

Efficient Determination of Cluster Boundaries for Analysis of Gene Expression Profile Data Using Hierarchical Clustering and Wavelet Transform

Harry Amri Moesa¹

hammus00@yahoo.com

Dukka Bahadur K.C.²

dukka@kuicr.kyoto-u.ac.jp

Tatsuya Akutsu²

takutsu@kuicr.kyoto-u.ac.jp

¹ NEC Soft Ltd, Platform System Division, Tokyo, Japan

² Bioinformatics Center, Institute for Chemical Research, Kyoto University, Kyoto 611-0011, Japan

Abstract

The existing methods for clustering of gene expression profile data either require manual inspection and other biological knowledge or require some cut-off value which can not be directly calculated from the given data set. Thus, the problem of systematic and efficient determination of cluster boundaries of clusters in gene expression profile data still remains demanding.

In this context, we have developed a procedure for automatic and systematic determination of the boundaries of clusters in the hierarchical clustering of gene expression data based on the ratio of with-in class variance and between-class variance, which can be fully calculated from the given expression data. After the determination of dendrogram based on agglomerative hierarchical clustering, this ratio is used to determine the cluster boundary. Except this ratio which can be completely calculated from the given expression profile data, unlike other existing approaches, our approach does not require any manual inspection or biological knowledge. Our results are favorably comparable and in some of cases better than existing method which does not utilize prior information or manual inspection. Moreover, gene expression profile data are often contaminated with various type of noise and in order to reduce this noise content, we have also applied image enhancing technique called discrete wavelet transform. We tested a number of mother wavelet functions to smooth the noise in the gene expression data set and obtained some improvements in the quality of the results.

Keywords: Gene expression analysis, micro-array technology, hierarchical clustering, group average method, image enhancing, wavelet transform

1 Introduction

With the advent of technologies for genomic scale data generation, monitoring gene expression levels in different developmental stages and biochemical conditions of every gene in an organism is possible. Especially, a global and simultaneous observation on the expression levels of thousands of genes can be easily obtained. A large amount of gene expression profile data has been accumulated by various research groups [3, 13] which has changed the paradigm of research from ‘one gene’ to ‘all genes’. This monitoring of large scale gene expression levels in different organisms can help us in characterizing gene function, gene networks, chemical pathways and eventually help in medical treatments. A natural and usual approach for analysis of the gene expression data profile is concerned about the detection of similar groups. Hence, the initial step in analyzing gene expression data is to group together genes of similar expression patterns. Hence, this problem of characterization of groups of genes is the clustering problem in the field of computer science.

Several approaches have been proposed for detecting similar patterns of gene expressions [2, 15]. Hierarchical clustering is very often used in the clustering of gene expression profiles. This is due to

the fact that the dendrogram enables one to visually understand the clustering of genes. However, in most of the cluster analysis approaches either manual inspection or biological knowledge is required to determine the number of clusters. Hence, the approach which can systematically determine the number of clusters from the gene expression profile itself is highly awaited.

Although, there exists a number of existing works [2, 15] for the clustering of gene expression profile data, to our knowledge, the only approach which does not require visual inspection or biochemical knowledge is the one by Horimoto *et al.* [8]. However, this method uses the variance inflation factor (VIF) in the multiple regression analysis in order to evaluate the cluster boundaries and thus, a certain cut off value ($= 10$) has to be applied in order to evaluate the cluster boundaries which cannot be calculated from the given gene expression data. Hence, with the motive of developing a novel technique which does not use any other knowledge except the given input data for the characterization of cluster boundaries, we have undertaken this present research.

On one hand, this large scale genomic expression tools like microarrays are helpful in the large scale understanding of gene function and gene networks. Nevertheless, the vast amount of gene expression profile data that is produced by these experiments is associated with a substantial amount of noise. This is due to the fact that distinguishing noise (false positive results) from the real expression data is a tedious and difficult task. It is obvious to think of reducing this noise in order to increase the efficiency of prediction.

Hence, in this paper we present an efficient and systematic determination of cluster boundaries based on the ratio of with-in class variance and between-class variance. We have compared the results of our approach with some of the existing approaches and our approach is comparable or better than the existing approaches which do not require manual inspection [8] or *a priori* to determine the cluster boundaries.

Moreover, in order to reduce the noise content in the expression data, we have also applied an image enhancement technique, widely used in the field of image processing, called discrete wavelet transform before the clustering procedure to smooth the noise. We have tested three different types of mother wavelet functions viz. Daubechies D4 wavelet, Haar mother wavelet and Symlet mother wavelet. The data enhancement by wavelet transform yielded better results for time series data which has periodicity. The details of the method and results are given in the sections below.

2 Materials and Methods

The aim of this research is to determine the cluster boundaries in the dendrogram of the gene expression profile data. Hence, our approach is comprised of two main parts: the dendrogram construction part which yields a dendrogram and the subsequent part which deals with the systematic and efficient determination of cluster boundaries of the obtained dendrogram.

2.1 Dendrogram Construction

We do not describe in detail the clustering algorithm and the distance metric used in this approach. Interested readers are requested to refer to a book [6]. In this approach, agglomerative hierarchical clustering is utilized.

The distance metric used for this research is based on Pearson Correlation Coefficients given by:

$$r_{is} = \frac{\sum_{k=1}^p (q_{ik} - \bar{q}_i)(q_{sk} - \bar{q}_s)}{\sqrt{\sum_{k=1}^p (q_{ik} - \bar{q}_i)^2 \sum_{k=1}^p (q_{sk} - \bar{q}_s)^2}} \quad (1)$$

where r_{is} is the Pearson Correlation Coefficient between data i and data s of the expression profiles measured at p points, q_{ik} ($k = 1, \dots, p$) is the k th vector element of data i , the data vector \mathbf{q}_i and the arithmetic average \bar{q}_i of q_{ik} over p points given as:

$$\mathbf{q}_i = \{q_{i1}, \dots, q_{ip}\} \quad (2)$$

and

$$\bar{\mathbf{q}}_i = \frac{1}{p} \sum_{k=1}^p q_{ik} \quad (3)$$

Hence, the final form of the distance metric between data i and data j in our work is given by

$$d_{ij} = \sqrt{\sum_{s=1}^N (r_{is} - r_{js})^2} \quad (4)$$

where N is the total number of genes and r_{xy} is the Pearson Correlation Coefficient between data x and data y . The lesser the distance between the two genes, the more similar the corresponding genes are in terms of their expression patterns.

In essence, the clustering algorithm works in the following steps:

1. Construct a list of clusters considering the data belonging to each gene as a separate cluster.
2. Repeat the process from (a) to (c) until all the data is included in a single cluster.
 - (a) Calculate the distance using the distance metric using equation (4) for all the pairs of the clusters in the list.
 - (b) Select the pair of clusters with the minimum distance and go on merging to get a larger cluster.
 - (c) Remove the clusters selected in the above step from the list.

In real computation, when C_k is the cluster obtained after combining cluster C_i and C_j then the expression of distance between the cluster k and the rest of the clusters C_l is given by

$$d_{kl} = \frac{d_{il}|C_i| + d_{jl}|C_j|}{|C_i| + |C_j|} \quad (5)$$

where $|C_x|$ is the number of members of cluster x .

The result of this clustering step is a dendrogram and this dendrogram is just the candidate set of the final clusters. So, in order to get the final clusters, one needs to divide this dendrogram into several sub-trees where each sub-tree represents a cluster and the leaves of the sub-tree become the member of that cluster.

In the determination of cluster boundaries, often the visual inspection which require intense prior knowledge is utilized. As the motive of our research is to propose an approach which does not require any prior-knowledge, the ratio of within-class variance and between-class variance, which can be fully calculated without any prior knowledge is utilized. In the field of statistics, between-class variance and with-in class variance are some of the statistical parameters which are often used in order to analyze the quality of the results of clustering [11].

2.2 Determination of Cluster Boundary

Clustering deals with partitioning N data into M number of clusters, however, determination of the value of M is a difficult task. Generally, some other standard is required to determine the clustering of the data. The fundamental property of all the data belonging to a cluster is that they share high similarity with each other in the same cluster and they are highly dissimilar with data in other clusters.

Hence, in this work the ratio of between-class variance and with-in class variance is used as the criteria for merging of clusters.

In other words, for the clustering to be successful, the with-in class variance (intra-cluster variance) should be as small as possible and in the same time the between-class variance (inter-cluster variance) should be as large as possible. The ratio of between class variance to within class variance is a class separability criterion in multiple discriminant analysis [4] and is called Fisher's ratio. Although, this ratio is a very popular criterion in pattern recognition, to our knowledge it has not been yet applied to the clustering of gene expression data. We use the inverse of Fisher's ratio in our work.

As the task of dividing the dendrogram complements to the task of determining clusters, for the cluster discrimination, S , the ratio of with-in class variance S_W and the between-class variance S_B is utilized which is defined as:

$$S = \frac{S_W}{S_B} \quad (6)$$

where S_W , the within-class variance is defined as:

$$S_W = \frac{1}{M} \sum_{f \in F} \sum_{t, j \in f} (d_{tj} - \bar{d}_f)^2 \quad (7)$$

where F is the set of clusters, f is one of the member of the set, d_{tj} is the distance between the member t and member j of the cluster, \bar{d}_f is the average distance between all the members of cluster f where, \bar{d}_f and the total number of data in the cluster M are defined as:

$$\bar{d}_f = \frac{1}{M_f} \sum_{t, j \in f} d_{tj}, \quad (8)$$

$$M = \sum_{f \in F} M_f, M_f = |f| \quad (9)$$

Similarly, for F , f , M , M_f as in equation (7) and the average distance between all the members \bar{d} given as

$$\bar{d} = \frac{1}{M} \sum_{f \in F} \sum_{t, j \in F} d_{tj}, \quad (10)$$

the between-class variance S_B is given by:

$$S_B = \frac{1}{M} \sum_{f \in F} M_f (\bar{d}_f - \bar{d})^2 \quad (11)$$

Hence, dividing the dendrogram into sub-trees corresponds to the merging of clusters such that the ratio S becomes small. The sub-trees of the dendrogram correspond to the clusters and the nodes of the tree correspond to the root of each sub-tree, hence in order to get the best possible clustering, it is required to calculate the ratio S for all possible combinations of nodes of the dendrogram and select the nodes with the minimum S as clusters. For the dendrogram of a large scale data like gene expression profile data, this procedure is computationally intense. Hence, following procedure is undertaken in order to reduce the computational complexity.

1. Two lists, named *temporary* and *result* are generated.

2. The root node of the tree is inserted in the S , *temporary* list.
3. The following steps are performed until the *temporary* list becomes null.
 - (a) Considering the first node of the *temporary* list as the parent node par , the ratio S_{par} is calculated.
 - (b) For the left children $left$ and right children $right$ of par , S_{left} and S_{right} are calculated.
 - (c) S_{left} and S_{right} is compared with S_{par} and if $S_{left}(S_{right})$
 - i. is smaller than S_{par} , then $left$ ($right$) is inserted in the end of the *temporary* list.
 - ii. is larger than S_{par} , then $left$ ($right$) is inserted in the *result* list.
4. Output each of the nodes in the *result* list as clusters, where the members of the clusters are the leaves of the sub-trees with these nodes as the root node.

Hence, instead of computing the ratio S for all possible combinations of nodes, only the ratio of the node and two of its children is required to be calculated, thus reducing the computational complexity of the problem. The performance of the approach is elucidated in the computational experiments section.

3 Discrete Wavelet Transform as Pre-Processing

Gene expression profile data is often contaminated with noise consisting of biological noise, experimental noise or image processing noise. Hence, we also tried to enhance the quality of the gene profile data using discrete wavelet transform [12] before the clustering step so as to increase the efficiency.

Besides, Klevecz [9] has found the time-series nature of the data in the dynamic architecture of the yeast cell cycle using wavelet decomposition. Continuous wavelet transform consumes significant amount of time and resources, hence, discrete wavelet transform which is easy to implement and reduces the computational time and resources is chosen for this work. It can be concluded that for the gene expression data with periodicity, performing wavelet transform in the pre-processing step would enhance the predictivity of the method. Thus, wavelet transform was applied to the data. In the next section, a brief description of wavelet transform and its application in our method is presented.

3.1 Discrete Wavelet Transform

The main advantage of wavelet transform is that unlike a sine wave as in Fourier transformation, wavelets have their energy concentrated in time i.e. wavelets are more suited for the analysis of transient, time varying signals as gene expression profile. Details about discrete wavelet transform is omitted here. Interested readers are requested to refer to [7, 12, 14] for the details. Moreover, a lot of studies has been done in Bioinformatics using wavelet transform [10].

A wavelet expansion is similar in form to the well-known Fourier series expansion with two parameter family of functions,

$$f(t) = \sum_k \sum_j a_{j,k} \psi_{j,k}(t), \quad (12)$$

where j and k are integers and the function $\psi_{j,k}(t)$ is the wavelet expansion function. Wavelet expansion function usually forms an orthogonal basis. The two-parameter expansion coefficients $a_{j,k}$ are called discrete wavelet transform (DWT) coefficients of $f(t)$. The coefficients are given by:

$$a_{j,k} = \int f(t) \psi_{j,k}(t) dt. \quad (13)$$

In real computations, function f and wavelet function ψ are given in discrete form. Hence, three set of vectors is required. The first set of vectors are the observed data, i.e. the vector \mathbf{F} of signal $f(t)$. If we consider signal $f(t)$ to be composed of p discrete data, then the observed data vector \mathbf{F} can be represented as:

$$\mathbf{F} = [f_1, \dots, f_p]. \quad (14)$$

The second set of vectors are the wavelet vectors \mathbf{W}_i . These are the vectors corresponding to wavelet transform function ψ in equation (12) and can be represented as:

$$\mathbf{W}_i = [\psi_{i,1}, \dots, \psi_{i,N}], \quad (15)$$

where it is considered that for all i the wavelet vectors $\psi_{i,j}$ for $j = 1, \dots, N$ are given.

Finally, the third set of vectors are the coefficient vectors which correspond to the discrete wavelet transform coefficients of $f(t)$. Hence, $a_{j,k}$ in equation (13) is calculated in the form of D_i as:

$$\mathbf{D}_i = [a_{i,1}, \dots, a_{i,K}]. \quad (16)$$

For the lateral shift(translation) of $k = 1, \dots, K$, the vector \mathbf{D}_i can be expressed in the similar manner as \mathbf{W}_i as:

$$\mathbf{D}_i = [d_1^{(i)}, \dots, d_K^{(i)}]. \quad (17)$$

Furthermore, the wavelet transform at compression parameter i can be calculated by convoluting discrete function \mathbf{F} with \mathbf{W}_i as:

$$\mathbf{D}_i = \mathbf{F} * \mathbf{W}_i.$$

If we introduce the wavelet coefficient matrix \mathbf{D} and wavelet matrix \mathbf{W} as:

$$\begin{aligned} \mathbf{D} &= [\mathbf{D}_1^t, \dots, \mathbf{D}_j^t]^t, \\ \mathbf{W} &= [\mathbf{W}_1^t, \dots, \mathbf{W}_j^t]^t, \end{aligned} \quad (18)$$

where $[\cdot]^t$ represents the transpose of the matrix, the final form of the discrete wavelet transform(DWT) is reduced to:

$$\mathbf{D} = \mathbf{F}\mathbf{W}. \quad (19)$$

For our experiments, three different types of mother wavelet function i.e. Daubechies d4 wavelet, Haar mother wavelet and Symlet mother wavelet were utilized [16].

In order to retain all the information, we did not filter out the coefficients. Hence, the input to the clustering in the computational experiments is the wavelet coefficient matrix itself. The reason for this comes from our preliminary observation of the results when some filtering functions were utilized. The data obtained after the DWT is then processed to get the dendrogram as described in Section 2 and then estimation of boundaries is performed.

4 Computational Experiments

The data set for the validation of the proposed algorithm is the data set first used in the work of Eisen [5]. This data set is obtained from <http://genome-www.stanford.edu/clustering>.

This data set consists of data obtained from DNA microarrays with elements representing nearly all the ORFs from the fully sequenced *S. Cerevisiae* genome and all the measurements were made against a time 0 reference sample except for cell-cycle experiments. While generating the data, all genes(2467), for which functional annotation was available in the *Saccharomyces* Genome Database, were included.

These data were driven from time courses during the following processes: the cell division cycle after synchronization by alpha factor arrest(ALPHA: 18 time points); centrifugal elutriation(ELU: 14 time points), and with a temperature-sensitive *cdc15* mutant (CDC15-15time points); sporulation(SPO, 7 time points plus four additional samples); shock by high temperature(HT, 6 time points); reducing agents(D, 4 time points) and the diauxic shift (DX, 7 time points). In Figure 2 of [5], it has been shown that the clustering produced nine representative clusters containing functionally related genes involved in various functions. The functional category is depicted in Table 1.

Table 1: Functional Classes of Genes from Eisen *et al.*

Category	Function
I	Spindle pole body assembly and function
II	Proteasome
III	mRNA splicing
IV	Glycolysis
V	Mitochondrial ribosome
VI	ATP synthesis
VII	Chromatin structure
VIII	DNA replication
IX	Tricarboxylic acid cycle and respiration

4.1 Results without Pre-Processing

We performed computational experiments of our approach without pre-processing and compared our results with the results of Eisen [5] and the results of Horimoto *et al.* [8]. The method of Horimoto *et al.* uses Variance Inflation Factor (VIF) for the determination of cluster boundaries. The category column under Clusters in Table 2 represents the functional category as mentioned in Table 1, the Num. column shows the number of members in each category as obtained by Eisen *et al.* [5]. Similarly, the Num. column under the VIF shows the cluster number for each functional group as in category and the Inc. shows the number of genes predicted by the methods of Horimoto *et al.* [8]. Finally, the Num. column under the BIR column and Inc. shows the cluster number for the respective functional groups and the number of genes included in each clusters obtained by using our approach.

It can be observed from Table 2 that our approach performs equally well as the method of Horimoto *et al.* and in some cases it even performs better. Especially, in the case of functional group V (Mitochondrial ribosome) and functional group IX (Tricarboxylic acid cycle and respiration) all the members are allocated to cluster V by our method where as only 20 members and 14 members are allocated by the method of Horimoto *et al.* respectively. Moreover, in the case of functional group III our method allocates 12 members whereas the method of Horimoto *et al.* only allocates 11 members. However, in the case of functional group VI the method of Horimoto *et al.* allocates 13 members whereas our method only allocates 9 members. For all other functional groups, the number of members

Table 2: Comparison of results

Clusters		VIF		BIR	
Category	Num.	Num.	Inc.	Num.	Inc.
I	11	7	11	14	11
II	27	10	25	15	25
III	14	28	11	9	12
IV	17	30	17	1	17
V	22	31	20	12	22
VI	15	31	13	12	9
VII	8	11	8	18	8
VIII	5	9	5	17	5
IX	16	22	14	16	16

allocated to each functional group by the method of Horimoto *et al.* and our method is the same. It is interesting to note here that unlike the methods of Eisen *et al.* [5] where the functional categories V and VI are assigned to different clusters, our method and the method of Horimoto *et al.* [8] assign both of them to the same cluster. Moreover, it has been elucidated by various experimental studies that the expression patterns of these two functional groups are very related. So, we think this result to be consistent with the experimental results. Prominently, it has to be noted here that unlike the methods of Horimoto *et al.* which requires some input values for determination of cluster boundaries, our approach does not require any *a priori*. Hence, it can be concluded that our method performs comparably well and in some cases performs better than the method of Horimoto *et al.* [8].

4.2 Results with Wavelet Transform as Pre-Processing

We also performed clustering of the gene expression profile data after the wavelet transform and the results of the computational experiments are shown in Table 3. In order to compare the boundary identification method without preprocessing, we also include the results of the clustering with our approach without preprocessing.(represented by BIR column in Table 3)

Table 3: Comparison of results using different mother wavelets

Cat.	#	BIR	DAUB4	HAAR	SY
I	11	11	11	11	11
II	27	25	25	25	26
III	14	12	14	13	14
IV	17	17	17	17	17
V	22	22	22	22	21
VI	15	9	14	11	9
VII	8	8	8	8	8
VIII	5	5	5	5	5
IX	16	16	16	8	10

The Cat. in the Table 3 shows the functional category, # shows the number of genes in each category obtained by Eisen *et al.* [5], BIR shows the number genes predicted by BIR (our approach without using wavelet transform), DAUB4 shows the results using Daubechies mother wavelet function,

HAAR represents the results using Haar mother wavelet function, and SY represents the results using Symlet mother wavelet function. From Table 3 it can be observed that the number of members allocated in the functional group III increases from 12 of BIR to 14 in case of DAUB4, 13 in case of HAAR and 12 in case of SY. The number of members decreases in case of functional group VI in case of wavelet-clustered methods. Although, Symlet mother wavelet does not perform well in the overall, it performs well in the case of functional group II.

Finally, it can be concluded that all the three mother wavelet function perform equally well. Although, no significant gain in the quality of the result is obtained using the wavelet transform for all functional groups, it can be seen that for function group III the number of cluster has increased. For other clusters, the numbers are almost the same.

5 Discussion and Future Works

This approach utilizes information only from the given data itself to determine the boundary of the clusters. Unlike the existing approaches which require some biological knowledge or some manual inspection, this approach uses only the ratio of within-class variance and between-class variance which can be fully calculated from the gene expression data itself, in order to determine the boundary of the clusters. Although, this work utilizes agglomerative hierarchical clustering approach along with the group-average method for the clustering, any distance measure or any clustering techniques can be adopted in this approach.

From the results, it can be noticed that our approach yields equally good results for all the nine functional category except the one for ATP synthesis in which case the method by Horimoto *et al.* produced better results. Prominently, it has to be noted here that unlike the methods of Horimoto *et al.* which requires some input values for determination of cluster boundaries, our approach does not require any *a priori*.

Gene expression profiles produced by different technologies are contaminated with errors. In order to reduce these errors, discrete wavelet transform was applied before the clustering step. In the case of wavelet-based clustering although, the number of members allocated for each functional category did not increase that much, it can be inferred that the clustering method produces almost the best results, so there is very little space for improvement. Especially in the case of functional category III, the original number of members from the work of Eisen [5], it can be seen that the total number is 14 whereas the clustering only allocated 12 members. So there was a place for the improvement and the wavelet based clustering performed well.

Furthermore, by using Daubechies the results were better than the clustering without wavelet pre-processing which suggests that we have achieved improvement in the quality of the results by using Daubechies mother wavelet. So, it can be said that the clustering using Daubechies mother wavelet as a pre-processing step enhances the quality of the results.

However, for proteasome functional group, glycolysis functional category, ATP synthesis category and DNA replication category the results of the hierarchical clustering without pre-processing by wavelet transform yielded better results than the clustering with pre-processing by wavelet transform.

An immediate future work could be the comparison of cluster numbers and number of genes included in each cluster produced by the combination of various clustering techniques and distance metrics. In conclusion, it can be said that by applying wavelet transform the number of allotted genes increases in the clusters. This is in consistent with the results of Klevecz *et al.* [9] which shows that the CDC 15, ALPHA, and ELU data are the time-series data and the wavelet transform helped in extracting this periodicity. Although, deep insight into these biological functional group may reveal some biological reasons, from the image processing point of view, these bad results for some functional groups can be due to the threshold value applied in order to distinguish noise from the real data. Another future work could be the selection of a better threshold value for discriminating noise with real data.

References

- [1] Bakshi, B.R., Multiscale analysis and modeling using wavelets, *Journal of Chemometrics*, 13:415–434, 1999.
- [2] Ben-Dor, A., Shamir, R., and Yakhini, Z., Clustering gene expression patterns, *J. Comp. Biol.*, 6:281–297, 1999.
- [3] DeRisi, J., Iyer, V., and Brown, P., Exploring the metabolic genetic control of gene expression on a genomic scale, *Science*, 278:680–686, 1997.
- [4] Duda, R. and Hart, P., *Pattern Classification and Scene Analysis*, John Wiley & Sons, 1973.
- [5] Eisen, M. B., Spellman, P. T., Brown, P.O., and Botstein, D. , Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. USA*, 95:14863–14868, 1998.
- [6] Everitt B. *Cluster Analysis*, Edward Arnold, London, third edition, 1993.
- [7] Gonzalez, R.C. and Woods, Richard E., *Digital Image Processing*, Prentice Hall, Second Edition, 2002.
- [8] Horimoto, T. and Toh, H., Statistical estimation of cluster boundaries in gene expression data, *Bioinformatics*, 17:1143–1151, 2001.
- [9] Klevecz, R.R., Dynamic architecture of the yeast cell cycle uncovered by wavelet decomposition, *Funct. Integr. Genomics*, 1:186–192, 2000.
- [10] Lilo, P., Wavelets in bioinformatics and computational biology: State of art and perspectives, *Bioinformatics*, 19:2–9, 2003.
- [11] Otsu, N., Optimal linear and nonlinear solutions for least-square discriminant feature extraction, *Proc. 6th Int. Con. on Pattern Recognition*, 557–560, 1982.
- [12] Shensa, M.J., Discrete wavelet transform: Wedding the a trous and Mallat algorithms, *IEEE, Trans. Signal Processing*, 40(2):464–482, 1992.
- [13] Spellman, P. T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Ander, K., Eisen, M.B., Brown, P.O., Botstein, D., and Futcher, B., Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization, *Mol. Biol. Cell*, 9:3273–3297, 1998.
- [14] Weeks, M. and Bayoumi M, Discrete Wavelet transform: Architectures, design and performance issues, *J. VLSI Signal Proc. Syst.*, 35(2):155–178, 2003.
- [15] Yeung, K.Y., Haynor, D.R., and Ruzzo, W.L., Validating clustering for gene expression data, *Bioinformatics*, 17:309–318, 2001.
- [16] <http://www.mathworks.com/access/helpdesk/help/toolbox/wavelet/wavelet.shtml>