

Clustering Time-Varying Gene Expression Profiles using Scale-space Signals

Tanveer Syeda-Mahmood
IBM Almaden Research Center
650 Harry Road, San Jose 95120
stf@almaden.ibm.com

Abstract

The functional state of an organism is determined largely by the pattern of expression of its genes. The analysis of gene expression data from gene chips has primarily revolved around clustering and classification of the data using machine learning techniques based on the intensity of expression alone with the time-varying pattern mostly ignored. In this paper, we present a pattern recognition-based approach to capturing similarity by finding salient changes in the time-varying expression patterns of genes. Such changes can give clues about important events, such as gene regulation by cell-cycle phases, or even signal the onset of a disease. Specifically, we observe that dis-similarity between time series is revealed by the sharp twists and bends produced in a higher-dimensional curve formed from the constituent signals. Scale-space analysis is used to detect the sharp twists and turns and their relative strength with respect to the component signals is estimated to form a shape similarity measure between time profiles. A clustering algorithm is presented to cluster gene profiles using the scale-space distance as a similarity metric. Multi-dimensional curves formed from time series within clusters are used as cluster prototypes or indexes to the gene expression database, and are used to retrieve the functionally similar genes to a query gene profile. Extensive comparison of clustering using scale-space distance in comparison to traditional Euclidean distance is presented on the yeast genome database.

1 Introduction

The analysis of time-varying nature of biological processes is becoming increasingly important. Common biological processes unfold at different rates and may show salient changes at time points that can signal important events. The events of DNA replication, chromosome segregation, and mitosis define a fundamental synchrony of genes in the eukaryotic cell cycle[2]. An analysis of such synchrony in gene profiles may reveal the cell cycle-based regulatory properties of genes. This can help identify fluctuating functional relationship between genes as has been observed for temporal expression patterns of genes in developing spinal cord [10]. By clustering similarly-varying gene profiles, co-regulated and functionally similar genes can be identified. Functions of newly discovered genes can be inferred from similarity to prototypes of existing gene clusters. Thus clustering and retrieving matching time profiles can be an important step in functional genomics and biological process characterization. In fact, developing robust methods of clustering and retrieving matching time series is an important problem, occurring in a number of other applications including economic forecasting, medical signal processing, and for recognizing actions described through temporal trajectories.

The current technique for measuring gene expression uses a gene chip in which target genes to be tracked are laid out on a substrate (silicon or glass), and the sample whose genes are to be characterized is poured as a probe over the chip. The corresponding genes present in the samples hybridize(i.e. form base pairs) with their respective counterparts, and this expression is recorded using a confocal microscope by fluorescent labeling the samples. Thus gene chips allow the expression of thousands of genes to be measured simultaneously. Using gene chips, we can conduct precise experiments to measure the expression of genes under a variety of conditions, including disease vs. healthy tissues, and to deduce co-regulation of genes. In the latter, the aim is to infer functional relationships between genes through clustering of expression patterns produced within consecutive cell-cycles of an organism.

The resulting data produced in this case consists of expression patterns of genes as a time series (after suitable data normalization and preprocessing).

In this paper we introduce a pattern recognition approach to clustering time-varying expression profiles to characterize functionally similar genes. Specifically, we regard expression patterns as curves, and capture the similarity of gene profiles by comparing the individual gene profiles to the multi-dimensional curve formed from the combined profiles. In particular, we note that additional sharp bends and twists are introduced in the multi-d profile in cases where the individual gene expression profiles exhibit lack of synchrony in their variations. Sharp changes in time profiles are characterized by a scale-space decomposition of the gene curves. This forms the basis of a similarity metric to compare different sets of gene profiles. The multi-dimensional curve also serves as a compact representation of clusters so that matching time profiles to a query gene can be easily determined.

The novelty of our contribution lies in the development of the scale-space distance metric, and the application of such pattern recognition techniques to the characterization of functional similarity of gene profiles. Previous approaches for finding functionally similar genes have primarily relied on the intensity of expression rather than its time-varying pattern. Although the clustering of time series is posed in the context of gene expression matching, it is clear that these techniques are also relevant for clustering time series in other application domains where data can be abstracted as a time series.

The rest of the paper is organized as follows. In Section 2 we motivate the need for scale-space distance for characterizing similarity in gene expression time series and review related work. In Section 3, we present the scale-space distance metric. In Section 4, we present an approach to clustering gene expression data using scale-space distance. Finally, in Section 5, we present results showing the effectiveness of the metric for the clustering of gene expression data from the yeast data set, where all the cell-cycle regulated genes have been identified.

2 Motivation

Consider the gene expression profiles of three genes shown in Figure 1 a-c. The data shown here is the expression of the gene recorded over two cell-cycles. The cell cycle can be divided into phases, called G1, G2, S and M phases, each corresponding to some step during the mitosis (cell-division) process. As we can see, the expression of different genes change in different phases of the cell cycle, signaling respective events. One way to characterize genes involved in a related function is to group those genes whose time series show change at similar time positions. Since the expression value is relative to a control, and is often corrupted with noise during gene chip scanning, it is better to rely on the change and the extent of change rather than signal correlation based on the absolute value of the expression.

The predominant approach in the case of gene expression data has been to cluster and classify the data based on the intensity of the expression pattern alone. Lately, a few approaches have emerged to analyze the time-varying pattern in gene expression, adapting ideas from previous work in the analysis of time series in economic forecasting[3], database searching [1], and signal processing [5], and spline interpolation [4]. The common approaches to time series clustering can be classified into three groups, based on projecting the time series data as (a) points in multi-dimensional space formed from the time instants (most common method available in data mining software), (b) into a space of distance-preserving transforms such as the Fourier transform[1], and (c) using probabilistic models, eg. HMMs[7]. Projecting the time series data as points in multi-d space and using the Euclidean distance for metric is equivalent to characterizing the similarity between time series using a mean-square error metric. Since direct intensity values are used in this analysis, clustering is dominated by the intensity levels rather than the pattern of expression of the signal. Figure 4 shows three gene expression profiles. While the similarity between the series of ORF (open reading frame) 1 and 3 can be seen, using the mean square error, we actually conclude that the expression profiles of 2 and 3 are similar (see row 2 of Table 1). A number of approaches to comparing time series are available in the database searching literature including the use of first few coefficients of the Fourier series[1], characterizing important features of the time series such as the maxima and minima [8], and slopes[1] and conversion of curve fragments to symbols. These approaches are rather heuristic in nature, and cannot characterize similarity in time series over long time periods (when the signal is non-stationary). Finally, the approaches to time series clustering using parametric models such as hidden Markov models work in a supervised fashion (need training data), and can capture overall properties of a stationary signal rather than instantaneous time correspondence.

3 Scale-space distance

In this paper, we take the approach that individual gene expression patterns can be modeled as 2d curves (gene expression vs. time) $(g(t), t)$. Modeling time series as curves helps emphasize the pattern of variation without requiring the normalization of intensity values (both positive and negative intensity values can be handled). If two curves $(g_1(t), t)$ and $(g_2(t), t)$ are similar, a 3d curve formed by projecting the two gene expressions profiles in a 3d space $(g_1(t), g_2(t), t)$ is similar to the two expression profiles. Generalizing this to multiple gene expression profiles, a single multi-dimensional curve can be formed to serve as a prototypical compact description of the underlying cluster/group of genes expression profiles.

A remarkable observation is that such higher-dimensional curves when composed from dis-similar curves, show relatively large amounts of sharp twists, bends and turns over and above the changes present in the component curves. Figure 2b-d shows 3d curves formed from pairs of gene expression profiles shown in Figure 2a. As can be seen, when the component curves are similar (for ORFs 18srRnaa and 18srRnac), their corresponding 3d curve (Figure 2b shows similar changes as in the original signal. On the other hand, when two dis-similar signals are composed (18srRnaa and 18srRnab) as in Figure 2d or profiles 18srRnab and 18srRnac as shown in Figure 2e, the sharp bends and twists are apparent in the 3d curve. In fact, the sharpness of turn is proportional to the mismatch between the two component curves from which the 3d curve is derived. Thus by comparing the sharpness of bends in the 3d curve to the underlying shape of the component curves at corresponding time points, we can obtain a measure of similarity between the two curves. The sharp change points are of particular interest in the context of cell-cycle gene expression data where the change (upward or downward) is more important than the intensity of expression itself.

3.1 Detecting salient change points in curves

We first observe that change points on curves correspond to inflection points i.e. places where there are zero-crossings of the second derivative. Salient change points are those that are preserved even after multiple levels of smoothing. In particular, by successively smoothing a gene profile $f(t)$ using a Gaussian of varying σ , we get the smoothed curve $\hat{C}(t, \sigma)$ as

$$\hat{C}(t, \sigma) = C(t) * G(t, \sigma) = \int_{-\infty}^{\infty} C(u) \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-u)^2}{2\sigma^2}} du \quad (1)$$

If we vary the scale σ , the above smoothed signal is a 2d signal of time and σ . Since we are interested in curvature change points, we consider extrema of slope or inflection points, i.e., points at which there is a zero-crossing in the second derivative as

$$\frac{\partial C^2}{\partial t^2} = 0 \quad (2)$$

To handle multi-dimensional signals, we take the derivative of the convolved multi-dimensional curve, and record the magnitude of the change. The original inflection points can then be recovered from *negative-going* zero-crossings of the second derivative.

Thus if we look for places where there is a change of sign in the second derivative of the signal as a function of scale, the resulting 2d image looks as shown in Figure 3b. Here the zero-crossing contours are the contours of the colored regions. In particular, the negative-going zero-crossings are the contours of the red to blue transition regions (light-to-dark -in-gray image renderings). This is the curvature scale-space as described in [11, 6]. In particular, it can be shown that in the case of Gaussian smoothing, the zero-crossing contours are always closed at the bottom (higher scale) and open at the top (U shaped curves). Also, the zero-crossings shift with increasing scale, so that the time location of a zero-crossing is found by starting from the peak of a contour and tracking the contour down to its finest scale location as described in [11]. The resulting signal is called the *scale-space signal*, and describes the location of sharp change points in the signal. In particular, the intensity at a time point in a scale-space signal is the highest scale at which the change disappears. Thus sharper changes are reflected as high intensity points in the signal. Figure 3c shows the scale-space signal for the curve in Figure 3a.

3.2 Scale-space similarity metric

As we said earlier, if two curves have a similar pattern of variation, the 3d curve formed from the component curves maintains that pattern. Since the scale-space signals can be formed for each curve (2d as well as 3d curves),

the distance between two curves $f_1(t)$, and $f_2(t)$ can be found by the *scale-space distance* as

$$D(f_1, f_2) = \sum_{i=1}^T (I_C(i) - (I_1(i) + I_2(i))/2)^2 \quad (3)$$

for T time points. Here I_1, I_2, I_C are the scale-space signals of the individual and combined curves respectively. This similarity measure remains a metric, since it can be interpreted as the Euclidean distance between the transformed curves in scale-space.

Since the scale-space signals for multi-dimensional curves are also one-dimensional, the above distance metric can be used to compute the similarity between multi-dimensional curves and one-dimensional curves, two multi-dimensional curves, etc. This will be found useful in clustering the gene profiles as discussed next.

4 Clustering genes in scale-space

All clustering algorithms use a notion of distance between the data points. By substituting the scale-space distance in place of the usual (Euclidean) distance between the time series we can use, in principle, any of the existing clustering methods. In this section, we show how K-means, a popular clustering algorithm, can be adapted using the scale-space distance. In particular, the concept of centroids will be replaced by prototypical multi-dimensional curves formed from a group of gene curves. Following the K-means algorithm, we proceed to do the clustering in 3 steps, namely, (1) initial prototype selection, (2) classification of curves, and (3) re-computation of the prototypes. Steps 2 and 3 are repeated until convergence is reached (when the prototypes do not change much). Different methods of initial selection of centroids can be used. Here we use the maximum scale-space distance between a randomly selected gene profile and all other gene profiles, to assemble initial cluster prototypes. That is, K gene curves whose distance to one another gene is greater than $0.9 * \text{maximum scale-space distance}$, are retained as the initial K prototypes. In the classification step, the minimum scale-space distance between a gene curve and all K prototypes is used to assign a gene curve to the corresponding cluster. The multi-dimensional curve formed from the gene curves in a cluster becomes the new prototype for the next iteration. The complexity of the algorithm remains $O(nk)$ for initialization of prototypes, and $O(mnK)$ for m iterations of data classification and $O(mn)$ for re-computation of prototypes.

4.1 Retrieving functionally similar genes

With the gene profiles in a database clustered using the above algorithm, the time-varying profile of a gene of unknown function can now be given as a query and matching genes retrieved. In particular, since the prototypical members serve as indexes to the database, the matching genes are retrieved using a two-stage search in which the nearest prototype is first identified using scale-space distance. The individual gene profiles within the cluster(s) identified are then searched based on scale-space distance again to find matching gene profiles.

5 Results

We now present results to illustrate the use of scale-space signals in capturing similarity of time-varying gene profiles, and the utility of such measure in retrieving functionally similar genes.

A system called Gene Expression Miner was developed to process gene profiles. Given a database of gene profiles, the scale-space signals are derived for each of the curves. Scale-space signals of higher-dimensional curves are formed during the iteration steps of clustering as cluster prototypes are assembled. The result is an indexed database with the prototype curves per cluster serving as indexes. Given a new gene curve as a query, the system retrieves matching prototypes from clusters and lists the constituent genes in a cluster along with links to their associated information in a public database to allow scientists to infer functional similarity of a newly discovered gene.

The database used for experiments is the cell cycle recording the expression of 6600 ORFs (some of which are genes) in the yeast genome. This dataset was assembled by Spellman et al.[9]who had found that several genes of the yeast genome are regulated by different phases of the mitotic cell-cycle using careful lab experiments. The data is available from Stanford (<http://cellcycle-www.stanford.edu>)[2] and depicts 17 time points of expression data for synchronized yeast cells with genes critical to the cell cycle reported in selected ORF (open reading frames).

Our goal in this work was to automate the task of predicting the common cell-cycle regulation of genes through automatic clustering and retrieving of functionally similar genes to a query gene. The candidate similar genes identified are validated from ground truth data recorded for these genes as obtained through wet-lab experiments (eg. gene knockout).

We begin by illustrating the scale-space similarity metric. Figure 4a-c shows three time-varying gene profiles, two of which are co-regulated. Figure 4d-f shows the 3d curves formed from pairing curves 1 & 2, 2 & 3, and 1 & 3 respectively. Figure 4g-l shows the scale-space signals derived from the scale-space images for the respective curves and their pairings. Here the scale-space profile of the combined curve is indicated in green. By comparing the aligned plots of Figure 4j-l with those of the individual curves in Figure 4g-i, it can be seen that the alignment of the zero-crossings both in time location and intensity in scale-space is best in Figure 4k corresponding to the pairing of functionally similar genes 18srRnaa and 18srRnac. The actual scale-space distance between the pairings are shown in Table 1. In comparison, the Euclidean distance between the paired curves is shown in row 2 of this table. Here the Euclidean distance is computed by projecting the individual time series as points in multi-dimensional space. As can be seen, the curves 1 and 3 are judged closer using the scale-space distance but not the Euclidean distance.

Next, we compare clustering using scale-space distance and Euclidean distance for different choices of the number of clusters. The algorithm for initialization of centroids remains the same in both case, using the farthest distance between a pair of genes as the seed distance for cluster separation. The results are tabulated in Table 2. Here Column 1 indicates the choice of K, column 2 & 3 indicates the number of members in corresponding clusters for the 10 largest clusters, and the percentage overlap between the corresponding cluster in Euclidean distance case. The corresponding clusters are based on the highest amount of overlap. From this, we can infer that the two methods produce different cluster distributions. Next, in Column 4 and 5, we list the average intra-cluster compactness for the two cases as the ratio of the average distance to the maximum distance between pairs of curves for the two choices of distance metrics. As expected, the cluster-compactness increases with the number of clusters.

To compare the performance of both metrics for clustering against the ground truth data, we repeated clustering on a smaller data set consisting of 104 cell-cycle regulated genes that have already been manually clustered into functionally similar based biological verification of their cell-cycle co-regulation patterns. These genes are listed in Table 4 Column 3. We ran the clustering algorithms on the reduced data set consisting of 104 genes isolated above. The percentage overlap with the ground truth clusters for the same value of K is recorded for both clustering methods, and is shown in Column 4 and 5 of Table 4. As can be seen, the scale-space distance-based metric is more effective in grouping genes known to be functionally similar.

Finally, we test the performance in retrieving functionally similar genes for chosen gene curve queries. For this, we select a value of K=20 and cluster the entire database of 6600 ORFs. We then use each of the 104 genes identified above as the query, and retrieve top n matches within the nearest cluster identified through the cluster prototype. The value of n was set to the number of genes known in the respective clusters of the ground truth data. The cluster size for each group is indicated in Column 3 of Table 3. The fraction of retrieved ORF that belong to the cluster are used as recall. Precision is recorded by noting the number of matches that need to be retained to ensure that all the class members of a particular cluster are retrieved. The precision and recall were averaged across queries per cluster, and the result is shown in Column 4-5 and 6-7 of Table 3. As can be seen from the table, both precision and recall are higher on the average using scale-space distance in comparison to Euclidean distance during clustering.

6 Conclusions

In this paper, we have introduced a pattern recognition technique based on scale-space characterization of curves to the problem of clustering and retrieving gene expression patterns to infer functional similarity in genes. Unlike techniques that favor dimensionality reduction, we actually promote projecting the curves in a higher-dimensional space to infer the similarity. This is done without an increase in complexity since a higher dimensional projection of curves remains a curve. Further, the use of scale-space for capturing salient changes in expression patterns has been shown to be a more appropriate metric for finding relationships between gene expression patterns in comparison to straight Euclidean distance based on intensity variation of expression. Finally, the scale-space distance can serve as a general metric for comparing time series data arising in a number of different applications.

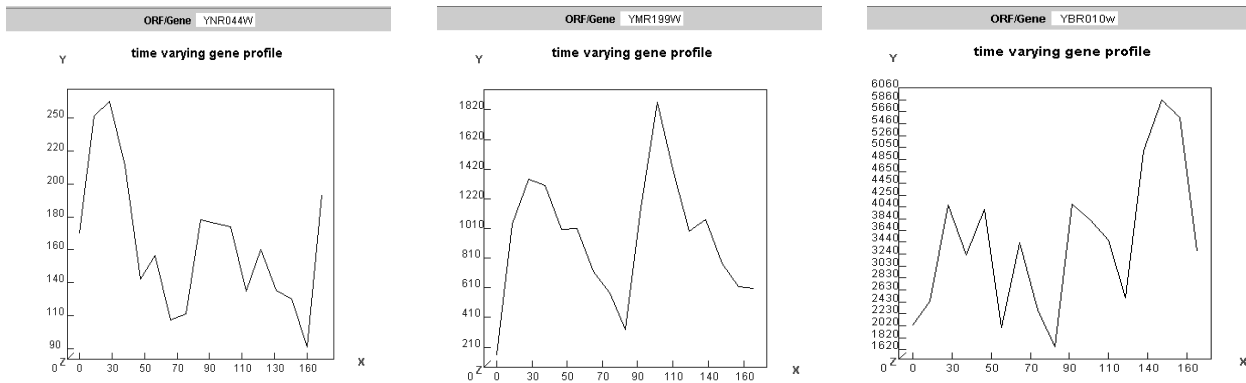


Figure 1: Illustration of time varying expression of genes active in different phases of the mitotic cell cycle.

| Distance Metric | 18srRnaa & 18srRnab | 18srRnaa & 18srRnac | 18srRnab & 18srRnac |
|----------------------|---------------------|---------------------|---------------------|
| scale-space distance | 69.67 | 35.17 | 54.03 |
| Euclidean distance | 17.03 | 45.98 | 24.23 |

Table 1: Illustration of the comparison of scale-space distance and mean square and Euclidean distance for three time-varying expression profiles shown in Figure 4a-c.

References

- [1] R. Agrawal et al. Fast similarity search in the presence of noise, scaling and translation in time series databases. In *Proceedings of Conf. on Very Large Databases*, pages 434–439, 1995.
- [2] R.J. Cho et al. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell.*, 2:65–73, 1998.
- [3] S. Focardi. Clustering economic and financial time series: Exploring the existence of stable correlation conditions. Technical report, Technical Report, The Intertek Group, August 2001.
- [4] Z-B. Joseph, D. Gifford, and T. Jaakkola. A new approach to analyzing gene expression time series. In *Proc. RECOMB*, pages 326–327, 2002.
- [5] Y. Kakizawa et al. Discrimination and clustering of multi-variate time series. *Jl. of American Statistical Association*, 201, 1998.
- [6] T. Lindeberg. Detecting salient blob-like image structures and their scales with a scale-space primal sketch: A method for focus of attention. *International Journal of Computer Vision*, 3(11):283–318, 1993.
- [7] T. Oates et al. Clustering time series with hidden markov models and dynamic time warping. In *Proceedings of the IJCAI-99 Workshop on Reinforcement Learning*, 1999.
- [8] C-S. Perng et al. Landmarks: A new model for similarity-based pattern queries in time series databases. In *Proceedings of the Int. Conf. on Data Engineering*, pages 66–70, 2000.
- [9] P. Spellman et al. Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of Cell*, 9:3273–3297, 1998.
- [10] X. Wen et al. Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl. Acad. Sci.*, pages 334–339, 1998.
- [11] A. Witkin. Scale space filtering: A new approach to multi-scale description. In *Proceedings Int. Joint. Conf. Artif. Intell.*, 1984.

| S.No | # clusters | Scale-space clustering | Euclidean clustering | Cluster Compactness (scale-space) | Cluster compactness (Euclid) |
|------|------------|--|---|-----------------------------------|------------------------------|
| 1. | 20 | 1542,983,878,624,358,213,178,156,110,98,76 | 60%,23%,45%,89%,12%,24%,12%,56%,11%,24% | 0.67 | 0.42 |
| 2. | 40 | 523,276,338,213,138,121,108,76,56,48,36 | 70%,43%,34%,24%,38%,54%,32%,12%,89%,20% | 0.73 | 0.62 |
| 3. | 60 | 358,203,213,192,163,158,136,110,98,76 | 40%,23%,45%,89%,12%,24%,12%,56%,11%,24% | 0.87 | 0.72 |

Table 2: Illustration of the comparison of clustering using scale-space distance and Euclidean distance for the gene expression data base consisting of 6600 ORFs from two cell cycles of budding yeast.

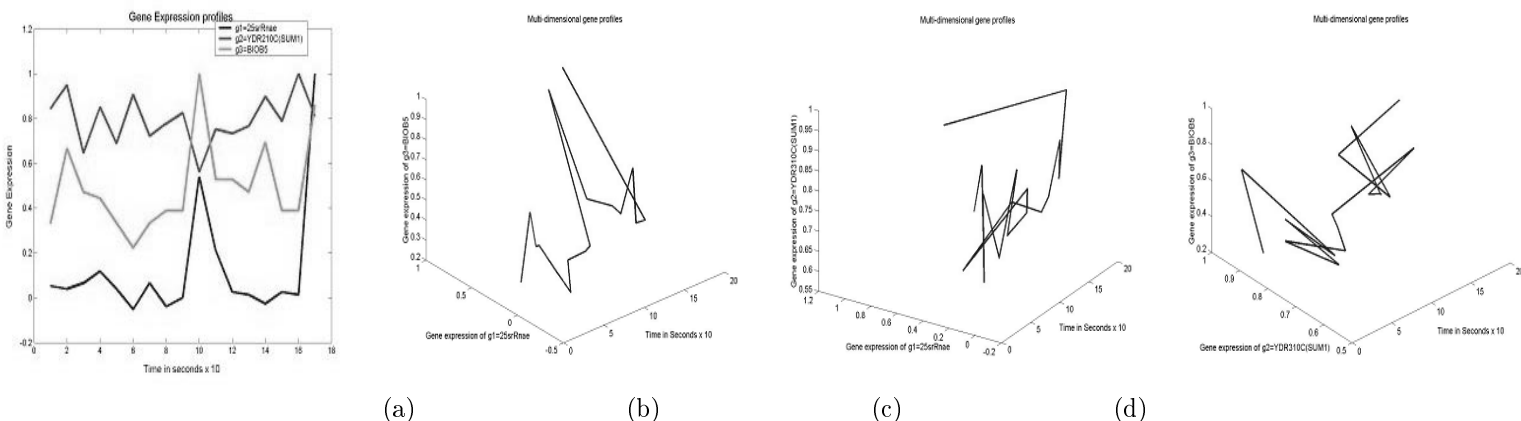


Figure 2: Illustration of higher-dimensional projection of curves to form a shape similarity metric. (a) Expression profiles of three gene profiles 25srRnae (blue), YDR 210C(red) and BIOB5(green). (b) The 3d curve from projecting the curves of 25srRnae and BIOB5 (known to be functionally similar). (c)The 3d curve from projecting the curves of 25srRnae and YDR210C. (d) The 3d curve from projecting the curves of YDR210C and BIOB5.

| S.No | clusters label | Number of genes | Avg. Precision | Avg. Recall | Avg. Precision | Avg. Recall |
|------|------------------------|-----------------|----------------|-------------|----------------|-------------|
| 1. | M/G1 Boundary | 19 | 0.49 | 0.73 | 0.18 | 0.67 |
| 2. | Late G1, SCB regulated | 14 | 0.31 | 0.71 | 0.23 | 0.5 |
| 3. | Late G1, MCB regulated | 39 | 0.5 | 0.76 | 0.54 | 0.67 |
| 4. | S-phase | 8 | 0.41 | 0.78 | 0.25 | 0.62 |
| 5. | S/G2-phase | 9 | 0.49 | 0.82 | 0.21 | 0.81 |
| 6. | G2/M-phase | 15 | 0.51 | 0.76 | 0.45 | 0.12 |

Table 3: Illustration of the comparison of retrieval using scale-space distance and Euclidean distance for the gene expression data base consisting 6600 ORF, using K=20, and querying with the reference ground-truth data set of 104 genes that are known to belong to 6 clusters of varying size. In each case, the avg. precision and recall performance of scale-space distance is better than the Euclidean distance, although the precision values are still worse than the recall values for both measures.

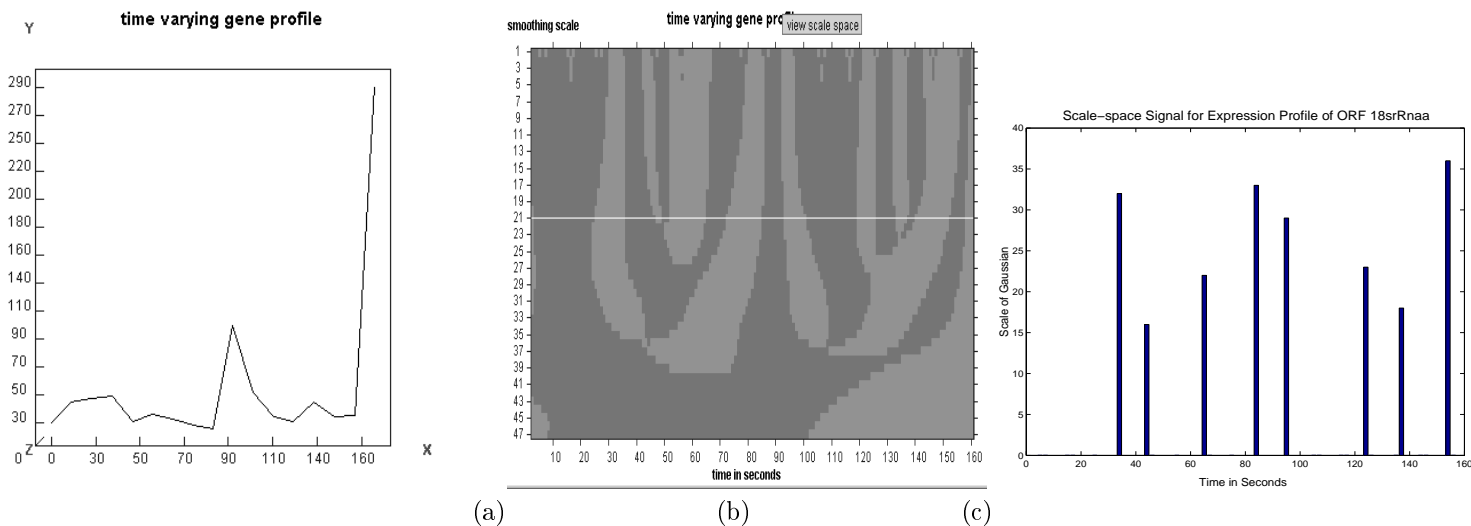


Figure 3: Illustration of scale-space signal computation. (a) Time-varying expression curve for ORF=18srRnaa. (b) Scale-space image reflecting the zero-crossings of the second derivative. Note that the zero-crossings are closed at higher scales (the scale is inverted in the figure). (c) The scale-space signal formed by tracing the peaks of zero-crossings to their lowest scale.

| S.No | clusters | Member | % found | % found |
|------|-------------------------|---|-------------|-----------|
| | Description | genes | scale-space | Euclidean |
| 1. | M/G1 Boundary | AGA1, ASH1, CDC46, CDC47, CDC6, CHS1, CLN3, CTS1, EGT2, FUS1, MFA2, PCL2 ,PCL9, RME1, SIC1, SST2, STE2, SWI4, TEC1 | 89 | 65 |
| 2. | Late G1, SCB regulated | CLN1, CLN2, CSD2/CHS3, FKS1/CWH53, GAS1, HO, KAR4, KRE6, MNN1, PCL1, PSA1, SWE1, TIP1, VAN2/GOG5 | 78 | 56 |
| 3. | Late G1 , MCB regulated | ASF1, ASF2, CDC21, CDC45, CDC8, CDC9, CLB5, CLB6, DBF4, DPB2, DPB3, GIC2, MCD1, MSH2, MSH6, NIK1/HSL1, PDS1, PMS1, POL1, POL12, POL2, POL3/CDC2, POL30, PRI1, PRI2, RAD17, RAD27, RAD51, RAD54, RFA1, RFA2, RFA3, RNR1, RNR3, SPC110/NUF1, SPC42, SPK1, SRS2/HPR5, UNG1 | 72 | 47 |
| 4. | S-phase | Histones: HHT1, HHT2, HHF1, HHF2, HTA1, HTA2, HTB1, HTB2 | 89 | 63 |
| 5. | S/G2-phase | CDC14, CIK1 ,CLB3, CLB4, CWP1, CWP2, KAR3, NUM1, TIR1 | 74 | 65 |
| 6. | G2/M-phase | ACE2, ASE1, CDC20, CDC5, CLB1, CLB2, DBF2, FAR1, KIN3, MOB1, YRO2(MST1), YDR033w(MST2), SED1, SPO12, SWI5 | 88 | 71 |

Table 4: Illustration of the clustering performance of scale-space distance and Euclidean distance-based K-means for the ground truth data set of clustered genes of yeast. The clustering was determined manually based on biological experiments. The cluster labels in Column 2 refer to the various phases in a cell cycle. For example, the S-phase is the cell-division phase. The time-varying profiles of genes active in respective cell cycle phase show sharp changes during the phases. Datasize tested: 104.

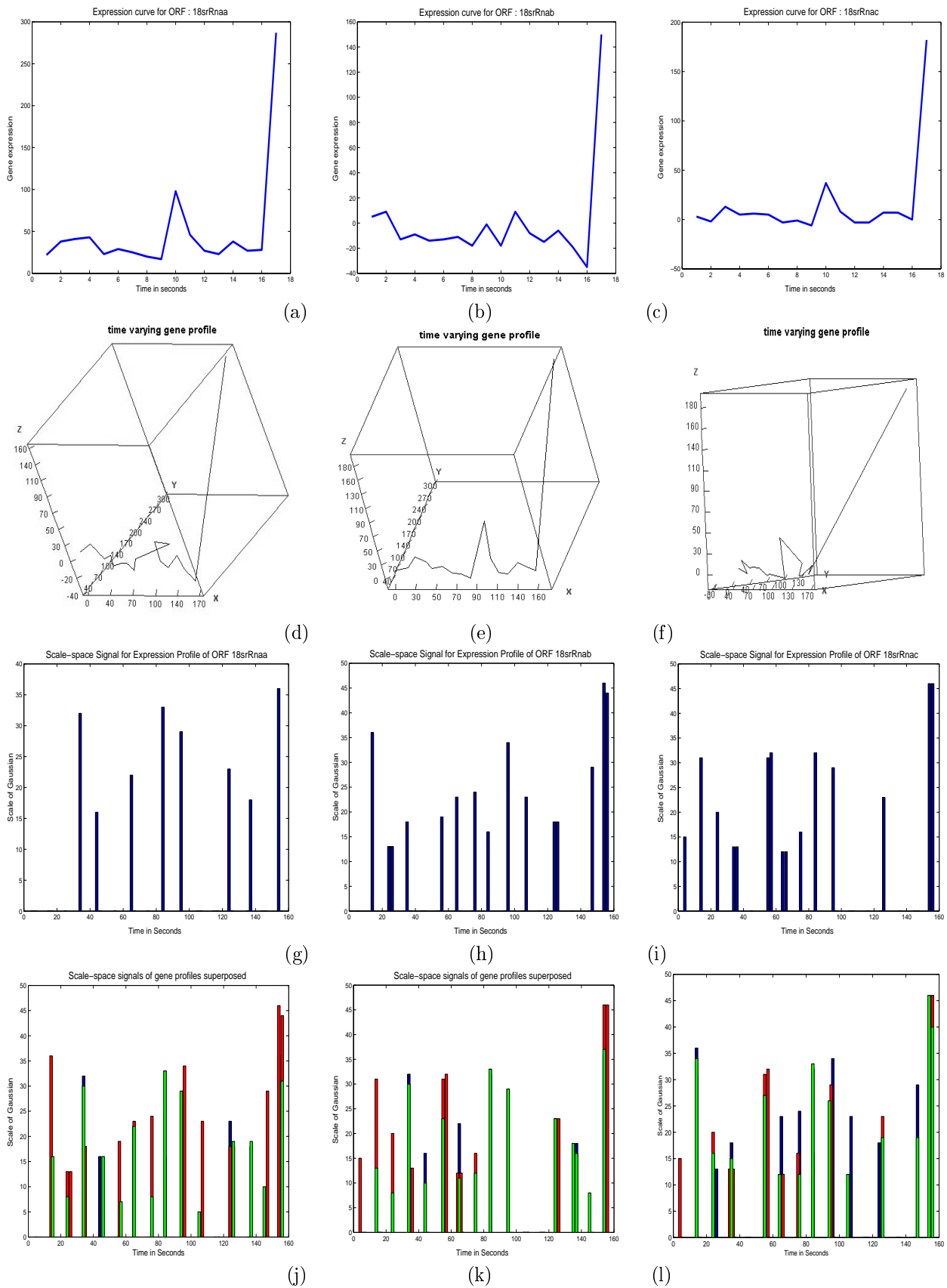


Figure 4: Illustration of scale-space distance metric computation. (a) - (c) Three time-varying expression profiles for ORFs 18srRnaa, 18srRnab, 18srRnac respectively. (d) - (f) 3D curve formed from the high-dimensional projection of curves of (a) & (b), (a) & (c), (b) & (c) respectively. (g) - (i) Scale-space signals for the individual ORFs 18srRnaa, 18srRnab, 18srRnac respectively. (j) - (l) Scale-space signals for the high-dimensional projection of curves of (a) & (b), (a) & (c), (b) & (c) respectively. Note the alignment of zero-crossings at the respective locations in (k) in comparison to the respective profiles. In each case, the scale-space profile of the combined curve is indicated in green.