J. R., Bowtell, D. D., Holloway, A. J., May 2005. An expression-based site of origin diagnostic method designed for clinical application to cancer of unknown origin. Cancer Res 65, 4031–4040.

Zhong, S., Tian, L., Li, C., Storch, K.-F., Wong, W., 2004. Comparative analysis of gene sets in the gene ontology space under the multiple hypothesis testing framework. In: CSB 2004. Proceedings. 2004 IEEE. pp. 425–435.