Comparing Functional Visualisations of Lists of Genes using Singular Value Decomposition

Hamid Ghous and Paul J. Kennedy
Centre for Artificial Intelligence,
School of Software, Faculty of Engineering and Information Technology,
University of Technology Sydney
PO Box 123, Broadway NSW 2007, Australia
Emails: Hamid.Ghous@alumni.uts.edu.au; Paul.Kennedy@uts.edu.au

Nicholas Ho and Daniel R. Catchpoole
Biospecimens Research Group and Tumour Bank, Children's Cancer Research Unit,
The Kid's Research Institute, The Children's Hospital at Westmead
Locked Bag 4001, Westmead NSW 2145, Australia
Emails: NicholaH@chw.edu.au; daniel.catchpoole@health.nsw.gov.au

Progress in understanding core pathways of cancer requires analysis of many genes. New insights are hampered due to the lack of tools to make sense of large lists of genes identified using high throughput technology. Data mining, particularly visualisation that finds relationships between genes and the Gene Ontology (GO), can assist in functional understanding. This paper addresses the question using GO annotations for functional understanding of genes. We augment genes with GO terms using two similarity measures: a Hop-based measure and an Information Content based measure, and visualise with Singular Value Decomposition (SVD). The results demonstrate that SVD visualisation of GO augmented genes matches the biological understanding expected in simulated and real-life data. Differences are observed in visualisation of GO terms, where the information content method produces more tightly-packed clusters than the hop-based method.

ACM Classifications: *H.2.8 [Database Management]: Database Applications-Data mining; J.3 [Life and Medical Sciences]; G.1.3 [Numerical Analysis]: Numerical Linear Algebra-Singular Value Decomposition*

Keywords: *Singular value decomposition, visualisation, genes, gene ontology*

1. Introduction

It is becoming clear that progress towards new insights in cancer treatment requires a thorough analysis of many genes (Jones *et al,* 2008). The routine use of microarray-based high-throughput technology has made more data available for interpretation and consideration by biologists. High-throughput microarray technology is a mechanism to simultaneously measure the activity level of thousands of genes in a biological sample. Running microarray experiments for all patients in a

Manuscript received: 30 September 2013
Communicating Editor: Jiuyong Li

cohort leads to the generation of datasets of hundreds of patients and tens of thousands of genes, where each entry is a positive number indicating the activity for that gene in that sample. Researchers typically select subsets of genes related to a target class, for example whether the sample comes from someone with a disease or not, using methods such as random forest. This results in lists of genes of interest where each entry in the list is the name of the gene. The two datasets explored in this paper are lists of gene names: one a list of genes with known functional relationships used to validate our approach and the other a list of genes selected from application of random forest on microarray data from paediatric cancer patients.

Biologists, then, are faced with the difficulty of making sense of lists of gene names, typically numbering in the hundreds. Adding to this complexity is the fact that since genes do not have a one-to-one mapping to phenotype, genes highlighted by experiments in one area of biology may have been discovered and annotated in a different area. Consequently, the gene name on its own may not assist in understanding gene function. For this reason, researchers have investigated ways of making sense of lists of genes by augmenting or enriching the data with functional information from databases such as the Gene Ontology (Ashburner *et al*, 2000).

The Gene Ontology is a structured vocabulary of gene products and functions curated by biologists, currently consisting of more than 28,000 terms, associated annotations and links to corroborating databases. It is composed of three sub-ontologies: molecular functions, cellular components and biological processes. Terms in these hierarchies relate to the biochemical activity, the physical location and the biological objective of gene products respectively. One or more terms are related to individual genes. Each term may have multiple parents in the sub-ontology using, predominantly, inheritance (or "is-a") and containment ("kind-of") relationships. The hierarchical structure between terms facilitates the construction of similarity measures between the genes by calculating the similarities between the terms associated with the genes.

The Gene Ontology project is a collaborative effort since 1998 that aims to address the need for consistent descriptions of gene products in different databases. The Gene Ontology structure is based on terms with each term consisting of (i) a unique alphanumerical identifier (GO:#######); (ii) a term name, e.g., cell, fibroblast growth factor receptor binding or signal transduction; (iii) synonyms (if applicable); and (iv) a definition. Each term belongs to one of the three hierarchies, which are structured as directed acyclic graphs. Each gene has one or more terms related to it and a term may have multiple parents in the hierarchy. Together these terms provide us with a description of the known functionality of a gene. One challenge with using terms from the Gene Ontology is that terms give different amounts of information. For example, some genes are associated with only very general terms shared by many other genes whereas others are associated with very specific terms. Also, some genes are not associated with many terms. In short, the information associated with genes in the Gene Ontology is of mixed quality.

The main issues surrounding the use of unsupervised learning methods to lists of genes enriched with Gene Ontology annotations are: how to deal with the hierarchical nature of the Gene Ontology in gene similarity measures, and what unsupervised approaches and tools are appropriate for this type of data.

Researchers have taken several approaches to the first of these problems: dealing with the structure of the Gene Ontology. That is, working with a directed acyclic structure of terms from three distinct sub-ontologies with the terms permitted to have multiple parents. There are three main approaches: those based on shared common ancestor terms, those that use information theory and those that differ from these.

The majority of researchers define a similarity measure based on the number of shared common ancestor terms associated with genes. The more common ancestor terms shared between two genes, the more similar they are. Sheehan *et al* (2008) describe several approaches for similarity measures between GO annotations including those based on sets, vectors, graphs and terms. They propose an algorithm that finds specific common ancestors between terms over the hierarchical GO structure. Mathur and Dinakarpandian (2007) use the hierarchical structure of GO to compute similarity between gene products on the basis of common GO terms. Mistry and Pavlidis (2008) take a slightly different approach and define a term overlap measure for gene functional similarity. They make a set of all the annotations related to a gene and all the parent terms, compare them to other genes and fetch the common terms. As before, the greater the number of common terms the higher the similarity.

The other main approach is to use information-theoretic measures to assess measure similarity. Richards *et al* (2010) assess functional coherence of a gene set using both a graph-based similarity measure and an information content similarity measure. Speer *et al* (2005) and Fröhlich *et al* (2007) take a kernel-based approach and cluster genes with an information-theoretic kernel function to calculate the similarity between genes over the GO. The motivation behind this approach as opposed to a distance measure over the GO graph is to better handle the variable branching and density of GO. They derive gene clusters by applying a dual k-means clustering algorithm.

The third set of methods takes a different approach than counting common ancestors or using information theory. Rather than working with the structure of the Gene Ontology, Lee *et al* (2004) transform the directed acyclic structure associated with annotations of genes of interest into a modified structure they call a GO tree. They then define a distance function on the GO tree to identify representative biological meanings and significance. Sanfilippo *et al* (2007) propose a cross-ontological approach that exploits similarity measures over the three GO sub-ontologies in two ways: firstly, by calculating similarity within a sub-ontology and secondly by finding inter-gene relationships across the three sub-ontologies. The latter method identifies gene annotations in a sub-ontology based on the annotations for similar genes. Yi *et al* (2007) take the idea that genes are physically situated on chromosomes and therefore also have a distance in base pairs along the DNA molecule. They combine this with the GO to identify functionally similar genes in close proximity on chromosomes. Finally, Popescu *et al* (2004) take a fuzzy approach and use GO terms to extract a functional summary of gene clusters. They identify the highest frequency terms by applying fuzzy methods to clusters of genes and produce a hierarchical clustering of genes that results in clusters labelled with the “most representative term” of the contained genes.

Regardless of the approach for measuring similarity between genes, it is applied in a specific tool. Huang *et al* (2008) evaluate tools for functional analysis of large gene lists. They classify tools according to key statistical methods and divide them into three categories: singular enrichment analysis, gene set enrichment analysis and modular enrichment analysis. Using these categories they give users a list of the strengths and limitations of tools. Huang *et al* (2007) describe the tool ‘DAVID’ for finding functional relationships between a set of genes using statistical methods such as heuristic fuzzy multiple-linkage partitioning. FuncAssociate (Berriz *et al*, 2009) has been developed to identify the enriched properties from a list of genes or proteins and uses the hierarchical structure of GO and the synergizer database (Berriz and Roth, 2008), a database developed from several different data sources. Similarly, GeneTrail (Backes *et al*, 2007) helps in finding functional enrichments in gene and protein data sets by using two statistical methods:

over-representation methods and gene set enrichment analysis. However, few of these reviewed methods are used in routine biomedical research.

In this paper, we apply singular value decomposition to visualise lists of genes. Our motivation for applying SVD compared to other dimensionality reduction methods such as Principal Component Analysis (PCA) is that both genes and terms may be visualised on the same graph. This allows improved understanding of the biological function of genes. We choose to apply SVD rather than factor analysis because factor analysis assumes that the data fits an underlying model where the observed variables are linear functions of a smaller set of common factors and an explicit noise term. Because GO terms, the variables in our data, are hierarchically related it is not appropriate to assume an underlying linear model. SVD identifies components that are linear combinations of the GO terms and we show in this paper that they can describe the data, but SVD is model-free.

Furthermore, we explore two similarity measures in this paper, one from each of the two main approaches outlined above. We explore a hop-based similarity measure from the set of approaches using shared common ancestors to measure similarity and an information-theoretic similarity measure (Fröhlich *et al*, 2006).

Singular Value Decomposition is applied to two data sets in this paper. The first data set has a known structure and is used to validate our approach. It is composed of genes selected from the KEGG database (Kanehisa *et al*, 2008). The second data set is composed of genes highlighted from biological experiments in childhood cancer. Our approach differs from those above by recognising that functionality needs to be described over several 'axes'. Rather than looking at only two or three functional dimensions, we find that it is valuable to also examine later dimensions that describe more subtle functional similarities between genes. Our approach differs from commercial products like GeneGo Metacore[1] and Ingenuity Pathway Analysis[2] by focusing on gene functionality rather than metabolic pathways. Whilst we agree that metabolic pathways are important, our motivation is to concentrate on full explication of functional interrelationships before augmenting data with pathway interconnectivity.

2. Data Sets

Two datasets are interrogated in this study: a validation set of genes selected from known classes and a data set of genes identified from an experiment in the cancer domain. In both cases, the datasets consist of a list of gene names that are annotated with associated terms from the Gene Ontology.

2.1 KEGG Data Set

A set of genes has been selected from the Kyoto Encyclopaedia of Genes and Genomes (KEGG) database (Kanehisa *et al*, 2008), which includes a functional classification of genes independent of the GO. The rationale is to validate our approach with genes of known functional similarity. KEGG links genomes to their biological systems and is a series of interconnected databases that interrelate (i) genes and proteins, (ii) chemical building blocks, (iii) molecular interaction pathways and (iv) hierarchies of biological objects. The last of these, KEGG BRITE, links genes into a functional hierarchy called the KEGG Orthology (KO). This hierarchy is different from the GO and has been constructed independently. We validate our approach by extracting genes from classes based on their KO terms and visualise them using GO terms. Our KEGG data set (see Table 1) contains

[1] http://www.genego.com
[2] http://www.ingenuity.com

Class	KO structure and list of genes used
1	genetic information processing : translation : ribosome rpsA, rpsB, rpsC, rpsD, rpsE, rpsF, rplB, rplC, rplD, rplE, rplF, RPS2J, RPS23, RPS24, RPS25, rpmB, rpmC, rpmD, rpmE, rpmF
2	genetic information processing : transcription : RNA polymerase FL/A, RPOA, RPOB, RPOZ, RPOH, RPON, RPOD, RPB2, RPBJ, RPB3, RPA49, RPAJ4, RPA34, RPA43, RPAJ2, RPCJ9, RPC25, RPB7, RPB4
3	genetic information processing : transcription GREA, GREB, NUSA, NUSB, NUSG, MBFJ, RclJ, RHO, ELP3, POLRMT, gtf2a2
4	metabolism : carbohydrate metabolism : pentose phosphate pathway pgl, zwf, edd, rpe, tktA, fbp, rpiA, gcd, rbsK, pgm, eda
5	metabolism : carbohydrate metabolism : pentose and glucoronate interconversions GUSB, galU, rpe, AKRJ, mtlY, mtlD, clpX

Table 1: Genes in the KEGG dataset listed by class identifier. Column 1: class number Column 2: KO terms describing class and associated genes.

genes (also in GO) from five KO classes: ribosome (ko03010, class 1), RNA polymerase (ko03020, class 2), transcription (ko01210, class 3), pentose phosphate pathway (ko00030, class 4) and pentose and glucoronate interconversions (ko00040, class 5). We expect genes in classes 1, 2 and 3 will be similar (with classes 2 and 3 more similar to each other than to class 1). Genes in classes 4 and 5 should be similar to one another but different from the other classes.

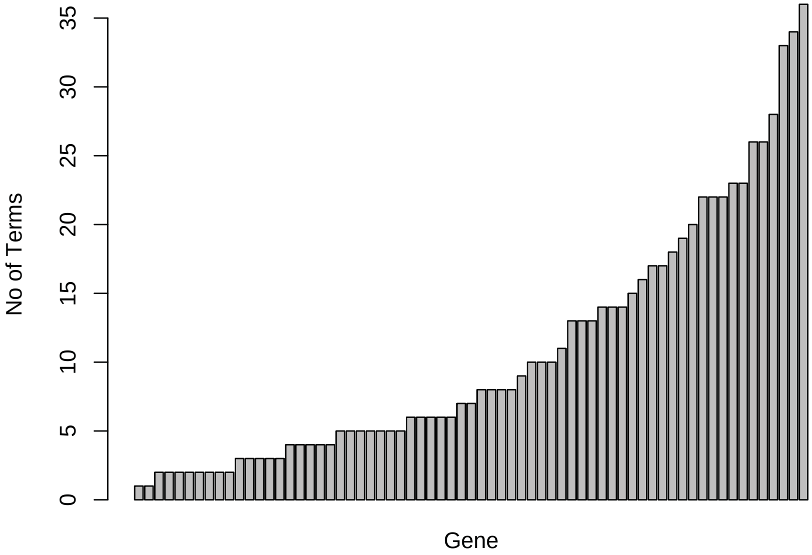


Figure 1: The number of terms associated with each gene in the KEGG dataset. Genes are ordered in increasing number of terms.

Our KEGG data set consists of a matrix of 67 rows, one for each gene, and 286 columns, one for each GO term. The set of 286 GO terms is the union of all GO terms directly associated with the genes. Each entry in the matrix is 1 if the gene is directly associated with the term, 0 otherwise. Figure 1 shows the number of terms associated with each gene in the KEGG dataset. Figure 2 shows the frequency of terms having direct association to various numbers of genes in the KEGG dataset. This shows that almost all terms are directly associated with very few genes and motivates our use of the proximity measures to relate terms using their relationships over the ontology.

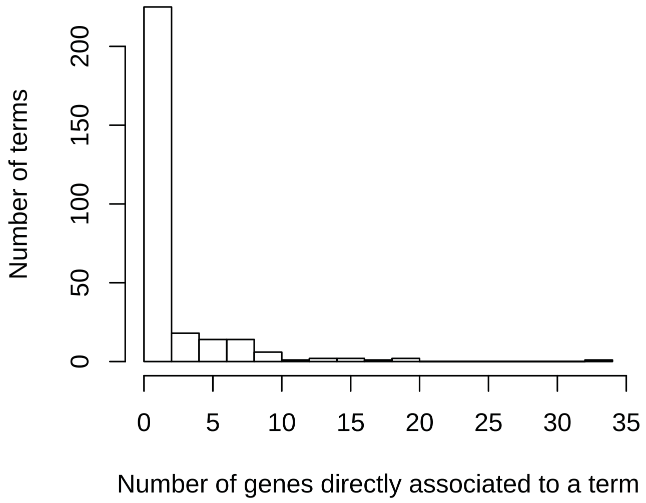


Figure 2: The frequency of terms having a direct association to various numbers of genes in the KEGG dataset.

2.2 Acute Lymphoblastic Leukaemia Data Set

Acute Lymphoblastic Leukaemia (ALL) is the most common childhood malignancy with around 250 children in Australia diagnosed annually. Microarray technology has been used extensively in attempts to identify markers that are predictive of treatment outcome in ALL.

The cancer dataset lists genes identified as important in Acute Lymphoblastic Leukemia (ALL). It was constructed based on results from Flotho *et al* (2007) and Catchpoole *et al* (2008). Flotho and colleagues identified a fourteen gene signature with expression values able to separate a cohort of ALL patients into two groups that agreed with minimal residual disease (MRD) results. Minimal residual disease refers to small numbers of cancerous cells remaining after treatment (in the order of one cancerous cell in a million normal cells). It is used in oncology to know when a cancer has been eliminated and to compare therapies.

Catchpoole *et al* (2008) examined these genes on a different cohort of ALL patients and also discovered a separation of patients but it did not agree with MRD results nor with clinical presentation. The data mining algorithm, Random Forest (Breiman, 2001), was used to identify other genes that supported the same separation of patients as achieved by Flotho's gene signature. Random Forest (RF) is an ensemble classifier algorithm that produces a "forest" of decision trees each constructed from feature subsets. RF can handle multidimensional data, which makes it increasingly popular in microarray and other high-throughput studies (e.g. Zhang *et al*, 2008; Hoffmann *et al*, 2006; Kim and Kim, 2007; Ward *et al*, 2006), and can provide a measure of the predictive performance of each feature, which allows for feature selection analyses.

Patients were clustered using hierarchical clustering based on the gene signature from Flotho *et al* (2007). The resulting two clusters formed the class labels for constructing our predictive model.

A RF model of 50,000 trees was constructed on a gene expression dataset of 127 ALL patients and 22,280 probesets. The model was generated using Affymetrix Human Genome U133-based chips on diagnostic bone marrow samples. Using this RF model, the 250 probesets with the largest mean decrease in Gini index (effectively the 250 probesets that contribute most to the RF model differentiating between the patients in the two clusters formed using Flotho's signature) were selected to form the "cancer dataset". Since some of these probesets referred to the same gene, we ended up with 195 unique genes.

After pairing the genes with their associated GO terms, the cancer data set, then, is a matrix of 195 rows, one for each gene, and 980 columns, one for each GO term. As before, the set of GO terms is the union of all the terms directly associated with the genes and entries in the matrix are 1 if the gene is directly associated with the term or 0 if not. Figure 3 shows the distribution of the number of terms for genes in the cancer dataset. Figure 4 shows the frequency of terms having direct association to various numbers of genes in the cancer dataset. As before, almost all terms are directly associated with very few genes.

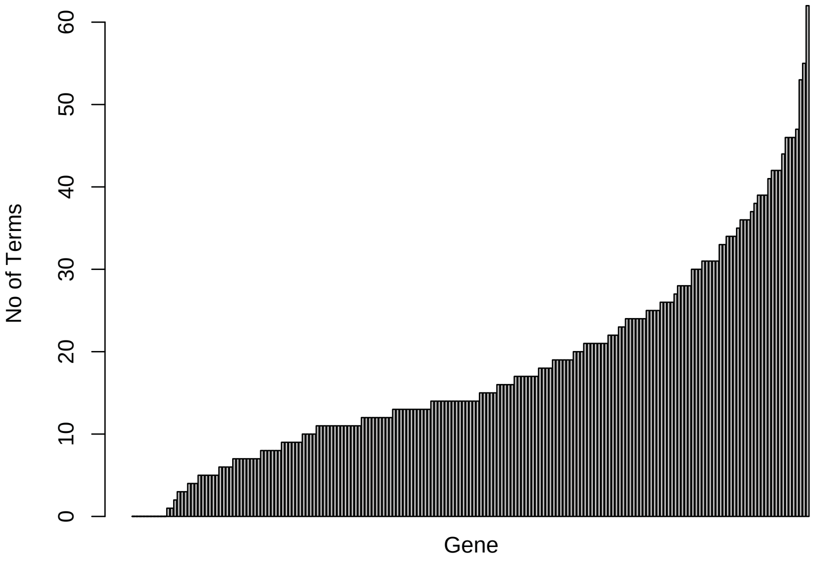


Figure 3: Distribution of the number of terms for genes in the cancer dataset. Genes are ordered by increasing number of terms.

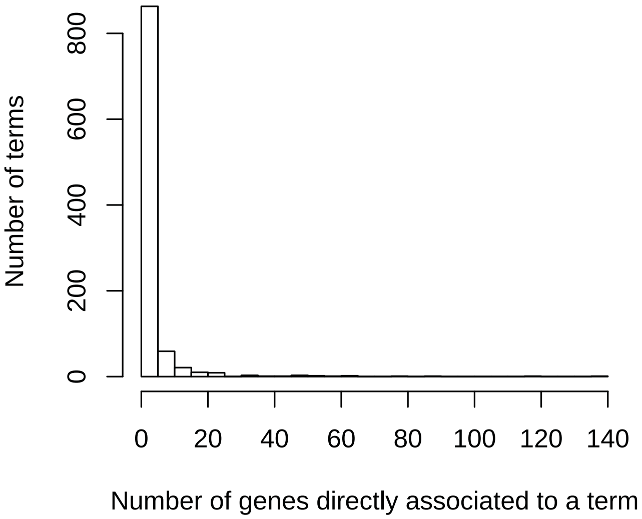


Figure 4: The frequency of terms having a direct association to various numbers of genes in the cancer dataset.

3. Methods

3.1 Singular Value Decomposition (SVD)

Singular value decomposition (Golub and Van Loan, 1996) is a method that transforms a data matrix $X \in R^{nxm}$ into the orthogonal matrices $U \in R^{nxr}$, $V \in R^{mxr}$ and a diagonal matrix $D \in R^{rxr}$ where $r \leq m$ is the rank of X.

$$X = UDV^T \tag{1}$$

Row vectors of U relate to the original data points (rows of X) and rows of V are associated with the data attributes (columns of X). The columns of U are called the left singular vectors of X and columns of V are called the right singular vectors. The elements of D are termed the singular values of X. Singular value decomposition has been used often in bioinformatics, for example, in visualisation of gene expression values (Tomfohr *et al*, 2005), but the novelty in our work is to augment lists of genes with knowledge from a domain ontology and to examine several dimensions (columns of U and V) to extract a better understanding.

In this study, we apply SVD to an augmented data matrix that reflects term similarities. Before applying SVD, the matrix is centred and scaled.

3.2 Incorporating Functional Information into the SVD

Given a set of genes, G define T as the set of GO terms directly associated with any of the genes. From G we create a matrix $X \in R^{nxt}$ where n is the number of genes |G| and t the number of GO terms |T|. Each element x_{ij} of X has the value 1 if the gene i is directly associated with term j otherwise 0. This is similar to computational linguistics where “genes” are replaced by “documents”.

This data matrix is augmented by information reflecting inter-term similarities. A symmetric proximity matrix $P \in R^{txt}$ is created with elements representing the proximity (or similarity) between GO terms i and j. In this paper we explore the use of two approaches to calculating values in the proximity matrix: a hop-based approach and an information-content based approach. Our interest in choosing these two specific measures is that, as discussed in Section 1, they are representatives of the two prevailing schemes for calculating similarities between genes, and hence we are interested in their effects on visualisation.

Hop-based approach: The proximity between GO terms is based on the number of links (or distance) between them and is defined as $p_{i\,j} = (d_{i\,j} + 1)^{-1}$ where $d_{i\,j}$ is the minimum distance between terms i and j over the hierarchy using "is-a" links which are more frequent than "kind-of" relationships, extracted from GO using SQL. Elements of P are $0 \leq p_{i\,j} \leq 1$. Terms i and j having a close relationship will have $p_{i\,j}$ with a value near 1. Diagonal elements of P are $p_{ij} = 1$.

Information-content approach: The information content (IC) based proximity (Fröhlich *et al*, 2006), on the other hand, uses information content theory (Resnik, 1995) to calculate the semantic similarity between GO terms. It is based on the probability of GO terms in the gene dataset X. The information content measure is defined as

$$IC(t\,) = -\log_2 P(t\,) \tag{2}$$

where P(t) is the probability of term t in the data matrix and is calculated as P(t) = freq(t)/N where N is the total number of GO terms in X and freq(t) is the number of occurrences of t or any of its child terms. Similarity between terms i and j is defined as

$$p_{ij} = -\log_2 \min_{\hat{t} \in Q_a(i,j)} P(\hat{t}) = -\log_2 P_{ms}(i, j) \tag{3}$$

where Q_a (i, j) is a function returning the set of common shared parent terms between terms i and j and P_{ms} , the probability of the minimum subsumer (Lord *et al*, 2003), is the minimum P(t) if there is more than one parent. Values of $p_{i\,j}$ using this scheme are of course not limited to the range [0, 1].

The augmented data matrix is defined as X' = XP with P being the term-to-term proximity matrix computed using either the hop-based or information-content based approach. SVD is applied to X' after centring and normalisation. Whilst proximity matrices have been used for text kernels, we are unaware of their use with GO terms.

We visualise the genes (rows of X') by plotting rows of U using various columns. Similarly, we visualise the GO terms (columns of X') on the same graph as the genes by plotting rows of V using the same columns as for U. The first column is the projection of the data into the axis of most variation of the data. This is called the projection of the data into the first principal component (PC1). The second column represents the projection into the axis related to the next largest amount of variation. That is, the projection of the data into the second principal component (PC2). A similar scheme applies for the other columns of U and V.

Finally, we calculate the Pearson correlation between each column of U (the data projected to a principal component) and columns of X representing specific GO terms, as well as to a new column containing the number of GO terms associated with each gene (i.e. the sum of each row of X). We assume that columns of X with high absolute correlation to a particular column of U are GO terms that explain, in some sense, the meaning of the respective principal component.

4. Results

4.1 Visualising KEGG Data Set

First, we present the distribution of the proximity matrices. Figure 5 shows box plots for the values in the upper triangular section of the hop-based and IC proximity matrices for the datasets.

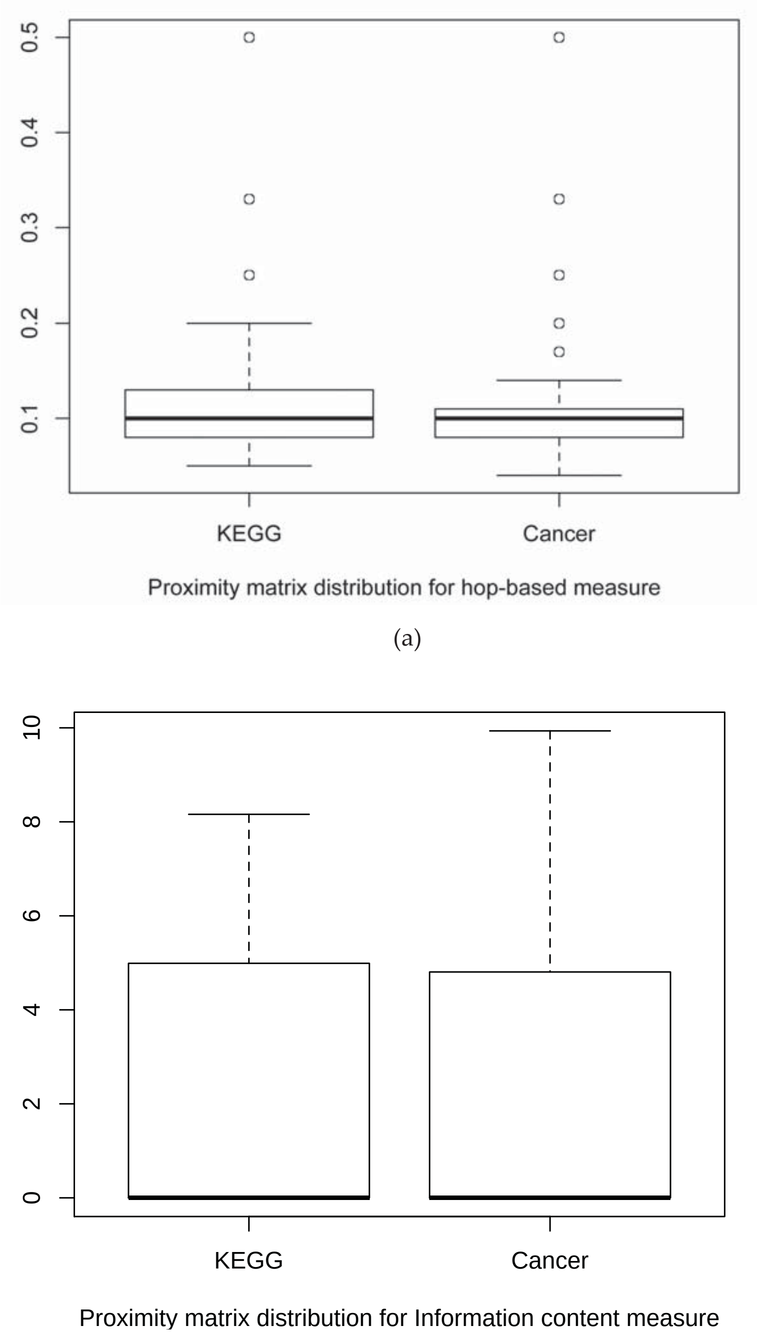
Next, we explore the KEGG dataset with and without using a proximity matrix, to show that the proximity matrix is required for visualisation. Figure 6(a) shows genes projected into PC1 and PC2 without using a proximity matrix. That is, simply applying SVD to the X matrix. The figure shows that because similar terms cannot be correctly determined then, apart from a few genes, all genes group together. This is in contrast with the clear spread of genes when using proximity matrices as shown in figures 6(b) and 7(a).

After transformation of the KEGG dataset with SVD we calculated the Pearson correlation between the data projected to principal components and to the association of GO terms to genes (i.e., X) with both similarity measures, the total number of GO terms for each gene and to the gene class. The strongest relationship expected in the KEGG data set is that comparing genetic information processing genes (those in classes 1, 2 and 3) with carbohydrate metabolism genes (those in classes 4 and 5).

There was however a very strong correlation of 0.995 between the data projected into principal component 1 (denoted as PC1 in this paper) and the number of terms associated with each gene suggesting that this principal component is a “size” component (Jolliffe, 2004). It seems reasonable that the most variation in the dataset is based on the number of terms for genes. When plotted, as in Figure 7(a), PC1 shows variation in the genes, but not much in the terms. This is due to the relative amounts of variation in genes compared to terms. When the terms are plotted on their own as in Figure 8 the relationships are clearer. Figure 8(a) shows that PC1 does not separate based on the sub-ontologies for the hop-based measure (although PC2 does to a small extent), but there is a separation for the information-content measure (Fig. 8(b)).

Principal component 2, associated with the next largest variance, generally contrasts the genetic information processing genes with the carbohydrate metabolism genes as can be seen in Figure 9(a) with the hop-based similarity measure, where PC2 denotes the axis for principal component 2. However, we acknowledge that it is not a completely clear division: there is some overlap. The outlier (circled) with high PC2 and PC3 values is the gene RHO which is associated with the largest number of terms in the data. Table 2 shows that the highest correlation to PC2 is with the class label, which represents the expected differences between genes in this validation dataset, followed by strong positive correlations to GO terms describing carbohydrate metabolism and negative correlations to terms associated with ribosomes. The IC method separates genetic information processing and carbohydrate metabolism genes only at PC5 (Figure 9(b)) demonstrating that both methods eventually find the expected functional relationship between genes, but that they are clearer with the hop-based approach.

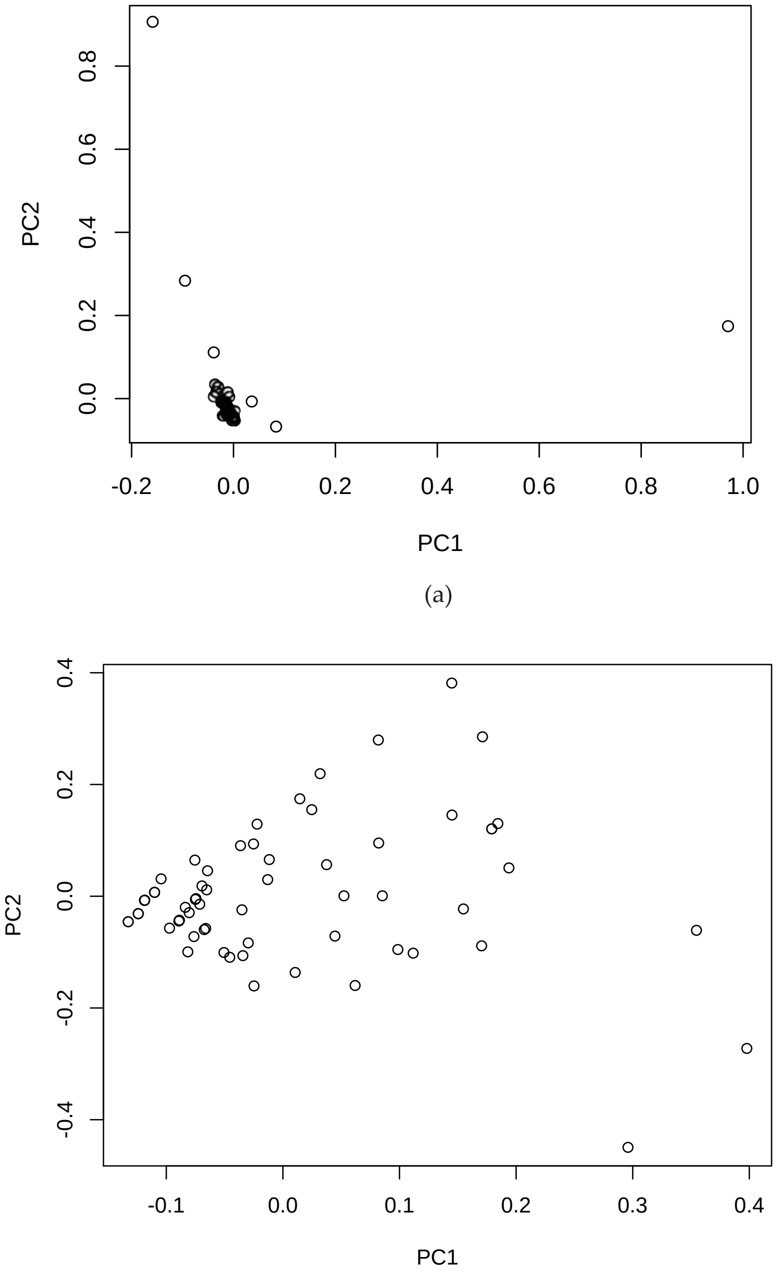
Apart from the outlier RHO in the top right hand corner, Figure 10 shows that PCs 2 and 4 separate the different kinds of genetic information processing genes as expected because there are more of these than the carbohydrate processing genes. Again, the separation involves some overlap between the classes. We do not present PCs 5 and greater for the hop-based approach because the expected structure has been explained using PCs 2–4. Correlations for PCs 2 through 4 are given in Table 3 showing a mixture of some expected (e.g. for PC3) and unexpected GO terms (e.g. those for PC2 and 4).



(a)

(b)

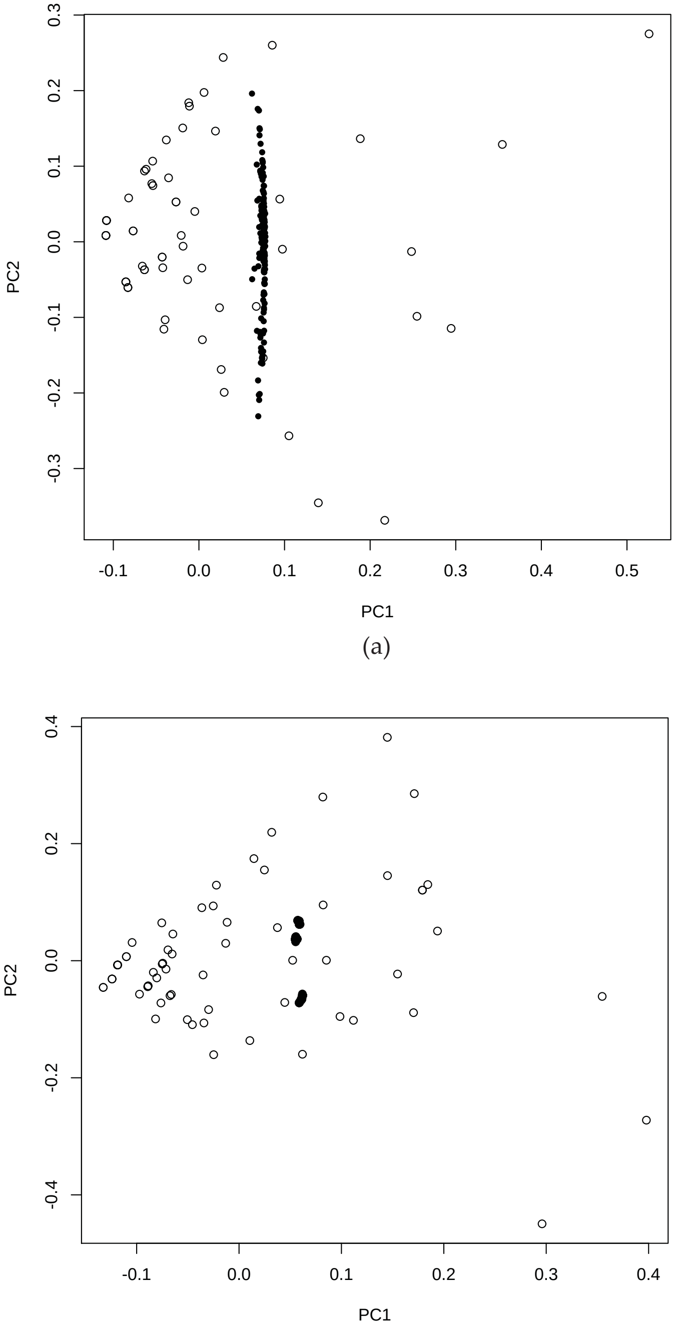
Figure 5: Distribution of values in the proximity matrices. (a) hop–based proximity matrix (b) information-content proximity matrix.



(a)

(b)

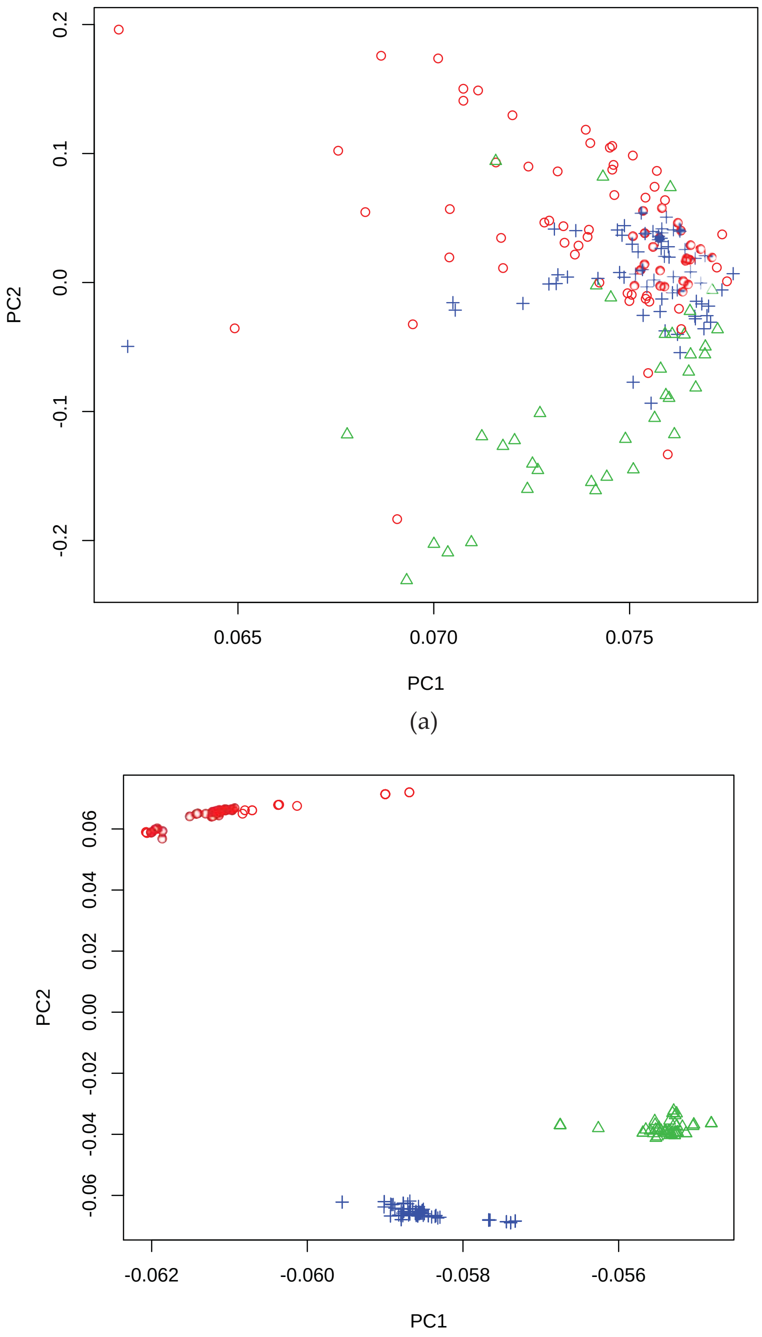
Figure 6: Plot of genes from KEGG dataset projected into PC1 and PC2 (a) without using a proximity matrix, (b) using the information-content proximity matrix. Legend: ○ is gene.



(a)

(b)

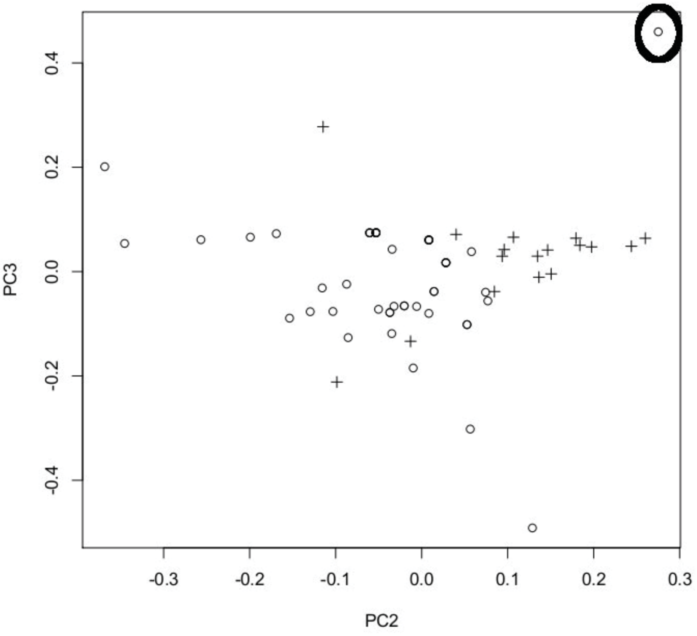
Figure 7: Plot for PC1 and PC2 for both methods using KEGG dataset. (a) Hop-based method. (b) IC similarity measure. Legend: ○ is gene, ● is term.



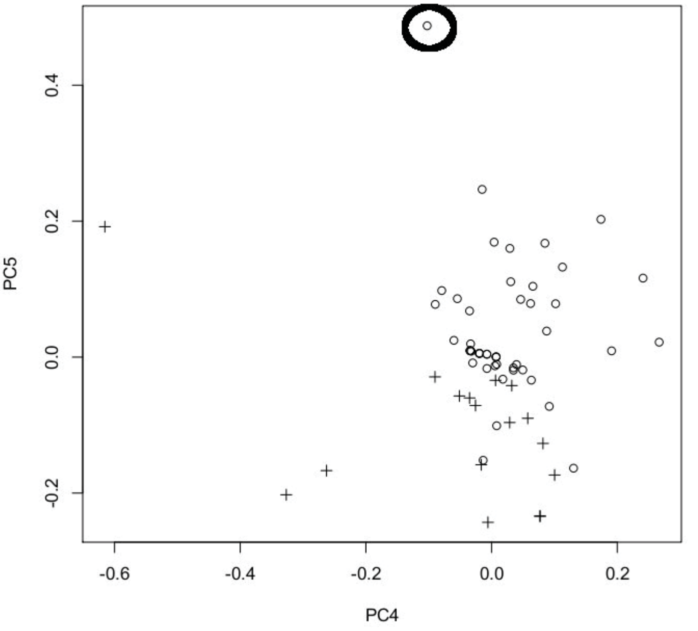
(a)

(b)

Figure 8: Plot of terms from the KEGG dataset projected into PC1 and PC2 (a) for the hop-based proximity matrix, (b) using the information-content proximity matrix. Legend: + is molecular function GO term, ○ is biological process GO term and △ is cellular component term.



(a)



(b)

Figure 9: (a) Principal components (PC) 2 and 3 for hop-based method and (b) PC4 and 5 for IC method. Legend: ○ is genetic information processing genes and (+) represent carbohydrate metabolism genes.

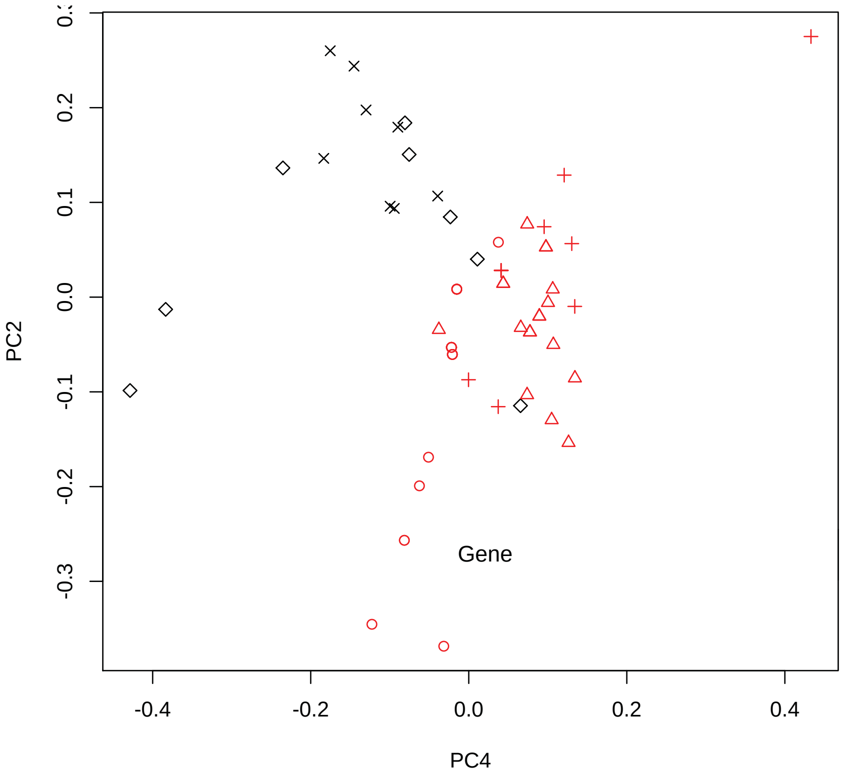


Figure 10: Plot of principal components 4 and 2 for U matrix (genes) for the KEGG dataset using the hop-based approach. Genes are related to KEGG categories for ribosomes (○), RNA polymerase (△), transcription (+), pentose phosphate pathway (x) and pentose and Glucoronate interconversions (◇).

4.2 Visualising Cancer Data Set

As with the KEGG dataset, we first examine the distribution of the proximity matrices for the cancer dataset. Figure 5 shows box plots for the values in the upper triangular section of each proximity matrix. Next, similarly to the KEGG data, we explore the cancer dataset with and without using a proximity matrix, to show that the proximity matrix is required for visualisation. Figure 11 shows genes projected into PC1 and PC2 without using a proximity matrix (top) and with using the IC similarity measure (bottom). That is, simply applying SVD to the X matrix. The figure shows that, because similar terms cannot be ascertained correctly, genes do not cluster meaningfully.

As with the KEGG dataset, the hop-based and IC approaches both show a strong correlation (0.997 and 0.988 respectively) between the number of GO terms and projected points to PC1. Overall, the distributions of GO terms for PC2 and PC3 (or PC1) make three clusters, associated with each GO subontology (see Figure 12) as evidenced by the distribution of intra- and inter-cluster distances shown in Figure 13. To untangle the relationships across the subontologies, we applied SVD to the terms from each subontology separately. We examined clusters through correlation and by listing the terms in each cluster.

Term name and accession	Correlation
Class	0.550
Carbon utilization by utilization of organic compounds (GO:0015978)	0.539
Cellular catabolic process (GO:0044248)	0.539
Ribosome (GO:0005840)	-0.626
Ribonucleoprotein complex (GO:0030529)	-0.626
Intracellular (GO:0005622)	-0.606
Structural constituent of ribosome (GO:0003735)	-0.606
Translation (GO:0006412)	-0.606
Cytosolic small ribosomal subunit sensu Eukaryota (GO:0005843)	-0.577

Table 2: GO term name and accession for terms with Pearson correlation > 0.5 to PC2 values for KEGG data with hop-based similarity measure. "Class" refers to the class identifier for the gene.

For the Cellular Component GO terms, the hop-based approach revealed a separation between cytoplasmic structure terms and DNA replication terms along PC3 axis as shown in Figure 14(a). It also highlighted a cluster of terms associated with the membrane on the negative end of PC2 and a cluster of tubulin and kinesin GO terms towards the positive end of PC3 (see Figure 14(a) clusters A and B respectively). PC2 in the IC approach reveals four small distinct clusters: a cluster of membrane and extracellular-matrix-related terms, a cluster of terms associated to organelles, protein-complex-related terms and a cluster of cell-division-apparatus-related terms as shown in Figure 14(b) as clusters A, B, C and D respectively. Some of the terms in these clusters for the IC measure are listed in Table 4.

For the Biological Process GO terms, PC3 in the hop-based approach reveals a cluster of terms associated with development (e.g. embryonic development, notochord development, forebrain development, embryonic axis specification) as shown in Figure 15(a). PC2 in the IC approach identifies five tight clusters (shown in Figure 15(b)): cluster A relates to morphogenesis and early development, cluster B to homeostasis and response to stimulus (within which is a subgroup related to molecular transport in the cell), cluster C relates to gene expression regulation and metabolism, cluster D to differentiation and cluster E to DNA metabolism and function along with a number of small subgroups e.g. vesicle transport. As before, some of the terms found in these clusters (IC measure) are listed in Table 5.

For the Molecular Function terms, PC2 in the hop-based approach identifies a cluster of terms associated with DNA helicase activity. In close proximity to this cluster is a loosely packed cluster of six genes that code for mini chromosome maintenance proteins (MCM2, MCM3, MCM4, MCM5, MCM6 and MCM7) as shown in Figure 16(a). Both MCM proteins and replicative helicase play integral roles in eukaryotic DNA replication. The IC approach also identified this loose cluster of MCM genes and the GO term for DNA helicase activity. Across PC2, the rest of the clusters relate to enzyme activity No. 1, enzyme activity No. 2 and non-enzymatic molecular interactions as shown in Figure 16(b) as A, B and C respectively. Some terms from these clusters (IC measure) are in Table 6.

Term name and accession	Correlation
PC2	
GO:0007165 (signal transduction)	0.452873
GO:0006955 (immune response)	0.452873
GO:0005102 (receptor binding)	0.452873
GO:0005164 (tumor necrosis factor receptor binding)	0.452873
GO:0008675 (2-dehydro-3-deoxy-phosphogluconate aldolase activity)	0.452873
GO:0005576 (extracellular region)	0.452873
PC3	
GO:0005886 (plasma membrane)	0.469644
GO:0005624 (membrane fraction)	0.469644
GO:0005622 (intracellular)	0.455098
GO:0030529 (ribonucleoprotein complex)	0.436264
GO:0005840 (ribosome)	0.436264
GO:0006414 (translational elongation)	0.424762
PC4	
GO:0016021 (integral to membrane)	-0.676252
GO:0007165 (signal transduction)	-0.619926
GO:0006955 (immune response)	-0.619926
GO:0005102 (receptor binding)	-0.619926

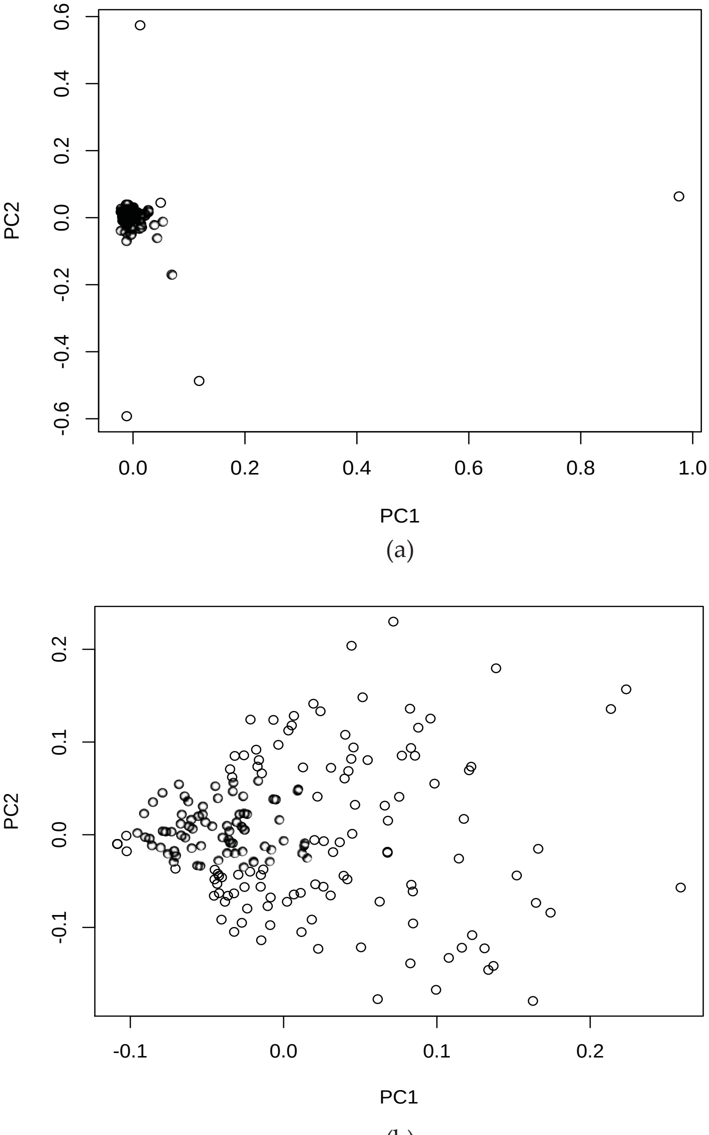
Table 3: GO term name and accession number for those terms with absolute value of Pearson correlation > 0.25 for PC2-PC4 values for the KEGG data set with information content similarity measure.

PC	GO term name and accession	Correlation
1	Number of terms	0.855
2	GO:0000777 (condensed chromosome kinetochore)	0.547
	GO:0000775 (chromosome, centromeric region)	0.503
	GO:0000776 (kinetochore)	0.466
	GO:0000778 (condensed nuclear chromosome kinetochore)	0.427
3	GO:0005856 (cytoskeleton)	0.465
	GO:0005874 (microtubule)	0.418
	GO:0005819 (spindle)	0.386
	GO:0031298 (replication fork protection complex)	-0.376
	GO:0042555 (MCM complex)	-0.359
4	GO:0005737 (cytoplasm)	0.383
	GO:0005730 (nucleolus)	0.372
	GO:0005634 (nucleus)	0.365

Table 4: GO terms from the cellular component sub-ontology with absolute value of Pearson correlation > 0.35 for PCl-4 values from the cancer data set for the IC measure.

A summary of the GO term clusters shown in Figures 14(b), 15(b) and 16(b) is presented in Table 7 using IC method for Cellular Components (CC), Biological Process (BP) and Molecular Function (MF) of GO based on correlation results with their biological interpretation.

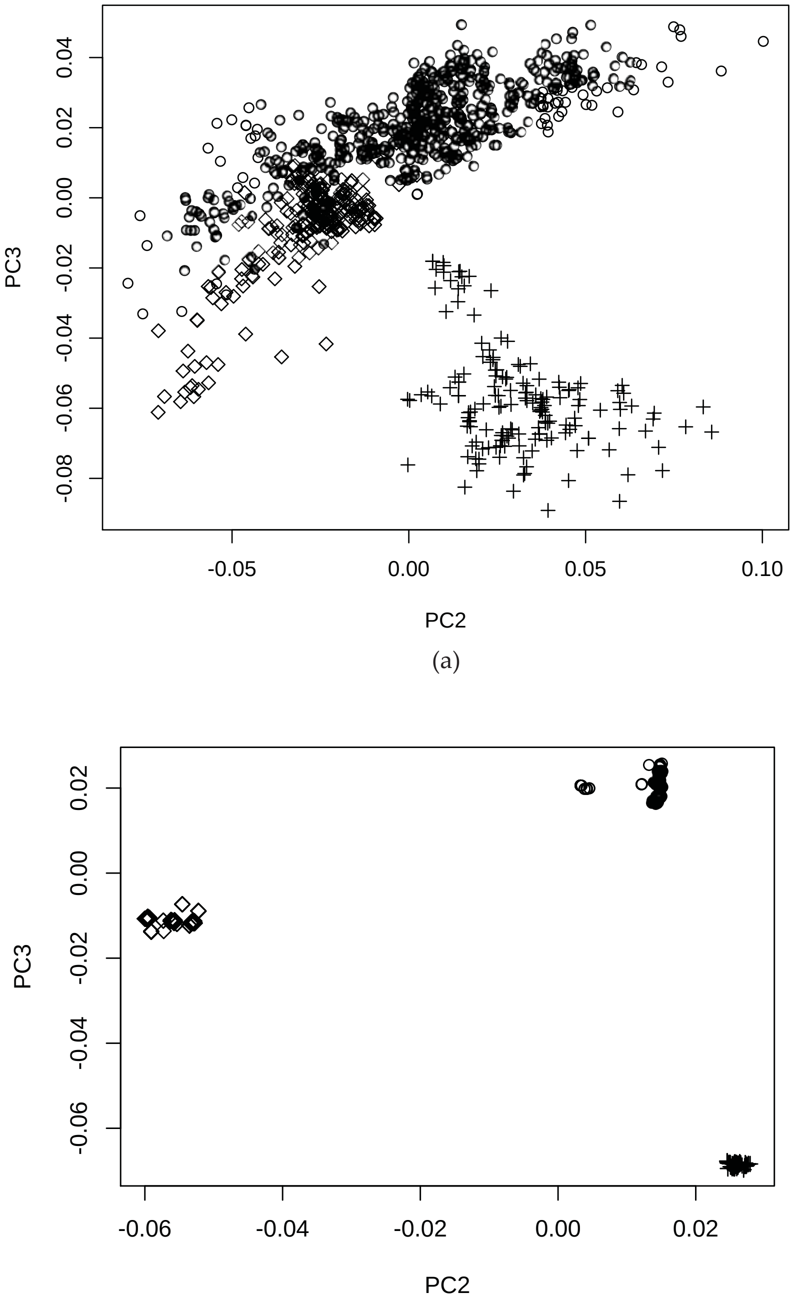
SVD visualisation of the cancer data results in a meaningful functional visualisation of the genes, particularly when limited to terms in sub-ontologies. Clusters of terms highlight functional groupings of genes and the genes themselves cluster "behind" the terms that describe them. Correlations describe the PC axes. Each PC describes a different functional aspect of the gene set.



(a)

(b)

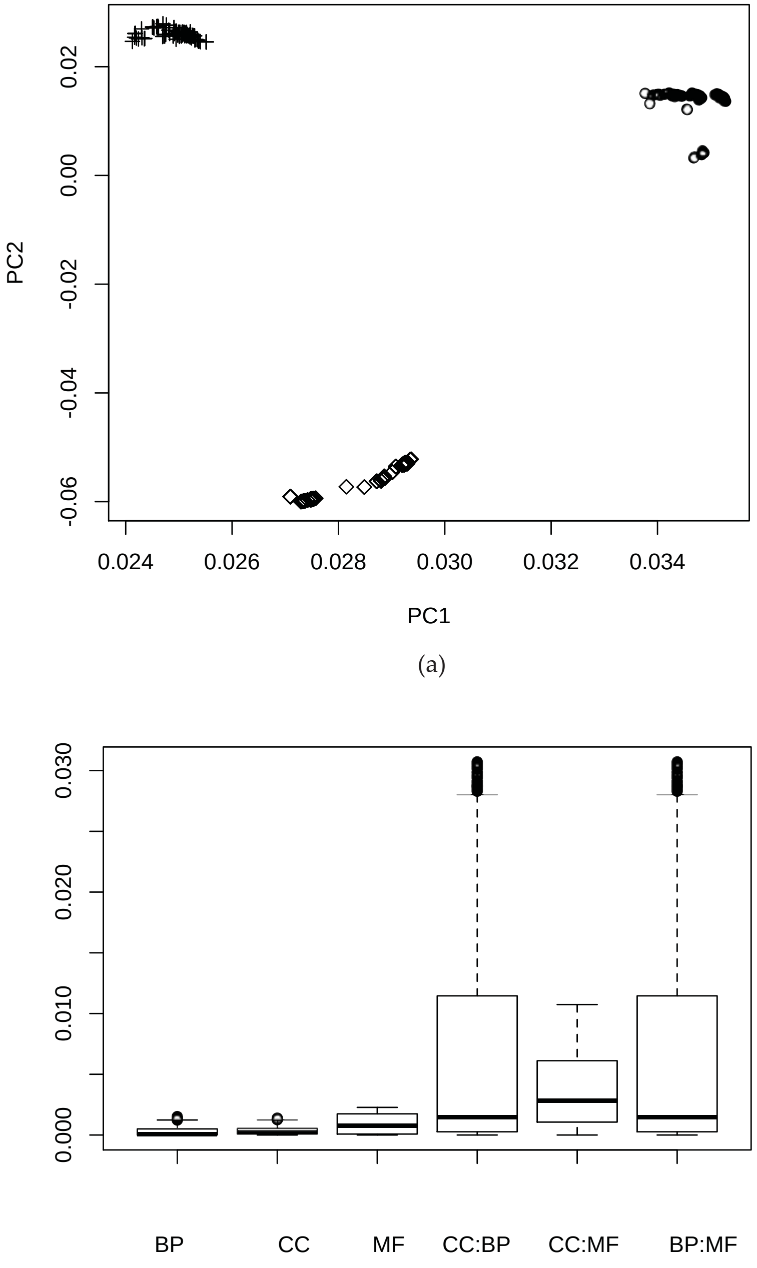
Figure 11: Plot of genes from the cancer dataset projected into PC1 and PC2 (a) without using a proximity matrix, (b) using the information-content proximity matrix. Legend: ○ is gene.



(a)

(b)

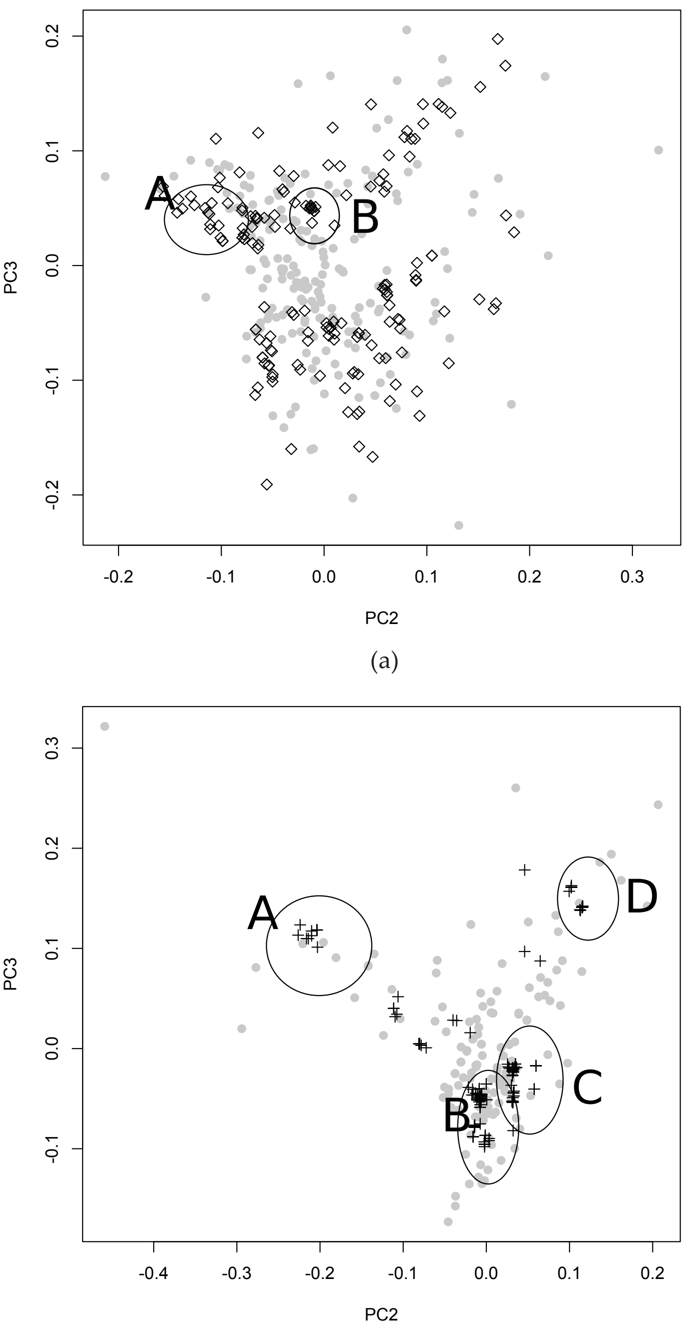
Figure 12: Plot of terms from the cancer dataset projected into PC2 and PC3 (a) using hop-based measure, (b) using the information-content based measure. Legend: ◇ is molecular function GO term, ○ is biological process GO term and + is cellular component terms.



(a)

(b)

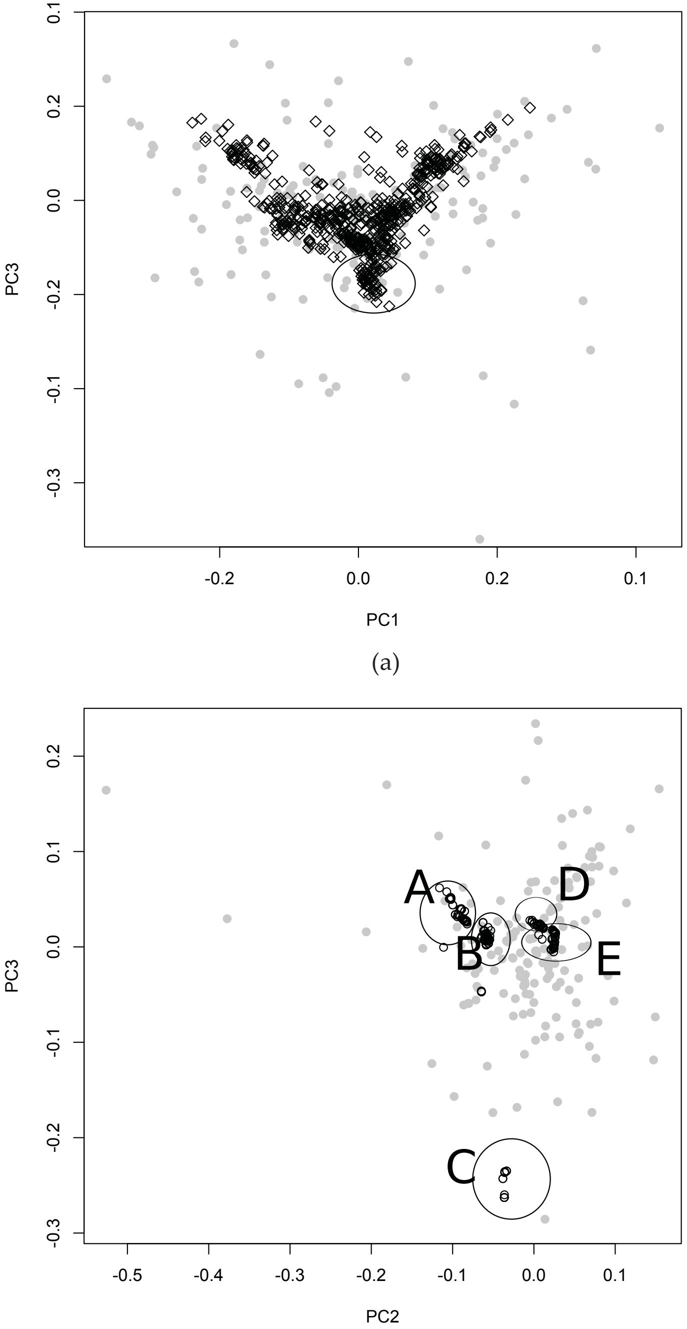
Figure 13: Terms from the cancer dataset projected into PC1 and PC2 using the IC similarity measure form clusters associated with the sub-ontology. (a) Plot of terms projected to PC1 and PC2, Legend: ◇ is molecular function GO term, ○ is biological process GO term and + is cellular component term (b) Distributions of distances inside and between clusters over PC1 and PC2.



(a)

(b)

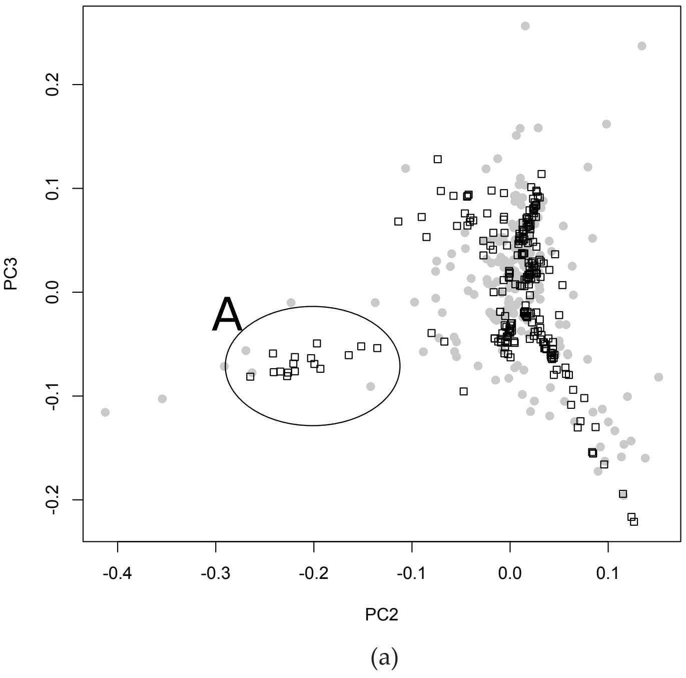
Figure 14: Plot of principal components 2 and 3 of cancer dataset with cellular component (CC) terms. (a) Hop-based similarity measure. Legend: (●) is genes and (◇) is CC terms. (b) IC similarity measure. Legend: (●) is genes and (+) is CC terms.



(a)

(b)

Figure 15: Plot of principal components 2 and 3 of cancer dataset with biological process (BP) terms. (a) Hop based similarity measure. Legend: (●) is genes and (◇) is BP terms (b) IC similarity measure. Legend: (●) is genes and (○) is BP terms.



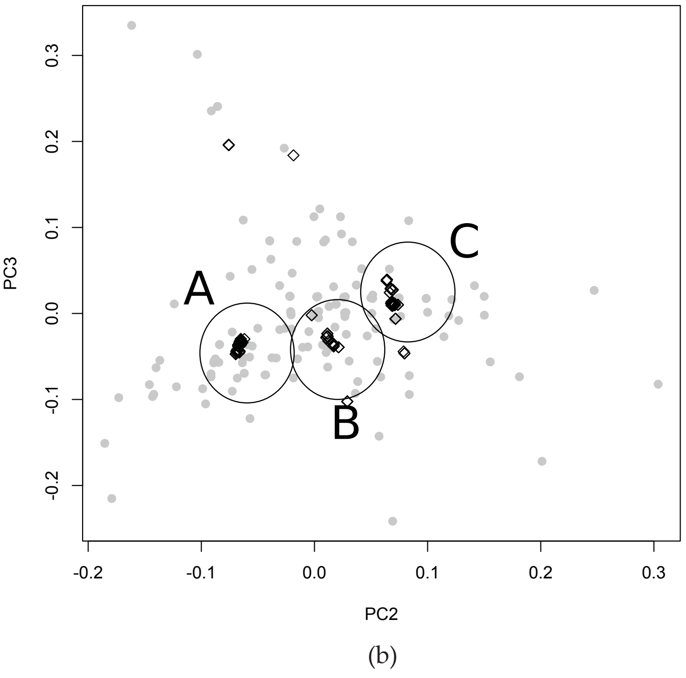


Figure 16: Plot of principal components 2 and 3 of cancer dataset with molecular function terms. Legend: (●) is genes and (◇) is molecular function terms. (a) Hop based similarity measure (b) IC similarity measure.

5. Conclusion

We applied SVD to lists of genes augmented with GO terms and inter-term similarities. Two datasets were visualised: validation data from KEGG and a set of genes identified experimentally. Results showed that principal component 1 measured the number of terms associated with genes. Later components allowed visualisation of genes according to their functional information, but the meaning of PCs varied depending on the underlying genes. For the KEGG data, PCs described gene functionality. For the larger cancer dataset, the early PCs simply identified known hierarchies. Separate visualisation using terms from the individual sub-ontologies was more informative. Correlation between GO terms and PCs improved understanding of the functional meaning of the PCs. These results show that our approach can bring meaningful biological interpretation to gene lists. Users should not expect that the meaning of the PCs should generalise from one gene list to another, apart from gross patterns such as the sub-ontologies. This is because different sets of GO terms will be associated with lists of genes and SVD will focus on those that explain the most variance. In practice, our approach should be applied to specific gene lists of interest to explore only the functional characteristics of those genes.

It is reasonable to suggest that rotation of components may find simpler factors. However, factor rotation requires the user to identify a priori how many components they wish to rotate. Rotating three components and then subsequently four components for the same data set does not mean that the first three components of the four will be the same. They could lie in different directions and therefore have different meanings. Because we do not know beforehand how many components we want to explore, we have not pursued factor rotation. However, this is something we intend to look at in the future.

We plan to address the bias towards genes with many terms by applying methods based on local distance measures. However, unlike the methods in this paper, those methods require parameter tuning, which in turn requires investigation of how to decide whether one visualisation is “better” than another. This will also involve comparing the visualisations derived using our approach more widely with other state-of-the-art methods. Variability of the quality of information throughout GO is an issue and we plan to investigate ways to deal with this.

We acknowledge that interpretation of our results is somewhat subjective. This is a problem generally with visualisation and unsupervised learning. We plan to investigate more informative and objective approaches to characterising clusters than simple Pearson correlation that can also take into account the level of GO terms in the hierarchies.

PC	GO term name and accession	Correlation
1	Number of terms	0.950
2	GO:0007067 (mitosis)	0.672
	GO:0051301 (cell division)	0.665
	GO:0007049 (cell cycle)	0.438
	GO:0006260 (DNA replication)	-0.498
3	GO:0009790 (embryonic development)	-0.353
4	GO:0006281 (DNA repair)	0.588
	GO:0006974 (response to DNA damage stimulus)	0.445
	GO:0000724 (double-strand break repair)	0.388
	GO:0006350 (transcription)	-0.488
	GO:0045449 (regulation of transcription)	-0.487

Table 5: GO terms from the biological process sub-ontology with absolute value of Pearson correlation > 0.35 for PCl-4 values for the cancer data set for the IC measure.

PC	GO term name and accession	Correlation
1	Number of terms	-0.872
2	GO:0043140 (ATP-dependent 3'-5' DNA helicase activity)	-0.604
	GO:0003678 (DNA helicase activity)	-0.575
	GO:0004003 (ATP-dependent DNA helicase activity)	-0.574
	GO:0009378 (four-way junction helicase activity)	-0.529
	GO:0003697 (single-stranded DNA binding)	-0.562
3	GO:0016301 (kinase activity)	-0.565
	GO:0004672 (protein kinase activity)	-0.533
	GO:0004674 (threonine kinase activity)	-0.571
4	GO:0004518 (nuclease activity)	0.670
	GO:0004527 (exonuclease activity)	0.650
	GO:0004523 (ribonuclease H activity)	0.589
	GO:0008409 (5'-3' exonuclease activity)	0.557

Table 6: GO terms from the molecular function sub-ontology with absolute value of Pearson correlation > 0.35 for PCl-4 values for the cancer data set for the IC measure.

Clusters	**Example Terms**	**Description**
CC Terms		
A	GO:0042175/nuclear envelope-endoplasmic reticulum network, GO:0005887/integral to plasma membrane	Cluster of membrane and extracellular matrix
B	GO:0005635/nuclear envelope, GO:0030117/membrane coat and GO:0000324/fungal-type vacuole	0rganelles
C	GO:0042719/mitochondrial inter-membrane space protein transporter complex, GO:0005760/gamma DNA polymerase complex and GO:0031588/AMP- activated protein kinase complex	Protein complexes
D	GO:0044430/cytoskeletal part, GO:0031616/spindle pole centrosome and GO:0000922/spindle pole	Cell division apparatus
BP Terms		
A	GO:0001658/branching involved in ureteric bud morphogenesis, GO:0048754/branching morphogenesis of a tube and GO:0001947 heart looping	Morphogenesis and Early Development (Stem Cells)
B	GO:0006974/response to DNA damage stimulus, GO:0007548/sex differentiation and GO:0007276/gamete generation	Response to Stimulus Transport or Homeostasis
C	GO:0010468/regulation of gene expression GO:0010628/positive regulation of gene expression and GO:0005975/carbohydrate metabolic process	Gene expression regulation and metabolism
D	GO:0048676/axon extension involved in development, GO:0045467/R7 cell development and GO:0007409/axonogenesis	Differentiation
E	GO:0000718/nucleotide-excision repair, DNA damage removal, GO:0000720/pyrimidine dimer repair by nucleotide-excision repair, GO:0000724/double strand break repair via homologous recombination	DNA metabolism and function with a number of small sub- groups e.g. vesicle transport

Continued on following page

MF Terms		
A	GO:0003678/DNA helicase activity, GO:0004003/ATP-dependent DNA helicase activity and GO:0008026/ATP-dependent helicase activity	Enzyme activity No.1
B	GO:0003777/microtubule motor activity, GO:0003774/motor activity and GO:0003924/GTPase activity	Enzyme activity No. 2
C	GO:0016853/isomerase activity, GO:0003689/DNA clamp loader activity and GO:0003916/DNA topoi- somerase activity	Molecular interactions non-enzymatic

Table 7: GO term clusters using IC method for Cellular Components(CC), Biological Process (BP) and Molecular Function (MF) of GO based on correlation results.

Acknowledgments

Hamid Ghous was partially supported by a University of Technology Sydney Challenge Grant. Daniel R Catchpoole holds a conjoint appointment with the School of Paediatrics and Child Health, Faculty of Medicine, University of Sydney. Nicholas Ho is funded by the Miles 2 Go Foundation. We wish to thank the anonymous reviewers for their comments, which have significantly improved our paper.

References

ASHBURNER, M., BALL, C.A., BLAKE, J.A., BOTSTEIN, D., BUTLER, H. and CHERRY, J.M. (2000): Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* 25(1), 25–29.

BACKES, C., KELLER, A., KUENTZER, J., KNEISSL, B., COMTESSE, N., ELNAKADY, Y.A. *et al* (2007): GeneTrail-advanced gene set enrichment analysis, *Nucleic Acids Research* 35 (suppl 2), W186–W192.

BERRIZ, G., BEAVER, J.E., CENIK, C., TASAN, M. and ROTH, F.P. (2009): Next generation software for functional trend analysis, *Bioinformatics* 25(22): 3043–3044.

BERRIZ, G. and ROTH, F. (2008): The Synergizer service for translating gene, protein and other biological identifiers, *Bioinformatics* 24(19): 2272–2273.

BREIMAN, L. (2001): Random forests, *Machine Learning* 45(1): 5–32.

CATCHPOOLE, D., GUO, D., JIANG, H. and BIESHEUVEL, C. (2008): Predicting outcome in childhood acute lymphoblastic leukemia using gene expression profiling: Prognostication or protocol selection? *Blood* 111(4): 2486–2487.

FLOTHO, C., SPEER, N., SPIETH, C. and ZELL, A. (2007): A set of genes that regulate cell proliferation predicts treatment outcome in childhood acute lymphoblastic leukemia, *Blood* 110(4): 1271–1277.

FRÖHLICH, H., SPEER, N., POUSTKA, A. and BEISSBARTH, T. (2007): GOSim-An R-package for computation of information theoretic GO similarities between terms and gene products, *BMC Bioinformatics* 8: 166.

FRÖHLICH, H., SPEER, N., SPIETH, C. and ZELL, A. (2006): Kernel based functional gene grouping, Neural Networks, 2006. *IJCNN'06. International Joint Conference*, 3580–3585.

GOLUB, G. and VAN LOAN, C. (1996): Matrix computations, Johns Hopkins University Press.

HOFFMANN, K., FIRTH, M., BEESLEY, A., DE KLERK, N. and KEES, U. (2006): Translating microarray data for diagnostic testing in childhood leukaemia, *BMC Cancer* 6(1): 229.

HUANG, D., SHERMAN, B.T. and LEMPICKI, R.A. (2008): Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists, *Nucleic Acids Research* 37(1): 1–13.

HUANG, D., SHERMAN, B., TAN, Q., COLLINS, J.R., ALVORD, W.G., ROAYAEI, J. *et al* (2007): The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists, *Genome Biology* 8(9): R183.

JOLLIFFE, I.T. (2004): *Principal Component Analysis*, second edn, Springer.

JONES, S., ZHANG, X., PARSONS, D.W., LIN, J.C., LEARY, R.J., ANGENENDT, P. *et al* (2008): Core signaling pathways in human pancreatic cancers revealed by global genomic analyses, *Science* 321(5897): 1801–1806.

KANEHISA, M., ARAKI, M., GOTO, S., HATTORI, M., HIRAKAWA, M., ITOH, M. *et al* (2008): KEGG for linking genomes to life and the environment, *Nucleic Acids Research* 36: 480-484.

KIM, Y-H. and KIM, H. (2007): Application of random forests to association studies using mitochondrial single nucleotide polymorphisms, *Genomics and Informatics* 5(4): 168–173.

LEE, S., HUR, H.U. and KIM, Y.S. (2004): A graph-theoretic modeling on GO space for biological interpretation of gene clusters, *Bioinformatics* 20(3): 381–388.

LORD, P., STEVENS, R., BRASS, A. and GOBLE, C. (2003): Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation, *Bioinformatics* 19(10): 1275–1283.

MATHUR, S. and DINAKARPANDIAN, D. (2007): A New Metric to Measure Gene Product Similarity, Bioinformatics and Biomedicine, 2007. *BIBM 2007. IEEE International Conference*, 333–338.

MISTRY, M. and PAVLIDIS, P. (2008): Gene Ontology term overlap as a measure of gene functional similarity, *BMC Bioinformatics* 9(1): 327.

POPESCU, M., KELLER, J., MITCHELL, J. and BEZDEK, J. (2004): Functional summarization of gene product clusters using Gene Ontology similarity measures, in *Proceedings of IEEE Intelligent Sensors, Sensor Networks and Information Processing Conference, IEEE*, 553–558.

RESNIK, P. (1995): Using information content to evaluate semantic similarity in a taxonomy, in *Int. Joint Conference on Artificial Intelligence*, 448-453.

RICHARDS, A., MULLER, B., SHOTWELL, M., COWART, A., ROHRER, B. and LU, X. (2010): Assessing the functional coherence of gene sets with metrics based on the Gene Ontology graph, *Bioinformatics* 26(12): i79.

SANFILIPPO, A., POSSE, C., GOPALAN, B., RIENSCHE, R., BEAGLEY, N. and BADDELEY, B. (2007): Combining Hierarchical and Associative Gene Ontology Relations With Textual Evidence in Estimating Gene and Gene Product Similarity, *IEEE Transactions on Nanobioscience* 6(1): 51–59.

SHEEHAN, B., QUIGLEY, A., GAUDIN, B. and DOBSON, S. (2008): A relation based measure of semantic similarity for gene ontology annotations, *BMC Bioinformatics* 9(1): 468.

SPEER, N., FRÖHLICH, H., SPIETH, C. and ZELL, A. (2005): Functional grouping of genes using spectral clustering and gene ontology, in *Proceedings of the IEEE International Joint Conference on Neural Networks*, 298–303.

TOMFOHR, J., LU, J. and KEPLER, T. (2005): Pathway level analysis of gene expression using singular value decomposition, *BMC Bioinformatics* 6(1): 225.

WARD, M.M., PAJEVIC, S., DREYFUSS, J. and MALLEY, J.D. (2006): Short-term prediction of mortality in patients with systemic lupus erythematosus: Classification of outcomes using random forests, *Arthritis and Rheumatism* 55(1): 74–80.

YI, G., SZE, S.H. and THON, M.R. (2007): Identifying clusters of functionally related genes in genomes, *Bioinformatics* 23(9): 1053–1060.

ZHANG, J., SOKAL, I., PESKIND, E., QUINN, J., JANKOVIC, J., KENNEY, C. *et al* (2008): CSF multianalyte profile distinguishes Alzheimer and Parkinson diseases, *American Journal of Clinical Pathology*, 129(4): 526–529.

Biographical Notes

Hamid Ghous

Hamid Ghous *is a PhD student at the University of Technology, Sydney. He is working on Visual Analytics methods for discovering and evaluating functional relationships between genes.*

Nicholas Ho

Nicholas Ho *is a computational biologist at The Children's Hospital at Westmead. He received his Bachelor of Science (Hons 1) in Bioinformatics from the University of Sydney. His current research areas include paediatric cancers, genomics, bioinformatics and machine learning.*

Paul Kennedy

Paul Kennedy *has a PhD (Computing Science) and joined UTS in 1999 where he is currently Director of the Knowledge Infrastructure Laboratory in the Centre for Quantum Computation and Intelligent Systems. Dr Kennedy's research interests are currently in the area of data analytics and visualisation of large complex data sets, particularly those in the biomedical domain, but also in customer sales and text mining. For the past nine years he has been developing approaches to better diagnose and treat childhood cancer sufferers (acute lymphoblastic leukaemia and neuroblastoma). Recently he has been developing bioinformatics approaches for reverse vaccinology to combat animal parasites.*

Daniel Catchpoole

Associate Professor Daniel Catchpoole *was appointed Head of the Tumour Bank at the Children's Cancer Research Unit within The Children's Hospital at Westmead in 2001. His research career has focused on the molecular basis of paediatric malignancies in which he has extensive experience exploring genomics in paediatric tumours. His scientific achievements and publications have focused on the assessment of childhood tumours, with specific attention given to acute lymphoblastic leukaemia and neuroblastoma. He has developed a panomics approach to the assessment of cancer patients which includes the implementation of the data-mining and visualisation of complex multidimensional biomedical data derived from various high-throughput applications.*