Assigning putative protein products to SAGE tags by determining their predisposition to associate with clusters which exhibit functional enrichment.

John Boyle

Abstract

The fact that genes associated with parts of cellular machinery can be collectively controlled means that it is theoretically possible to assign phenotypic function to certain genes by analysis of their expression profiles. However the multi-faceted nature of genetic function, and the complexities of the signal and noise interactions that are inherent in a cell, means that whilst illustrations of such mediated control (for example the repression of translational machinery) have been well studied it is difficult to use them to reliably ascertain genetic function for individual genes. In this paper the Gene Ontology is used to classify and then compare clusters from different sets of SAGE samples to discover which tags have a strong predisposition to reside in a particular type, or types, of cluster. This predisposition is then used to resolve ambiguities in tag assignment that have arisen by using sequence homology to predict protein products.

Introduction

A number of techniques have been proposed for the analysis of gene expression data sets (for reviews see [Sharan 2002, Szabo 2002, Pan 2001]). As has been suggested previously [Li 2002, Gat-Viks 2003], it is the biological significance of the results that is important; particularly where expression patterns can be related to identifiable biological phenomena, such as disease pathology [Getz 2000, Golub 1999]. As gene expression profiles will be influenced by a variety of signal factors (be it their response to transcription factors, environmental conditions, the level of RNA degradation, cell cycle variability, translational pauses or just their chromosome location) the desired underlying biological behaviour can be hard to differentiate from 'noise' (due to individual variability, multi-functional genes, differing experimental conditions, different levels of annotations and our lack of understanding). To enable understanding of an experiment, underlying patterns within the data need to be identified and then correctly mapped to known functional behaviour.

To aid in such understanding, it is possible to assign putative functionality to clusters of genes based on their expression profiles; this is often done by measuring their 'functional category enrichment' [Tavazoie 1999]. Such functional assignment can be performed by automatically correlating sets of genes with similar functionality (e.g. part of ribosomal machinery, cyclin activity) to statistically defined components which are elucidated by their expression profile. It is then possible to assign a series of putative functionalities to genes with unknown functionality depending on their level of involvement in these calculated components. Such a method depends on being able to generate components from a matrix of gene expression values which are able to define the correct granularity of functionality: too specific and group based functionality can not be assigned; too general and the functionality ranges will be too broad. With the advent of the Gene Ontology [Gene Ontology Consortium 2000], there exists a standard classification system for such genetic functionality. Work on the use of expression profiles to automatically generate GO annotations for unknown genes has generally focused on supervised learning techniques [Hvidsten 2002], and has mainly used gene chip time course experiments [Lagreid 2003, Hvidsten 2003]. Gene functionality assignment has also been undertaken using techniques based on proteomics data [Deng 2004] and literature mining [Perez 2004, Chiang 2003].

Little work has been done on assignment of functionality to SAGE tags using their expression profiles. Unsurprisingly, the principle technique for determining the functionality of a tag is homology searching [Lash 2000]. However, due to the inherent error rate with ESTs and SAGE itself, there exists a large number of tags which do not have 'reliable' mappings, but either have ambiguities or there is no known assignment.

In this paper the Gene Ontology is used to classify and then compare clusters from different sets of SAGE experiments to discover which tags have a strong predisposition to reside in a particular type, or types, of cluster. This analysis allows for the possibility of deducing functionality of a SAGE tag based on its expression profile, as well as the ability to resolve ambiguities in homology-based SAGE tag assignment. Using such ontology defined clusters has a number of advantages:

Biological relevance. The 'meta data', rather then the tag identity, about the sets of genes is used to define the function of the cluster. This definition is determined using the GO defined function enrichment of the cluster, so that the comparison of clusters is assigned based on biological relevance. Therefore, the functional definition of the clusters attempts to represent a range of biological behaviours which describes the complexities of the interactions that occur within a cell. If the clustering method is able to distinguish between different genetic controls, then similar clusters will appear within the different samples. That is not to say that other types of information could not be used to describe the clusters (e.g. proteomics data, genome location), however the abstract nature of GO makes it suitable for such functional definitions. Ontology terms have been used previously to identify components within a system; such approaches have been used in supervised learning based analysis, either through ontology mapping [Midelfart 2001, Laegreid 2003] or disease classification [Dudoit 2002]

Robustness. The complexities of analysing SAGE data have been well reported [Man 2000, Baggerly 2003]. The large number of tags and the incomplete assignments, coupled with the non-parametric distribution of the experimental results, undermines the successful use of this powerful exploratory technique. The use of functional definitions on the generated clusters provides a level of immunity from errors that arise due to both the absence of tags from certain experiments and incorrect annotations.

Additionally, as the analysis uses a procedure of sampling from a number of collections of SAGE experiments, underlying biases that reside within the data can be identified separately to noise factors.

The next section outlines the methods that were used to calculate the predisposing of tags to reside within different clusters. The results section gives the results from a 100 SAGE experiment analysis, and shows how the technique can be used to predict function for SAGE tags.

Methods

To assign protein function to a SAGE tag, based on its expression profile, a methodology must be developed which demonstrates that the assignment has biological evidence and has a significant probability of being correct. The methodology adopted in this paper for such an analysis of the tags involves four steps:

1)Subdivide a number of SAGE experiments into equal size samples (in this case 100 SAGE experiments were divided into 10 samples of equal size). The data items will then be pre-processed using both ranking and projection techniques.

2)For each sample find clusters using appropriate techniques. Three techniques were used, these were based on finding clusters using Euclidian distances, Cosine distances and a Semi-discrete decomposition [O'Leary 1983].

3)Assign functional enrichment scores to the clusters. The enrichment scores for the clusters were based upon the ontology terms of their members.

4)Finding the tags that occur within similar clusters in a significant number of samples. This was done by scoring each tag by the frequency with which it exists within a cluster with a given functional enrichment in each of the ten samples. The significance of the result was determined by using both a model for the distribution of results (presuming no underlying bias within the data) and by using random samples.

The following sections discuss in detail the techniques that were used in each of the steps. More details about the samples used can be found on the web site.

Processing the data

In this paper, both projection and ranking [Wilcoxon 1945] were used to process the data before cluster analysis was performed. The use of such techniques has become increasingly popular in microarray analysis [Wall 2003, Troyanskaya 2002]. One of the more commonly used projection techniques to identify components in gene expression involves the identification of Principle components [Dysvik 2001, Sturn 2000, Raychaudhuri 2000]. Singular-value decompositions (SVD) have been used to find Principle components, and have been applied extensively in expression analysis (for robust analysis [Liu 2003], cross species comparison [Alter 2003], as well as (genome based) principle component identification [Alter 2000]). Principle components are orthogonal vectors which represent the maximum amount of variance within the chosen data set. By projecting out the data using these vectors the similarity between the vectors and each of the genes can be readily identified. In this paper, SVDs were used to project the data so that it was possible to find clusters of genes, with each group exhibiting a degree of variance with the others.

The effects of the ranking and projection techniques were exhaustively explored, so that for each sample four different levels of pre-processing were performed: 1) no pre-processing used; 2) the data underwent a global ranking procedure; 3) the data was projected using SVDs; and 4) the data was ranked and then projected.

Cluster Discovery

A number of different filtering and distance functions have been suggested for comparing SAGE data [Ng 2001]. In this paper two distance functions (Euclidian distance and Cosine distance) are used to explore the phenotype predictive power of gene expression profiles. Additionally the Semi-Discrete matrix decomposition method is used to identify sub parts of the gene expression profiles which contribute towards areas which have values above or below threshold values (areas of high density or 'bumps'). The clustering techniques that are described are customised to deal with the issues of outliers and non-normal distributions that arise when analysing SAGE data. These three different clustering techniques are discussed below:

Euclidian Distance. The K-Means algorithm was altered to make it suitable for working with the sparseness of SAGE data. A large number of initial centroids were chosen by randomly selecting data points within the data set, in a manner similar to that suggested by [Forgy 1965]. By using the data points themselves as the initial centroids (rather than the points with the furthest distance) a large number of clusters were generated for the lower expression levels. If during an iteration cycle a cluster had less than a minimum threshold of members it was removed. Such a removal of clusters causes a larger decrease in the number of clusters within areas where there is little variation or sparsely populated areas, reflecting the distribution of data points within a typical SAGE experiment. Additionally an anchoring procedure was introduced, so that the resulting centroids within a sample were adjusted at each iteration to match those of a 'real value'. Such an anchoring technique is used to minimise the effect of outliers. The use of a Euclidian distance function attempts to identify clusters of genes which are not only regulated in a similar manner, but are also expressed to approximately the same extent within the different experiments.

Cosine Distance. To identify groups that exhibit co-expression a cosine similarity test was implemented. This measures the difference in the angles between the expression profiles rather than the Euclidian distance. Genes are grouped by iteratively calculating and fitting the centroids, with the distance model being the cosine between the vectors, that is to say: $\cos(\alpha) = (a \bullet b) / |a| * |b|$. Convergence is reached when the size and shape of the components within the system remains unaltered. As with the Euclidian distance measure, values were seeded from a set of real vales, and during iterations the resulting model was altered to be the same as the nearest matching result. Such an approach attempts to find genes that whilst they may not be

expressed in the same amount, they are regulated in a similar manner.

Semi-Discrete Decomposition

SDD was originally developed for image compression. SDD attempts to identify the most significant 'bump' (area of local density) in a matrix, and then calculates which items are affected by it. It then removes the bumps from the matrix and attempts to find the set of next most significant bumps. This process is continued until a pre-defined number of bumps have been found, and the items which contribute towards them have been assigned. An SDD iteration involves:

- Find the nth cluster
- Construct a matrix (A') which represents the previous cluster by taking the product of X[n-1]*D[n-1]*Y[n-1]. This matrix represents the previous concepts contribution to the local densities in the matrix.
- Subtract A' from A, giving a matrix (B) with the density approximations removed.
- Find the next set of densities, and approximate the local max/min

In a similar manner to SVDs, the semi discrete decomposition will decompose a matrix into three separate matrices (e.g. a rotation, a scale and a rotation).
e.g. Anm=XnrDrYrm' (where A is the original matrix.). The entries in X and Y are limited to members of the vector {1, 0,-1} (see Figure 1)

$$\begin{Bmatrix} 5 & 5 & 0 & 0 \\ 5 & 5 & -3 & -3 \\ 0 & 0 & 3 & -3 \end{Bmatrix} \approx \begin{Bmatrix} 1 & 0 & 0 \\ 1 & -1 & -1 \\ 0 & 1 & -1 \end{Bmatrix} \bullet \begin{Bmatrix} 5 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 3 \end{Bmatrix} \bullet \begin{Bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{Bmatrix}$$

$$A = X \bullet D \bullet Y^T$$

Figure 1: The decomposition has been used to identify three factors. Examination of the X matrix shows how the different genes are involved in the 'bumps' that occur within the matrix. The first factor (column) is contributed towards by the first and second genes (rows) of A, the D matrix shows the first factor has a significance (height) of 5, and the Y matrix shows that the first and second experiment are significant in contributing towards the factor. The second factor (which has a significance of 3) is contributed towards by genes 2 and 3, although they contribute orthogonally. The final factor is made up of genes 2 and 3, and they both contribute in the same manner.

SDD finds a range of areas within the matrix and approximates the average of these values into one factor, thereby combining a range of complex factors into one concept. Therefore, SDD will find small clusters which represent genes which contribute towards 'bumps' within the expression data. Whilst the other clustering methods used in this analysis clusters find similarities in the whole of the expression pattern. SDD is useful in finding small clusters of genes that exhibit similar behaviour within subsets of expression data. SDD has been shown to work with high dimensional non-normal data; in particular involving sparsely populated data sets with discrete values (thus, is suitable for SAGE data).

Local Density. An alternative to using overall distance between the different expression profiles is to compare local sub-spaces within the data. In this paper the use of a semi-discrete decomposition (SDD) is proposed, which approximates the contribution that genes make to local densities within the gene/expression matrix. As the derived decomposition approximates the areas of local density that occur within a matrix, the resulting calculation is less effected by outliers and extreme values which are typically found within a matrix of gene expression values. More information is available in the Semi-Discrete Decomposition Box.

Calculating the Functional Enrichment

Where possible each tag is mapped to an associated set of ontology terms. In the SAGE data sets this is done using the NCBI SAGE to Unigene Unigene to Locuslink, and then the GOA Locuslink to GO mappings (the mappings were current as of March 2004). Where multiple mappings are provided no preference is given. The exception to this is the SAGE mappings, where the most probable mapping (or the two best mappings) is used to obtain the Unigene annotations.
Each gene is associated with its set of ontology terms and their ancestors; this is done by following the ISA relationship only (see Figure 2). Such a use of the relationships in the Gene Ontology allows for a richer definition of clusters. For example a cluster which contains a high proportion of histone proteins, which have been identified by their expression profiles due to their involvement in transcription, would be described as being involved in a number of related biological processes (chromatin assembly/disassembly, chromosome organisation) and having a number of molecular functions (DNA binding, nucleic acid binding) and being located in particular cellular components (nucleus). In this paper each ontology term that describes a cluster is referred to as one of the clusters *facets*. The multiple facets of a cluster enables the description of the contained genes in their entirety, unfortunately this also means that only a subset of each of the facets will accurately describe an individual Tag within a cluster. This means that whilst a complete description of a cluster is a useful tool, it does introduce erroneous results when using such descriptions as an indicator of individual tag's protein function.

Using the ontology terms that are associated with each cluster the functional enrichment can be modelled as a binomial distribution, with the probability of x genes marked up with a specific ontology term occurring (with replacement) in a sample of size n being: $\left(n!/\left(x!*(n-x)!\right)\right)* p^x\left(1-p\right)^{n-x}$, where p is the probability of a given gene having a specific ontology term. For each refined cluster the ontology terms that have been mapped to the member genes are examined to see if the probability of this event occurring without bias is less than 0.01, if so then the cluster is marker as being *significant*. Due to the size of the data sets a binomial, rather than hypergeometric, probability is used.

Terms	No of Genes	Genes
Term 1	4	1,2,5,6
Term 2	3	1,2,5
Term 3	2	1,8
Term 4	2	1,2
Term 5	2	3,4
Term 6	1	1

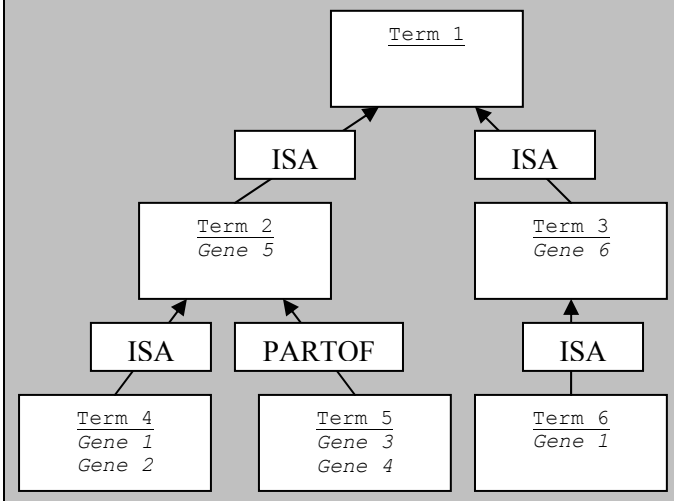

Figure 2: The relationships within the ontology are explored to give a fuller description of each SAGE tag. For every ontology term associated with a tag all the parent terms (found by following the ISA relationship) are also associated.

The significance of the results.

When matching the number of times a particular tag appears within a cluster with a particular functional enrichment in each of the 10 sets of SAGE experiments, a reference as to what constitutes a significant trend needs to be determined. In the experiments such a reference is determined by the construction of a probability model which can be used to determine relevance cut off level for the results that were obtained. The number and size of the clusters has a direct effect on the number of facets that cluster will have and therefore the number of matches that are observed between the different samples. It is thus possible to build a probability model using information about the number of terms used to describe each tag, and the number of tags that reside within clusters. This information can be used to determine the probability of an event occurring presuming that there is no underlying bias. For example, if {P1,P2,P3} are the probabilities of tags with specific ontology terms occurring in 3 samples, then the probability of the tag occurring in only one of the three sample can be modelled as: $(P1*(1-P2)*(1-P3))+((1-P1)*P2*(1-P3))+((1-P1)*(1-P2)*P3)$.

The above probability model assumes that there are no other significant underlying biases that exist within the data. Such an assumption is a fallacy, as the regulation of gene expression is only one of a number of factors which affects the results of such predictions. As ontology annotations are used to describe each cluster, one of significant issues is the bias that can be observed with these annotations. The ontology defined facets of a cluster are influenced by:

The uneven distribution of annotation on the SAGE Tags. Terms are used by annotators to a lesser and greater degree which results in a bias in the ratios of terms used (families of genes that are better studied have more frequent and uniform annotations). The use of automatic annotation systems means that those genes for which there exists a body of literature will have significantly more detailed (and possible inaccurate) annotations.

The lack of known function for the majority of tags. Additionally a large number of Tags have more than one putative function which, by their nature, will contain anomalies.

GO is not a model of gene expression. As the GO relationships represent a formal understanding of certain processes, rather than reflecting the underlying cell behaviour, the match between sets of GO terms and clusters can (at best) be thought of as an approximation (or decomposition), rather than a proper description, of the behaviour that the different expression profiles represent.

To study the effects of the annotations bias on each analysis a random sample was simultaneously used for comparison. The randomisation procedure was designed so that only tags with known gene ontology terms were randomly exchanged. The results were clusters of the same size, and with the same proportion of genes with assigned ontology terms, but with different (random) functional enrichment.

The bias for certain ontology terms is not always based on uneven distributions of annotations, that is to say certain ontology terms describe functionality which is more readily distinguishable by examination of expression profiles than others (e.g. terms describe translation machinery). This bias for certain ontology terms to define functional enrichment is known [Gat-Viks 2003], although the exact bias will change depending on the types of experiment. To identify the terms which best describe the members of a cluster 50 SAGE experiments were analysed (see web site for experiment details). Each cluster was analysed to examine the significance of its 'functional enrichment' ontology terms. This was done by sequentially removing tags from the cluster and comparing the real phenotypic functionality against the clusters predicted functionality.

The tags were analysed sequentially in each cluster. This was done by calculating the functional enrichment for that cluster without the specific tag. The generated enrichment for the cluster was then compared against the specific tag, if the tag contained an ontology term that had been recognised as significant for the cluster then this was scored (see Table 1 for subset of results). When comparing the tag with the cluster no ontology navigation was performed, the comparison was for the exact ontology term. If the ontology graph is navigated the number of significant terms increases, however there is a corresponding loss of specificity. The resulting data shows the significance of different ontology terms in describing the genes within specific clusters.

Term	Freq	Pre
nucleus	5957	0.294
integral to membrane	2380	0.230
regulation of transcription, DNA-	2376	0.193
DNA binding	2214	0.210
signal transduction	1442	0.209
RNA binding	1363	0.163
ATP binding	1198	0.133
hydrolase activity	1179	0.188
transferase activity	968	0.159
protein binding	951	0.165
cytoplasm	930	0.112
integral to plasma membrane	928	0.110
metabolism	811	0.495
transport	729	0.188
intracellular	679	0.123

Table 1: All the tags within each data set were analysed to see how closely they fitted the cluster in which they resided. The frequency column shows the number of times the ontology term was used to match between a cluster and a tag and the precision shows the average precision with the clusters (with regard to this term).

Implementation

The analyses carried out in this paper were performed using SeqExpress 1.1.7 and the SeqExpress SDK, which is available at http://www.seqexpress.com. SeqExpress is a desktop analysis and visualisation tool for gene expression experiments [Boyle 2004].

Results

The results are presented in three parts:

- The first section shows the predisposition of tags to appear in clusters which have a certain facet.
- The second section shows the precision of the different techniques.
- The third section shows how the technique can be used to resolve ambiguities in SAGE tag assignment.

Predisposition of Tags

Euclidian distance, cosine distance and SDD clustering were performed on all 10 sets of SAGE experiments (see Figures 3, 4, 5). The functional enrichment for each cluster was analysed, and then the results across all the sets were compared to see the number of times a specific facet (ontology term) of a cluster was used to describe a specific tag. These results show the frequency with which facets were used to describe tags, and do not indicate whether the particular facet is true for the tags it is describing. However, the results do show that there is a significant trend towards tags residing in clusters with the same facets.

Figure 3 shows the results for the Euclidian distance clustering technique. As can be seen the probability model predicts that the number of facets which describe the same tag in 6 or more sets of experiments is approximately zero. More interestingly the samples which randomise 'like with like' have a minimum of an order of magnitutde lower number of facets which correctly describe tags in 6-10 of the samples. The use of a Ranking procedure has a profound effect on the results; as such a ranking changes the distribution within the data set.

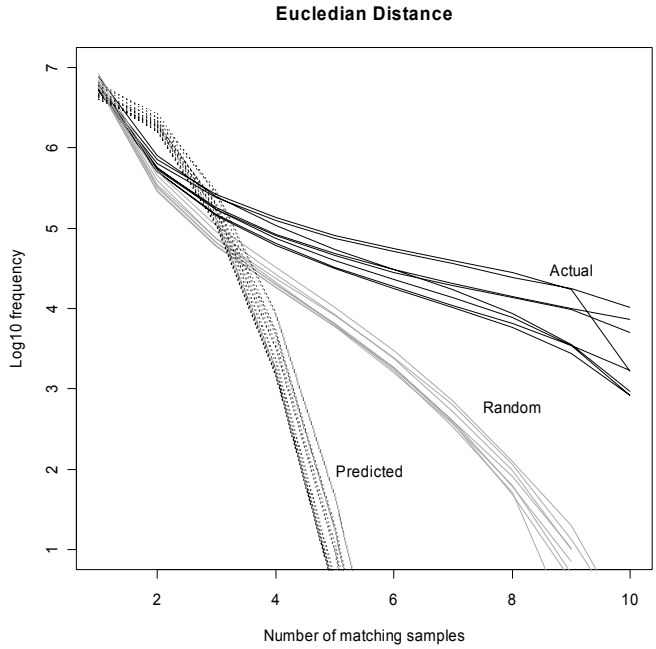

Figure 3: The Euclidian distance function was used to calculate clusters in 10 different sets of experiments. Each tag was then analysed to see the number of times it appears in a cluster with the same functional enrichment. The log plot on the y-axis shows the number of times a facet of a cluster described the same tag. The three distinct groups of lines show the difference between the predicted results (show as a dashed line), the semi-random results (rendered in grey) and the actual results (rendered in black). The highest actual results are those relating to the samples that were pre-processed using a global ranking, and those that were pre-processed with a global ranking and the projected using their Principle components.

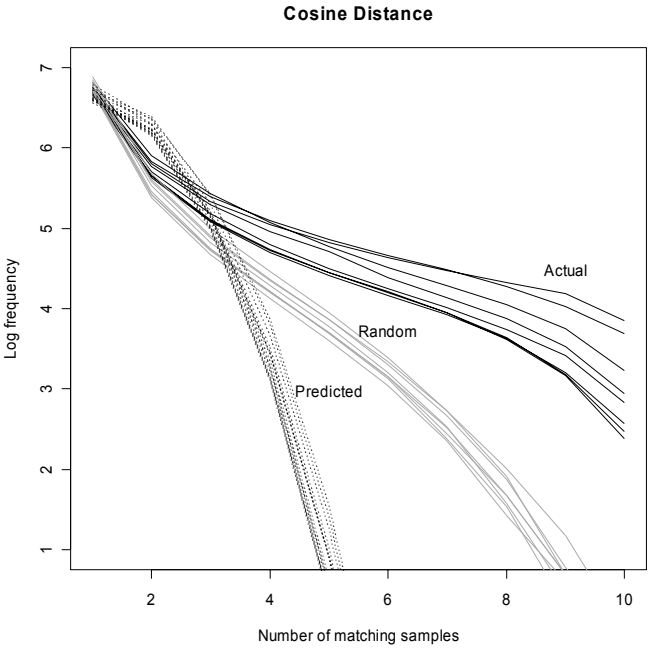

Figure 4: The Cosine distance function was used to calculate clusters in 10 different sets of experiments. Each tag was then analysed to see the number of times it appears in a cluster with the same functional enrichment. The highest actual results are those relating to the samples that were pre-processed using a global ranking, and those that were pre-processed with a global ranking and the projected using their Principle components. The Cosine distance function results are similar to those of the Euclidian distance clustering.

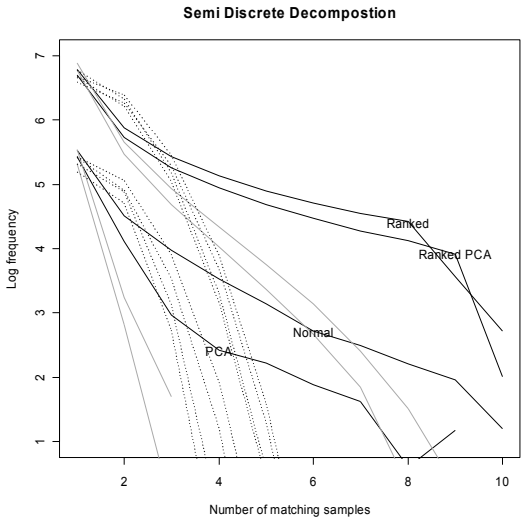

Figure 5: The SDD was used to calculate clusters in 10 different sets of experiments. Each tag was then analysed to see the number of times it appears in a cluster with the same functional enrichment. The highest actual results are those relating to the samples that were pre-processed using a global ranking, and those that were pre-processed with a global ranking and the projected using their Principle components. The black lines are the actual results labelled with the pre-processing technique that was used (broken lines are the predicted results, grayed lines are the randomised results)

The cosine distance results were similar in both proportion and population to the Euclidian distance results (see Figure 4). As with the Euclidian distance results the pre-processing of the data by Ranking, and to a lesser extent Projection, had a direct effect on the frequency with which facets described individual tags.

The SDD defined clusters had a different distribution of results than the distance clustering methods (see Figure 5). The number of significant tags (with the cut off for significance being set to matches in 6 or more sets of SAGE experiments) is approximately the same. As with the other methods ranking affects the results as it alters the distance between the points in the gene/expression matrix to better reflect the underlying differences.

The two methods that are able to generate the largest numbers of different *significant* tags are the Euclidian distance measure on the ranked and the (SVD) projected data set, and the SDD on the ranked data.

It is interesting to note that the SDD is commonly used as an alternative to SVD in latent semantic indexing (LSI) [Deerwester 1990] based information retrieval. In effect the Euclidian distance clustering of the projected data is finding multi-functional 'concepts' [Kolda 1999] in a manner indicative of LSI. Whilst both methods are known to exhibit similar behaviour in data sets that have large numbers of areas of high density ('bumps') [McConnell 2001], the items and relationships they discover are known to differ.

For all the cluster identification methods the initial cluster size (or the number of factors used) has a direct effect on the number of clusters that are found. In general, the higher the number of initial clusters the smaller the average cluster size. However, as smaller clusters will have less facets, the higher the initial cluster count the more precise a description of the tags contained within a cluster will be gained.

Precision of the clusters

As the purpose of this work was to be able to assign function to tags for which sequence homology was unable to reliably determine a mapping, an indication as to how well the facets of the cluster represented their constituent tags was needed. The precision measure needed to be two fold: tag precision is the proportion of the tags that match one or more of the facets in its cluster; and facet precision is a measure of how well the facets describe the tags. The facet precision will always be equal or less than the tag precision, as it is the proportion of facets which are "true" averaged over all the tags in the cluster, that is to say: $\sum_{1}^{t}(f'/f)$, where f' is the number of times the specific facet is actually one of the tags ontology terms, f is the number of facets for the cluster and t is the number of tags in the cluster.

The results in Figures 7 and 8 show both the tag precision and the facet precision. As can be seen in Figure 7, when the data is rank-transformed and then projected (and then analysed using the Euclidian distance clustering technique) there is a steady decline in the number of matched tags, although the precision measures goes up. Generally the higher the starting cluster size that is used (and therefore the size of resulting clusters is smaller and the precision of the prediction results is better) the more general the ontology term(s) that are found for the facets, as the ability to find functional enrichment is dependent on the clusters having unusual concentrations of terms. This means that whilst we can get high precision values, their usefulness may be limited if they are only available for a few tags and their granularity is such that it is difficult to differentiate between tags (e.g. only able to differentiate between tags that are involved in a 'physiological process' and those that are not).

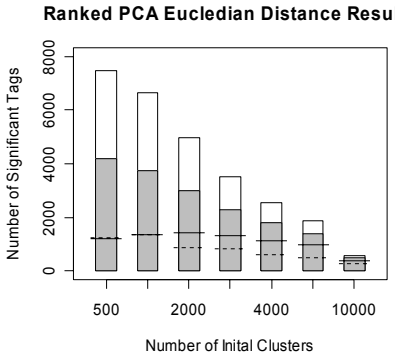

Figure 7: Showing the number of tags which occur in clusters with the same functional enrichment in 6 or more samples. Anchored Euclidian distance clusters were found by seeding the algorithm with a varying number of initial clusters (shown on the x-axis). If a Tag was found to be in 6 or more clusters which exhibit the same type of functional enrichment, it was scored as significant. The number of tags which were found to be significant is shown by the height of each of the histograms. The grayed portion of the histograms shows the proportion of Tags which had a GO ontology term which matched the functional enriched clusters in which they were found (the 'tag precision' of the technique). The two horizontal lines along each bar give a measure of the proportion of facets which are correct (the 'facet precision'), the solid line is the proportion for results that were Ranked and then projected using their Principle components, the dashed line is for the results that were only Ranked.

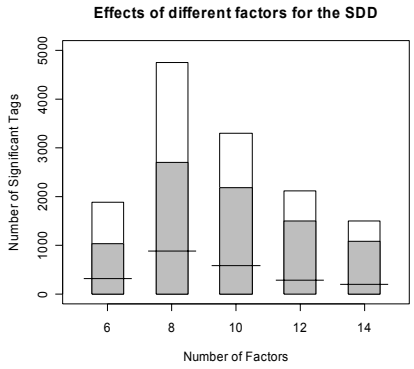

Figure 8: Showing the number of tags which occur in clusters with the same functional enrichment in 6 or more samples. The clusters were generated using SDD, the x-axis shows the number of factors that were determined. The number of tags which were found to be significant is shown by the height of each of the histograms. The grayed portion of the histograms shows the proportion of Tags which has a GO ontology term which matched the functional enriched clusters in which they were found (the 'tag precision'. The horizontal lines along each bar give a measure of the proportion of facets which are correctly described the tags in its cluster (the facet precision).

Resolving Ambiguities.

For tags which have more than one 'reliable' protein function, this technique could be used to indicate which of the possibilities is exhibited by the tag's expression profile. In this case, the SAGE tags which had more than one 'reliable' mapping to Unigene clusters (using Locuslink ids) were examined to see if they had a predisposition to (re)occur in clusters which had a particular facet. Only one facet for each cluster was compared (this was the facet which had the lowest probability of occurring due to chance assuming a binomial distribution). In the cases where this facet could not distinguish between the possible phenotypes for the facet (as the facet described all of the possibilities), the next highest scoring facet was used.

For the SDD analysis the number of factors was set to 8, and for the Euclidian distance clustering algorithm the number of seeded clusters was set to 2000. These starting parameters would enable the categorisation of a large number of tags, with a reasonable level of precision (as the homology searches already provide a good 'hint' as to the functionality).

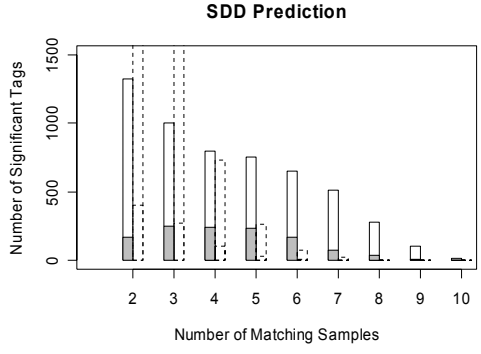

Figure 10: If a phenotype for a tag is favoured above the other possible choices in 6 or more of the samples then it is marked as being significant. By using such a cut off, the SDD clustering finds and resolved 1,600 tag ambiguities. The gray boxes show the number of tags which whilst the majority of tags suggested one phenotype, the evidence was not clear (there was evidence for more than one of the possible putative functions). The dotted line shows the number of tags which were matched using the random sample.

As can be seen in Figures 10 and 11, at the predicted cut off of 6 or more tags the number of random samples containing a predisposition towards one of the possible phenotypes is low. By using such a procedure it is possible to assign expression indicated functionality to approximately 1600 (using SDD) and 1300 (using Euclidian distance) of the SAGE tags which have ambiguities. By taking the union of the two sets of results for all tags that are matched in 6 or more sample the proportion of tags for which a phenotype can be identified rises to 1900 tags.

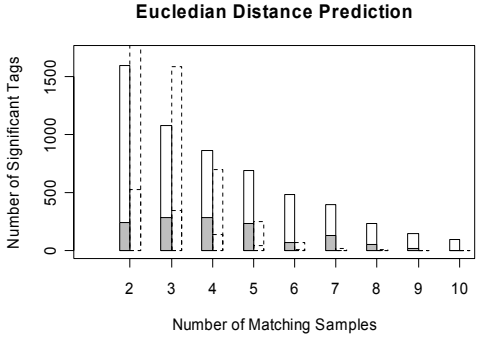

Figure 11: With a cut off of 6 the Euclidian distance clustering technique can resolve 1,300 tag ambiguities. Details as per Figure 10.

Whilst the results are strong indicators as to function, examination of the results highlights a number of drawbacks to the technique:

Differentiation of tags with similar annotations: Functional enrichment of the clusters will not be able to distinguish between putative phenotypes where there are only minor differences in the ontology terms used to describe each of the possibilities. Even though the proteins may have different functionality there may still be strong overlap in their ontology annotations, making the likelihood of differentiating between them small. Conversely if one of the products does not have any ontology terms (for example if it is a 'hypothetical' protein) then this possibility will never be assigned to a functional predisposition score. For example, the *GGTGTGAGCC* tag has three reliable Unigene mappings: a hypothetical protein, a transcription factor, or a histone deacetylase. The hypothetical protein has no matching ontology terms, and so it is not possible to match this to any functionally enriched cluster. The M1 transcription factor and histone deacetylase share a number of ontology terms as they are both associated with DNA transcription (being an RNA polymerase activator, and the other altering chromosome structure by deacetylation of the histones). In most cases this tag appears in clusters which contain enrichment of genes involved in the 'nucleus'. However, as this cannot be used to distinguish these two possibilities it is ignored. Other facets of the cluster have to be used to attempt to identify if the tags expression profile suggests one of the possible protein functionalities. In this case the other facets for the clusters in which the tag appears (involved in cell cycle regulation, has hydrolase activity, involved in chromation modification etc.) are indicative of

the histone deacetylase phenotype. However, these alternative facets are of higher probability than those discarded, and so are inherently less accurate in describing the functional enrichment characteristics of the cluster.

Accuracy of the annotations: There are a number of mechanisms for associating ontology terms with annotations. GO provides evidence codes which indicate the source and mechanism through which the annotation was derived. The use of automatic annotations will lead to inconsistencies within the definitions for the tags, which will result in inaccurate definition of the clusters. For example, the tag *CTGGAAATAA* has two reliable mappings: a mitochondrial flavoprotein, or Plasminogen. In the majority of cases the tag can be shown to have an expression profile similar to other genes identified as having chymotrypsin activity (like Plasminogen). However, this tag is also matched to 'DNA-binding'. This is caused by the Unigene entry for this gene having been automatically annotated with DNA binding/regulation of transcription terms (inferred from electronic annotation: IEA). Whilst Plasminogen itself does not directly exhibit this functionality it has still been found in groups which have been marked with the 'DNA-binding' term (presumably because either they have been marked up using similar automatic methods, the group has other types of enrichment with lower scores or due to chance). This suggests that a mechanism for filtering or scoring based on annotation type will increase prediction accuracy.

Ontology terms not reflecting behaviour: The annotations not only reflect the functionality of the tag, but in some cases also reflecting important biological phenomena associated with its genetic product which may not be seen in the majority of experiments. For example, the tag *TCAAAAAAAG* has two reliable mappings: an actin filament capping protein or an RNA splicing enzyme. The evidence strongly supports the RNA splicing variant, as the majority of clusters in which this tag is found exhibit a strong enrichment in nucleotide acid binding. However, one of the clusters in which this tag resides has the 'apoptosis' enrichment as the best match for the possible phenotypes. The apoptosis term has been automatically matched with this gene because when it is over expressed it causes apoptosis (due to its disease associations), this does not define the proteins normal behaviour. Whilst this annotation is not a mistake, it is not one that can generally be used to correctly identify the tags phenotype. Whilst such annotations are rare, they will lead to incorrect identification of the tags general functionality.

Discussion

This paper outlines a technique for assigning putative phenotypes to SAGE tags by exploring co-expression of tags through analysis of their predisposition to re-occur in clusters which exhibit levels of functional enrichment. It would be naïve to believe that such a methodology could capture the complex semantic of cellular interactions required to divine protein function solely from an expression profile (with the degree of certainty required for scientific analysis). The methodology proposed will only be able to identify tags whose expression is regulated in a similar manner to other tags which have the same (ontology defined) functionality. However, as has been shown the technique can be used effectively with homology information to resolve ambiguities in certain SAGE tag assignments. The presence of these ambiguities is a major limitation of the SAGE technique.

The techniques used to select the clusters do so using different mechanisms: the decomposition (SDD) selects items relevant to a subspace within the n-dimensional matrix; whilst the distance measures compare complete vectors. It may be possible to improve the methodology by combining these techniques using a hybrid EM based solution which uses both a semi-discrete decomposition and a distance measure to select initial clusters and then refines the solution. Such a technique could be used to alter both the size and number of the clusters, so that they favour classification using functional enrichment. Additionally, the use of a different ontology (in particular a more disease specific ontology) will affect the results considerably.

This initial work present here could be expanded to provide a means for detecting both incorrect functional assignment of tags and possible sequencing errors. Such error detection would involve the identification of tags whose ontology assignments are different from those of the clusters for which they exhibit a strong predisposition. The success of such error detection would depend on the type of facets which were being examined, as different ontology terms are better at defining the contents of clusters than others. To accurately describe clusters by using their constituent members' GO terms relies strongly on: correct and inclusive gene annotations; suitable mappings from identifiers to specific GO terms; and the parts of the Gene Ontology graph accurately reflecting the relationships that occur in gene co-expression. Genes that exhibit unusual behaviour, which include those that reside within 'bumps', are more likely to have been assigned GO terms as these are likely to have been studied in greater detail. The use of automated annotations increases the level of annotation, but introduces a higher level of error into the system. The possibilities for weight based navigation, based on both relationship type and on how GO relationships are reflected in cell expression levels (for example, categorised by tissue type), has not been explored. Such navigation would enable better usage of the high level of knowledge that is found in GO.

The majority of tags that do show a predisposition to appear in clusters with specific facets do so without actually being annotated with the specific ontology term. Tags which are co-expressed and collectively have a specific molecular function or are involved in a specific biological process are often additionally co-expressed with a number of other tags which exhibit a different functionality. This is due to the facets which are used to describe the cluster not having the semantic richness to describe the actual cellular mechanism that the co-expressed cluster exhibits. The collection of facets is only able to give an indication as to the cellular machinery which resulted in the group having a statistically identifiable expression profile. This leads to the possibility of using the facets as a means to generate a more 'expression orientated' ontology.

All the analyses discussed in this paper were run on a 1.9GHz (with 1GB RAM) desktop computer. The SDD technique was considerably faster than the distance searching, as the generation of the factors is a sequential (rather than cycle based) technique.

Further experimental details and results can be found at: http://www.seqexpress.com/bioinformatics_may

References

[Alter 2000] Alter O, Brown P, and BotsteinD (2000). *Singular value decomposition for genome-wide expression data processing and modelling.* PNAS, vol. 97(18), pp 10101–10106

[Alter 2003] Alter O, Brown P, and BotsteinD (2003). *Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms.* PNAS, vol. 10(6), pp 3351–3356

[Baggerly 2003] Baggerly K, Deng L,, Morris J and Aldaz C. Marcelo (2003). *Differential expression in SAGE: accounting for normal between-library variation*. Bioinformatics Vol. 19(12) pp 1477–1483

[Boyle 2004] Boyle J. (2004), *SeqExpress: Desktop analysis and visualisation tool for gene expression experiments.* Bioinformatics

[Chiang 2003] Chiang JH, Yu HC. 2003. *MeKE: discovering the functions of gene products from biomedical literature via sentence alignment.* Bioinformatics 19: 1417-1422.

[Deng 2004] Deng M, Tu Z, Sun F, Chen T. 2004. *Mapping gene ontology to proteins based on protein-protein interaction data.* Bioinformatics 20: 895-902.

[Deerwester 1990] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R (1990). *Indexing by latent semantic analysis*. Journal of the American Society for Information Science 41(6), pp. 391-407.

[Dudoit 2002] Dudoit S., Fridlyand J., and Speed T. (2002). *Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data.* Journal of the American Statistical Association, vol 97(457), pp77-87.

[Dysvik 2001] Dysvik B., Jonassen I. (2001). *J-Express: Exploring Gene Expression Data using Java*. Bioinformatics, vol 17, 369-370

[Forgy 1965] *Forgy E (1965) Cluster analysis of multivariate data: Efficiency vs interpretability of classifications.* Biometrics, *21,768*.

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

[Gene Ontology Consortium 2000] The Gene Ontology Consortium (2000). *Gene Ontology: tool for the unification of biology*. Nature Genetics. vol 25, pp 25-29

[Getz 2000] Getz G., Levine E., and Domany E.(2000). *Coupled two-way clustering analysis of gene microarray data.* PNAS ,vol. 97(22) pp 12079-12084

[Golub 1999] Golub T, Slonim D, Tamayo P, Huard C, Gassenbeek M, Mesirov J, Coller H, Loh M, Downing J, Caligiuri M, Bloomfield C and Lander E (1999). *Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring.* Science vol 286, pp 531-537.

[Hvidsten 2002] Hvidsten TR, Komorowski J, Sandvik AK, Laegreid A. 2002. *Predicting gene function from gene expressions and ontologies. Pac Symp Biocomput:* 299-310.

[Hvidsten 2003] Hvidsten TR, Laegreid A, Komorowski J. 2003. *Learning rule-based models of biological process from gene expression time profiles using Gene Ontology.* Bioinformatics 19: 1116-1123.

[Kolda 1999] Kolda T , O'Leary P.,(1999) *A semidiscrete matrix decomposition for latent semantic indexing information retrieval,* ACM Transactions on Information Systems. TOIS, vol.16(4), pp.322-346.

[Lagreid 2003] Lagreid A, Hvidsten TR, Midelfart H, Komorowski J, Sandvik AK. 2003. *Predicting Gene Ontology Biological Process From Temporal Gene Expression Patterns.* Genome Res 13: 965-979.

[Lash 2000] Lash AE, Tolstoshev CM, Wagner L, Schuler GD, Strausberg RL, Riggins GJ, Altschul SF (2000), *SAGEmap: a public gene expression resource.e* Genome Res. vol 10(7), pp 1051-60

[Li 2002] Li M., B. Wang, Z. Momeni, and F. Valafar. 2002. *Pattern Recognition Techniques in Microarray Data Analysis.* Proc Int Conf on Mathematics and Engineering Techniques in Medicine and Biological Sciences 2002 (METMBS'02), Las Vegas, Nevada, pp 610-616.

[Liu 2003] Liu L, Hawkins D, Ghosh S and Young S (2003). *Robust singular value decomposition analysis of microarray data.* PNAS, vol. 100(23), pp 13167–13172.

[Man 2000] Man M, Wang X, Wang Y (2002). *POWER_SAGE:comparing statistical tests for SAGE experiments.* Bioinformatics, vol 16(11), pp 953-959.

[Midelfart 2001] Midelfart H, Lægreid A, and Komorowski J. (2001) *Classification of Gene Expression Data in an Ontology.* ISMDA 2001, LNCS 2199, pp. 186–194.

[McConnell 2001] McConnell S and Skillicorn D. (2001). *Outlier detection using semi-discrete decomposition.* Technical Report 2001-452, Dept. of Computing and Information Science, Queen's University, 2001

[Ng 2001] Ng R., Sander J., and Sleumer M. (2001). *Hierarchical Cluster Analysis of SAGE Data for Cancer Profiling.* Workshop on Data Mining in Bioinformatics.

[O'Leary 1983] O'Leary D and Peleg S (1983). *Digital Image Compression by Outer Product Expansion.* IEEE Transaction on Communications, vol 31, pp 441-444.

[Pan 2001] Pan, W. 2002. *A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments.* Bioinformatics 18(4), pp 546-54.

[Perez 2004] Perez AJ, Perez-Iratxeta C, Bork P, Thode G, Andrade MA. 2004. *Gene annotation from scientific literature using mappings between keyword systems.* Bioinformatics

[Raychaudhuri 2000] Raychaudhuri, S., Stuart, J. M. and Altman, R. B. (2000) *Principal components analysis to summarize microarray experiments: application to sporulation time series.* In Pacific Symposium on Biocomputing, vol. 5.

[Sharan 2002] Sharan,R., Elkon,R. and Shamir,R. (2002) *Cluster analysis and its applications to gene expression data.* Bioinformatics and Genome Analysis. Springer, Berlin, pp. 83–108.

[Sturn 2000] Sturn A, Quackenbush J, Trajanoski Z (2002). *Genesis: cluster analysis of microarray data.* Bioinformatics. vol 18(1), pp 207-8.

[Szabo 2002] Szabo A., Boucher K., Carroll W.L., Klebanov L.B., Tsodikov A.D., Yakovlev A.Y. (2002). *Variable Selection and Pattern Recognition with Gene Expression Data Generated by the Microarray Technology.* Mathematical Biosciences, vol 176, pp. 71–98.

[Tavazoie 1999] Tavazoie S., Hughes J., Campbell M., Cho R. and Church G. *Systematic determination of genetic network architecture.* Nature Genetics, vol 22, pp 281-285.

[Troyanskaya 2002] Troyanskaya O, Garber M, Brown P., Botstein D and Altman R (2002). *Nonparametric methods for identifying differentially expressed genes in microarray data.* Bioinformatics, vol 19(11), pp 1454-1461.

[Wall 2003] Wall, Michael E., Andreas Rechtsteiner, Luis M. Rocha. (2003). *Singular value decomposition and principal component analysis.* A Practical Approach to Microarray Data Analysis. D.P. Berrar, W. Dubitzky, M. Granzow, eds. pp. 91-109.

[Wilcoxon 1945] Wilcoxon, F. (1945). *Individual comparisons by ranking methods*. Biometrics, 1, pp. 80-83.