Functional Visualisation of Genes using Singular Value Decomposition

Hamid Ghous[1] **Paul J. Kennedy**[1] **Nicholas Ho** [2] **Daniel R. Catchpoole** [2]

[1] Centre for Quantum Computation and Intelligent Systems,
School of Software, Faculty of Engineering and Information Technology,
University of Technology, Sydney, PO Box 123, Broadway NSW 2007, AUSTRALIA
Email: Hamid.Ghous@student.uts.edu.au, Paul.Kennedy@uts.edu.au

[2] 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.
Email: Nicholah@chw.edu.au, DanielC@chw.edu.au

Abstract

Progress in understanding core pathways and processes of cancer requires thorough analysis of many coding regions of the genome. 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), has the potential to assist in functional understanding. This paper addresses the question of how well GO annotations can help in functional understanding of genes. We augment genes with associated GO terms and visualise with Singular Value Decomposition (SVD). Meaning of derived components is further interpreted using correlations to GO terms. The results demonstrate that SVD visualisation of GO–augmented genes matches the biological understanding expected in the simulated data and presents understanding of childhood cancer genes that aligns with published results.

Keywords: singular value decomposition, visualisation, genes, gene ontology.

1 Introduction

It is becoming clear that progress towards new insights in cancer treatment require 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. However, the sheer scale of this data makes understanding by humans challenging. Also, as integration of multiple datasets becomes commonplace, for example using single nucleotide polymorphisms or the proteome, making sense of the data becomes even more difficult. 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 may not assist in understanding gene function. For these reasons, 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.

There has been much recent work to explore the problem of applying unsupervised learning methods to lists of genes. Work generally falls into two main areas: defining similarity measures using GO annotations and applying unsupervised methods to visualise the functional relationship between genes. 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. Richards et al. (2010) assess functional coherence of a gene set using both a graph–based similarity measure and an information content similarity measure. Mistry & Pavlidis (2008) 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. The greater the number of common terms the higher the similarity. Mathur & Dinakarpandian (2007) use the hierarchical structure of GO to compute similarity between gene products on the basis of common GO terms. Sanfilippo et al. (2007) propose a cross–ontological approach that exploits similarity measures over the 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) find functionally similar genes in close proximity on chromosomes. Lee et al. (2004) find clusters of genes according to significant biological features using the hierarchical GO structure. They define a similarity measure by transforming the directed acyclic graph structure of GO into a distance function, which results in clusters of genes with similar terms or functionality. Similarly Popescu et al. (2004) 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.

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 based on singular enrichment analysis, gene set enrichment analysis and modular enrichment analysis. These categories 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 & 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. Speer et al. (2005) and Fröhlich et al. (2007) cluster genes with an information–theoretic kernel function to calculate the similarity between genes using GO. The motivation behind this approach as opposed to a distance measure using the distance 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. However few of these reviewed methods are used in routine biomedical research.

In this paper we apply singular value decomposition to visualisation of genes. Our motivation for applying SVD compared to other dimensionality reduction methods such as Principal Component Analysis (PCA) is that genes and terms may be visualised on the same graph. This allows improved understanding of the biological function of genes. The approach is applied to two data sets: a data set used to validate the approach composed of genes selected from the KEGG database (Kanehisa et al. 2008) and a data set 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 Metacore and Ingenuity 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 Methods

2.1 Singular Value Decomposition

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

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T \tag{1}$$

Row vectors of $\mathbf{U}$ relate to the original data points (rows of $\mathbf{X}$) and rows of $\mathbf{V}$ are associated with the data attributes (columns of $\mathbf{X}$). The columns of $\mathbf{U}$ are called the left singular vectors of $\mathbf{X}$ and columns of $\mathbf{V}$ are called the right singular vectors. The elements of $\mathbf{D}$ are termed the singular values of $\mathbf{X}$. 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 use the later principal components to extract superior understanding.

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

2.2 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 $\mathbf{X} \in \mathbb{R}^{n\times t}$ where n is the number of genes $|G|$ and t the number of GO terms $|T|$. Each element x_{ij} of $\mathbf{X}$ 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 $\mathbf{P} \in \mathbb{R}^{t\times t}$ is created with elements $0 \leq p_{ij} \leq 1$ representing the proximity (or similarity) between GO terms i and j. Terms with a close relationship have values close to 1, with the diagonal elements $p_{ii} = 1$. The proximity between GO terms is based on the number of links (or distance) between them and is defined as $p_{ij} = (d_{ij} + 1)^{-1}$ where d_{ij} 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. The augmented data matrix is defined as $\mathbf{X}' = \mathbf{X}\mathbf{P}$. SVD is applied to $\mathbf{X}'$ after centring and normalisation. Whilst proximity matrices have been used for text kernels, we are unaware of their use with GO terms. Pearson correlation between GO terms and data projected into PC space is calculated and important terms are those with higher absolute values of correlation.

2.3 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.

2.3.1 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 to 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 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 than to class 1). Genes in classes 4 and 5 should be similar to one another but different to the other classes.

2.3.2 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 ALL dataset was constructed by building on previous work by Flotho et al. (2007) and Catchpoole et al. (2008). Flotho et al. reported a fourteen gene signature (encompassed by fifteen Affymetrix expression probesets) that separated a cohort of ALL patients treated at the St. Jude Children's Research Hospital into two distinct groups. The observed separation was associated with relapse potential, leading the investigators to conclude the fourteen gene signature as predictive of relapse. Catchpoole and colleagues examined these fourteen genes and found that the signature produced a separation in their cohort of ALL patients treated at The Children's Hospital at Westmead. However, the separation observed in this cohort was not associated with relapse nor treatment outcome.

To further explore this separation observed by both groups of investigators, Ho et al. (submitted) applied Random Forest to identify other probesets that further support this separation. The authors identified the 250 most important probesets that underlie this patient separation and found that the genes encompassed by these probesets are heavily involved in the cell cycle, mitosis, DNA replication, apoptosis and DNA damage repair mechanisms. Our study will use these 250 probesets for further analysis by SVD and Expectation Maximisation clustering to explore their findings.

3 Results

3.1 Visualising KEGG data set

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., $\mathbf{X}$), the total number of GO terms for each gene and the gene class. There was 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.

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 1, 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 followed by strong positive correlations to GO terms describing carbohydrate metabolism and negative correlations to terms associated with ribosomes.

Apart from the outlier *RHO*, Figure 1 shows that PC3 separates 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.

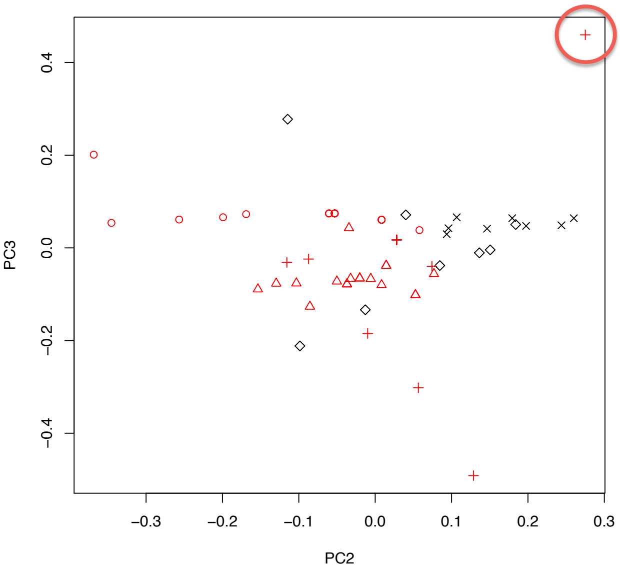

Figure 1: KEGG genes by PC2 and PC3. Ribosome ○; RNA polymerase △; transcription +; pentose phosphate pathway ×; pentose/glucoronate interconversions ◇. PC2 and PC3 are the axes for principal components 2 and 3 respectively.

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

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, RPS21, RPS23, RPS24, RPS25, rpmB, rpmC, rpmD, rpmE, rpmF*
2 △	genetic information processing : transcription : **RNA polymerase** *FLIA, RPOA, RPOB, RPOZ, RPOH, RPON, RPOD, RPB2, RPB1, RPB3, RPA49, RPA14, RPA34, RPA43, RPA12, RPC19, RPC25, RPB7, RPB4*
3 +	genetic information processing : **transcription** *GREA, GREB, NUSA, NUSB, NUSG, MBF1, Rcl1, 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, AKR1, mtlY, mtlD, clpX*

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

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

3.2 Visualising cancer dataset

As with the KEGG dataset, there is a strong correlation between the number of GO terms associated with the genes and principal component 1 (PC1). The second, third and fourth PCs separate GO terms by their respective subontologies as shown in Figure 2. This suggests unsurprisingly that most of the variance in the dataset is based on technicalities rather than biological factors. Consequently we split the GO terms according to the three sub–ontologies and performed SVD on each individually.

Results for the Cellular Component GO terms, shown in Figure 3 (top) highlight two clusters of terms, separated along the PC3 axis. Pearson correlation between GO terms and the PCs (see Table 3) reveals that the separation between terms is associated with the cytoplasmic structure (e.g. GO:0005856 *cytoskeleton* and GO:0005874 *microtubule*) and DNA replication (e.g. GO:0031298 *replication fork protection complex* and GO:0042555 *MCM complex*).

For the Biological Process terms in Figure 3 (middle) PC2 separates terms associated with cell division (e.g. GO:0007067 *mitosis* and GO:0051301 *cell division*) from those related to DNA replication (cluster A). PC3 reveals a tight group of terms (cluster B in Figure 3 middle) associated with development (e.g. GO:0009790 *embryonic development* and GO:0030903 *notochord development*). See also Table 4.

For the Molecular Function terms in Figure 3 (bottom), PC2 shows a cluster of terms separate from the main grouping (cluster C) that is related to DNA helicase activity (see Table 5). Located in close prox-

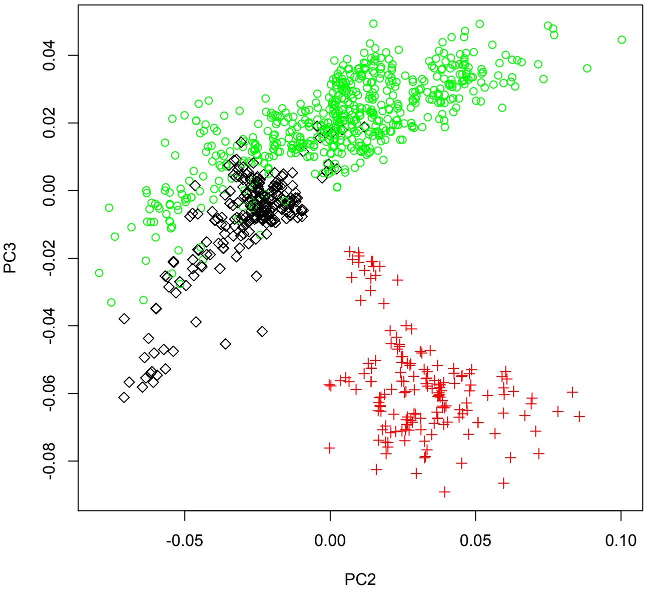

Figure 2: Plot of GO terms by PC2 and PC3 for cancer data. Terms labelled by sub-ontology: cellular component (red +), molecular function (black ◇) and biological process (green ○). PC2 and PC3 denote axes for principal components 2 and 3 respectively.

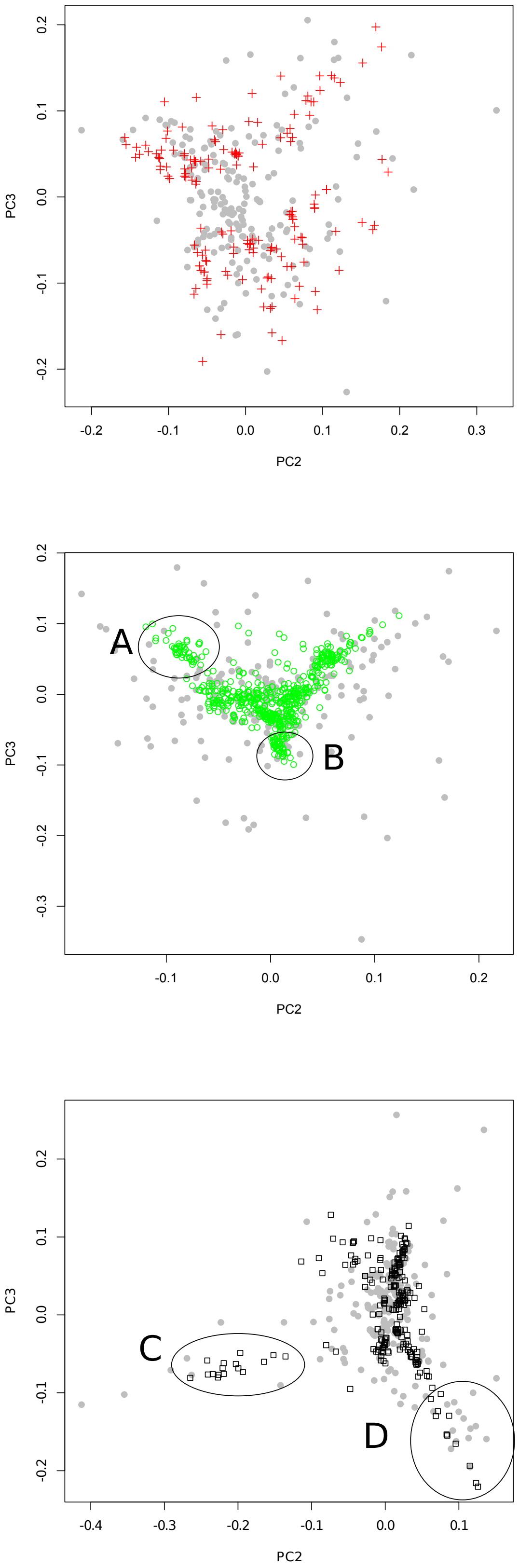

Figure 3: Plot of principal components 2 and 3 for **U** matrix (genes = •) and **V** matrix (terms: red + for cellular component, green ∘ for biological process and black □ for molecular function) for the cancer dataset. Top: cellular component terms, Middle: biological process terms, Bottom: molecular function terms. Circled clusters A, B, C and D are described in the text.

imity is a loose cluster of six genes (the grey circles) that code for mini chromosome maintenance proteins (*MCM2–MCM7*). Both *MCM* and replicative helicase play integral roles in eukaryotic DNA replication: the *MCM* protein complex formed by *MCM2–MCM7* is involved in initiation (Costa & Onesti 2008) and replicative helicase is an enzyme that plays an role in unwinding the strands (Johnson et al. 2007). *MCM10*, the remaining MCM gene in the data is found in the main group of genes rather than the cluster. Whilst *MCM10* is involved in DNA replication (Chattopadhyay & Bielinsky 2007) and interacts with *MCM2–MCM7*, it is not part of the *MCM2–MCM7* family (Merchant et al. 1997) and our visualisation can highlight this.

Cluster D in Figure 3 (bottom) shows Molecular Function terms associated with kinase activity. This groups genes with roles in mitosis and, in particular, in mitotic spindle checkpoint signalling and includes *NEK2*, which has been shown to be an important protein in mitotic checkpoint signalling (Lou et al. 2004) and *BUB1*, which functions as a regulator of spindle assembly and has been shown to lead to aneuploidy in leukemic cells lines if mutated (Ru et al. 2002). Also within this cluster is thymidine kinase 1 (*TK1*), which has been reported to be predictive of remission duration (Jahns-Streubel et al. 1997) and relapse (Votava et al. 2007) in acute leukemias, and is essential to DNA synthesis.

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.

4 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 subontologies 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.

We plan to explore other similarity measures, specifically an information–theoretic one (Speer et al. 2005). We will address the bias to 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 gener-

Table 3: GO terms from the cellular component sub-ontology with absolute value of Pearson correlation > 0.35 for PC1–4 values from the cancer data set.

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 biological process sub-ontology with absolute value of Pearson correlation > 0.35 for PC1–4 values for the cancer data set.

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 molecular function sub-ontology with absolute value of Pearson correlation > 0.35 for PC1–4 values for the cancer data set.

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

ally 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.

References

Ashburner, M. et al. (2000), 'Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium.', *Nature Genetics* **25**(1), 25–9.

Backes, C. et al. (2007), 'GeneTrail–advanced gene set enrichment analysis', *Nucleic Acids Research* **35**(suppl 2), W186–W192.

Berriz, G. & Roth, F. (2008), 'The Synergizer service for translating gene, protein and other biological identifiers', *Bioinformatics* **24**(19), 2272.

Berriz, G. et al. (2009), 'Next generation software for functional trend analysis', *Bioinformatics* **25**(22), 3043.

Catchpoole, D. et al. (2008), 'Predicting outcome in childhood acute lymphoblastic leukemia using gene expression profiling: Prognostication or protocol selection?', *Blood* **111**(4), 2486.

Chattopadhyay, S. & Bielinsky, A. (2007), 'Human Mcm10 regulates the catalytic subunit of DNA polymerase-α and prevents DNA damage during replication', *Molecular Biology of the Cell* **18**(10), 4085.

Costa, A. & Onesti, S. (2008), 'The MCM complex: (just) a replicative helicase?', *Biochemical Society Transactions* **36**, 136–140.

Flotho, C. et al. (2007), 'A set of genes that regulate cell proliferation predicts treatment outcome in childhood acute lymphoblastic leukemia', *Blood* **110**(4), 1271.

Fröhlich, H. et al. (2007), 'GOSim–An R-package for computation of information theoretic GO similarities between terms and gene products', *BMC Bioinformatics* **8**, 166.

Golub, G. & Van Loan, C. (1996), *Matrix computations*, Johns Hopkins University Press.

Huang, D. 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.

Huang, D. et al. (2008), 'Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists', *Nucleic Acids Research* **37**(1), 1–13.

Jahns-Streubel, G. et al. (1997), 'Activity of thymidine kinase and of polymerase alpha as well as activity and gene expression of deoxycytidine deaminase in leukemic blasts are correlated with clinical response in the setting of granulocyte-macrophage colony-stimulating factor-based priming before and during TAD-9 induction therapy in acute myeloid leukemia', *Blood* **90**(5), 1968–1976.

Johnson, D. et al. (2007), 'Single-molecule studies reveal dynamics of DNA unwinding by the ring-shaped T7 helicase', *Cell* **129**(7), 1299–1309.

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

Jones, S. et al. (2008), 'Core signaling pathways in human pancreatic cancers revealed by global genomic analyses', *Science* **321**(5897), 1801–1806.

Kanehisa, M. et al. (2008), 'KEGG for linking genomes to life and the environment', *Nucleic Acids Research* **36**, 480–484.

Lee, S. et al. (2004), 'A graph-theoretic modeling on GO space for biological interpretation of gene clusters', *Bioinformatics* **20**(3), 381–388.

Lou, Y. et al. (2004), 'NEK2A interacts with MAD1 and possibly functions as a novel integrator of the spindle checkpoint signaling', *Journal of Biological Chemistry* **279**(19), 20049.

Mathur, S. & Dinakarpandian, D. (2007), 'A New Metric to Measure Gene Product Similarity', *Bioinformatics and Biomedicine, 2007. BIBM 2007. IEEE International Conference on* pp. 333—338.

Merchant, A. et al. (1997), 'A lesion in the DNA replication initiation factor Mcm10 induces pausing of elongation forks through chromosomal replication origins in Saccharomyces cerevisiae', *Molecular and Cellular Biology* **17**(6), 3261.

Mistry, M. & Pavlidis, P. (2008), 'Gene ontology term overlap as a measure of gene functional similarity', *BMC Bioinformatics* **9**(1), 327.

Popescu, M. et al. (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, pp. 553–558.

Richards, A. et al. (2010), 'Assessing the functional coherence of gene sets with metrics based on the Gene Ontology graph', *Bioinformatics* **26**(12), i79.

Ru, H. et al. (2002), 'hBUB1 defects in leukemia and lymphoma cells.', *Oncogene* **21**(30), 4673–4679.

Sanfilippo, A. et al. (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. et al. (2008), 'A relation based measure of semantic similarity for gene ontology annotations', *BMC Bioinformatics* **9**(1), 468.

Speer, N. et al. (2005), Functional grouping of genes using spectral clustering and gene ontology, *in* 'Proceedings of the IEEE International Joint Conference on Neural Networks', pp. 298–303.

Tomfohr, J. et al. (2005), 'Pathway level analysis of gene expression using singular value decomposition', *BMC Bioinformatics* **6**(1), 225.

Votava, T. et al. (2007), 'Changes of serum thymidine kinase in children with acute leukemia', *Anticancer research* **27**(4A), 1925.

Yi, G. et al. (2007), 'Identifying clusters of functionally related genes in genomes', *Bioinformatics* **23**(9), 1053.