# DETECTING SALIENT CHANGES IN GENOMIC SIGNALS

*Tanveer Syeda-Mahmood*

IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120. stf@almaden.ibm.com

**ABSTRACT**

The functional state of an organism is determined largely by the pattern of expression of its genes. Salient changes in variation in expression of genes can give clues about important events, such as the onset of a disease. In this paper, we address the problem of detecting salient inflection points in genomic signals. These are detected using a scale-space decomposition, to be the negative-going zero-crossings of the second derivative of smoothed signal that are preserved over an automatically selected scale. The utility of salient change detection is demonstrated in the automatic identification of regulatory phase for genes active in the mitotic cell cyle of budding yeast.

## 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. For instance, different individuals affected by a common disease may progress at varying rates with salient changes signaled at different times. Similarly, the events of DNA replication, chromosome segregation, and mitosis define a fundamental periodicity in the eukaryotic cell cycle[1]. An analysis of such periodicity in gene profiles using salient change detection 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 [3]. Thus the analysis of time-varying profiles, and in particular, methods for detecting salient change points in time-varying profiles can be an important step in functional genomics and biological process characterization.

In this paper, we present a novel method of salient change detection in genomic signals. Specifically, we regard gene profiles as curves, and characterize salient changes as those arising from salient inflection points in a multi-dimensional curve formed from individual gene profiles. Robust detection of inflection points is achieved using a scale-space representation[4]. Automatic scale selection is applied to derive a scale threshold for determining the change points.

## 2. Characterizing salient changes in gene expression

Consider Figure 1a that shows the time-varying profile of a signal, in this case, the average velocity curve of a moving object. Although the signal rises and falls at many time points, perceptually, the change at time point marked 1 in Figure 1c appears to ”pop out” or be salient. This is also borne out in the actual physical experiment where this constitutes a major change in direction of movement of the object and is perceived as a distinctly different action. Similarly, in the time-varying gene expression profile of gene YLR304C shown in Figure 1b, the change at time point marked 1 in Figure 1d can be seen to be significant. From these examples, it appears that both the amplitude of change and the sharpness of change seem to be important factors in determining saliency. In other words, salient change points can be defined as those time points where there is an extrema in the slope or inflection points, i.e. points where the second derivative of the signal has a negative-going zero-crossing.

To determine the salient inflection points, we use the rationale that a change is salient if it is preserved over multiple levels of smoothing. Consider a gene profile $f(t)$, we can regard it as a curve parameterized by t. Using a Gaussian scale-space decompostion of the signal[4], the smoothed curve $\hat{C}(t,\sigma)$ can be given 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 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 correspond to inflection points in the signal and are indicated by the red to blue transition regions (light to dark in grey image renderings). Since the zero-crossings shift with increasing scale, the actual location of salient change points in time are found by tracking their coarse extrema to their fine scale locations[4].

Automatic scale selection

Determining appropriate scale of description has been

active area of research in scale-space theory[2, 4] with methods ranging from stability heuristics[4] to finding a maxima in the combined derivative of the smoothed signal as a function of scale[2]. Our approach to scale selection exploits the observation that there is an inherent tradeoff between increasing the approximation error with increasing $\sigma$ and decreasing the deviation from smoothness. If we express the approximation error as a monotonically increasing function, and the deviation from smoothness as a monotonically decreasing function, then the optimal scale corresponds to a minimum in this objective function where the two curves cross-over. With appropriate normalization, the crossover point can be ensured to lie within the allowed scale range.

Specifically, we measure the deviation from smoothness by $f_1(\sigma) = \frac{N(\sigma)}{N(0)}$, where $N(\sigma)$ is the number of zero-crossings at level $\sigma$, and is taken as an indication of smoothness. Due to the Gaussian smoothing property, this is guaranteed to be a monotonically decreasing function in the range [1 ... 0]. The error in approximation, on the other hand, is given by the mean square error

$$f_2(\sigma) = \frac{\sum_t (C(t) - \hat{C}(t,\sigma))^2}{T} \quad (2)$$

The overall objective function is given by

$$F(\sigma) = f_1(\sigma) + \frac{f_2(\sigma)}{max_\sigma(f_2(\sigma))} \quad (3)$$

The optimal value of scale is determined by finding the minimum of this function using standard gradient descent methods. The minimum is caused by the linear combination of a monotonically decreasing and monotonically increasing function. In practice, since $F(\sigma)$ can have multiple minima (due to noise), we select the scale at the cross-over point as a reliable indicator of optimcal scale.

Figure 2 shows the result of automatic scale selection for the gene curve of Figure 2a. The scale-space representation is shown in Figure 2b. Here the curve $f_1(\sigma)$ is shown colored in black and $f_2(\sigma)$ is shown colored in red. The combined curve has a minima at the cross-over point as shown by the curve in blue. The resulting chosen scale = 24 is superposed as a threshold line in the scale-space image of Figure 2b.

## 3. Results

We now illustrate the scale-space-based salient change detection with a few examples. Figure 3a-c shows examples of gene profiles of three genes AGA1,CLN1,HHT1 in S. Cerevesiae data set available from Stanford. Each of the genes selected is active in different phases of the cell cycle. For example, HHT1 is a S-phase regulated gene. The scale-space images corresponding to the gene profile curves of Figure 3(a)-(c) shown in Figure 3d-f. The optimal scale threshold detected using the above method is indicated by the thick horizontal line superposed on each figure. The salient change points were obtained by tracking the zero-crossing contours above the threshold to their locations at the lowest scale. The location of the salient change points, as well as their ranking is shown in Figure 3g-i for the gene curves in Figure 3a-c. As can be seen, the most salient change in each case is predicted in the correct correponding regulatory phase.

Next, we analyzed the time-varying profiles of 106 genes in the over 6000 ORF budding yeast data set available from Stanford (http://cellcycle-www.stanford.edu)[1]. The data 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) . The data has been used to show the cell-cycle regulation of genes. Using our salient change point detection, we were able to record significant change points in these time profiles. The phase within which the most significant change point fell was then taken as the corresponding regulation phase for the gene. This was validated against the cell-cycle regulation ground truth data available for the 104 genes for this dataset as posted on their website. The resulting validation recorded for a few genes is shown in Table 1. Here the time durations for the various phases of the cell cycle are approximate. As can be seen in the table, even in cases where the phase prediction is incorrect, a salient change point is found in the correct phase and/or the predicted phase is adjacent, in most cases, to the actual phase. Overall, of the 104 genes tested, we found the salient change detection-based regulation phase identification was accurate for 89 of them, giving an overall accuracy of 86%.

## 4. Conclusions

We have present a new method for automatic detection of salient change points in genomic signals. The method is generalizable to multi-dimensional signals as well as general time-varying signals.

### 1. REFERENCES


[1] R.J. Cho et al. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell.*, 2:65–73, 1998.

[2] 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.

[3] X. Wen et al. Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl. Acad. Sci.*, pages 334–339, 1998.

[4] A. Witkin. Scale space filtering: A new approach to multi-scale description. In *Proceedings Int. Joint. Conf. Artif. Intell.*, 1984.

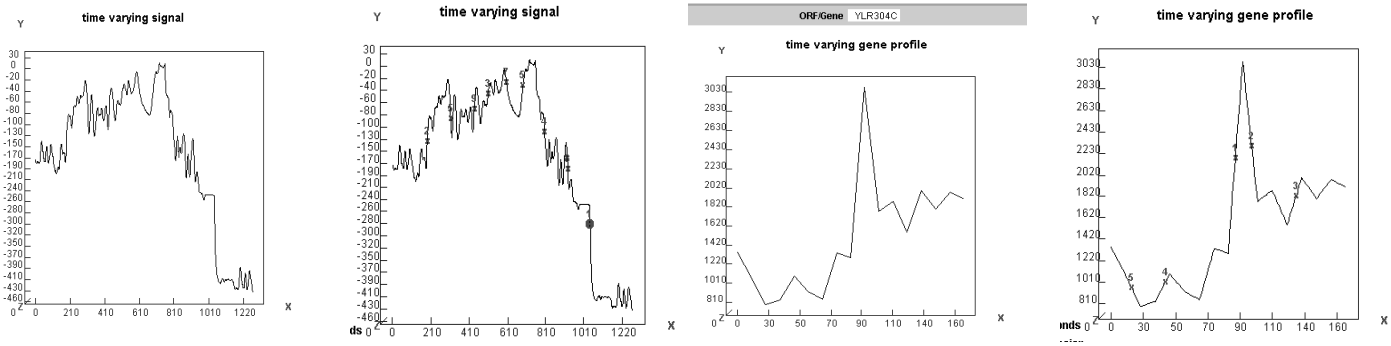


**Fig. 1**. Illustration of saliency for time-varying profiles. (a) Time varying profile of a moving object(velocity curve). (b) Time-varying profile of gene expression for gene YLR304C of budding yeast. (c) Salient change points detected on the profile of (a). (d)Salient change points on the curve of (b).

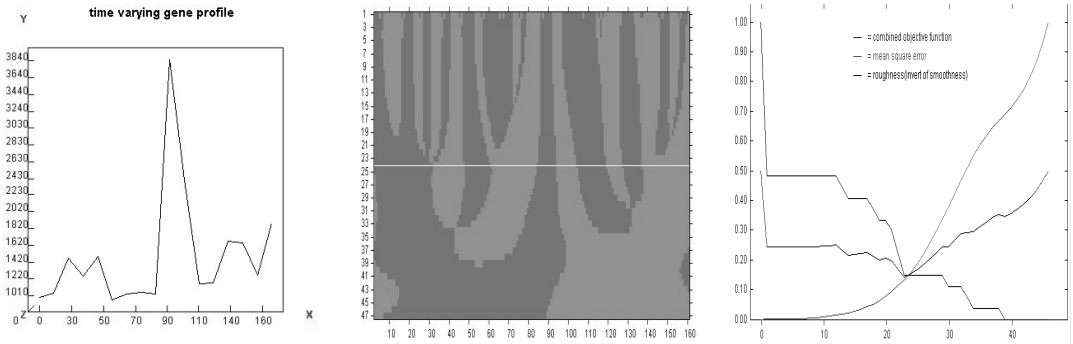


**Fig. 2**. Illustration of automatic scale selection. (a) Time-varying profile of ORF YPR132W. (b)Scale-space image corresponding to (a). (c) The smoothness and approximation error profiles. The automatic scale is chosen at the cross-over point.

| S.No. | Gene Name | Ground-truth phase | Time Duration | Predicted phase | Predicted time |
|---|---|---|---|---|---|
| 1. | AGA1 | M/G1 boundary | 80-90sec | M/G1 | 35,55,**85**,95,115 |
| 2. | CLN1 | Late G1(SCB regulated) | 20-25sec,100-105sec | Late G1 | 45,75,90,**105**,125 |
| 3. | HHT1 | S-Phase | 25-45sec,105-125sec | S/G2 boundary | **45**,65,90,105,115,125,135,145 |
| 4. | CDC14 | S/G2-phase | 40-50sec,120-130sec | G2/M phase | **65**,95,225,135,155 |
| 5. | ACE2 | G2/M phase | 60-90sec,140-150sec | G2/M boundary | 25,55,**90**,110,125,145,165 |
| 6. | ASF1 | Late G1(MCB regulated) | 20-25sec,100-105sec | late G1 | 45,55,**105**,125,132 |

**Table 1**. Illustration of regulatory phase prediction using salient change detection. The most salient change point is indicated in bold.

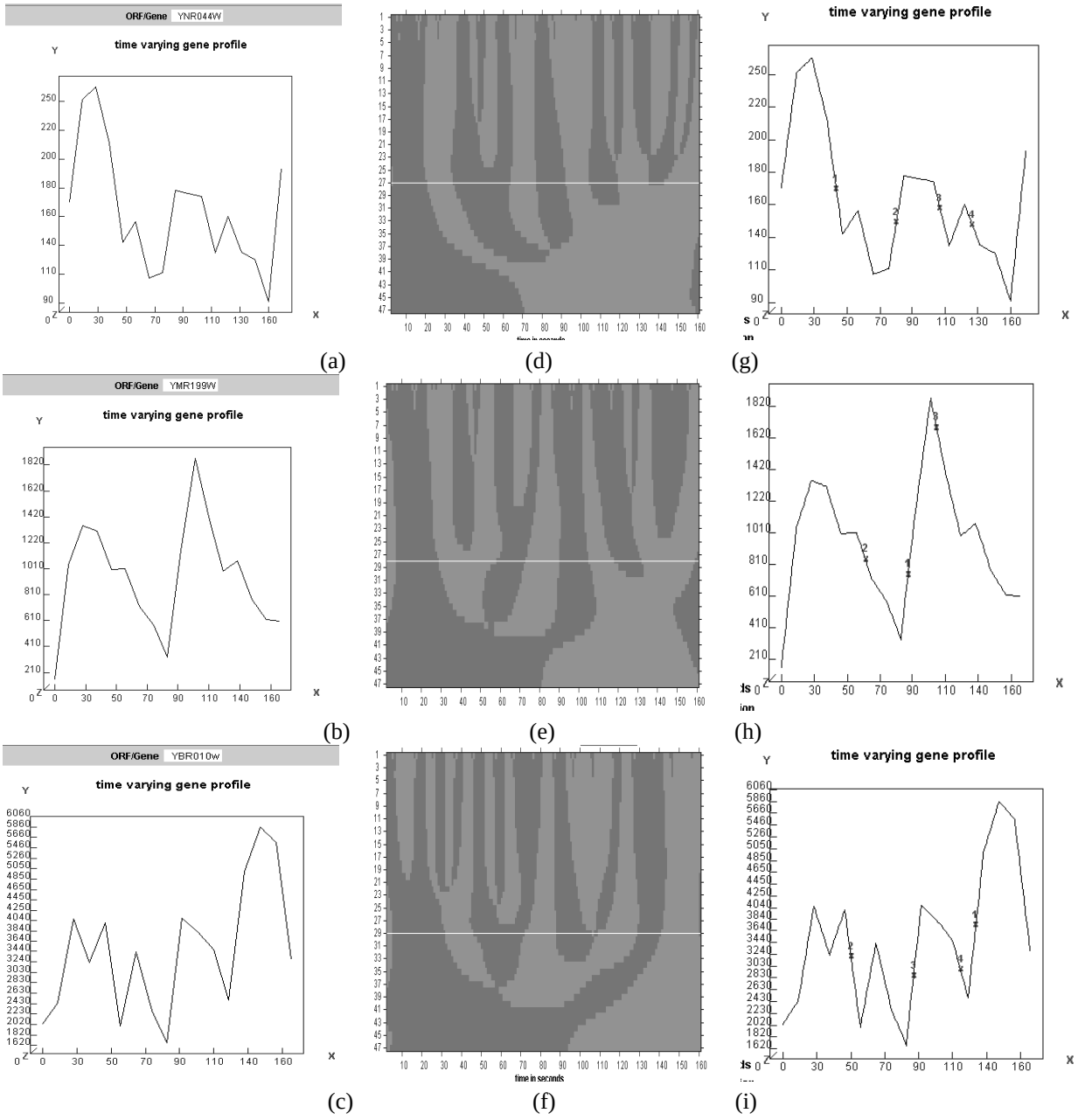


**Fig. 3**. Illustration of saliency detection in gene profiles. (a)-(c) Time varying profiles of sample genes in the mitotic cell cycle of budding yeast. Genes regulated by different phases of the cell cycle are shown here. (b)-(d) scale-space gene profile images corresponding to the gene profile curves of (a) -(c). The optimal scale threshold in each case is shown by the thick white line. (g)-(i) Selected salient change points on the gene profile based on the choice of optimal threshold.