Identifying Non-random Patterns from Gene Expression Profiles

Radhakrishnan Nagarajan[1], Meenakshi Upreti[2], and Mariofanna Milanova[3]

[1] Department of Biostatistics
rnagarajan@uams.edu
[2] Department of Biochemistry and Molecular Biology,
University of Arkansas for Medical Sciences, USA
mupreti@uams.edu
[3] Department of Computer Sciences, University of Arkansas at Little Rock, USA
mgmilanova@ualr.edu

Abstract. There has been considerable interest in identifying biologically relevant genes from temporal microarray gene expression profiles using linear and nonlinear measures. The present study uses two distinct approaches namely: classical order zero-crossing count (ZCC) and Lempel-Ziv (LZ) complexity in identifying non-random patterns from temporal gene expression profiles. While the former captures the linear statistical properties of the time series such a power-spectrum, the latter has been used to capture nonlinear dynamical properties of gene expression profiles. The results presented elucidate that ZCC can perform better than LZ in identifying biologically relevant genes. The robustness of the findings are established on the given gene expression profiles as well as their noisy versions. The performance of these two techniques is demonstrated on publicly available yeast cell-cycle gene expression data. A possible explanation for the better performance of the ZCC over LZ complexity may be attributed to inherent cyclic patterns characteristic of the yeast cell-cycle experiment. Finally we discuss the biological relevance of new genes identified using ZCC not previously reported.

Keywords: Gene expression, Time series, Zero-crossing count, Lempel-Ziv complexity.

1 Introduction

Microarrays have proven to be useful tools in capturing simultaneous expression of a large number of genes in a given paradigm. These high-throughput assays provide system-level understanding of the given paradigm. Recently, microarrays have been used to generate temporal expression profiles. Such profiles capture the transcriptional activity of genes as a function of time, hence their dynamics. There has been considerable interest in developing appropriate techniques to understand functional relationships and network structures from temporal gene expression profiles (see [1] and references therein). These include global analysis techniques such as PCA, hierarchical clustering and self-organizing maps (see [2] and references

M. Chetty, A. Ngom, and S. Ahmad (Eds.): PRIB 2008, LNBI 5265, pp. 299–310, 2008.

therein). As pointed out by [3], such techniques treat the expression across the time points as independent entities, hence immune to the temporal structure/dynamics. Such global analyses may also be susceptible to inherent transcriptional delays. Therefore, it might be necessary to explore alternate techniques that are sensitive to the temporal structure of the data. Temporal expression profiles of biologically relevant genes orchestrating a specific paradigm follow characteristic and reproducible patterns. This in turn supports their inherent non-random nature. While genetic networks are undoubtedly nonlinear dynamical systems, these dynamical nonlinearities need not necessarily manifest in the external recording such as microarray expression profiles. This can be attributed to noisiness and nonlinearities at the measurement and dynamical levels [4]. On a related note, nonlinear dynamical systems can also give rise to simple cyclic/periodic behavior (limit cycle) for certain choice of the system parameters which can be modeled as linear stochastic processes. Therefore, measures sensitive to linear as well as nonlinear statistical properties have been used to investigate patterns in gene expression profiles. In the present study, we compare the performance of two measures in identifying biologically relevant genes, namely. (a) *zero-crossing count* (ZCC), [5] which is related to the linear statistical properties of temporal data such as power-spectrum and (b) *Lempel-Ziv (LZ) complexity measure*, [6,7] which is sensitive to linear as well as nonlinear dynamical properties in the given data. We show that ZCC can prove to be a better choice than LZ complexity in identifying biologically significant genes. The better performance of ZCC may be due to inherent cyclic patterns in the yeast cell-cycle data. Such patterns can be modeled as linear stochastic processes with Gaussian innovations. Therefore, measures sensitive to the linear correlation structure may be sufficient to describe them.

Techniques such as power-spectral analysis which capture the linear statistical properties of the stationary time series are a natural choice for investigating experimental time series [8]. Power-spectrum is related to the auto-correlation function (Wiener-Khinchin theorem) which in turn is used to estimate the optimal process parameters of linear stochastic processes (Yule-Walker equations) [9, 10]. Alternatively, auto-correlation function is sufficient statistics for describing normally distributed linear stochastic processes. Uncorrelated noise is characterized by a flat power-spectrum representing equal power across all frequencies. A significant skew in the power-spectrum towards lower-frequencies is indicative of correlation or non-random signatures in the given time series. These in turn may be indicative of biologically relevant genes. Classical spectral estimation may not be possible due to the small length of the temporal gene expression profiles. However, the spectral properties of a given time series can also be captured by zero-crossing count [5]. The latter overcomes some of the caveats encountered in classical spectral estimation and its interpretation is fairly straight forward. In the present study, biologically relevant genes will be identified from ZCC. In order to establish statistical significance, ZCC estimates on the given data are compared to those obtained on random shuffled surrogates. The random shuffled surrogates represent uncorrelated counterpart of the given data.

Information theoretic approaches have also been popular in capturing patterns in gene expression profiles. These include entropy-based approaches [11]. However, entropy estimates are governed solely by the probability distribution of the expression values, hence immune to the temporal expression profile. For instance a gene exhibiting a periodic pattern across 9 time points (001001001) has the same Shannon entropy as gene exhibiting seemingly random pattern (101000100). The latter is obtained by randomly shuffling the former. Complexity measures which overcome some of the caveats of entropy have recently been proposed to investigate gene expression profiles [12, 15, 16]. Traditionally, uncorrelated noise is considered to be maximally complex or random. Any deviation from randomness ensures correlation and accompanied by a decrease in complexity. From the perspective of gene expression analysis, genes with low complexity are hypothesized to be biologically relevant. A recent study, proposed several measures of complexity for the analysis of temporal gene expression profiles in yeast cell-cycle experiment [12]. Such an analysis was carried out in an unbiased manner in the absence of any prior knowledge about the given data. The authors successfully identified biologically relevant genes in addition to those that exhibited characteristic cyclic behavior [13, 14]. Complexity measures have also been successfully used to gain insight into the dynamical aspects of genetic networks from the temporal expression profiles [15, 16]. In [15], the distribution of the Lempel-Ziv complexity from experimental gene expression profiles was compared to those generated from synthetic random Boolean networks (RBN) using Kullback-Leibler divergence and Euclidean distance. Subsequently, the dynamics of the genetic network governing the paradigm was found to lie in the ordered regime or between order and chaos. More recently, a variant of the Kolmogorov-complexity (i.e. normalized compression distance) was used to argue in favor of criticality in macrophage dynamics. In the present study, Lempel-Ziv (LZ) complexity is used identify biologically relevant genes. Unlike ZCC, LZ is sensitive to linear as well as nonlinear correlations in the given data. For instance, qualitative behavior of LZ complexity has been found to mirror invariants such Lyapunov exponents in nonlinear dynamical systems [7, 17]. Periodic time series such as sine-waves are highly compressible and result in low values of LZ complexity. Ideally, uncorrelated noise cannot be compressed hence accompanied by large values of LZ complexity. As in the case of ZCC, LZ values obtained on the given data are compared to those obtained on random shuffled surrogates in order to establish statistical significance. There is no direct relation between LZ complexity and ZCC. However, it should be noted that ZCC as well LZ is likely to increase monotonically with increasing noise in the given data. Noise being reflected by high-frequency components in the power-spectrum.

2 Methods

2.1 Spectral Analysis by Zero-Crossing

Consider a zero-mean normally distributed stationary process $x_t, t = 1 \dots N$. The corresponding binary sequence of the differenced series $z_t = x_t - x_{t-1}$ is generated as

$$\begin{aligned} y_t &= 1 \quad \text{if} \quad z_t > 0 \\ &= 0 \quad \text{otherwise} \end{aligned} \tag{1}$$

Expression (1) essentially quantizes the given input signal z_t onto a coarse-grained binary sequence. The sequence y_t is subsequently differenced and passed through a memoryless nonlinear transform as

$$\nabla_t = (y_t - y_{t-1})^2 \tag{2}$$

A zero-crossing is said to occur if $\nabla_t = 1$. Subsequently, the zero-crossing count is given by

$$D_1 = \sum_{t=1}^{N-1} \nabla_t \tag{3}$$

The zero-crossing count is related to the linear correlation, hence the power-spectrum [see 5 and references therein]. For certain class of colored Gaussian noise, zero-crossing analysis may be sufficient to describe the process dynamics [5]. Examples include cyclic patterns such as those encountered in the Spellman alpha-factor synchronization yeast cell-cycle experiment. At this point, it might be necessary to point out certain resemblance between the zero-crossing analysis and the complexity maps proposed in [12]. The first-order crossing $\delta_t = y_t - y_{t-1}$ shares resemblances to the map γ_{Δ_1} proposed recently in [12]. On a similar note the maps γ_{Δ_2} and γ_{Δ_3} in [12] share resemblance to the expressions δ_t^2 (i.e.$\delta(\delta_t) = \delta y_t - \delta y_{t-1}$) and δ_t^3 (i.e.$\delta(\delta_t^2)$) for higher order crossings obtained repeated application of the difference operator (high-pass filter) [5]. On closer observation, we note subtle differences in their definitions. The complexity maps [12] use (*i*) ranks of the values as opposed to the sign of the mean-subtracted values (1). (*ii*) Discontinuous memoryless function as opposed to continuous memoryless function (2). While the maps in [12] capture the complexity of the given process, ZCC captures the spectral characteristics of the process.

2.2 Lempel-Ziv Complexity

Lempel-Ziv complexity (LZ) [6] and its extensions [7] have been used widely to understand the dynamics in biomedical [18, 19] and genomic signals [15]. An elegant implementation of the LZ algorithm for binary sequences was proposed in [7]. In the present study, binary sequence of the expression profiles was generated by thresholding about the mean. The objective of the LZ algorithm is to reconstruct the given sequence *s* of length *n*, using two fundamental operations, namely: *copy* and *insert* by parsing it from left to right. This information in turn is used for estimating the algorithmic complexity of that. The working principle of the LZ algorithm is shown below for completeness. Prior to the discussion of the example we introduce the notation *v*(*s*) in the following example corresponds to the vocabulary set [6]. Consider *s* = 00, then *v*(s) represents all possible words that can be reconstructed from s when scanning from left to right, i.e. *v*(*s*) = {0, 00}. It is important to note that while 0 can be generated from *v*(*s*), 1 cannot be generated from *v*(*s*).

Consider a period 3 sequence s_0 = 001001001……. as before

(a) The first digit 0 is unknown hence have to be inserted, resulting in *c*(*n*) = 1 and s^* = 0.

(b) Consider the second digit 0. Now *s* = 0, *q* = 0; *sq* = 00; *sq*π = 0; *q*∈ *v*(*sq*π); therefore copying is sufficient resulting in no change in the complexity i.e. *c*(*n*) = 1 and s^* = 0.0

(c) Consider the third digit 1. Now $s = 0$, $q = 01$; $sq = 001$; $sq\pi = 00$; $q \notin v(sq\pi)$; therefore insertion is required, resulting in $c(n) = 2$ and $s^* = 0.01$.
(d) Consider the fourth digit 0: $s = 001$; $q = 0$; $sq = 0010$; $sq\pi = 001$; $q \in v(sq\pi)$; therefore copying is sufficient, resulting in $c(n) = 2$ and $s^* = 0.01.0$
(e) Consider the fifth digit 0: $s = 001$; $q = 00$; $sq = 00100$; $sq\pi = 0010$; $q \in v(sq\pi)$; therefore copying is sufficient, resulting in $c(n) = 2$ and $s^* = 0.01.00$
(f) Consider the sixth digit 1: $s = 001$; $q = 001$; $sq = 001001$; $sq\pi = 00100$; $q \in v(sq\pi)$; therefore copying, is sufficient resulting in $c(n) = 2$ and $s^* = 0.01.001$

It is important to note that subsequent addition of symbols from s does not change $c(n)$. This can be attributed to the inherent periodicity of the sequence s. Since s^* does not end in a dot (**.**) we add one to the complexity, resulting in $c(n) = 3$. In the present study, we consider the normalized complexity measure (γ) given by the expression

$$\gamma = \frac{c(n)}{b(n)}, \text{ where } b(n) = \frac{n}{\log_2 n} \tag{4}$$

The normalized complexity (γ) tends to unity in the asymptotic limit for sequences whose Shannon entropy is unity [6, 7].

2.3 Random Shuffled Surrogates

Resampling techniques are encouraged in literature for establishing statistical significance where the null distribution is unknown. Resampling without replacement is used widely within the context of correlated data analysis [20, 21]. Such an approach retains certain statistical properties of the given empirical sample in the surrogates. For the same reason these surrogates are termed as *constrained realizations* [21]. In the present context, the objective is to argue in favor of correlation in the given gene expression profile, i.e. the statistics considered namely: D_1 (3) and γ (4) are sensitive to correlation in the given data. Therefore, the objective is the reject the null that the given data is uncorrelated noise. The surrogates under the above null are generated by randomly shuffling (RS) the temporal expression profile of that gene. The constraint here is on retaining the distribution of the gene expression profile in the surrogates. Alternatively, the distribution of the gene expression profile is treated as a nuisance variable [21]. The discriminant statistics (3) and (4) are sensitive to the temporal structure, hence are expected to exhibit a significant discrepancy between the empirical sample and the surrogate counterpart in the case of correlated expression profiles. More importantly, estimates of (3) and (4) on the surrogate realizations will be higher than those estimated on the empirical sample for correlated gene expression profiles. Therefore, a one-sided test is sufficient to establish statistical significance. In the present study, the number of surrogates were chosen as ($n_s = 99$) corresponding to a significance level $\alpha = 1/(99+1) = 0.01$ [21]. It is important reiterate that the null is rejected only when the estimate of (3) and (4) is lesser than each of the 99 surrogate realizations.

3 Data

In order to establish reproducibility and direct comparison, we have used the same data sets as those investigated recently by [12] using a battery of complexity measures. The final list of genes was obtained from the Ahnert et al., (personal communication). The authors in [12] identified 150 genes as top-ranked by one of their proposed complexity measures $k(f/\gamma_{\Delta 3})$. 52 of these 150 genes were listed as biologically relevant either by Spellman (104 genes, *http://genome-www.stanford.edu /cellcycle/data/rawdata/-Knowngenes.doc*) or Simon (140 genes, *http://web.wi.mit. edu/young/cellccycle/-Table of Regulated genes*). Their set of 150 genes also included genes not listed in either Spellman (104) or Simon (140).

LIST OF 133 GENES: Since identifying the missing data is not one of the objectives of the proposed study, we first eliminate all genes whose values are missing even at a single time point in the Spellman alpha-factor synchronization yeast cell-cycle experiment (6178 genes across 18 time points) [14]. The reduced set consisted of 4491 genes. Therefore, all subsequent discussions will be restricted to this set of 4491 genes. Spellman et al., 1998 [14] identified 104 genes as well-documented through extensive literature survey. Out of these 104 genes, 72 genes had values across all 18 time points (i.e. overlapped with the list of 4491 genes). The ORFs of these 72 genes were subsequently identified. In a related study, Simon et al., 2001 [13] identified 140 genes as being relevant to yeast cell-cycle. The gene IDs of these 140 genes were retrieved by comparing their ORFs to those spotted on the array. Only 125 of the 140 identified had gene IDs. Out of these 125 genes, 93 genes had values across all 18 time points (i.e. overlapped with the set of 4491 genes). Thus in essence we have 72 out of the 104 genes from study [14] and 93 out of the 140 genes from study [13]. The union of the above sets resulted in 133 genes relevant to yeast cell-cycle.

LIST OF 40 GENES: The performance of the measures (3) and (4) are also determined using the set of 40 genes as ground truth from Ahnert et al., (personal communication). Of the 52 genes identified by Ahnert et al., 2006 40 had values across all 18 time points (i.e. overlapped with the 4491 genes).

The objective of the present study is to determine the effectiveness of the measures (3) and (4) on retrieving biologically relevant genes from the set of 4491 genes. Of interest is to investigate the false-positive and false-negative rates using the 133 genes and 40 genes as ground truths. The effect of noise on the performance of the two measures (3), (4) is also investigated. Finally the usefulness of the ZCC (3) in identifying new genes not previously reported in either Simon [13] or Spellman [14] is discussed. The temporal expression profiles of the 4491 genes were mean-subtracted prior to the analysis. The present study also assumes the gene expression profiles to be generated from stationary stochastic processes.

4 Results

Prior to investigating the gene expression profiles we demonstrate the effectiveness of the proposed measures in quantifying regularity in patterns across synthetic linear and nonlinear dynamical processes.

4.1 Synthetic Data

Periodic sine-wave

Consider a periodic sine-wave given by $s(t) = \sin(2\pi f_1 \omega)$ with (f_1 = 2, N = 200), Fig 1a. The sampling frequency (F_s = 15) was chosen so as to satisfy the Nyquist criterion, i.e. $F_s > 2f_1$. Periodic signatures can be modeled as linear stochastic processes such as auto-regressive moving average processes (ARMA), hence linear statistical properties are sufficient to describe them. However, periodic signatures can also be generated by nonlinear dynamical systems (limit cycles). The power-spectrum of the periodic sine-wave exhibits a dominant frequency, Fig. 1b. The results of the zero-crossing analysis (3) and the normalized complexity (4) are shown in Fig. 1c and 1d respectively. The statistical measures (3) and (4) of the periodic sine-wave is clearly lesser than those obtained on the random shuffled surrogates (n_s = 99, α = 0.01) rejecting the null as expected, Figs. 1c and 1d.

Chaotic logistic map

Consider a chaotic logistic map given by $x_{n+1} = 3.8.x_n.(1-x_n)$, Fig 1e. Unlike periodic sine-wave, chaotic logistic map is nonlinearly correlated and is accompanied by a broad-band power-spectrum characteristic of random noise, Fig. 1f. The results of zero-crossing analysis (3) and the normalized complexity (4) are shown in Fig. 1g and 1h respectively. The time series was mapped onto a binary sequence by thresholding about the superstable fixed point (0.5). This particular choice has been shown to capture the dynamics faithfully [17]. Zero-crossing estimate (3) on the chaotic process was considerably higher than those on the random shuffled surrogates failing to reject

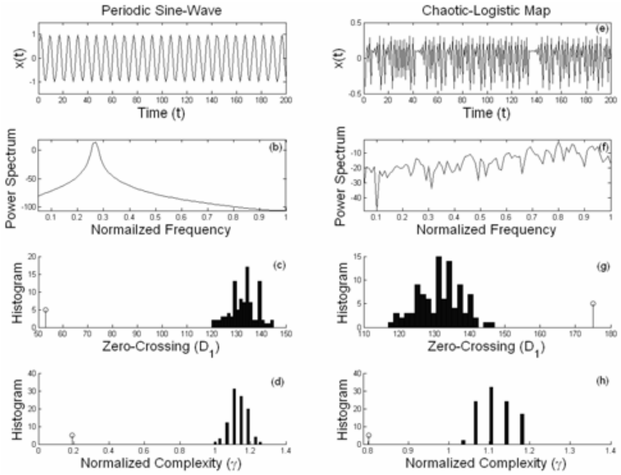

Fig. 1. Time series generated from periodic sine-wave (left) and chaotic logistic map (right) are shown in (a) and (e) respectively. The corresponding power-spectra are shown in (b) and (f) respectively. Zero-crossing crossing analysis (c and g) and normalized complexity (d and h) estimates of the empirical samples are shown by hollow circles. The histogram of their estimates on (n_s = 99) random shuffled surrogates are also shown in the corresponding subplots (black bars).

the null ($n_s = 99$, $\alpha = 0.01$), Fig. 1g that the given process is uncorrelated noise. The one-step auto-correlation estimated using (7) and directly from the data was negative failing to provide any meaningful insight into the dynamics. However, analysis using the normalized complexity clearly rejected the null, Fig. 1h. Thus the zero-crossing analysis may have clear limitations when analyzing from nonlinear processes.

4.2 Yeast Cell-Cycle Experiment

Zero-crossing analysis (3) of the 4491 genes (Sec. 3) identified 101 genes as being significant ($\alpha = 0.01$). A similar analysis using normalized complexity (4) identified 133 genes as being significant ($\alpha = 0.01$). Prior to a detailed analysis we investigated two genes namely: YBR010W (Fig. 2a) and YPR150W (Fig.2b) using D_1 and γ. The choice of these two genes can be attributed to a recent study which investigated their biological relevance using a battery of complexity measures [12]. Visual inspection of the temporal expression profiles of YBR010W (Fig. 2a) revealed characteristic low-frequency non-random signatures unlike YPR150W (Figs. 2b). Zero-crossing (D_1), Fig. 2c, as well as normalized complexity (γ), Fig. 2e, estimates on YBR010W were considerably less than those obtained on the surrogate realizations rejecting the null that YBR010W is uncorrelated noise. The results of a similar analysis of YPR150W using (D_1) and (γ) are shown in Figs. 2d and 2f respectively. Both the measures failed to reject the null in the case of YPR150W. Earlier studies have reported YBR010W (or HHT1) to be actively involved in cell-cycle [13]. In contrast, YPR150W is an open-reading frame (ORF) with no documented evidence of its role in yeast cell-cycle. The performance of the two measures (D_1, γ) were subsequently investigated by introducing noise in the zero-mean temporal expression profile for a gene x as

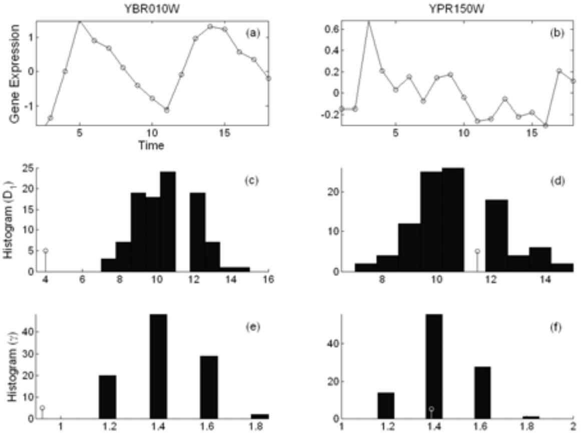

Fig. 2. Temporal expression profiles of two genes YBR010W and YPR150W are shown in (a) and (b) respectively. The corresponding zero-crossing (D_1) estimates (circle) along with the distribution of the estimates on the 99 random shuffled surrogates (black bars) for each of the three genes are shown right below them in (c) and (d) respectively. The results of a similar analysis using normalized complexity (γ) for the genes are shown right below them in (e) and (f) respectively.

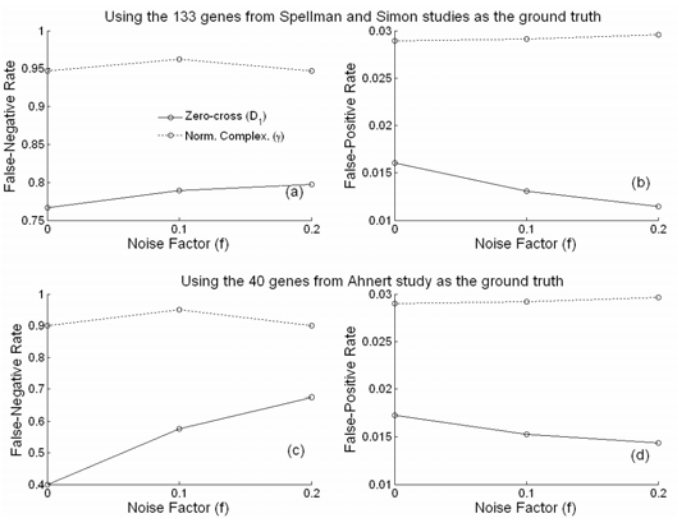

Fig. 3. False-negative rate and false positive rate obtained assuming the 133 genes [13,14] as ground truth using zero-crossing (D_1, solid lines) and the normalized complexity (γ, dotted lines) for noise factors (f = 0, 0.10, and 0.20) is shown in (a) and (b) respectively. False-negative rate and false-positive rate obtained by a similar analysis assuming the 40 genes [12] as ground truth is shown in (c) and (d) respectively.

($y_t = x_t + f.e_t$), where e_t is i.i.d Gaussian noise. Introducing noise to the observed expression falls under measurement noise ubiquitous in microarray studies.

Assuming the 133 genes identified by the union of 140 genes and 104 genes [13, 14] whose values are known across all time points as the ground truth.
Zero-crossing analysis (D_1) identified 37 of the 133 biologically relevant genes (i.e. 31/133 ~ 23%). The normalized complexity (γ) identified only 7 out of the 133 genes (i.e. 7/133 ~ 5%). These have to be compared to those of [12], who identified 52 genes from an union set of 195 genes (52/195 ~ 26.7%) using complexity measure $k(f/d3)$. The discrepancy in the number of genes between [12] and the present study (133) can be attributed to the fact that we considered only genes whose expression values are known across all time points. The false-positive rate (FPR) and false-negative rate (FNR) for the normalized complexity (γ) and zero-crossing analysis (D_1) across noise factors (f = 0, 0.10, 0.20) is shown in Figs. 3a and 3b respectively. From Fig. 3a it is evident that the FPR and FNR of the normalized complexity are considerably higher than that of the zero-crossing analysis.

Assuming the 40 genes from (52) [12] whose values are known across al the time points as the ground truth.
It should be noted that these 40 genes are present in the union set of 133 genes discussed above. While zero-crossing analysis (D_1) identified (24/40) genes, the normalized complexity (γ) identified only (4/40) genes from [12]. The FPR and FNR for the two measures assuming these 40 genes as the ground truth across noise factors (f = 0, 0.10, 0.20) is shown in Figs. 3c and 3d respectively. The results are similar to

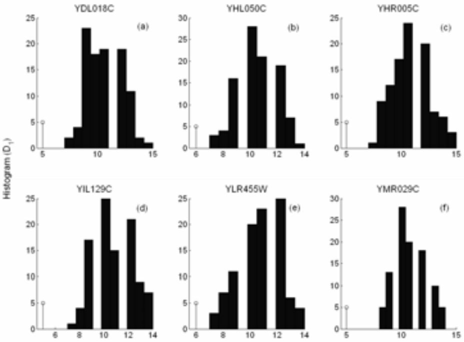

Fig. 4. Six genes YDL018C, YHL050C, YHR005C, YIL129C, YLR455W and YMR029C identified by D_1 and not present in the list of 133 genes. The estimate of D_1 on the gene expression profile (circle) and their ($n_s = 99$) random shuffled surrogates (black bars) are shown in each of the subplots.

those obtained for the 133 genes with considerably higher FNR and FPR for (D_1) compared to γ.

New genes discovered by the zero-crossing analysis

As noted earlier, out of the 101 genes detected as being statistically significant by zero-crossing analysis, 37 exhibited overlap with the documented 133 genes. The remaining 64 genes consisted of YDL018C (ERP3) [22], YHL050C, YHR005C (GPA1) [23], YIL129C (TAO3) [24, 25], YLR455W (Uncharacterized ORF) [26], YMR029C (FAR8) [27], see Fig. 4. YDL018C or ERP3 had been shown to be involved in the G1/S phase of the cell cycle [22]. YHL050C is an uncharacterized open-reading frame but we believe its non-random nearly periodic pattern is compelling for us to hypothesize it as a likely candidate in yeast cell cycle regulation. There has been evidence [23] on the role of YHR005C (GPA1, G-protein alpha subunit 1) on the mating-factor mediated cell cycle arrest. YIL129C (TAO3) is a member of the RAM (Regulation of Ace2p activity and cellular Morphogenesis) [24] signaling network which governs cell separation, integrity and progression [25]. There has been evidence of YLR455W being involved in S-phase of the cell-cycle [26]. YMR029C (FAR8) has been documented to be involved in G1 cell cycle arrest in response to pheromone [27].

5 Discussion

There has been considerable interest in understanding temporal gene expression profiles using suitable techniques. Biologically relevant genes are hypothesized to exhibit non-random and reproducible temporal expression profiles. Understanding the correlation in these patterns has been of great interest. The present study investigated two distinct measures namely: zero-crossing count and the normalized Lempel-Ziv complexity in identifying biologically relevant genes from their expression profile. While the former is sensitive to only the linear correlation, the latter is sensitive to linear as well as nonlinear correlations in the given data. These techniques implicitly assume that the given temporal expression profiles are sampled from a stationary process. From the results presented, it is evident that zero-crossing count which mimics the spectral content of the given data may prove to be useful in determining biologically relevant genes from their expression patterns. Its performance was found to be better than that of the normalized complexity. The results were demonstrated on yeast cell cycle expression profiles with and without noise. The better performance of the zero-crossing count may be attributed to cyclic patterns which can be modeled as linear stochastic processes.

Acknowledgments. R.N. was partially supported by NIH Grant # P20 RR-16460 from the IDeA Networks of Biomedical Research Excellence (INBRE) Program of the National Center for Research Resources. R.N would like to thank Xiaoyan (Iris) Leng for useful comments on gene annotation and Sebastian Ahnert for sharing the list of genes identified in his study.

References

1. Bar-Joseph, Z.: Analyzing time series gene expression data. Bioinformatics 20(16), 2493–2503 (2004)
2. Butte, A.: The use and analysis of microarray data. Nat. Rev. Drug Discov. 1(12), 951–960 (2002)
3. Leng, X., Müller, H.G.: Classification using functional data analysis for temporal gene expression data. Bioinformatics 22(1), 68–76 (2006)
4. Nagarajan, R., Aubin, J.E., Peterson, C.A.: Modeling genetic networks from clonal analysis. Journal of Theoretical Biology 230(3), 359–373 (2004)
5. Kedem, B.: Time Series Analysis by Higher Order Crossings. IEEE Press, Los Alamitos (1994)
6. Lempel, A., Ziv, J.: On the Complexity of Finite Sequences. Information Theory, IEEE Transactions 22(1), 75–81 (1976)
7. Kaspar, F., Schuster, H.G.: Easily calculable measure for the complexity of spatio-temporal patterns. Phys. Rev. A. 36, 842–848 (1987)
8. Butte, A.J., Bao, L., Reis, B.Y., Watkins, T.W., Kohane, I.S.: Comparing the similarity of time-series gene expression using signal processing metrics. J. Biomed. Inform. 34(6), 396–405 (2001)
9. Proakis, J.G., Manolakis, D.G.: Digital Signal Processing, Principles Algorithms and Applications. Prentice-Hall, Englewood Cliffs (1996)
10. Papoulis, A., Pillai, S.U.: Probability, Random Variables and Stochastic Processes, 4th edn. McGraw-Hill, New York (2002)
11. Fuhrman, S., Cunningham, M.J., Wen, X., Zweiger, G., Seilhamer, J.J., Somogyi., J.: The Application of Shannon Entropy in the Identification of Putative Drug Targets. Biosystem 55(1-3), 5–14 (2000)
12. Ahnert, S.E., Willbrand, K., Brown, F.C.S., Fink, T.M.A.: Unbiased pattern detection in microarray data series. Bioinformatics 22, 1471–1476 (2006)
13. Simon, I., et al.: Serial regulation of transcriptional regulators in the yeast cell cycle. Cell 106, 697–708 (2001)

14. Spellman, P.T., et al.: Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol. Bio. Cell 9, 3273–3297 (1998)
15. Shmulevich, I., Kauffman, S.A., Aldana, M.: Eukaryotic cells are dynamically ordered or critical but not chaotic. Proc. Nat. Acad. of Sci (USA) 102(38), 13439–13444 (2005)
16. Nykter, M., Price, N.D., Aldana, M., Ramsey, S.A., Kauffman, S.A., Hood, L., Yli-Harja, O., Shmulevich, I.: Gene Expression Dynamics in the Macrophage Exhibit Criticality. Proc. Nat. Acad. of Sci (USA) 105(6), 1897–1900 (2008)
17. Rasband, N.: Chaotic Dynamics of Nonlinear Systems. Wiley-Interscience, Chichester (1997)
18. Nagarajan, R.: Quantifying physiological data with Lempel-Ziv complexity - certain issues. IEEE Trans Biomed. Engg. 49(11), 1371–1372 (2002)
19. Nagarajan, R., Szczepanski, J., Wajnryb, E.: Interpreting non-random signatures in biomedical signals with Lempel–Ziv complexity. Physica. D. 237(3), 359–364 (2008)
20. Theiler, J., Eubank, S., Longtin, A., Galdrikian, B., Farmer., J.D.: Testing for nonlinearity in time series: the method of surrogate data. Physica D 58, 77–94 (1992)
21. Schreiber, T., Schmitz, A.: Surrogate time series. Physica D 142, 346–382 (2000)
22. Bean, J.M., et al.: High functional overlap between MluI cell-cycle box binding factor and Swi4/6 cell-cycle box binding factor in the G1/S transcriptional program in Saccharomyces cerevisiae. Genetics 171(1), 49–61 (2005)
23. Miyajima., I., et al.: GPA1, a haploid-specific essential gene, encodes a yeast homolog of mammalian G protein which be involved in mating factor signal transduction. Cell 50(7), 1011–1019 (1987)
24. Jorgensen, P., et al.: High-resolution genetic mapping with ordered arrays of Saccharomyces cerevisiae deletion mutants. Genetics 162, 1091–1099 (2002)
25. Bogomolnaya, L.M., Pathak, R., Guo, J., Polymenis, M.: Roles of the RAM signaling network in cell cycle progression in Saccharomyces cerevisiae. Curr. Genet. 49(6), 384–392 (2006)
26. de Lichtenberg, U., et al.: New weakly expressed cell cycle-regulated genes in yeast. Yeast 22(15), 1191–11201 (2005)
27. Kemp, H.A., Sprague Jr., G.F.: Far3 and five interacting proteins prevent premature recovery from pheromone arrest in the budding yeast Saccharomyces cerevisiae. Mol. Cell. Biol. 23(5), 1750–1763 (2003)