A Time Series Analysis of Microarray Data

Selnur Erdal[1,2], Ozgur Ozturk[3], David Armbruster[1,2], Hakan Ferhatosmanoglu[3,4], William C. Ray[1,2]

[1]Columbus Children's Hospital Research Institute (CCRI), Columbus, Ohio.
[2]Department of Pediatrics, College of Medicine, The Ohio State University.
[3]Department of Computer Science and Engineering, The Ohio State University.
[4]Department of Biomedical Informatics, College of Medicine, The Ohio State University.
Emails: {erdals,rayw,armbrusd}@pediatrics.ohio-state.edu, {ozturk,hakan}@cse.ohio-state.edu

Abstract

As the capture and analysis of single-time-point microarray expression data becomes routine, investigators are turning to time-series expression data to investigate complex gene regulation schemes and metabolic pathways. These investigations are facilitated by algorithms that can extract and cluster related behaviors from the full population of time-series behaviors observed. Although traditional clustering techniques have shown to be effective for certain types of expression analysis, they do not take the biological nature of the process into account, and therefore are clearly not optimized for this purpose. Moreover, the current approaches provide internal comparisons for the experiments utilized for clustering, but cross-comparisons between clustered results are qualitative and subjective. We present a combination of current and novel methods for the analysis of time series gene expression data. We focus on an actual study we have performed for Haemophilus influenzae which is a major cause of otitis media in children. We first perform a discretization of the gene expression data that takes both positive and negative correlations into consideration and then develop a clustering algorithm optimized for such data that allows elucidation and searching of time-series patterns. The resulting approach allows time-series data to be usefully compared across multiple experiments. We demonstrate the success of our algorithm by showing some of the genes that it finds to be co-regulated are not detected by current methods. As a result we are able to identify several signal pathways that initiate competence development, and to characterize the transcriptomes of wild-type and an adenylate cyclase mutant (cya) strains under both nutrient-limiting and nutrient-complete growth conditions.

1. Introduction

Genes are the code of proteins that are fundamental components of all living cells and carry out vital organism functions. Before being translated into protein, this code must be transcribed from chromosomal DNA into messenger RNA (mRNA). The rate of transcription by the cell for some genes can be varied, and therefore the amount of certain mRNAs in the cell cytoplasm is a measure of the production speed of corresponding protein in the cell. Depending on the environment of the cell (and other factors), different amounts of some proteins are required; hence different concentrations of mRNAs for different genes exist in the cell. The relationship between the amount of an mRNA observed under experimental conditions, versus the amount observed under control conditions is called the expression level. Immobilized DNA microarrays (aka. probe arrays) are a tool for high-throughput gene expression studies. In microarray studies, probes (i.e. oligonucleotide sequences) with known identity are placed on glass or nylon substrates in a grid and used to determine expression levels through hybridization to bulk unknown populations of sequences (11). In the results we see the relative expression levels of genes.

As the capture and analysis of single-time-point microarray expression data becomes routine, investigators have started examining time-series expression data to investigate complex gene regulation schemes and metabolic pathways. The current approach is basically to cluster the time-series sequences based on common methods such as k-means. These algorithms provide internal comparisons for the experiments utilized for clustering, but cross-comparisons between clustered results are qualitative and subjective. In this paper, we present a combination of current and novel methods for the analysis of time series gene expression data. We first perform a discretization of gene expression data that takes both positive and negative correlations into consideration and then develop a clustering algorithm optimized for such data that allows both elucidation and searching of

time-series patterns. The resulting approach allows time-series data to be usefully compared across multiple experiments.

The proposed technique can be used as a decision support tool for a researcher who is searching for candidate genes in the process of identifying co-expressed genes, namely operons and regulons, based on microarray time-series data. An operon is a group of genes that are co-localized in the genome, and for which the mRNA is created as a single transcriptional event. A regulon is a group of operons or genes whose expression is coordinately regulated by a global regulatory mechanism (41). This can occur even though their locations may be unrelated in the genome. There is an urgent need for a decision support tool that provides intuitive ways of posing queries to discover meaningful patterns. Even though current methods reveal a significant insight into the data, in many cases they either produce more correlations than necessary and/or classify unrelated results as similar.

It cannot be strongly enough stated that in microarray experiments, the fundamental assumption on which all applications of the technique rely, is that the transcriptional events observed are related to the experimental conditions. The implications of this fundamental assumption however, are rarely brought to bear on the analysis of microarray expression data. This assumption demands that expression patterns from an experiment be treated as a completely related collection of data, and the relationships between all patterns examined, rather than treated as a collection of data containing disparate groupings of related patterns. Additionally, the relationships that might be observed in the data can occupy several biological axes, and the factors that a researcher may be interested in on one axis may have no correlation to values along another. This implies that any individual distance metric or clustering algorithm is incapable of capturing the full detail of microarray expression relationships, and also that any clustering algorithm may capture unique relationships that are unavailable through other distance/clustering methodologies. In this research we wished to examine the hypothesis that a distance metric that captured similarities in the changes of expression level between expression patterns (as opposed to metrics that capture similarities in the magnitude of expression), would be of assistance in clustering genes with coupled regulatory mechanisms.

Case Study on Haemophilus influenzae Although the proposed techniques have been implemented and evaluated for several microarray expression data sets, we have focused on analyzing our own microarray experiments for *Haemophilus influenzae*, a major cause of otitis media in children. *Haemophilus influenzae* is a particularly interesting application since it is known to be capable of natural DNA transformation, and this capability is strongly con-

trolled by known environmental factors. Many of the requirements and environmental factors that lead to this transformation competent state have been identified, however, additional signal pathways that initiate competence development have yet to be elucidated (27). Induction of competence absolutely requires the catabolite regulator protein CRP and the cofactor cyclic-AMP (cAMP) (7). This CRP-cAMP complex binds CRP regulatory elements proximal to various promoters resulting in the increased transcription of those genes. Although, the mechanisms by which CRP-cAMP binds the CRP site to promote transcription is fairly well understood, the global nature of competence has remained elusive due to the one gene at a time approach used to identify competence genes. Here, using microarray technology and a unique clustering algorithm over discretized data, we are able to characterize the transcriptomes of wild-type and an adenylate cyclase mutant(*cya*) strains under both nutrient-limiting and nutrient-complete growth conditions. The details of our biological findings are summarized in the Experiments section.

2. Background

Microarray technology is an evolutionary descendent of well-known nucleic-acid (NA) hybridization techniques such as Southern hybridization, scaled to genome-scale numbers of immobilized probes, and micron-scale interprobe spacing on the substrate (34; 35). These scales allow (potentially) every sequence in an organism to be simultaneously used as a probe for hybridization to an unknown population of NA sequences, in an attempt to determine the composition of that population. Probes are localized by a variety of methods into arrays of "spots" with known locations on a specially treated 1x3 microscope slide. Each slide can contain on the order of a few tens of thousands of spots using current technology. For a bacterial genome this allows several replications of the genome probes to be printed onto each slide, providing some internal controls and consistency checks. For human-sized genomes, the whole genome must be spread over several arrays, preventing the current use of complete internal controls.

The primary interest in determining population compositions is in comparing the composition of differing populations (32; 30). These populations are frequently bulk cellular RNA under differing environmental conditions. However, applications to other populations are being developed, such as the analysis of bulk genomic DNA, or the analysis of mass microbial populations (36; 5; 45). Applications of microarray technology span the biological gamut and include: the basic science elucidation of coordinately regulated genes (38); drug discovery (4; 13); discovery of functionally related genes (38; 15); determination of organ-related tissues (9); differential classification of cancers (2);

determination of certain sampling parameters (temporal ordering) in uncategorized samples (28); detection of chromosomal abnormalities (36; 5); surveying bacterial clones with random knock-out mutations for the gene locus or loci affected by the mutation[1]; and even screening for bioterrorism agents (45). It is expected to become a critically important tool in both clinical diagnostics as well as basic biological and health-sciences research (5; 24).

Populations are uniquely labeled and hybridized to the immobilized probes. Initial data processing involves the reduction of each scanned slide image into a table of values related to the label counts for each spot. Typically these include raw values, as well as the results of a number of statistical analysis of the spots themselves. Each spot is typically subsampled and internal statistics calculated on the sample values to detect sub-spot imperfections in the substrate and spot-morphology defects that might affect the results (16).

The primary form of information resolved as a result of this initial processing is a series of ratios detailing the (probable) relative abundance of the target for each probe, in each population (8). After ratios are calculated between the inferred experimental and control expression levels (or other differential effectors), simple questions may be asked, such as "What genes appear to be up-regulated in the experimental group as compared to the Control?". This is commonly expressed in terms of "fold change" between the experiment and control. There are however, significant difficulties due to the character of the data, with producing an unambiguous answer to even such a simple query (12; 26).

The next, and more interesting level of analysis is to ask the question of what expression behaviors are similar, or otherwise appear to be related between a set of experiments. Though some argue that there are fundamental flaws in attempting to use this information to infer global regulatory networks (10), this question is being asked, and potential ways of answering it are being proposed in the literature at an increasing rate. At its simplest, the question is asked with respect to the replicates in an experiment themselves (23; 43, and unpublished results). This is an implicit admittance of understanding that the reproducibility of results is sometimes insufficient for what should be replicate experiments to be statistically detectable as similar. More interesting applications include the detection of individual genes with behaviors that are similar across a varied set of environmental conditions or tissue types, or of groups of genes that have similar patterns of behavior across some set of conditions or timepoints in a time series (15).

A number of common statistical methods have been applied to the discovery of microarray data clusters. Agglomerative hierarchical clustering (14) is one of the most common. K-means clustering (18), and Self-Organising Maps (SOMs) (25) are also popular. Singular Value Decomposition (1) is applied to generate truncated descriptions of the entire sample/expression-level matrix and perform clustering in this lossily-compressed space (19). Their success validates the potential for capturing biologically relevant features with a dramatically simplified description. An even more severe data reduction is successfully applied when they decrease the expression details to a binary description of fully expressed or not expressed for each gene (29).

Many of the proposed applications of clustering techniques to microarray data explicitly note that there is a significant problem with the clusters that unbiased statistical methods predict, in that the clusters frequently have statistical but not biological relevance (39; 31; 3). Keogh et al obliquely recognize this when they observe that a large percentage of all statistical clustering studies are critically flawed due to the size of the variance in the data (20). Despite these observations however, few attempt to incorporate biological biases into the methodology, opting instead to investigate the performance of different unbiased methods.

Recently, the possibility of additional complexity on the time axis has been admitted in some queries, resulting in the question of groups of genes that have similar behaviors for an environmental condition, across multiple timepoints. For example timeseries data is searched for repetitive patterns to detect periodically expressed genes (44). These questions lead to additional complexities with respect to pattern scaling and shifting.

Many of the methods for matching (and therefore calculating distance measures between) timeseries datasets make use of some sort of lossy compression method for the data, and actually perform clustering in this reduced data space. For example in (17) maxima and minima are extracted from the timeseries data, and each timeseries is described as a vector of maxima and minima values. Keogh et al reduce the data to a series of linear approximations of the actual signal (22) and in a later study to a bitstring (binary vector) (21).

We therefore propose a data reduction method that reduces microarray results into a compact form that can be used directly for clustering and for specifying search patterns against existing clusters. While any known clustering method can be applied on this reduced description, we further propose a clustering algorithm. It explicitly contains certain relevant biological biases regarding the manner in which gene expression levels may cluster.

3. Discretization of Gene Expression Data

Trimming raw data: Minor details in the raw data should be filtered out because non-biological variability exists in

---

1   Personal Correspondance

microarray experiment results. Many factors may contribute to such non-biological variability, including differences in the process for obtaining and storing samples; differences in experimental practices; differences in adjustment of equipment, and so on (11). In addition, there is unavoidable random behavior inherent in any biochemical process. We want to identify the overall movement and the points where significant changes have occurred. This can be accomplished by discretizing the data. The user may set a threshold that corresponds to a significant change between the consecutive time-points. This threshold also reflects the confidence in measurement accuracy. Any change below that threshold is considered to be negligible.

Looking at negative correlation: An important point to take into account is that negative correlation between two patterns is not zero correlation between the patterns. Negative regulatory effects exist as well as positive regulatory effects, and so negative correlations can be a clue to members of regulatory networks that are completely overlooked by methods that can only cluster like-signed trajectories. Tools using distance metrics such as Euclidean distance or inner-product techniques that could take correlation values as measures of similarity do not take this into consideration. Because we do not want to exclude negative regulatory events from consideration, we require the capability to recognize similarity in expression patterns even when expression level changes are inversely related. As a solution, we use absolute values of change in expression levels. For our algorithm this means that any change in expression level above the positive threshold or below the negative threshold will both be recorded as a "1" (implying "there is significant change") and any change between the positive and negative thresholds will be recorded as a "0" (implying "there is no significant change").

In a sense, this is a classification algorithm with bins for each possible combination of change, or no-change at a timepoint - that is, a maximum of $2^{(\#timepoints)}$ bins. In theory, there will be no more than N populated bins, where N is the total number of genes under consideration. In practice, there are many fewer, as some genes typically display identical patterns in terms of their expression level changes, and many genes in most experiments display no interesting changes in expression, and such transform into a zero vector.

At this point we have a binary vectors description for each expression pattern, some of which may be (and probably are) identical. The next step in our algorithm is to further cluster these vectors based on an agglomerative method that combines nearest neighbors based on a Longest-Common-Subsequence-Length (LCSL)distance measure(40).

Scan through the data, look at each time point and get the statistics.
T = Threshold for a change between timepoints.
Std = Standard Deviation of Time Point 0
Mul = Number to multiply Std //user defined
$Std = standard\ deviation.$
$T = (std * Mul)$
$P = TimePoint$
Ch = Change between each time point
Find the ones that have changed by given threshold in the first time point
For (each gene)
{
    Ch= P(i)[gene]-P(i-1)[gene]
    If ($|Ch| > T$)
        append value 1
    Else append value 0
}
//Output: A binary vector accumulated for each gene by appending 1s and 0s
Treat the accumulated vector as a bit-string encoding a base-10 number and sort in base-10
//Allows for an efficient, already implemented sort.
//The bit-strings with the same base-10 value are the ones that match exactly in terms of pattern.

Generation of the bitstring representation of the expression level time series.

---

4. Threshold Clustering

Our clustering algorithm, Threshold Clustering, TC, is a bottom-up approach. The merging criterion is defined by the user as a similarity value. This threshold defines what the accepted minimum inner-cluster similarity, as calculated by LCSL is. When two candidate clusters are considered for a merge, the distance from all elements in a cluster to all elements of the other is measured. If there is no element violating the threshold then the two clusters are merged into one. This guarantees that everything in the cluster is at least as similar as the given similarity threshold. It should be noted that this is costlier than simply comparing the distances between the centers of clusters. Comparing center similarity unfortunately leads to error propagation. As new elements are added into a cluster, the center shifts and this might lead in later iterations to addition of elements that aren't similar to initial entries in the cluster, which causes further shifts and so on.

4.1. Similarity Distance Function

In order to cluster similar time-series patterns together we need to define a similarity metric. Some of the patterns that we wish to capture are the result of regulatory cascades, wherein the product of one gene influences the rate of transcription for a second, the product of the second influences the rate of transcription for a third, and so on. These interactions are in fact more interesting from the standpoint of inferring global regulatory networks than simple first-order interactions that might be detected by traditional clustering methods.

To capture the similarity even in the presence of shifts, we utilize the Longest Common Subsequence as our similarity metric. This allows us to cluster patterns even if there is a "shift effect". Given a pair of timeseries that are identical except for a temporal offset, none of the popular metrics such as Manhattan, or Euclidean or other geometry-based distances would capture the similarity. The LCSL distance measure is also tolerant of scattered single-bit errors, which are very common due to the nature of microarray experiments.

4.2. Clustering Algorithm

Our clustering algorithm is incrementally agglomerative, and builds clusters by successively combining subclusters. Rather than asking the user for a target number of clusters to create, we ask for a similarity threshold, and combine subclusters until none can be grouped under that threshold. Candidate clusters to possibly be merged are determined based on the distance calculated between the cluster medians. Expression patterns in candidate clusters that pass the threshold cutoff for consideration are then iterated over to determine whether all members of the potentially combinable clusters are consistent with the merge.

To initialize the process, distances are calculated between all patterns in the initial discretized dataset. The individual patterns are then treated as single-pattern clusters and the algorithm enters a recursive stage where nearest-neighbor clusters are merged until there are no more nearest neighbors with medians that are closer than the threshold cutoff. Because the algorithm describes each cluster by its median, which is a real pattern from the input dataset, and the inter-pattern distances are all precalculated, inter-cluster distances are automatically available without further calculation at each step of the recursion.

To facilitate cluster joining, we maintain a "merge candidate list" (MCL) of clusters that have similarity at or above the given threshold (ex: if the desired in-cluster similarity is 80%, the list will only hold candidate clusters that have 80% or higher similarity between their medians). Cluster medians are assigned such that they have the maximum similarity to the rest of the cluster. During each comparison between clusters on the MCL we rigorously check the distance between every pattern pair to determine if the distance breaks the threshold limit. If there are members of the candidate clusters with distances greater than the threshold allows, the clusters are not merged, and are removed from the MCL. If the distances between all patterns in the merge candidates are consistent with the merge, every distance regarding the two merging clusters is removed from the MCL, and new distances are entered based on the new cluster. The MCL is sorted such that clusters with minimum distances to each other are on the top of the list as candidates for the next recursive iteration. The process recurses until there are no candidate clusters remaining in the MCL.

The algorithm below ensures that each cluster holds only the genes that are at least similar by given threshold, if not more similar.

As expected the number of resulting clusters does decrease when the constraint (similarity threshold) is relaxed (Figure 1).


```
While(MCL.size() > 0)
{
    MCL.pop_front(candidate_clusters)
    Compare each gene in candidate_clusters;
    If (if all patterns within threshold)
    {
        merge
        calculate the new median
        change the size of total number of clus-
        ters
        update MCL
        MCL.sort()
    }
    Else abort the merge
}
```

TC Algorithm

---

5. Experiments and Biological Findings

To validate our algorithm we applied it to the Stanford Yeast Database (37). Many expected findings in this dataset have been characterized by researchers. We observe in our results many similarities to previously observed relationships in the expression data, but due to the significant difference in our distance metric and the discretization applied, we observe a number of differences as well. Because we treat microarray analysis as a decision support process to indicate possible areas which are worthy of more directed
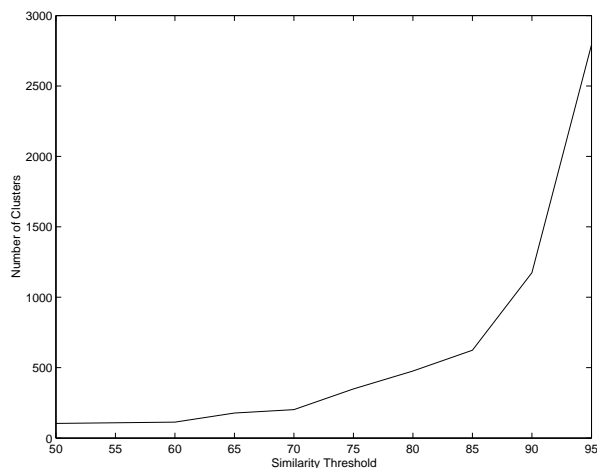
Figure 1. Similarity vs. number of clusters in The Stanford Yeast Dataset with 24 time-points based on cdc-15 experiments

biological examination, these differences provide potential clues to relationships overlooked by other clustering methods.

5.1. Biological Findings on Haemophilus influenzae

Our initial goal was to develop a novel tool for investigating $Haemophilus\ influenzae$, a major cause of otitis media in children, using microarray expression data. Utilizing the proposed techniques we are able to identify several signal pathways that initiate competence development. The microarray experiments on which this research is based were performed in the Columbus Children's Research Institute Microarray core. Here we summarize our findings on this using the proposed techniques.

As a result of the proposed time series analysis, a significant number of genes (240) were determined to be transcriptionally regulated in a *cya* mutant strain of *H. influenzae* grown in minimal culture medium upon the addition of cAMP. This represents approximately 15% of the transcriptome. In these controlled studies, we were able to identify four unique genes clusters, three of which contain CRP-cAMP regulated genes that have not previously been reported. The fact that some of CRP-cAMP regulated genes are parsed into separate clusters is consistent with the notion that additional factors such as PurR modulate the expression of some CRP-cAMP regulated genes (Macfadyen et al., 2001). Indeed, based on the fact that there are several identifiable expression behaviors (clusters) suggests that there are probably additional factors other than PurR and CRP involved in competence. Most sig-

nificantly, one cluster containing nine elements (HI0098, HI0099, HI0109, HI0113, HI0251, HI0252, HI1432, and HI1564) was generated based on a late-time point expression change. Four of these genes encode proteins involved in iron transport. HI0098 and HI0099, encode ABC iron transport proteins. Genomic evidence, including gene proximity, direction of transcription, and intergenic spacing between HI0098 and HI0099, indicates that these genes are part of the same transcriptional unit (operon). Likewise, genes HI0251 and HI0252 are apparently part of a separate and distinct operon, and encode proteins involved in iron transport. Both operons lack identifiable CRP regulatory sequence elements, yet appear to be regulated by CRP-cAMP. This indicates that these operons, and perhaps also the other genes found in this same cluster, might be part of a yet uncharacterized regulon.

5.2. Findings on Yeast Microarray

We have tested our algorithm on the extensively studied Stanford Yeast Database (38; 37). As we mentioned earlier, popular distances applied in microarray studies like Euclidean and Manhattan Distances do not allow certain correlations to be detected. Our Distance function based on LCSL suggests many interesting patterns others reject. For example in Figures 2, 3 and 4 we see a similar pattern with a phase difference, in Figure 2 YOR074C ("*CDC21*") is repeating the pattern that YBR202W ("*CDC47*") has followed, and these two are suggested to be biologically correlated in the literature (33). Similarly the genes in Figure 3 have a similarity with phase and scale difference, which is a clue for a causal relation, and they are also shown to be related biologically. Another example where TC algorithm finds a correlation is between *PH085* and *PCL2* whose co-regulation is known biologically (42).

Another case where hierarchical clustering misses, but the TC detects the similarity is depicted in Figure 5. In this case YDR159W(*SAC3*) and YPL031C(*PH085*) are following similar patterns after the third time point. Other algorithms reject this pair because of the scale difference and possibly the level of apparent noise. Microarray clustering should be an error tolerant process. It should tolerate both biological variance and non-biological noise as pointed out in section 3.

In Figure 6 we see two groups of expression patterns with similar early and late behavior, that follow opposite patterns in mid time points. Both negative and positive correlation maybe important in recognizing gene regulation.
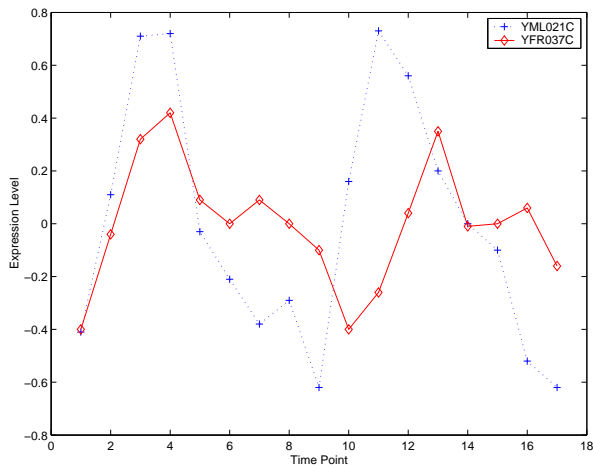
Figure 7 contains three genes, that don't appear correlated visually because of the shifts and scales. Typical algorithms don't cluster them together, while our algorithm locates them in the same cluster. Yeast GRID (6) suggests
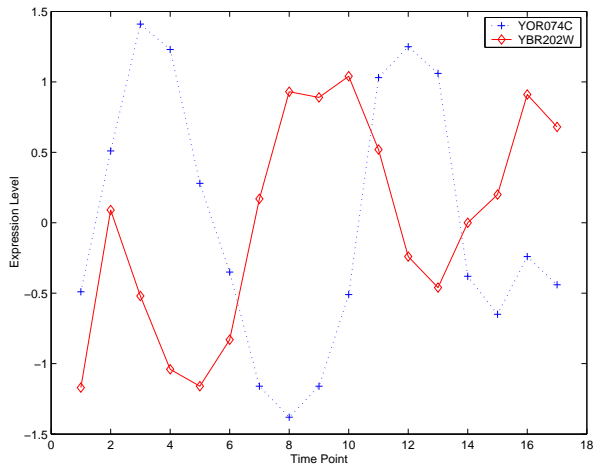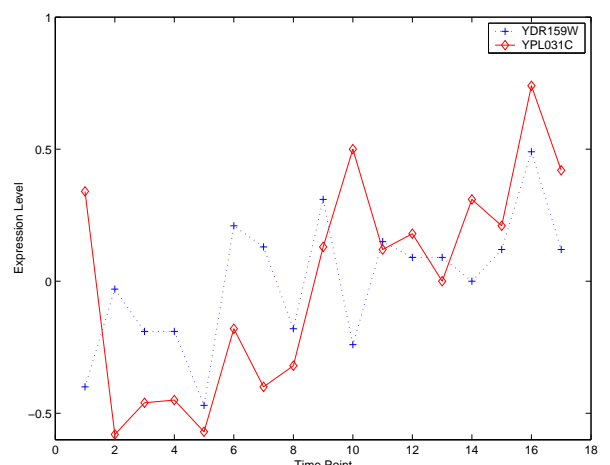
Figure 2. Expression levels of YOR074C and YBR202W
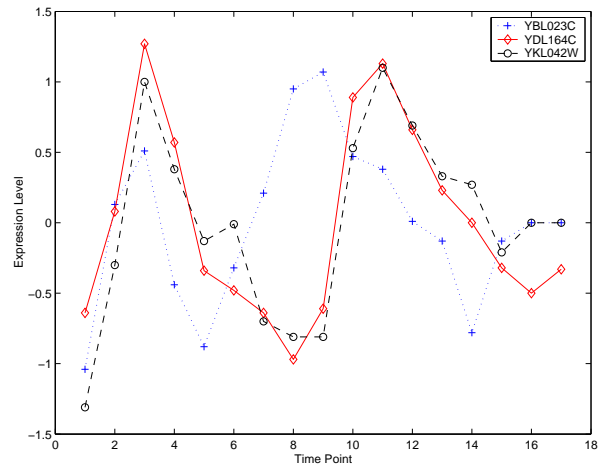
Figure 4. Expression levels of YBL023C,YDL164C and YKL042W

**Figure 3. Expression levels of YML021C and YFR037C**

Figure 5. Expression levels of YDR159W and YPL031C

these three genes are involved in protein translation and translocation.

6. Conclusion and Future Work

We proposed an effective process for time series analysis of microarray data and applied it to the specific problem of *Haemophilus influenzae* competence regulation. Since tolerance for error and biological variance is important when working on microarray experiment results, we adapted Longest Common Subsequence Length distance for this purpose. As expected, it detects distant similarities in sequences with non-matching regions or shifts. The data reduction technique that we have ap-

plied to the expression ratios, before applying our clustering algorithm, also helps in ignoring noise. TC has been shown to have several advantages over other algorithms. First, it does not force the data into an arbitrary and non-biologically related number of clusters. This is a significant advantage over algorithms that subdivide data into a predefined number of clusters regardless of the actual patterns found. In our algorithm, the clustering is decided based on the fact that, everything in a cluster is at least as similar as the given similarity threshold, which is a natural question for a researcher in biological domains. Since our algorithm is a bottom up approach, it is similar to Hierarchical Clustering. However in hierarchical clustering there is only a relative notion of in-class sim-
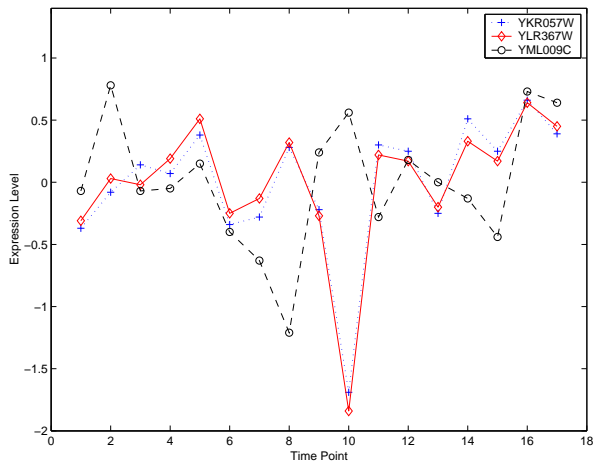
Figure 6. Expression levels of YKR057W,YLR367W and YML009C
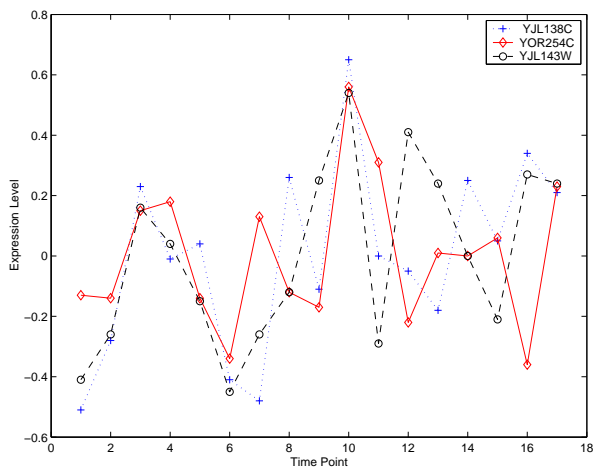
Figure 7. Expression levels of YJL138C,YOR254C and YJL143W

ilarity, whereas our algorithm guarantees an objective and measurable similarity. An examination of published experimental findings indicates that many clusters that were found by TC, and not predicted by other clustering algorithms, have been verified by *in vivo* experiments. Our findings suggest that in *H. influenzae* there are a significant number of CRP-regulated genes that have insufficient promoter and regulator homology to known members of the CRP regulon to be detected by sequence searches. Ongoing experiments in CCRI's microarray core lab have verified a number of these predictions, demonstrating the applicability of our algorithm to detecting non-trivial regulatory relationships.

7. Acknowledgements

References

[1] O. Alter, P. O. Brown, and D. Botstein, *Singular value decomposition for genome-wide expression data processing and modeling*, Proceedings of the National Academy of Sciences U S A **97** (2000), 10101–10106.

[2] Arindam Bhattacharjee, William G. Richards, Jane Staunton, Cheng Li, Stefano Monti, Priya Vasa, Christine Ladd, Javad Beheshti, Raphael Bueno, Michael Gillette, Massimo Loda, Griffin Weber, Eugene J. Mark, Eric S. Lander, Wing Wong, Bruce E. Johnson, Todd R. Golub, David J. Sugarbaker, and Matthew Meyerson, *Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses*, Proceedings of the National Academy of Sciences U S A **98** (2001), no. 4, 13790–13795.

[3] David R. Bickel, *Robust cluster analysis of microarray gene expression data with the number of clusters determined biologically*, Bioinformatics **19** (2003), no. 7, 818–824.

[4] S. Braxton and T. Bedillion, *The integration of microarray information in the drug development process*, Current Opinion in Biotechnology **9** (1998), no. 6, 643–649.

[5] C. J. Breen, Lynn Barton, Aiveen Carey, Adam Dunlop, Mary Glancy, Keara Hall, Anne Marie Hegarty, M. Tariq Khokhar, Monica Powe, Karen Ryan, Andrew J. Green, and Raymond L. Stallings, *Applications of comparative genomic hybridisation in constitutional chromosome studies*, Journal of Medical Genetics **36** (1999), 511–517.

[6] Bobby-Joe Breitkreutz, Chris Stark, and Mike Tyers, *yeast grid*.

[7] MS Chandler, *The gene encoding camp receptor protein is required for competence development in haemophilus influenzae rd*, Procedings of National Academy of Sciences **5** (1992), no. 89, 16261630.

[8] Yidong Chen, Vishnu Kamat, Edward r. Dougherty, Michael L. Bittner, Paul S. Meltzer, and Jeffrey M. Trent, *Ratio statistics of gene expression levels and applications to microarray data analysis*, Bioinformatics **18** (2002), no. 9, 1207–1215.

[9] Yangrae Cho, John Fernandes, Soo-Hwan Kim, and Virginia Walbot, *Gene-expression profile comparisons distinguish seven organs of maize*, Genome Biology **3** (2002), no. 9, research0045.1–0045.16.

[10] Tianjiao Chu, Clark Glymour, Richard Scheines, and Peter Spirtes, *A statistical problem for inference to regulatory structure from associations of gene expression measurements with microarrays*, Bioinformatics **19** (2003), no. 9, 1147–1152.

[11] Terence Critchlow and Zo Lacroix (Editors), *Bioinformatics - managing scientific data*, Morgan Kaufmann, 2003.

[12] Xiangqin Cui and Gary A. Churchill, *Statistical tests for differential expression in cdna microarray experiments*, Genome Biology **4** (2003), no. 4, 210.

[13] C. Debouck and P. N. Goodfelloe, *DNA microarrays in drug discovery and development*, Nature Genetics **21** (1999), no. 1 Suppl, 48–50.

[14] R. O. Duda and P. E. Hart, *Pattern classification and scene analysis*, Wiley, New York, NY, 1973.

[15] Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, and David Botstein, *Cluster analysis and display of genome-wide expression patterns*, Proceedings of the National Academy of Sciences U S A **995** (1998), no. 25, 14863–14868.

[16] M. R. Fielden, R. G. Halgren, E. Dere, and T. R. Zacharewski, *GP3: GenePix post-processing program for automated analysis of raw microarray data*, Bioinformatics **18** (2002), no. 5, 771–773.

[17] Eugene Fink and Kevin B. Pratt, *Indexing of compressed time series*.

[18] Edward W. Forgy, *Cluster analysis of multivariate data: Efficiency versus interpretability of classifications*, Biometrics **21** (1965), 768.

[19] D. Horn and I. Axel, *Novel clustering algorithm for microarray expression data in a truncated svd space*, Bioinformatics **19** (2003), no. 9, 1110–1115.

[20] E. Keogh and S. Kasetty, *On the need for time series data mining benchmarks: A survey and empirical demonstration*, Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining., 2002.

[21] E. J. Keogh and M. Pazzani, *An indexing scheme for fast similarity search in large time series databases*, 11th International Conference on Scientific and Statistical Database Management, SSDBM'99 (Cleveland, OH), IEEE Computer Society, 1999, pp. 56–67.

[22] Eamonn Keogh and M. Pazzani, *An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback*, Fourth International Conference on Knowledge Discovery and Data Mining (KDD'98) (New York City, NY) (R. Agrawal, P. Stolorz, and G. Piatetsky-Shapiro, eds.), ACM Press, 1998, pp. 239–241.

[23] M. K. Kerr, M. Martin, and G. A. Churchill, *Analysis of variance for gene expression microarray data*, Journal Comput Biol **7** (2000), no. 6, 819–837.

[24] J. Khan, M. L. Bittner, Y. Chen, P. S. Meltzer, and J. M. Trent, *Dna microarray technology: the anticipated impact on the study of human disease*, Biochim Biophys Acta **1423** (1999), no. 2, M17–28.

[25] T. Kohnen, *Self-organized formation of topologically correct feature maps*, Biological Cybernetics **43** (1982), 59–69.

[26] William J. Lemon, Sandya Liyanarachchi, and Ming You, *A high performance test of differential gene expression for oligonucleotide arrays*, Genome Biology **4** (2003), no. 10, R67.1–67.11.

[27] Leah P. Macfadyan, *Regulation of competence development in haemophilus influenzae proposed competence regulatory elements are crp-binding sites*, J. theor. Biol. (2000), no. 207, 1–11.

[28] Paul M. Magwene, Paul Lizardi, and Junhyong Kim, *Reconstructing the temporal ordering of biological samples using microarray data*, Bioinformatics **19** (2003), no. 7, 842–850.

[29] Hiroyuki Nakahara, Shin ichi Nishimura, Masato Inoue, Gen Hori, and Shun ichi Amari, *Gene interaction in dna microarray data is decomposed by information geometric measure*, Bioinformatics **19** (2003), no. 9, 1124–1131.

[30] Wei Pan, *A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments*, Bioinformatics **18** (2002), no. 4, 546–554.

[31] Jie Qin, Darrin P. Lewis, and William Stafford Noble, *Kernel hierarchical gene clustering from microarray expression data*, Bioinformatics **19** (2003), no. 16, 2097–2104.

[32] Daniel R. Rhodes, Jeremy C. Miller, Brian B. Haab, and Kyle A. Furge, *CIT: identification of differentially expressed clusters of genes from microarray data*, Bioinformatics **18** (2002), no. 1, 205–206.

[33] WU Jia Rui, *Regulation of eukaryotic dna replication and nuclear structure*, Cell Research **9** (1999), no. 3, 163–170.

[34] M. Schena, D. Schalon, R. W. Davis, and P. O. Brown, *Quantitative monitoring of gene expression patterns with a complementary DNA microarray*, Science **270** (1995), no. 5235, 467–70.

[35] M. Schena, D. Shalon, R. Heller, A. Chai, P. O. Brown, and R. W. Davis, *Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes*, Proceedings of the National Academy of Sciences U S A **93** (1996), no. 20, 1061410619.

[36] A. M. Snijders, N. Nowak, R. Segraves, S. Blackwood, N. Brown, J. Conroy, G. Hamilton, A. K. Hindle, B. Huey, K. Kimura, S. Law, K. Myambo, J. Palmer, B. Ylstra, J. P. Yue, J. w. Gray, A. N. Jain, D. Pinkel, and D. G. Albertson, *Assembly of microarrays for genome-wide measurement of dna copy number*, Nature Genetics **29** (2001), no. 3, 263–264.

[37] Paul T. Spellman, Gavin Sherlock, Michael Q. Zhang, Vishwanath R. Iyer, Kirk Anders, Michael B. Eisen, Patrick O. Brown, David Botstein, and Bruce Futcher, *Yeast cell cycle analysis project*.

[38] Paul T. Spellman, Gavin Sherlock, Michael Q. Zhang, Vishwanath, R. Iyer, Kirk Anders, Michael B. Eisen, Patrick O. Brown, David Botstein, and Bruce Futcher, *Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization*, Molecular Biology of the Cell **9** (1998), 3273–3297.

[39] M. Sultan, D. A. Wigle, C. A. Cumbaa, M. Maziarz, J. Glasgow, M. S. Tsao, and I. Jurisica, *Binary tree-structured vector quantization approach to clustering and visualizing microarray data*, Bioinformatics **18** (2002), no. sup1, s111–s119.

[40] Ronald L. Riverst Thomas H. Cormen, Charles E. Leiserson and Clifford Stein, *Introduction to algorithms*, McGraw-Hill Book Company., 2001.

[41] Nancy Trun and Janine Trempy, *Fundamental bacterial genetics*, Blackwell Publishing., 2003.

[42] Measday V., McBride H., Moffat J., Stillman D., and Andrews B, *Interactions between pho85 cyclin-dependent kinase complexes and the swi5 transcription factor in budding yeast.*, Molecular Microbiolgy **35** (2000), no. 4, 825–834.

[43] L. Wernisch, S. L. Kendall, S. Soneji, A. Wietzorrek, T. Parish, J. Hinds, P. D. Butcher, and N. G. Stoker, *Analysis of whole-genome microarray replicates using mixed models*, Bioinformatics **19** (2003), no. 1, 53–61.

[44] Sofia Wichert, Konstantinos Fokianos, and Korbinian Strimmer, *Identifying periodically expressed transcripts in microarray time series data*, Bioinformatics **20** (2004), no. 1, 5–20.

[45] W. J. Wilson, C. L. Strout, T. Z. DeSantis, J. L. Stilwell, A. V. Carrano, and G. L. Andersen, *Sequence-specific identification of 18 pathogenic microorganisms using microarray technology*, Molecular and Cellular Probes **16** (2002), no. 2, 119–127.