BALBOA: Extending Bicluster Analysis to Classify ORFs using Expression Data

Kenneth Bryan
Department of Computer Science and Informatics
University College Dublin
kenneth.bryan@ucd.ie

Pádraig Cunningham
Complex and Adaptive Systems Laboratory
University College Dublin
padraig.cunningham@ucd.ie

Abstract

1 Motivation:

Microarrays have the capacity to measure the expressions of thousands of genes in parallel over many experimental samples. The unsupervised technique of bicluster analysis has been employed previously to uncover gene expression correlations over subsets of samples with the aim of modelling the natural gene functional classes. However the bicluster model also has the potential to shed light on the functions of unannotated open reading frames (ORFs). This aspect of biclustering has been under-explored. In this work we illustrate how the bicluster representation of expression data may be extended to enable putative functional classification of unannotated ORFs. We develop an ORF annotation approach, referred to as BALBOA, in which classifiers are constructed from the class specific expression patterns discovered by bicluster analysis. We demonstrate the efficacy of this approach via cross validation and carry out a comparative evaluation with kNN classification across three yeast expression datasets. Finally, we assign putative functions to unannotated ORFs and attempt to corroborate the best supported annotations with external experimental and protein sequence information.

2 Introduction

Gene expression microarrays allow one to measure the expressions of thousands of genes in parallel over many experimental samples (growth conditions, time points, cell types etc.). Typically, the results from such experiments are presented as an expression data matrix in which rows represent genes and columns represent samples. Analysis of such data matrices has revealed that functionally related genes often have correlated expression profiles [9]. Sample profiles too, such as cell or disease types, often exhibit characteristic expression profiles [3].

From an analytical perspective, a sample or gene profile may be thought of as an object and its expression values as features. This gives rise to the prospect of classifying genes or samples based on their expression profiles. In the case of samples, such analysis is most often employed in the molecular characterisation and classification of cancer types [3, 13]. In this paper, however, we will focus on the problem of classifying functionally unannotated genes using expression data. Hereafter unannotated gene profiles will be referred to as ‘open reading frames’ (ORFs) as a functional protein product has yet to be experimentally verified.

Gene expression data may be analysed in various ways. Statistical methods such as differential analysis of gene expression over samples may be used to identify genes whose expression differs significantly between sample classes. This can lead to the ab initio elucidation of gene function as well as the identification of ‘marker’ genes that best characterise sample classes [22]. Where object class labels exist supervised methods may be applied to learn the characteristic expression patterns of a class. Techniques such as k-nearest neighbour and support vector machines have been used to classify of both genes and samples [4, 10, 21].

When class labels are unavailable, or perhaps debatable, unsupervised methods may be used to model the class structure by analysing the object similarities with reference to their features alone. Cluster analysis has been one of the most prevalent unsupervised approaches employed to model both sample classes and gene functional groups using profiles from gene expression data [9, 11, 15]. This technique typically separates the data into k disjoint groups of objects that have high similarity within groups and low similarity between groups. Ideally this partitioning of the data will closely reflect the natural underlying classes. In the context of expression data, similarity is best computed in terms of a correlation based distance measure, such as Pearson’s r, rather than an absolute measure, such as Euclidean distance. Genes showing similar expression patterns may be co-regulated to perform a common function in vivo. Thus, in this context, cluster analysis of gene profiles attempts to model the functional modules within the expression data.
Standard cluster analysis measures gene expression similarity over the full set of experimental sample features. However, as datasets increase in size it becomes increasingly unlikely, due to noise, measurement error and the nature of gene regulation, that even functionally related genes will retain expression similarity over all samples. As a result, calculating gene expression similarity exclusively over all samples has the potential to miss significant ‘local’ signals that may only be apparent over subsets of samples.

To address this drawback, the two-way clustering technique of bicluster analysis was proposed [7]. In this domain, biclustering involves grouping genes whose expression correlates over a subset of experimental samples. Apart from guarding against the prospect of missing significant local signals within the data, biclustering allows for genes to belong to more than one group (or to none at all). This added aspect allows better rendering of the natural class structure in which a gene may belong to more than one functional class. The principal application of bicluster analysis is in revealing this underlying class structure, i.e. the gene functional modules, within the expression data. Once this higher level class structure is modelled however, one may employ it to further elucidate the functions of genes or annotate unclassified ORFs. The success of this clearly depends on the accuracy of the primary class model.

The accuracy of the class model revealed by bicluster analysis may be assessed in terms of ‘functional enrichment’ [23]. The functional enrichment ($E$) of a bicluster may be given by $E = M/N$, where $M$ is number of genes from the most dominant functional category and $N$ is the total number of genes in the bicluster. This simple metric may be extended to account for the prevalence of the dominant class and the total number of genes in the dataset by calculation of a $p$-value from the hypergeometric distribution (or its binomial approximation). However, in some cases, i.e. when the bicluster is large or it captures part of a relatively small class, a bicluster may have a seemingly significant $p$-value but a low, and uninformative, functional enrichment.

The utilization of the bicluster model of the gene functional modules to aid in the functional classification of ORFs has been under-explored. One possible method of assigning putative functions to unannotated ORFs is by examining biclusters that contain unannotated genes. If these also have a high enrichment of genes from one functional category we might surmise that these unannotated ORFs may also belong to this category. This guilt by association method for ORF classification is discussed in section 3.2.1.

In this paper we present a novel extension to bicluster analysis that focuses on this ORF classification aspect rather than performing it ancillary to the main class analysis. Our Bicluster Analysis Based Orf Annotation (BALBOA) approach allows for a more directed and systematic classification of unannotated ORFs using expression data.

BALBOA begins by first dividing the expression dataset into its annotated genes and unannotated ORF subsets. Bicluster analysis is performed on the annotated gene set. The resulting class information is then used to classify ORF from the unannotated set. As BALBOA contains a bicluster analysis step we briefly review biclustering within this gene expression analysis domain in section 3.1. We then describe our BALBOA ORF classification strategy in section 3.2.2.

An advantage of the BALBOA approach is that it may be evaluated via cross validation. In section 4.3.1 we compare BALBOA to the standard classification approach of $k$-nearest neighbour (kNN), our specific multi-class kNN implementations are described in section 4.2. In this evaluation we use three independent gene expression datasets derived from Saccharomyces cerevisiae or budding yeast, detailed in 4.1. In section 4.3.2 we then attempt to functionally classify yeast’s unannotated ORFs using BALBOA.

Importantly, an advantage of using three independent datasets that refer to the same transcriptome is that any putative functional classifications can be cross referenced across datasets. This has the potential to significantly increase the support for annotations. We also attempt to support our putative annotations via external experimental and protein sequence information from the Saccharomyces Genome Database (SGD). We first give a brief review of bicluster analysis in the gene expression analysis domain.

3 Methods

3.1 Biclustering of Expression Data

Conventional cluster analysis of the gene profiles in expression data involves calculating similarity over all experimental samples. As briefly discussed in section 2, this approach may not be the most suitable method of analysing the class structure within large gene expression datasets. Such data may contain a significant amount of noise due to variations within the biological system being studied and errors within the experimental measurement process itself. Another aspect of gene expression data is that genes may only correlate over a subset of samples. For example, a functional module of co-regulated genes may only be active at certain stages in the cell cycle or under certain environmental conditions. As a result, especially within large datasets, even related gene profiles may not retain similarity over all experimental samples. Cluster analysis, therefore, has the potential to miss local but significant subspace similarities within the data that may contribute to the modelling of the natural set of functional classes.

To address this problem, the concept of biclustering was introduced to gene expression analysis by Cheng and Church (2000). In a gene expression data matrix of genes and samples, a bicluster is defined as a subset of genes that
show similar expression over a subset of samples. In examining different sample subsets one can also see that different behaviours of the same gene can be captured. This aspect of biclustering gives it the potential to model ‘overlapping’ functional modules between which genes are shared.

Cheng and Church proposed a bicluster scoring metric called the mean squared residue \((H_v)\) to score the correlation of the rows and columns within selected sub-matrices in the expression data matrix. This is given by:

\[
H(I, J) = \frac{1}{|I||J|} \sum_{i \in I, j \in J} (a_{ij} - a_{Ij} - a_{ij} + a_{IJ})^2
\]  

(1)

where \(a_{ij}\) is the entry at position \(ij\) in the sub-matrix \((I, J)\), \(a_{Ij}\) is the mean of the \(i\)th row, \(a_{ij}\) is the mean of the \(j\)th column and \(a_{IJ}\) mean of the whole sub-matrix.

Given that the number of sub-matrices increases exponentially with both the rows (genes) and columns (samples) of the gene expression data matrix, it is evident that an exhaustive search for biclusters is unfeasible. Cheng and Church therefore developed a greedy node deletion heuristic that utilized the means squared residue to locate good scoring biclusters within the expression data matrix. This seminal work spawned numerous related biclustering approaches \([24, 8, 2]\) as well as strategies dependent on alternate metrics \([16, 20]\). However, the mean squared residue was subsequently shown to contain some biases that directed searches towards less interesting biclusters containing genes with low expression variance \([1]\).

Recently, we proposed the Bottom-Up Biclustering By Locality Expansion (BUBBLE) approach \([5]\). Here we demonstrated improved results over the graph theoretic approach, SAMBA, proposed by Tanay et al. \((2002)\). BUBBLE searches for small regions of high correlation, bicluster seeds, within expression data via a simulated annealing based search method. These seeds are expanded in a deterministic fashion by adding the most similar gene profiles to produce a full bicluster. Importantly, BUBBLE utilizes a new metric, the \(H_v\)-score, which is unencumbered by the bias affecting the mean squared residue. This is defined by:

\[
H_v(I, J) = \frac{1}{|I||J|} \sum_{i \in I, j \in J} \left( \frac{a_{ij} - a_{Ij} - a_{ij} + a_{IJ}}{a_{IJ}} \right)^2
\]  

(2)

where \(a_{ij}\) is the entry at position \(ij\) in the sub-matrix \((I, J)\), \(a_{Ij}\) is the mean of the \(i\)th row, \(a_{ij}\) is the mean of the \(j\)th column and \(a_{IJ}\) mean of the whole sub-matrix.

BALBOA’s bicluster generation step is based upon BUBBLE with some important additions, discussed in the next section, designed to improve the subsequent ORF classification. Firstly, however, we shall examine the ‘co-occurrence’ method of ORF classification that has thus far been employed with cluster and bicluster analyses.

### 3.2 ORF Annotation via Class Modelling

#### 3.2.1 Classification via Co-occurrence Analysis

It is evident from previous cluster and bicluster analyses of gene expression data that functionally related genes tend to group together \([9, 6]\). A simple extension to this logic would suggest that, should unannotated ORFs be grouped with annotated, that these annotations might shed some light on the functions of the unannotated ORFs. Clearly the support for these annotations increases with the functional enrichment of the group. This co-occurrence analysis has also been referred to as the guilt by association approach \([17]\).

Wu et al. applied several gene clustering techniques to expression data \([23]\). Using the above premise, they inferred putative functions for unannotated ORFs and labelled the clusters using the MIPS functional database \((see \ section \ 4.1)\). To assess the precision of this strategy they iteratively re-labelled each annotated gene in turn as ‘unannotated’ and re-assessed all the cluster enrichments (and resultant ORF classifications). They achieved an accuracy of up to 80% with a recall of 40% in the case of the Protein Synthesis functional category. Although Wu et al. did some filtering on the basis of \(p\)-values they did not take into account the functional enrichment of the clusters, only whether the gene was present or not. Also, a significant number of the clusters contained many unannotated genes and therefore added limited information to the model. Lastly, as this annotation approach is built upon conventional clustering only, it does not utilize input from the gene relationships that only occur over significant subsets of experimental samples.

In another study biclusters generated by the SAMBA biclustering algorithm were subsequently analyzed to produce a so called ‘naive functional annotations’ of unclassified ORFs \([20]\). Again \(p\)-value filtering was carried out but in this case the evaluation involved hiding 30% of the known annotations and then assessing the precision. Here the authors used the 10 Gene Ontology (GO) categories (rather than the more specific 17 MIPS categories). It was also noted that genes could be assigned more than one annotation as a result of this process. However, as \(p\)-values were used to select significant biclusters \((p > 10^{-4})\) and no weight was given to functional enrichment of the biclusters, we can see that many annotations, some poorly supported, may be assigned to an ORF. It is also difficult to gauge the success of this approach as the authors only provide the specificity (percentage of true negatives correct) and the selectivity (recall) of their annotations.

Both of the above classification approaches use a co-occurrence approach to putatively label unannotated ORFs. An alternative strategy might be to borrow from the ‘labelled training set’ approach used in supervised learning.
3.2.2 The BALBOA strategy for ORF Classification.

A further way in which we may attempt to functionally classify unannotated ORFs is to first divide our expression dataset into a labelled set (annotated genes) and an unlabelled set (unannotated ORFs). We can then develop bicluster ‘classifiers’ using our annotated genes and apply these to classify our unannotated ORFs. We label our gene profiles via the MIPS database (see section 4.1). This approach mirrors the standard approach taken in supervised learning and has several advantages. Firstly, all the biclusters generated will be fully labelled, containing no unannotated genes, and therefore all will be informative. We also note that the biclusters generated in this manner tend to have higher functional enrichments than those generated over the total expression dataset. Furthermore, by dividing our dataset we can easily carry out standard cross validation to evaluate the performance of our bicluster classifiers.

Although our bicluster generation step is based upon BUBBLE we attempt to increase the diversity within our bicluster population by masking bicluster seeds as they are discovered. This masking involves replacing the entries of the bicluster seed with random values from the inter-quartile range of the dataset. We also increase the number of bicluster solutions generated. As in ensemble clustering it is in our interest to increase the size and diversity of our bicluster set as this will enable us to determine our most ‘stable’ relationships [19]. In our case however this stability adds increased support to our subsequent ORF classifications.

BALBOA is illustrated in Figure 1. In step 1 each bicluster is labelled with its dominant functional class. Where two or more classes are equally dominant all labels are applied. In step 2, biclusters for each functional class are selected. These represent the principal expression signals for each class. In step 3 we then select the ‘best’ biclusters for each functional category. In practice this set contains all biclusters with functional enrichment \( E \geq \alpha E_{max} \), where \( \alpha \) is user defined and \( E_{max} \) is the maximum enrichment for the biclusters of this functional category.

A simple way to perform classification using these bicluster sets is to assign their labels to ORFs with similar expression profiles. In practice these ORF profiles are first standardized (divided by their standard deviation). They are then sorted according to their mean squared residue score relative to the bicluster classifier. This relative profile scoring was first used by Cheng and Church in their node deletion approach. In this case, the stopping point for the annotation process is determined retrospectively by the largest ‘jump’ in dissimilarity that occurs as the ordered ORFs are presented to the classifier.

We see in step 4 of Figure 1 that several bicluster classifiers may contribute annotations for each functional category. Also in some cases, such as for YOR175C, bicluster classifiers may agree. To capture this added information on label support a ‘vote’ for each bicluster classifier is tallied within an ORF frequency list for a functional class. The inherent support for each vote is also taken into account by weighting each vote by the functional enrichment of the bicluster that casts it. This is achieved by multiplying each vote by the enrichment of the bicluster. Step 5 involves firstly creating this ‘weighted’ frequency list for each functional class. To reduce the potential for returning false positives the ‘best’ supported labels from each class list are then selected. We define a selection threshold \( \beta F_{max} \), where \( F_{max} \) highest weighted frequency for a class. This relative threshold takes into account the maximum weighted frequency in each class making it more suitable than an absolute threshold.

As BALBOA involves separate ‘training’ and ‘labelling’ steps we may evaluate it via standard cross validation. Furthermore, since we have three yeast datasets we have the option of cross referencing ORF annotations for consistent labelling across multiple datasets, as in step 6. Although the set of unannotated ORFs are all officially designated as ‘unclassified’ by MIPS, for many there does exist some ‘wet lab’ experimental and/or protein sequence information. Although, evidently, this information is not sufficient to support an official functional annotation it may still agree with our putative functional classifications. We carry out these internal and external evaluations in the next section.
4 Evaluation of BALBOA

4.1 Datasets

We use three yeast expression datasets in our evaluation. The ‘Eisen’ dataset contains 6,221 genes and 80 samples related to yeast cell-cycle, sporulation, and diauxic shift [9]. The ‘Hughes’ dataset contains 6,316 genes and 300 samples from an extensive functional analysis expression study [14]. The ‘Gasch’ dataset contains 6,129 genes and 150 samples from a yeast cell stress study [12]. Missing entries were randomly imputed from the inter-quartile range of each dataset. To reduce the impact of this imputation we removed all rows and columns containing extensive (≥ 25%) missing values.

ORF profiles were annotated via the MIPS (Munich Information Centre for Protein Sequences) database [18]. Approximately 1500 ORFs in yeast are assigned category 99 (Unclassified). These MIPS labels were used to evaluate the functional enrichment after our bicluster generation phase, allowing selection of good bicluster ‘classifiers’.

Lastly, to avoid classification of dubious (non-coding) ORFs those specified as ‘dubious’ (∼ 500) by the Saccharomyces Genome Database (SGD) were removed from each dataset. Datasets and algorithms are available on-line at http://mlg.ucd.ie/balboa.html.

4.2 Multi-class k-NN Implementations.

Supervised approaches such as k-nearest neighbour (kNN) and support vector machines (SVM) have been used to classify unlabelled ORFs within expression data [4, 10, 21]. Recently, the ‘local’ kNN approach has been shown to outperform the ‘global’ SVM approach in the gene expression domain [17]. In KNN an unlabelled ‘query’ object is compared with its k nearest labelled objects in feature space. This ‘nearness’ depends on its feature values and is defined by a suitable similarity metric. In our context the query consists of an unannotated ORF and its feature values. This ‘nearness’ depends on its feature values and is defined by a suitable similarity metric. We use Pearson’s correlation as our similarity metric. We use k=10, which was previously used in this domain [17], and is also the number of genes present in the initial bicluster seed in BALBOA.

Also important in kNN is the voting process. Firstly, we use a majority voting scheme in which we assign the most prevalent label from our k nearest neighbours. We also implement a unanimous voting scheme in which we only assign a label to the query object if all k neighbours ‘agree’ or contain that label. This will enable higher precision and reduce the potential for false positives. To cater for the possibility of a gene having multiple functional labels, we must implement these kNN in a multi-class manner. Therefore when two (or more) labels are equally prevalent among the k nearest neighbours we assign all labels to our query.

4.3 Results

4.3.1 Comparative Cross Validation with kNN.

Before applying BALBOA to classify the unannotated yeast ORFs we must first determine its prediction accuracy via cross validation. To do this we divide the annotated gene set into a training set and test set. We then train BALBOA on the training set and use the resultant bicluster classifiers to predict the functional labels of the test set. We perform two rounds of 4-fold cross validation. This 3:1 split reflects the natural annotated to unannotated ratio in yeast. In practice our thresholds, described in section 3.2.2, are set as follows. Our set of best biclusters for each functional class are chosen by selecting biclusters with functional enrichment (E) ≥ αE_{max}, where α = 0.9 and E_{max} is the maximal enrichment for a functional class. We combine the resultant labelled ORFs for each class into a weighted frequency list and select the best supported ORFs, such that F ≥ βF_{max}, where β = 0.9 and F_{max} is the label with the highest weighted frequency. Although we could finely tune these parameters to achieve better results for each functional class these values allowed for best average predictions across all functional categories and all datasets. This fact may also reduce the possibility of over-fitting our training data.

Cross validations for each yeast dataset given in section 4.1 can be seen in Table 1. Here we compare BALBOA’s classification precision (P) and recall (R) with majority and unanimous voting multi-class kNN. In column one we see that majority kNN has a high recall but low precision and as a result its use as an ORF classifier is unfeasible. In column two we can see that unanimous voting markedly improves precision at the expense of recall. In the last column we can see that BALBOA outperforms unanimous kNN by achieving a higher precision on average over all functional classes.

As opposed to kNN, BALBOA allows for the capturing of signals that occur over subsets of the experimental features in the dataset. In the Transcription (11) functional category in particular this facet seems to allow for a greater classification precision to be achieved across all three datasets. In other categories such as Protein Synthesis (12) UkNN is more competitive. This is not surprising as it is well known that genes in this functional category are strongly co-expressed over most conditions.

4.3.2 Classification of Unannotated ORFs.

Now that we have determined BALBOA’s classification precision for each functional class we can apply it to classify unannotated yeast ORFs. Heretofore our evaluation has focused on the top level MIPS categories (01, 02 etc.). However in some cases it may be possible to assign a more specific and informative functional label. A classic example is that of Protein Synthesis (12). Expression correlations
of genes from this category are most often due to the highly co-regulated sub-category of ribosomal proteins (12.01.01). In this study we have also observed that this ribosomal co-regulation is also evident in the ribosomal RNA (rRNA) processing genes. This sub-category, specifically rRNA processing (11.04.01), often dominates the biclusters that are functionally enriched for Transcription (11). These sub-category labels are provided where available and allow for a more specific putative annotation.

Of the 954 unannotated genes, BALBOA made putative functional predictions for 135 from the Eisen expression data, 113 from the Gasch data and 119 from the Hughes data. These span 13 of the 17 MIPS functional categories. Interestingly, despite the different experimental conditions investigated in the three independent expression studies, 21 annotations, spanning 7 MIPS functional categories, actually agreed across 2 or more datasets. We will now focus on these consistent annotations, presented in Table 2.

We can see from this table that some of BALBOA’s putative functional annotations seem to be well supported given the external experimental and sequence information. One case which stands out in particular is that of YCR072C. This ORF was labelled as unclassified by MIPS when we annotated our data. However this ORF has subsequently been labelled by MIPS as being involved in transcription of ribosomal RNA (rRNA) and assigned MIPS category rRNA processing (11.04.01). This functional label actually agrees with our predicted function. In fact six of the seven ORFs assigned to the Transcriptional functional category seem to have good supporting external evidence. This includes YDL167C and YJR003 both of which are supported by previous computational evidence defined as Reviewed Computational Evidence (RCA) by the SGD.

Another notable annotation is that of YIL060W. This ORF is putatively assigned to Transposable Elements (38) functional category and is the only ORF given the same annotation across all three datasets. Interestingly, the translation of this ORF seems to have a significant similarity (p-value = 2.2e−21) to retrotransposons TyA Gag and TyB Pol genes. This would seem to corroborate our putative annotation. The two ORFs annotated as ribosomal proteins (12.01.01) also have additional annotations supported over two datasets. In the case of YLR196W both predicted functions appear to be supported given that this ORF is already thought to be involved in rRNA processing.

Of BALBOA’s other putative annotations, the ORF assigned to amino acid metabolism (01.01), YJR154W, has been localized to the cytosol by green fluorescent protein (GFP) fusion localization experiments. Furthermore, the translation of YJR154W has some small similarity to (p-value = 2.8−5) cytosolic L-asparaginase. Two ORFs were putatively assigned to respiration (02.13) functional category. Of these the YDL072C null mutant shows decreased levels of secreted invertase. Invertase is an enzyme involved in converting sucrose into glucose and fructose, a required step if sucrose to be used as an energy source. Of the ORFs assigned to the transport routes (20.09) category the best supported seems to be YIL039W. The fact that deletion of this ORF confers sensitivity to the drug GSAO might suggest a possible involvement in the export of this compound. Lastly, the two ORFs putatively annotated as Biogenesis of Cellular Components: mitochondrion (42.16) are actually both localized to the mitochondrion, YIL157C by co-purification and YML030W by GFP-fusion. The YIL157C null mutant also shows some defect in cytochrome oxidase, a mitochondrial enzyme involved in generating the proton gradient needed for ATP synthesis.

Intuitively, annotations that correspond across two in-

### Table 1. Comparative cross validation of BALBOA with majority & unanimous voting kNN.

<table>
<thead>
<tr>
<th>MIPS Category</th>
<th>Eisen</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Hughes</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Metabolism (01)</td>
<td>0.46</td>
<td>0.43</td>
<td>0.50</td>
<td>0.00</td>
<td>0.54</td>
<td>0.01</td>
<td>0.42</td>
<td>0.35</td>
<td>0.53</td>
<td>0.05</td>
<td>0.70</td>
<td>0.09</td>
</tr>
<tr>
<td>Energy (02)</td>
<td>0.48</td>
<td>0.15</td>
<td>0.63</td>
<td>0.01</td>
<td>0.57</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cell Cycle (10)</td>
<td>0.48</td>
<td>0.41</td>
<td>0.86</td>
<td>0.03</td>
<td>0.89</td>
<td>0.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transcription (11)</td>
<td>0.47</td>
<td>0.36</td>
<td>0.13</td>
<td>0.00</td>
<td>0.58</td>
<td>0.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein Synthesis (12)</td>
<td>0.51</td>
<td>0.52</td>
<td>0.97</td>
<td>0.24</td>
<td>1.00</td>
<td>0.08</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein Fate (14)</td>
<td>0.38</td>
<td>0.26</td>
<td>0.94</td>
<td>0.01</td>
<td>0.91</td>
<td>0.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transp. Elements (38)</td>
<td>0.29</td>
<td>0.57</td>
<td>0.25</td>
<td>0.09</td>
<td>0.50</td>
<td>0.43</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cell Fate (43)</td>
<td>0.30</td>
<td>0.09</td>
<td>0.00</td>
<td>0.00</td>
<td>0.58</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biogen. Cell. Comp. (42)</td>
<td>0.32</td>
<td>0.15</td>
<td>0.41</td>
<td>0.01</td>
<td>0.74</td>
<td>0.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metabolism (01)</td>
<td>0.47</td>
<td>0.50</td>
<td>0.85</td>
<td>0.03</td>
<td>0.85</td>
<td>0.04</td>
<td>0.43</td>
<td>0.31</td>
<td>0.60</td>
<td>0.03</td>
<td>0.70</td>
<td>0.09</td>
</tr>
<tr>
<td>Energy (02)</td>
<td>0.59</td>
<td>0.16</td>
<td>0.98</td>
<td>0.03</td>
<td>0.76</td>
<td>0.08</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cell Cycle (10)</td>
<td>0.49</td>
<td>0.24</td>
<td>0.85</td>
<td>0.01</td>
<td>0.57</td>
<td>0.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transcription (11)</td>
<td>0.44</td>
<td>0.43</td>
<td>0.62</td>
<td>0.01</td>
<td>0.81</td>
<td>0.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein Synthesis (12)</td>
<td>0.52</td>
<td>0.47</td>
<td>0.95</td>
<td>0.18</td>
<td>1.00</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein Fate (14)</td>
<td>0.39</td>
<td>0.24</td>
<td>0.50</td>
<td>0.00</td>
<td>0.47</td>
<td>0.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prot. Bind. Func. (16)</td>
<td>0.29</td>
<td>0.14</td>
<td>0.00</td>
<td>0.00</td>
<td>0.57</td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cell Transport (20)</td>
<td>0.60</td>
<td>0.20</td>
<td>0.80</td>
<td>0.02</td>
<td>0.50</td>
<td>0.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transp. Elements (38)</td>
<td>0.38</td>
<td>0.48</td>
<td>0.00</td>
<td>0.00</td>
<td>0.77</td>
<td>0.47</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biogen. Cell. Comp. (42)</td>
<td>0.35</td>
<td>0.12</td>
<td>0.38</td>
<td>0.00</td>
<td>0.60</td>
<td>0.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- P: Predicted category
- R: Real category
- KNN: k-Nearest Neighbors
- UkNN: Unanimous k-Nearest Neighbors
- BALBOA: Our best prediction

The kNN, UkNN and BALBOA values correspond to the percentage of predictions made in each category, respectively.
dependent datasets should have more support than those from a single dataset only. One method of quantifying this cross dataset support is the union of probability rule for two corresponding independent events, given by $P_1 \cup P_2 = P_1 + P_2 - (P_1 \cap P_2)$. As a result, even the least supported annotations, i.e., those of Cellular Transport (20) with precisions 0.50 and 0.66, become more significant when consistently supported over two datasets, increasing to 0.83. With better supported classes, such as Transcription (11), such cross referencing achieves precision values close to 1.

The $\alpha$ and $\beta$ values we use in BALBOA may be somewhat stringent, especially if we include the dataset cross referencing step. However, looking at the final list, we see that most functional annotations seem to be supported by external evidence. One may consider this final cross referencing step a more valid selection procedure than the $\beta$ selection threshold, as it is derived from the data. Interestingly, we find that upon reducing $\beta$ to 0.5, 56 ORF annotations then agree across two or more datasets. This includes 12 that agree across all three. However, that being said, our chosen $\alpha$ and $\beta$ values are perhaps more conducive to minimizing false positives and also proving the efficacy of the BALBOA classification approach via external corroboration.

5 Discussion

BALBOA represents a novel, systematic approach for ORF classification using expression data. BALBOA's novelty lies both in the use of biclustering for training and classification and also in the subsequent voting based prediction strategy. Although such ensemble techniques are conceptually simple, resulting predictions tend to be more robust.

In section 4.3.1, cross validation showed that, on average, BALBOA achieves improved results over multi-class implementations of majority kNN and the more competitive unanimous kNN. Unlike kNN, BALBOA caters for 'local' correlations over feature subsets. This seems to allow for markedly improved classification precision for certain functional categories such as Transcription (11). This thus demonstrates the necessity for classification approaches in this domain to provide for the prospect of sub-space similarities. kNN is still a powerful technique when features are relevant and it may prove interesting in future work to perform kNN with those features selected by bicluster analysis.

In section 4.3.2 we saw that BALBOA's ORF predictions were well supported by external functional evidence from the Saccharomyces Genome Database (SGD). Strong support, in particular, was provided by the recent official
MIPS annotation of YCR072C. Although these annotations are putative they may still aid in the design of improved ‘wet lab’ experiments that may in turn lead to official annotation.

It is common to combine expression data into ‘global’ datasets prior to analysis. However, we have shown that maintaining multiple perspectives allows dataset cross referencing and potentially increases support for findings. Also, sample sets from different sources may reveal different ORF functions. Maintaining multiple perspectives may also prove useful in standard bicluster and cluster analyses.

Lastly, this work demonstrates the benefit of applying newly developed data mining techniques to re-assess expression data after initial studies and the potential new insights, in this case into ORF function, that this may provide.

Acknowledgements

Many thanks to Dr. Geraldine Butler for helpful comments and discussion. Research supported by Science Foundation Ireland Grant No. 05/IN.1/I24 and EU funded Network of Excellence Muscle - Grant No. FP6-507752.

References