Bottom-Up Biclustering of Expression Data

Kenneth Bryan and Pádraig Cunningham

***Abstract*— In a gene expression data matrix a bicluster is a sub-matrix of genes and conditions that exhibits a high correlation of expression activity across both rows and columns. The premise behind biclustering is that even related genes may only be expressed in a synchronized fashion over certain conditions. Conventional clustering groups over all features and may not capture these local relationships. Biclustering has the potential to retrieve these local signals and also to model overlapping groups of genes. These factors allow better representation of the natural state of functional modules in the cell.**

The mean squared residue is a popular measure of bicluster quality. One drawback however is that it is biased toward flat biclusters with low row variance. In this paper we introduce an improved bicluster score that removes this bias and promotes the discovery the most significant biclusters in the dataset. We employ this score within a new biclustering approach based on the bottom-up search strategy. We believe that the bottom-up search approach better models the underlying functional modules of the gene expression dataset.

We evaluate our new score against the mean squared residue score using a yeast cell cycle expression dataset. We then carry out a comparative analysis of our biclustering technique against previously published clustering and biclustering approaches. Lastly, we use the biclusters discovered by our method to attempt to putatively annotate unclassified genes.

I. Introduction

Advances in gene expression microarray technologies over the last decade or so have made it possible to measure the expression levels of thousands of genes over many experimental conditions (e.g. different patients, tissue types and growth environments). The data produced in these experiments are usually arranged in a data matrix of genes (rows) and conditions (columns). Results from multiple microarray experiments may be combined and the data matrix may easily exceed thousands of genes and hundreds of conditions in size.

Depending on the aims of the experiment in question there may be one or more objectives when analyzing gene expression datasets. If genes exhibit similar expression activity across experimental conditions this may be indicative of an *in vivo* functional relationship i.e. a common enzymatic pathway or cellular structure. This premise enables both the putative classification of unknown genes and the higher level grouping of genes into classes which may reflect in vivo system organization [1]. These objectives are the focus of this paper. Expression profiles of conditions may also be compared enabling disease types such as cancers to be organized and classified at a molecular level [2].

Uncovering the relationships between genes and their corresponding class information from such large volumes of data presents a far from trivial task. The unsupervised learning technique of cluster analysis was one of the first computational techniques applied to gene expression data [3]. This technique aims to group genes into distinct clusters based on their expression similarities across multiple experimental conditions. For expression data the most suitable similarity metric is one that computes correlation, rather than distance, such as Pearson's correlation coefficient.

As datasets increase in size however, it becomes increasingly unlikely that genes will retain correlation across the full set of conditions making clustering problematic. The gene expression context further exacerbates this problem as it is not uncommon for the expression of related genes to be highly similar under one set of conditions and yet independent under another set [4]. Given these issues it is perhaps more prudent to cluster genes over a significant subset of experimental conditions. This two-way clustering technique has been termed *biclustering* and was first introduced to gene expression analysis by Cheng and Church [5]. They developed a two-way correlation metric called the *mean squared residue score* to measure the bicluster quality of selected rows and columns from the gene expression data matrix. They employed this metric within a top-down greedy node deletion algorithm aimed at discovering all the significant biclusters within a gene expression data matrix. Following this seminal work other metrics and biclustering frameworks were developed [6],[7],[8]. However approaches based on Cheng and Church's mean squared residue score remain most prevalent in the literature [9],[10],[11],[12].

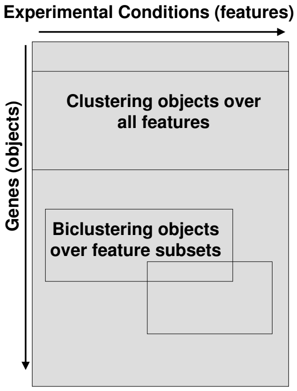
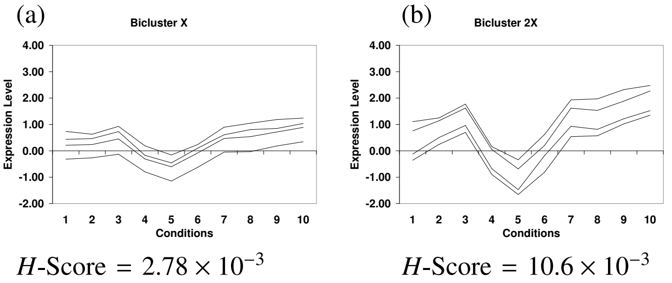

Experimental Conditions (features)
Genes (objects)
Clustering objects over all features
Biclustering objects over feature subsets

Fig. 1. A gene expression data matrix. Unlike conventional clustering which computes object similarity over all experimental conditions, biclustering may group similar objects over a subset of similar conditions.

One notable drawback however of the mean squared residue score is that it is also affected by variance, favouring correlations with low variances. Furthermore, because variance

This research was sponsored by Science Foundation Ireland under grant number SFI-02/IN1/I111.
1-4244-0623-4/06/$20.00 2006 IEEE.

changes by the square of the change in scale, the score tends to discover correlations over lower scales. These effects culminate in a bias toward 'flat' biclusters containing genes with relatively unfluctuating expression levels within the lower scales (fold changes) of the gene expression dataset. This issue has been articulated previously in [13]. We illustrate the mean squared residue (H-Score) bias in Figure 2. In this paper we introduce an improved bicluster scoring metric which compensates for this bias and enables the discovery of biclusters throughout expression data, including those potentially more interesting correlations over the higher scales (fold changes).



H-Score $= 2.78 \times 10^{-3}$ H-Score $= 10.6 \times 10^{-3}$

Fig. 2. Bias of the H-Score over different scales. The relative correlation is approximately the same yet the difference in H-Score is 4 fold. With a lower (better) H-score bicluster X (a) is favoured over bicluster 2X (b).

The original Cheng and Church node deletion algorithm was based on a top-down search approach. It began with the entire dataset and iteratively deleted rows and columns gradually improving the mean squared residue score of the sub-matrix. This top-down global approach is more likely to discover the best scoring bicluster within a dataset. However this approach may not be ideal when one wishes to discover a heterogeneous set of biclusters that reflects the local underlying trends in the dataset. Our new biclustering approach, termed BUBBLE (Bottom-Up Biclustering By Locality Expansion), is based upon the bottom-up search strategy and utilises our improved scoring function.

We discuss the biclustering approach in detail in section II. We detail our improved scoring metric in section III-A and the two steps of our algorithm in sections III-B and III-C In our evaluation, section IV, we first evaluate our improved scoring metric against the mean squared residue score. We then evaluate our biclustering technique by comparison with previously published clustering and biclustering techniques. Finally, through examining our biclustered genes, we attempt to putatively annotate unclassified genes.

II. Biclustering Gene Expression Data

A. The bicluster model of gene expression data

In general biclustering refers to the 'simultaneous clustering' of both rows and columns of a data matrix [14]. Hartigan pioneered this type of analysis, which he termed *direct clustering*, in the 1970s using two-way analysis of variance to locate constant valued sub-matrices within datasets. Biclustering is quite similar in concept to sub-space clustering. Sub-space clustering aims at improving the object similarities by selecting subsets of attributes. Biclustering however aims at improving similarity in both directions, within a subset of objects (rows) as well as a within a subset of features (columns). This approach suits the gene expression context as related genes are thought to be regulated in a synchronised fashion under certain cellular states (conditions) [4]. Biclustering attempts to identify both these related genes and the states in which they function together. Discovered biclusters may represent modules of genes which act together to carry out a specific function required by a specific cellular state. For example a bicluster could represent a group of genes which are only co-regulated under certain diseased states such as cancer. Discovery of such a set would provide possible drug targets in the treatment of this cancer type. In normal cells biclustering may also aid in elucidation of normal functional modules such as genes involved in the cell cycle or genes involved in transcription. Unclassified genes, or more correctly unclassified open reading frames (ORFs), may also be annotated from such a model if they are grouped within a bicluster with a predominant functional category.

B. The Cheng and Church Approach

Cheng and Church defined a bicluster to be a subset of genes and a subset of conditions with a high similarity score, where similarity is a measure of the coherence of genes and conditions in the subset. A group of genes are said to be coherent if their levels of expression react in parallel or correlate across a set of conditions. Similarly, a set of conditions may also have coherent levels of expression across a set of genes. Cheng and Church developed a measure, called the *mean squared residue score*, which takes into account both row and column correlations and therefore makes it possible to simultaneously evaluate the coherence of rows and columns within a matrix.

They thus defined a bicluster to be a matrix composed of a subset of genes and a subset of conditions with a low mean squared residue score (the lower the score the better the correlation of the rows and columns). The residue score of an entry a_{ij} in a bicluster $B(I, J)$ (where I is the subset of rows and J the subset of columns) is a measure of how well the entry fits into that bicluster. It is defined to be:

$$R(a_{ij}) = a_{ij} - a_{Ij} - a_{iJ} + a_{IJ} \tag{1}$$

where a_{iJ} is the mean of the ith row in the bicluster, a_{Ij} is the mean of the jth column and a_{IJ} mean of the whole bicluster. The overall *mean squared residue score* is:

$$H(I, J) = \frac{1}{|I||J|} \sum_{i \in I, j \in J} R(a_{ij})^2 \tag{2}$$

The next problem to be tackled is how to locate these low scoring biclusters within a parent data matrix. The exhaustive approach is to sequentially run through all the possible combinations of rows and columns of the data matrix and find the sub-matrices which satisfy a predefined low score, δ (the set of δ-biclusters). The most significant biclusters, the largest δ-biclusters, would be of most interest as they capture the relationships between the largest number of objects. However the number of possible sub-matrices increases exponentially with the size of the parent matrix making this task practically impossible when the matrix exceeds the fairly modest size of a few hundred elements. Cheng and Church likened the maximum bicluster search to that of locating a maximum biclique (largest complete sub-graph) within a parent bipartite graph which has been shown to be NP-Hard [15].

Cheng and Church designed a set of heuristic algorithms to efficiently locate these δ-biclusters. The search proceeds in a top-down manner with the initial solution being the parent matrix. Initially groups of ill-fitting rows and columns are deleted in a multiple node deletion phase. This is followed by a more refined single node deletion phase which continues until a predefined δ-score threshold is surpassed. If this single node deletion phase is carried out alone it is in fact possible to locate larger δ-biclusters but less efficiently. Upon reaching the chosen score (the δ threshold) a node addition phase is then carried out to add rows/columns which may have been missed. Cheng and Church achieve this by adding rows with mean squared residue scores less than the δ-threshold when compared with the discovered bicluster. Inversely correlated rows, which may represent negatively regulated genes, are also added at this stage by calculating the inverse mean squared residue score for all rows and adding those less than the δ-threshold. These relationships are also referred to as anti-correlated genes and 'diametrical' clustering has previously been employed in discovery of such relationships [16]. The set of δ-biclusters are discovered sequentially in a deterministic manner so solutions need to be masked to avoid rediscovery. This is achieved by replacement of these entries with random numbers generated from the same range of the dataset. The Cheng and Church approach is intuitive and powerful, finding significant bicluster signals within both human and yeast datasets. However in their seminal paper no effort is made to interpret these results from a biological perspective by looking for gene functional group correspondence within the discovered biclusters. Also one may argue that the top-down global approach may find an artificially large bicluster rather than reflect accurately the set of natural bicluster signals.

C. Subsequent Biclustering Approaches

Following the work above several alternative biclustering approaches have been taken within gene expression analysis. One approach taken by Tanay *et al.* [6] likens biclustering to the search for complete sub-graphs within a bipartite graph. They developed a statistical model of the expression data matrix and propose a heuristic algorithm, called SAMBA, that discovers statistically significant sub-graphs. Lazzeroni and Owen [7] developed what they termed a *plaid model* in which the dataset is represented by a linear function of variables or *layers* which correspond to biclusters. Yet another approach taken by Kluger *et al.* [8] involved decomposing the data matrix into its principle components by singular value decomposition. The resulting eigenvectors are then used to reorder the data matrix to reveal the set of biclusters as a *checkerboard* structure. These approaches are less intuitive and theoretically quite different from that of Cheng and Church.

Several biclustering approaches using metrics based on Cheng and Church's residue scoring have also been pursued. The approach by Yang *et al.* avoids random masking, and what they refer to as the *random interference* it may cause, by locating biclusters simultaneously rather than sequentially. Interestingly they also suggest the possibility of an additional row variance criterion to the search but do not pursue this point. In their evaluation Yang *et al.* only demonstrate improvements over Cheng and Church's technique within two discovered biclusters.

Cho *et al.* used the *sum* squared residue as a biclustering score rather than the mean squared residue. This is perhaps a more sensitive to individual row and column changes. Cho *et al.* do not improve on Cheng and Church's results in terms of bicluster size and quality but succeed in capturing some significant signals in the data [10]. Both of the above approaches generally fail to find biclusters as significant, in terms of size and quality as Cheng and Church. Bleuler *et al.* use a stochastic approach in an attempt to improve on Cheng and Church's greedy search by implementing an evolutionary algorithm (EA) [12]. They improve on the original Cheng and Church algorithm results in terms of bicluster size and quality discovering larger δ-biclusters. However they do not improve on the solutions achievable when one implements Cheng and Church's refined single node deletion search alone (see section II-B). Interestingly, the above approaches neglect to evaluate their bicluster models from a biological perspective by assessing the functional relationships of the genes in the biclusters. For a detailed review of the above biclustering approaches the reader is directed to [17]. Another stochastic approach by Bryan *et al.* was based on the global search technique of simulated annealing. This biclustering approach achieved improved results over the original Church and Cheng algorithm. Furthermore it also showed improvements over an augmented version of Cheng and Church's algorithm which dealt with single node deletion only over a defined minimum number of conditions. This was also the first mean squared residue biclustering technique to use a fully annotated expression dataset to biologically validate discovered biclusters [11].

III. Bottom-Up Biclustering By Locality Expansion

In general, as with conventional clustering, biclustering techniques fall into top-down and bottom-up approaches. Top-down approaches begin by finding an initial approximation of the solution over the full set of objects and features, in this case all the rows and columns of the gene expression data matrix. There is a possibility that this method may however discover an artificially large bicluster solution which may be a combination of parts of two or more local solutions.

Top-down methods tend to produce a more uniform set of clusters in terms of size and shape. This representation may not accurately model the diverse set of functional modules that may be present in expression data. Also because they deal initially with the full set of dimensions, top-down approaches may not scale well as the dataset increases in size.

Our framework is built around the bottom-up search approach. It is founded on the premise that searching for interesting structure in higher dimensions can be first reduced to a search for structure in lower dimensions. In general this search method is more adept at discovering several local solutions rather than global solutions. In this way we hope to discover a more natural set of biclusters that capture the diversity and organization of functional groups in the cell. This approach also has the advantage that it is computationally more efficient to search for solutions over a smaller set of dimensions.

In the bicluster model every sub-matrix within a bicluster is itself a bicluster. If we can locate the most highly correlated sub-bicluster then it is likely that we can expand this region, by adding correlating rows and columns, to reveal the larger bicluster. Even if biclusters partially overlap, i.e. share some of the same rows and columns, they must still contain a significant sub-bicluster which is unique to that bicluster and justifies the partition.

Our approach, termed BUBBLE, can be divided into two phases, firstly we perform a stochastic search for the set of highly correlated sub-biclusters. These are then used as a set of seeds for the next phase, that of a deterministic expansion into higher dimensions by adding the best fitting rows and columns. These algorithms will be discussed in this section but first we will introduce our improved bicluster scoring metric that is essential in locating these significant bicluster seeds.

A. An Improved Bicluster Scoring Metric

As already mentioned, the mean squared residue (H-Score) is affected by the scale of the variance within biclusters. This point was illustrated in Figure 2 and may bias a search toward low fold (possibly less interesting) gene correlations. To address this problem our bicluster scoring metric, the Hv-Score, takes into account each entry's distance from its row mean. The sum of the squares of these distances are computed and used as the denominator in a new bicluster scoring measure given in equation 3. The numerator is simply the sum of the squared residues in the matrix. Minimizing this function for the selected sub-matrix solution minimizes the sum of the squared residues while maximizing the sum of the distances from the row means. The Hv-Score is defined as:

$$Hv(I, J) = \frac{\sum_{i \in I, j \in J} R(a_{ij})^2}{\sum_{i \in I, j \in J} (a_{ij} - a_{iJ})^2} \tag{3}$$

where $R(a_{ij})$ is the residue score of each entry, a_{ij}, (see equation 1) and a_{iJ} is the row mean for each entry. We evaluate this new metric in section IV-A and use it to locate highly correlated bicluster seeds across the scales of the dataset in the *seed search* step of our approach.

B. Seed Search

The goal of the initial seed search algorithm is to locate a diverse set of highly correlated bicluster seeds throughout the gene expression data matrix. If these seeds are significant in size it is more probable that they represent part of a larger bicluster of rows and columns. For example a seed that consists of 10 genes actively correlating over 10 conditions is likely to be a part of a larger set of related genes and would be a good base on which to expand.

Our seed search is based on the well known stochastic search technique of simulated annealing. It has been shown that simulated annealing search can discover improved bicluster solutions over the best-first greedy search method [11]. Unlike greedy search techniques, simulated annealing allows the acceptance of reversals in fitness (worse solutions) and backtracking which gives it the capability to bypass locally good solutions until it converges on a solution near the global optimum. The acceptance of reversals is probabilistic as defined by Boltzman's equation:

$$P(\Delta E) \propto e^{-\frac{\Delta E}{T}} \tag{4}$$

From the equation it can be seen that this probability depends on two variables, the size of the reversal ΔE and Temperature of the system, T. Generally T is first given a high value to initially allow a large percentage of reversals to be accepted, this helps the search explore the entire search space. Gradually this T is lowered and the potential search space shrinks until it converges on the global optimum. The number of solutions evaluated (attempts) and the number of solutions accepted (successes) at each temperature are also input parameters. These determine how extensive a search is carried out before the T is lowered.

If we reduce both the initial temperature and the depth of search at each temperature we can confine the search to a more local area of the data. Sequential searches will be able to uncover several local optima rather than one solution close to the global optimum. With regard to gene expression data, these regional optima may correspond to the diverse set of bicluster seeds.

Beginning each seed search at a random starting point will then hopefully be able to locate a diverse set of local optima spanning all localities. These optima are then used as seeds and expanded in size to model the larger set of relationships within the full bicluster. This seed expansion phase is described in detail in the following section.

C. Seed Expansion

Upon locating a good set of seeds the deterministic phase of seed expansion involves adding the best fitting rows and columns to each seed. Prior to seed expansion the correlation of the rows and columns in the remainder of the dataset is measured relative to the seed. The following formulae based on the residue score, Equation (1), are used to score these rows and columns.

```
Algorithm I: Seed Search.

Variable definitions:
$x$ : current solution,
$t_0$ : initial temperature, $t$ : current temperature,
$rate$ : temperature fall rate,
$a$ : attempts, $s$ : successes,
$a_{count}$ : attempt count, $s_{count}$ : success count,
$Hv$ : $Hv$ fitness function, $M$ : data matrix,
$row$ : seed row size,
$col$ : seed column size,

Seed Search ($t_0, rate, row, col, s, a, M$)
$x \leftarrow$ randomSolution($row$,$col$)
$t \leftarrow t_0$
while($x$ not converging)
    while($a_{count} < a$ AND $s_{count} < s$)
     $x_{new} \leftarrow$ GenerateNewSolution($M$, x)
    if $Hv(x_{new}) < Hv(x)$
      then $x \leftarrow x_{new}$
    else if $\exp(-\frac{\Delta E}{T})$> random(0,1)
      then $x \leftarrow x_{new}$
$t \leftarrow$ Cool($t$,$rate$)
return $x$
```

All rows (genes) not in Seed(I, J) are scored as follows:

$$H(I) = \frac{1}{|J|} \sum_{i \in I, j \in J} (a_{ij} - a_{Ij} - a_{iJ} + a_{IJ})^2 \quad (5)$$

where $i \notin I$. All columns (conditions) not in the Seed(I, J) are similarly scored:

$$H(J) = \frac{1}{|I|} \sum_{i \in I, j \in J} (a_{ij} - a_{Ij} - a_{iJ} + a_{IJ})^2 \quad (6)$$

where $j \notin J$.

An important attribute of this expansion phase is that it adds correlated rows and columns regardless of scales. Rows and columns are standardized by subtracting the mean and dividing by the standard deviation. These scores are then sorted and the best fitting column or row is added in each iteration. As our main objective in this study is to capture gene (row) correlations we only add rows during seed expansion.

A key question we now encounter is when to halt this seed expansion process. One approach is a score or size threshold. However this method would be somewhat arbitrary as biclusters representing gene functional modules should be of different sizes and scores. A method of stopping can be derived from observing the trends in the scores of the added rows. It can be seen in Figure 3 that this trend is not a smooth gradient and is interrupted by steps as the dissimilarity increased abruptly as a more ill-fitting row is added. These steps may be viewed as partitions in similarity. Our stopping method involves recording all growth steps and stopping expansion at the largest step or partition.

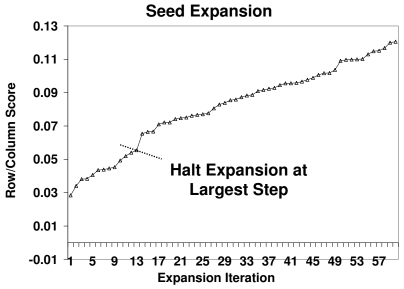


Fig. 3. The trend in row scores as the are added to the seed at each expansion iteration. Steps in the gradient may be observed retrospectively.

```
Algorithm II: The Seed Expansion Phase.

Variable definitions:
$r$ : row, $c$ : column,
$R$ : rows in seed,
$C$ : columns in seed,
$r$VarT : row variance threshold,
$c$VarT : column variance threshold,
$r$T : seed row threshold,
$c$T : seed column threshold.

SeedExpansion(Seed($R, C$),$r$VarT,$c$VarT,$r$T,$c$T)
Score all $r \notin R$
Score all $c \notin C$
Sort $r$ scores
Sort $c$ scores
      select best scoring $r$ or $c$
      if (variance of $r/c$ > $r/c$VarT)
         if ($r/c$ number < $r/c$T)
            add $r/c$ to Seed($R, C$)
      re-score Seed($R, C$)
return expanded Seed($R, C$)
```

D. The Yeast Cell Cycle Dataset

In our evaluation we used the yeast cell cycle dataset from Cho *et al.*, 1998 [18]. This dataset has been used in previous biclustering [5],[12] and clustering studies [19]. Cheng and Church used a subset of 2,884 of the most variable genes from Cho's dataset. Unlike expression data from the larger human genome, yeast data gives a more complete representation in which all the ORFs and functional modules are covered.

Cho's dataset contains 6,178 genes, 17 conditions and 6,453 missing values. Rows containing many missing values were removed giving 6,145 rows with 5,846 missing values. Missing values were replaced by random values generated between the first and third quartiles of the data range. This reduces the impact of these imputed values on the structure of the dataset. We annotated all genes in our dataset using the MIPS (Munich Information centre for Protein Sequences) online functional catalogue [20]. Over 1500 genes in the dataset were labelled as category 99 (Unclassified). These annotations were used to evaluate the correspondence of the biclusters to gene functional modules.

TABLE I

Comparison between the H-Score and the Hv-Score. 10 bicluster seeds discovered by *Seed Search* with each metric. The percentage of the Dominant Functional Category (D.F.C.) and MIPS code and name are also listed.

	H−Score		Hv-Score	
Seed	D.F.C.	MIPS Category	D.F.C.	MIPS Category
1	40%	99: Unclass. ORFs	70%	12: Protein Synth.
2	50%	14: Protein Fate	60%	10: Cell Cycle
3	30%	01: Metabolism	40%	99: Unclass. ORFs
4	40%	14: Protein Fate	90%	11: Transcription
5	60%	14: Protein Fate	60%	12: Protein Synth.
6	40%	32: Cell Rescue	50%	99: Unclass. ORFs
7	40%	99: Unclass. ORFs	50%	14: Protein Fate
8	20%	12: Protein Synth.	40%	99: Unclass. ORFs
9	50%	14: Protein Fate	50%	12: Protein Synth.
10	40%	14: Protein Fate	70%	10: Cell Cycle

IV. Evaluation

In the first part of our evaluation we evaluate our improved metric, the Hv-Score, against Cheng and Church's original H-score. We then carry out comparative evaluations with recent clustering and biclustering approaches and end by using our model to putatively annotate several unclassified yeast genes.

A. Evaluation of Metrics

The first part of our evaluation involves a comparison of the H-Score and the Hv-Score metrics and their ability to discover good seeds within expression data. We achieve this by implementing the seed search algorithm with both scores. In Figure 4 we graphically illustrate the difference between the seeds found with each metric. The Hv-Score discovers seeds with significant correlations, 4(b), missed by the biased mean squared residue, 4(a). The next step is to discern whether this more significant correlation translates into discovering seeds with higher functional enrichment i.e. seeds with a predominance of one functional category. In Table I we present the first 10 seeds found in the yeast dataset using each metric. We see that seeds found using the mean squared residue (H-Score) have, on average, 40% functional enrichment whereas those found using the Hv-Score have, on average, 60%.

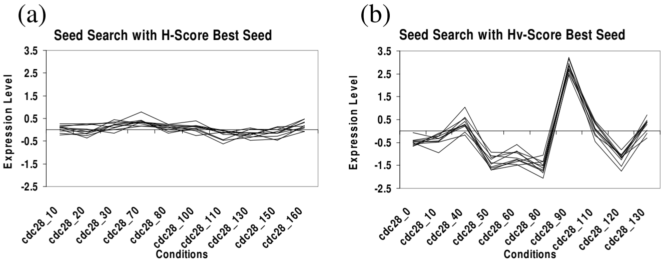


Fig. 4. Here we present the best scoring seed found with the H-Score (a) and the Hv-Score (b). The Hv-Score discovers correlations missed by the low variance biased H-Score.

B. Comparative Evaluation with Clustering

In this section of the evaluation we compare our biclustering method, BUBBLE, to the published results of a recently developed clustering method called CLARITY[19]. This approach was also evaluated using the full yeast cell cycle dataset from Cho *et al.* labelled from the MIPS database.

Biclustering has the potential to discover gene relationships only evident over a subset of conditions however this subset still needs to be significant in size. We chose a minimum condition size of 10 for our bicluster seeds. As BUBBLE has a stochastic seed search step we retrieve 100 biclusters to achieve good coverage within the dataset. The clusters and biclusters were evaluated by examining their functional enrichment i.e. the percentage belonging to the same functional category. The dominant functional category in each cluster or bicluster was used to label the group.

We present the results of this evaluation in Table II. The best clusters from each functional category discovered by CLARITY are given on the left. It can be seen that for each cluster discovered by CLARITY our method finds biclusters with higher functional enrichment. Furthermore BUBBLE also discovers biclusters with dominant functional categories not found by clustering, these are presented in Table III in the next section.

C. Comparative Evaluation with SAMBA

A biclustering method developed by Tanay *et al.* called SAMBA (Statistical-Algorithmic Method for Bicluster Analysis) is mentioned in section II-C. This is one of the best developed biclustering methods and is implemented within a software package called Expander [21]. In this section we compare our biclustering approach to SAMBA. We again use the yeast cell cycle dataset in this evaluation. SAMBA discovers 24 biclusters within this dataset. To compare these two differing biclustering methods we selected the biclusters with the highest enrichments for each of the 18 functional categories in MIPS. In Table III we list the most significant biclusters found and their corresponding MIPS categories. The largest functional enrichments for each category obtained by SAMBA and BUBBLE are also listed. For this particular yeast dataset BUBBLE wins in 14/18 categories in terms of functional enrichment, loses in 1 and draws in 3. Interestingly both methods discover a similar distribution across the functional categories. This supports the inference that this is the natural distribution in the dataset. Both methods also discover biclusters in which the dominant functional category is a group of unclassified genes or ORFs. In such groups the function of such genes may be inferred by other known genes in the bicluster. However the support for such inferences increases greatly when unclassified genes are grouped within a bicluster with a high functional enrichment for a known MIPS category. Such methods of classifying unclassified genes are examined in the next section.

D. Putative annotation of unclassified genes

The goal of unsupervised data analysis such as clustering and biclustering is to try to discover the underlying patterns within the gene expression data that accurately reflect the functional modules within the cell. Accurate generation of a global model of the cell's functional modules from expression

TABLE II

A comparison between the clusters and biclusters found in the yeast cell cycle dataset. The total genes or Open Reading Frames (ORFs) within each grouping (k) and number of ORFs belonging to the MIPS functional category are listed.

	Clustering (CLARITY)				**Biclustering (BUBBLE)**			
Dominant Functional Category (M.I.P.S. Code)	k	**ORFs in Cluster**	**ORFs in Category**	**Functional Enrichment**	k	**ORFs in Bicluster**	**ORFs in Category**	**Functional Enrichment**
Protein Synthesis (12)	0	43	33	77%	78	45	40	**89%**
Cell Cycle & DNA(10)	8	113	64	57%	31	36	25	**69%**
Transcription(11)	5	61	24	39%	91	15	9	**60%**
Energy(02)	21	201	33	16%	46	15	7	**47%**
Protein Fate(14)	6	86	19	22%	56	15	6	**40%**
C-Compound Metabolism (01.05)	21	201	32	16 %	53	21	6	**40%**
Amino Acid Metabolism (01.01)	9	77	9	12%	84	17	3	**18%**

TABLE III

Comparison between SAMBA and BUBBLE on the yeast cell cycle dataset. The best functional enrichments for each category are marked in bold.

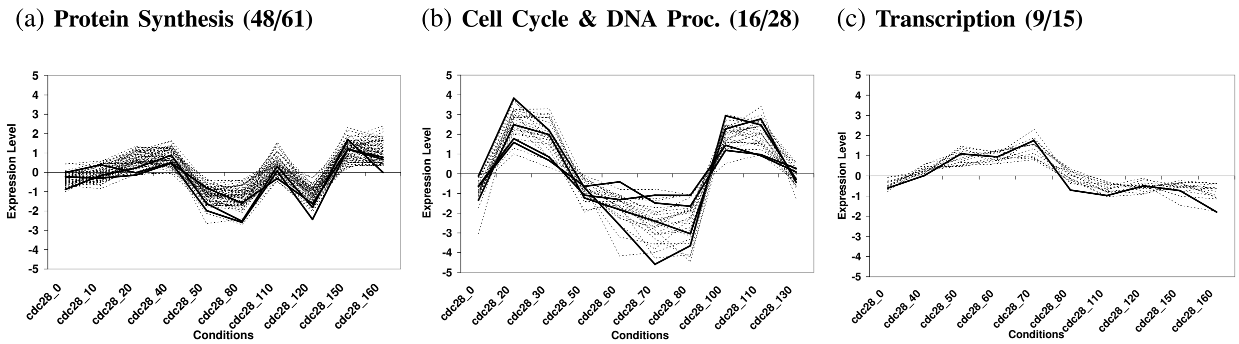
	SAMBA			**BUBBLE**		
Functional Category (M.I.P.S. Code)	**ORFs in Bicluster**	**ORFs in Category**	**Functional Enrichment**	**ORFs in Bicluster**	**ORFs in Category**	**Functional Enrichment**
Metabolism (01)	20	12	60%	15	9	60%
Energy (02)	60	11	18%	15	7	**47%**
Cell Cycle & DNA (10)	16	11	69%	36	25	69%
Transcription (11)	29	6	21%	15	9	**60%**
Protein Synthesis (12)	43	21	49%	45	40	**89%**
Protein Fate (14)	20	4	20%	15	6	**40%**
Protein with Binding Function (16)	20	5	25%	17	7	**42%**
Cellular Transport (20)	17	6	35%	17	11	**67%**
Cell Rescue,Defence & Virulence (32)	32	6	19%	20	6	**30%**
Biogenesis of Cellular Components (42)	27	8	30%	16	7	**44%**
Cell Type Differentiation (43)	16	3	19%	15	6	**27%**
Unclassified Genes (99)	17	7	41%	16	10	**63%**

data has not yet been achieved. Indeed it might be argued that the use of gene expression data alone may not be sufficient to elucidate such a model. As we have seen in the previous sections certain functional modules may be more easily modelled from their expression data than others. However even using this partial representation it may be possible to make inferences about the nature of some unclassified genes or ORFs.

In the biclustering of gene expression data we notice that unclassified ORFs are often grouped within biclusters that have high enrichments for one particular functional category. Essentially these unclassified ORFs have an expression similar to an expression signature of a group of related genes. Using this information we may attempt to putatively annotate functions to unclassified ORFs. Ideally we would then corroborate this inference with so called ‘wet lab’ experiments but it may also be possible to garner addition functional evidence from other sources such as nucleotide or protein sequence information. In this section we examine three biclusters with high functional enrichments that contain one or more unclassified ORFs in an attempt to putatively annotate these ORFs. Additional supporting evidence for the suggested functions is also presented when available. In Figure 5 we present three expression biclusters in (a) Protein Synthesis, (b) Cell Cycle & DNA Processing and (c) Transcription followed by the ratio of labelled ORFs in each bicluster. They contain 3, 4 and 1 unclassified ORFs respectively, highlighted in bold. Beneath each graph we list the names of the ORFs and any available external evidence to support our putative annotation. We found some additional protein sequence similarities between ORFs YDR154C, YDL009C and YLR073C and genes from the functional category to which they were putatively assigned. YDR210W was found to localize in cell periphery, where protein synthesis occurs. We also found that YPL267W was already postulated to be a substrate for the cell cycle regulator Cdc28p in budding yeast. This available evidence supports some of our putative annotations which in turn aids in validation of our model. More substantial evidence is needed to actually classify these unclassified ORFs, in the form of *in vitro* biological experimentation, however these putative annotations may aid in the design of such experiments.

V. Discussion & Future Work

In this paper we have demonstrated improvements to the popular mean squared residue scoring function by removing the bias it shows toward so called ‘flat’ biclusters. Furthermore our *Hv*-Score was shown to be a more effective objective function from both qualitative graphical illustrations and quantitative functional analysis of biclusters. Our BUBBLE algorithm which incorporates this scoring function performs well against recent clustering and biclustering techniques on the yeast cell cycle dataset. Our chosen method of bicluster validation, that of functional enrichment, seems the most appropriate measure when one wishes to apply the model to putatively annotate unclassified ORFs. Other significance measures for gene groupings, such as hypergeometric probability scores and


(a) **Protein Synthesis (48/61)**

(b) **Cell Cycle & DNA Proc. (16/28)**

(c) **Transcription (9/15)**

Unclassified ORFs in (a)	**Supporting Evidence**	**Unclassified ORFs in (b)**	**Supporting Evidence**	**Unclassified ORFs in (c)**	**Supporting Evidence**
YDR133C	-	YBR089W	-	YLR073C	Protein sequence similarity to gene (YOR290C) in this class.
YDR154C	Protein sequence similarity to gene (YNR038W) in this class.	YDL009C	Protein sequence similarity to gene (YKL189W) in this class.		
YDR210W	Localizes in cell periphery.	YNL303W	-		
		YPL267W	Potential Cdc28p substrate.		

Fig. 5. Three functionally enriched biclusters containing unclassified ORFs (bold lines) and supporting protein sequence evidence for these annotations.

p-values, often fail to render much meaning when one wishes to examine the nature of the bicluster results from real datasets.

The final section on annotation suggests some possible functions for unclassified ORFs. In several cases this annotation is supported by other bioinformatic methods such as primary sequence analysis of the hypothesized protein product or the ORF in question. Accumulating multiple diverse sources of evidence greatly strengthens such *in silico* classification.

In future work it is our hope to compound these and further putative classifications by cross validation across multiple expression datasets. Cho's yeast cell cycle dataset is a well studied dataset and a good benchmark but may lack the condition number to fully test a subspace clustering method such as biclustering. Using larger datasets we hope to further validate our biclustering method and also increase the support for gene relationships and putative classifications by discovering biclusters over a larger number of conditions.

References

[1] D. Berrer, W. Dubitzky, and S. Draghici, *A Practical Approach to Microarray Data Analysis*. Kluwer Academic Publishers, 2003, ch. 1, pp. 15–19.

[2] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E. McLaughlin, J. Y. Kim, L. C. Goumnerova, P. Black, C. Lau, J. C. Allen, D. Zagzag, J. M. O. andT Curran, C. Wetmore, J. A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, D. Stolovitzky, D. Louis, J. Mesirov, E. Lander, and T. Golub, "Prediction of central nervous system embryonal tumour outcome based on gene expression," *Nature*, vol. 24, no. 415, pp. 436–42, 2002.

[3] I. S. Dhillon, S. Mallela, and D. S. Modha, "Information theoretic co-clustering," in *Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003.

[4] A. Ben-Dor, B. Chor, R. Karp, and Z. Yakini, "Discovering local structure in gene expression data: the order-preserving submatrix problem," *Journal of Computational. Biology*, vol. 10, no. 3-4, pp. 373–84, 2003.

[5] Y. Cheng and G. M. Church, "Biclustering of expression data," in *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB)*, 2000, pp. 93–103.

[6] A. Tanay, R. Sharan, and R. Shamir, "Discovering statistically significant biclusters in gene expression data," *Bioinformatics.*, vol. 18, pp. 36–44, 2002.

[7] L. Lazzeroni and A. Owen, "Plaid models for gene expression data." *Statistica Sinica*, vol. 12, pp. 61–86, 2002.

[8] Y. Kluger, R. Basri, J. T. Chang, and M. Gerstein, "Spectral biclustering of microarraydata: Coclustering genes and conditions," *Genome Research*, vol. 13, pp. 703–716, 2003.

[9] J. Yang, H. Wang, W. Wang, and P. Yu, "Enhanced biclustering on expression data," in *IEEE Third Symposium on Bioinformatics and Bioengineering*, 2003.

[10] H. Cho, I. S. Dhillon, Y. Guan, and S. Sra., "Minimum sum squared residue co-clustering of gene expression data." in *SIAM international conference on datamining*, 2004.

[11] K. Bryan, P. Cunningham, and N. Bolshakova, "Biclustering of expression data using simulated annealing," in *Proceedings of the eighteenth IEEE Symposium on Computer Based Medical Systems*, 2005.

[12] S. Bleuler, A. Prelić, and E. Zitzler, "An EA framework for biclustering of gene expression data," in *Congress on Evolutionary Computation (CEC-2004)*. Piscataway, NJ: IEEE, 2004, pp. 166–173.

[13] J. Aguilar-Ruiz, "Shifting and scaling patterns from gene expression data," *Bioinformatics*, vol. 21, no. 20, pp. 3849–3845, 2005.

[14] B. Mirkin, *Mathematical Classification and Clustering*. Dordrecht: Kluwer, 1996.

[15] D. S. Johnsen, "The np-completeness column: an ongoing guide." *Journal of Algorithms*, vol. 8, pp. 438–448, 1987.

[16] I. S. Dhillon, E. M. Marcotte, and U. Roshan, "Diametrical clustering for identifying anti-correlated gene clusters," *Bioinformatics*, vol. 19, no. 13, pp. 1612–1619, 2003.

[17] S. C. Madeira and A. L. Oliveira, "Biclustering algorithms for biological data analysis: A survey," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 1, no. 1, pp. 24–45, 2004.

[18] R. Cho, M. Campbell, E. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. Wolfsberg, A. Gabrielian, D. Landsman, and D. Lockhart, "A genome-wide transcriptional analysis of the mitotic cell cycle." *Molecular Cell*, vol. 2, pp. 65–73, July 1998.

[19] R. Balasubramaniyan, E. Hullermeier, N. Weskamp, and J. Kamper, "Clustering of gene expression data using a local shape-based similarity measure," *Bioinformatics*, vol. 21, no. 7, pp. 1069–1077, 2005.

[20] H.-W. Mewes, D. Frishman, U. Güldener, G. Mannhaupt, K. F. X. Mayer, M. Mokrejs, B. Morgenstern, M. Münsterkötter, S. Rudd, and B. Weil, "Mips: a database for genomes and protein sequences." *Nucleic Acids Research*, vol. 30, no. 1, pp. 31–34, 2002.

[21] R. Shamir, A. Maron-Katz, A. Tanay, C. Linhart, I. Steinfeld, R. Sharan, Y. Shiloh, and R. Elkon, "Expander-an integrative program suite for microarray data analysis." *BMC Bioinformatics.*, vol. 21, no. 6, p. 232, 2005.