

Statistical Detection of Short Periodic Gene Expression Time Series Profiles

Alan Wee-Chung Liew^a, N.F. Law^b, and Hong Yan^{c,d}

^a*School of Information & Communication Technology, Griffith University, Brisbane, Australia*

^b*Centre for Signal Processing, Department of Electronic and Information Engineering,
The Hong Kong Polytechnic University, Hong Kong.*

^c*Department of Electronic Engineering, City University of Hong Kong, Hong Kong*

^d*School of Electronic and Information Engineering, University of Sydney, NSW2006, Australia*

Email: a.liew@griffith.edu.au, ennflaw@polyu.edu.hk, h.yan@cityu.edu.hk

Abstract. Many cellular processes exhibit periodic behaviors. Hence, one of the important tasks in gene expression data analysis is to detect subset of genes that exhibit cyclicity or periodicity in their gene expression time series profiles. Unfortunately, gene expression time series profiles are usually of very short length and highly contaminated with noise. This makes detection of periodic profiles a very difficult problem. Recently, a hypothesis testing method based on the Fisher g -statistic with correction for multiple testing has been proposed to detect periodic gene expression profiles. However, it was observed that the test is not reliable if the signal length is too short. In this paper, we performed extensive simulation study to investigate the statistical power of the test as a function of signal length, SNR, and the false discovery rate. We found that the number of periodic profiles can be severely underestimated for short length signal. The findings indicated that caution needs to be exercised when interpreting the test result for very short length signals.

Keywords: Gene expression profiles, periodicity detection, Fisher exact test, g -statistic, short signal.

PACS: 87.16.A, 87.18.Wd, 87.16.Yc

INTRODUCTION

Periodic phenomena are widely studied in biology and there are numerous biological applications where periodicities must be detected from experimental data. Gene expression data from microarray experiments are commonly used to measure the cell-cycle activities. An important task in gene expression data analysis is to detect subset of genes that exhibit cyclicity or periodicity in their gene expression time series profiles. However, this is a challenging problem due to the typically small number of measurements per gene (for example, in the human cancer cells study [1] on <http://genome-www.stanford.edu/Human-CellCycle/Hela/>, the smallest number of measurements per gene is 9). Moreover, the data is usually contaminated with high level of noise and missing values.

Recently, several methods for detecting periodic gene expression based on statistical hypothesis testing have been proposed [2-4]. In [2], Wichert et al. proposed to use the periodogram-based Fisher g-statistic test to determine whether or not a sequence is periodic. In [3], Chen proposed a statistical inference approach, the C&G procedure, to effectively detect statistically significant periodically expressed genes based on two statistical hypothesis testing procedures, one of which is the g-statistic. In [4], Ahdesmäki et al. proposed a robust spectral estimator for the Fisher g-statistic test that has better performance when the noise deviates from Gaussian. In our recent paper [5], we [have](#) proposed a new spectral estimation algorithm for unevenly sampled gene expression data and use the g-statistic to perform a ranking of the genes with respect to their periodicity tendency. We did not attempt to give the number of periodic genes in the gene expression datasets since we observed that the g-statistic hypothesis testing procedure based on the false discovery rate (FDR) framework has low statistical power for short length signals. This lack of statistical power is also observed in [2] and [4].

In this paper, we perform a systematic experimental study of the Fisher g-statistic periodicity test to investigate the accuracy of the test with respect to signal length, signal-to-noise ratio (SNR), and chosen level of statistical significance using the FDR. This is done by performing extensive simulation experiments with simulated signals that model the real time series profiles that arise from microarray gene expression data.

GENE EXPRESSION TIME SERIES PROFILES

Cellular processes can be studied by measuring the gene expression patterns through DNA microarray experiments. If the expression patterns of a group of genes are measured over a number of time points, we obtain a time series data describing the dynamic behaviors of the genes under study. A well known set of gene expression time series datasets is that of the Yeast *Saccharomyces cerevisiae* from Spellman et al. [6]. In this set of datasets, the genome-wide mRNA levels for 6178 yeast ORFs are monitored simultaneously using several different methods of synchronization including an alpha-factor-mediated G1 arrest which covers approximately two cell-cycle periods with measurements at 7 min intervals for 119 minutes with a total of 18 time points, a temperature-sensitive *cdc15* mutation to induce a reversible M-phase arrest (24 time points taken every 10 minutes covering approximately 3.5 cell-cycle periods), and a temperature-sensitive *cdc28* mutation to arrest cells in G1 phase reversibly (17 time points taken every 10 minutes covering approximately 2 cell-cycle periods), and finally, an elutriation synchronization to produce the elutriation dataset of 14 time points taken every 30 minutes covering approximately 1 cell-cycle period. The left and right panels of Fig. 1 show some example profiles from these 4 datasets that exhibit highly periodic and random behaviors, respectively. Although periodicity can readily be observed in the left panel of Fig.1, in many cases the distinction between periodic and random profiles cannot be easily made.

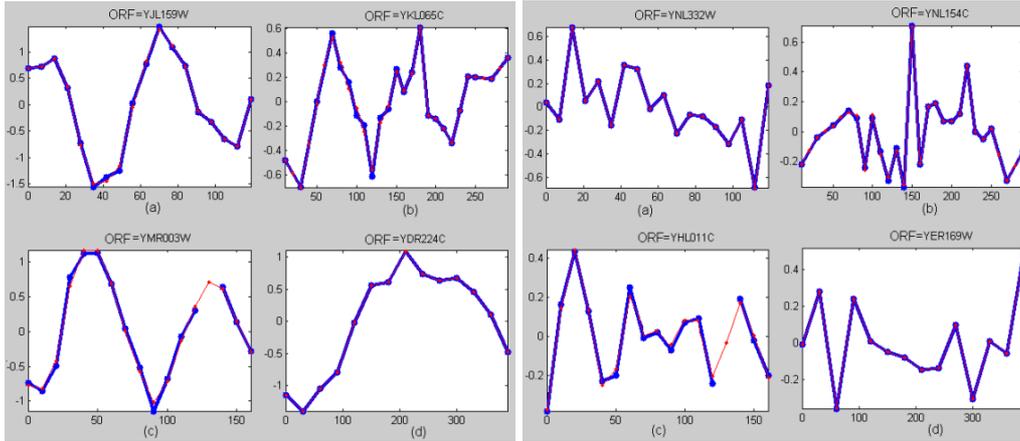


Figure 1. Left: highly periodic expression profiles, and Right: random profiles from Yeast datasets of Spellman et al. [6]. Profiles (a), (b), (c), (d) correspond to the alpha, cdc15, cdc28, elutriation datasets, respectively. The red curves in the figure are the interpolated profiles with missing values filled in (see [5]).

FISHER EXACT TEST OF PERIODICITY

The problem of deciding whether a time series is random or periodic can be cast as a statistical decision problem using hypothesis testing. In [2], Wichert et al. proposed to use the Fisher g -statistic to detect periodic sequence. The test determines whether a peak in the periodogram is significant or not. The test proceeds as follows. Given a time series of length N , the periodogram $I(\omega)$ is first computed as

$$I(\omega) = \frac{1}{N} \left| \sum_{n=1}^N y_n e^{-j\omega n} \right|^2, \quad \omega \in [0, \pi] \quad (1)$$

Furthermore, the periodogram is evaluated at the discrete normalized frequencies

$$\omega_l = \frac{2\pi l}{N}, \quad l = 0, 1, \dots, a \quad (2)$$

where $a = [(N-1)/2]$ and $[x]$ denotes the integer part of x . If a time series has a significant sinusoidal component with a frequency ω_k , then the periodogram will exhibit a peak at that frequency. Fisher derived an exact test of the significance of the spectral peak by introducing the Fisher g -statistic [7]

$$g = \frac{\max_l I(\omega_l)}{\sum_{l=1}^a I(\omega_l)} \quad (3)$$

Under the Gaussian noise assumption, the exact distribution of the g -statistic under the null hypothesis (that the spectral peak is insignificant) is given by

$$P(g > x) = \sum_{k=1}^b (-1)^{k-1} \frac{a!}{k!(a-k)!} (1-kx)^{a-1} \quad (4)$$

where b is the largest integer less than $1/x$ and x is the observed value of the g -statistic. Equation (4) yields a p-value that allows [us](#) to test whether a given time series behaves like a random sequence. Large value of g indicates a strong periodic component and leads us to reject the null hypothesis.

Since there are multiple gene expression profiles to be tested, there is a possibility that a profile can have a small p-value by chance even though it is a random sequence. To correct for multiple testing, Wichert et al. use the method of False Discovery Rate (FDR) [8] to control the expected proportion of false positives at a given rate q . The FDR procedure for the ordered set of p-values in ascending order $p_{(1)}, p_{(2)}, \dots, p_{(G)}$ with corresponding genes $g_{(1)}, g_{(2)}, \dots, g_{(G)}$ is as follows:

- (1) Let i_q be the largest i for which $p_{(i)} \leq \frac{i}{G}q$,
- (2) then reject the null hypothesis for all genes $g_{(1)}, g_{(2)}, \dots, g_{(i_q)}$

STATISTICAL POWER OF TEST

Simulated Signals

To investigate the statistical power of the Fisher exact test, we perform the test for dataset of simulated signals. The simulated signal is given by

$$y(n) = A\cos(2\pi n/T + \phi) + \varepsilon(n) \quad (5)$$

where T is the period, A is the amplitude of the sinusoid, $\phi \in [-\pi, \pi]$ is the phase, $n = 1, \dots, N$, and $\varepsilon(n)$ is Gaussian noise sequence of zero mean and unit variance. This signal model is frequently used to model gene expression profiles that exhibit cyclic behavior. Since the periodogram is invariant to ϕ , we set ϕ to zero for the simulated signals. For Spellman's Yeast cell-cycle datasets [6], the number of time points per period ranges between 8 to 14 and the number of periods covered by the gene expression profiles ranges from 1 to 4. This characteristic is taken into consideration in our simulation experiments.

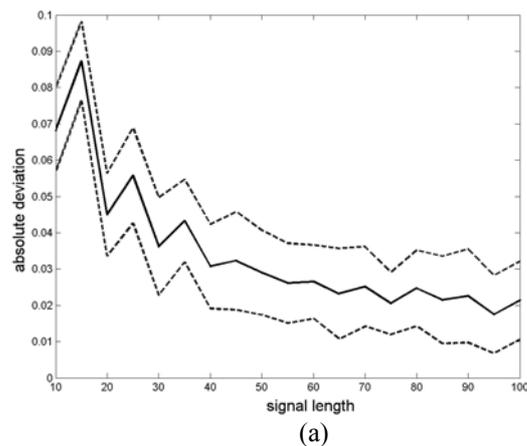
Empirical versus Exact Distribution

Fisher derived the exact distribution in (4) for the g -statistic computed from the periodogram defined in (1). To see whether the empirically calculated pdf of the g -statistic agrees with the exact distribution given by (4), we generated a test dataset consisted of 10000 random signals using (5), where A is set to zero. Multiple test datasets are generated for random signals of length $N = 10$ to 100 with increment of 5, and period $T = 10$. The maximum absolute error between the empirically obtained $\hat{P}(g > x)$ and the exact distribution of (4) is shown in Fig.2a. For each signal length, 30 simulation runs are performed and the mean (μ) \pm 2 standard deviations (σ) are plotted. We see that for short length signals ($N < 40$), the deviation from the exact distribution tends to be significant. In addition, we see that if the signal length is not an integer number of complete periods, the deviation tends to be larger too. However,

this larger deviation due to non-integer number of periods becomes less significant when N is large (i.e. $N > 40$). Figure 2b shows the exact distribution and the empirical distribution for $N = 10$. The deviation from exact distribution can be seen clearly. For larger value of N (i.e. $N > 40$), no significant deviation can be seen from the visual plot.

Although good agreement is obtained between the empirical and exact distribution for large N , we also notice that if a different method of calculating a signal's power spectrum is used, the deviation from the exact distribution can be very significant. Fig.2c shows the exact distribution, the empirical distribution using the periodogram of (1) and from the robust spectral estimator proposed in [4], we see that in the later method, the exact and the empirical distribution deviate significantly. In the situation that N is small or when a different spectral estimator is used, it is necessary that an empirically computed null distribution obtained from a large number of simulated random signals be used in place of the exact distribution during hypothesis testing. In the case of a real dataset, the dataset of random signals can be obtained by random permutation of the time points of each signal to destroy any periodicity in the signal.

For each of the experiments describe below, each test dataset consists of 4000 random signals and 1000 periodic signals (i.e. 20% of total number of signals) embedded in zero mean unit variance iid Gaussian noise. Due to the large deviation from the exact distribution for very short signal length, we resort to empirically generate the null distribution by performing random permutation of the time points of each signal in the test set. Periodicity is then determined by hypothesis testing using the empirical distribution. To correct for multiple testing, the FDR framework as described in [2] are adopted. For each simulation, three quantities are determined: (1) the number of periodic signals detected by the test (L), (2) the number of true positives (TP), and (3) the number of true positives in the top 1000 signals (Z) as ranked by their p-values.



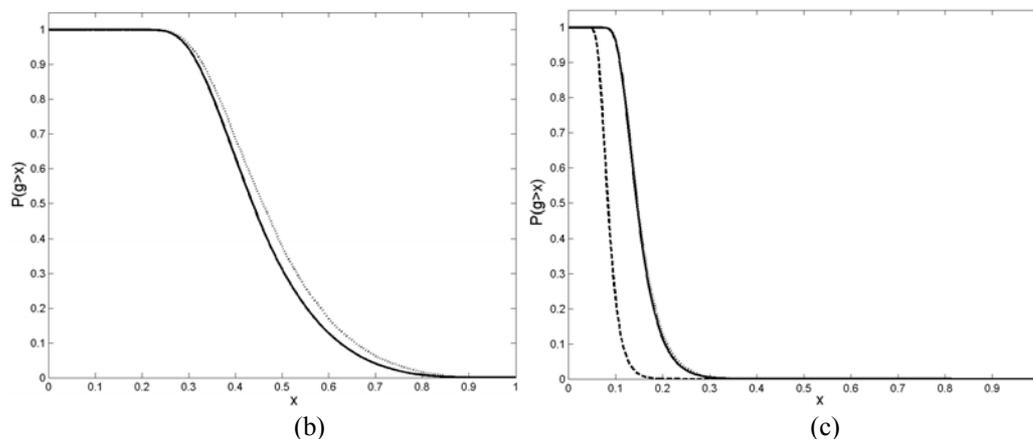


Figure 2. (a) The maximum absolute error between the empirically obtained distribution and the Fisher exact distribution. The dashed curves are $\pm 2\sigma$ around the mean μ (solid curve). (b) Empirically computed distribution (dashed curve) versus theoretical distribution (solid curve) for short length signal ($N = 10$). (c) For $N = 50$, the empirical distribution (dotted line) and the exact distribution (solid curve) is almost indistinguishable. If the robust spectral estimator of [4] is used, the empirical distribution (dashed curve) can deviate significantly from the exact distribution (solid curve).

Power of Fisher Test versus Signal Length

For this experiment, we investigate the effect of signal length on the power of the Fisher test. We set $T = 10$, $A = \sqrt{2}$ (giving a SNR of 0dB), and the number of signal time points N is varied from 10 to 100 (in step of 5) giving a signal spanning 1 to 10 periods. The false discovery rate is set to $FDR = 0.05$. For each N , we performed 30 simulation runs and calculated μ and σ . The results for L , TP , and Z versus N are plotted in Fig. 3. We see that the test is poor if signal length is short and/or signal is not integer number of periods. If the signal length is an integer number of periods, the test is good when $N > 40$. The poor result for signal length of non-integer number of periods is due to the truncation effect (windowing) of finite length signal when computing the periodogram using (1). This truncation effect is still very significant even when $N > 40$. This has important implication when testing for periodicity in real signals of short finite length with unknown period – severe underestimation can occur. The TP versus N plot of Fig. 3(b) is very similar to Fig. 3(a), indicating that most of the signals detected as periodic are indeed periodic. If we count the number of true positives within the top 1000 signals ranked by the g -statistics (or equivalently the p -values), we can see from Fig. 3(c) that even when the test fails, the ranking still returns many of the periodic signals (i.e., at $N = 10$ and at $N = 20$). This ranking has been exploited by us for periodicity detection in gene expression time series data in a recent paper where the Fisher test cannot be reliably performed [5].

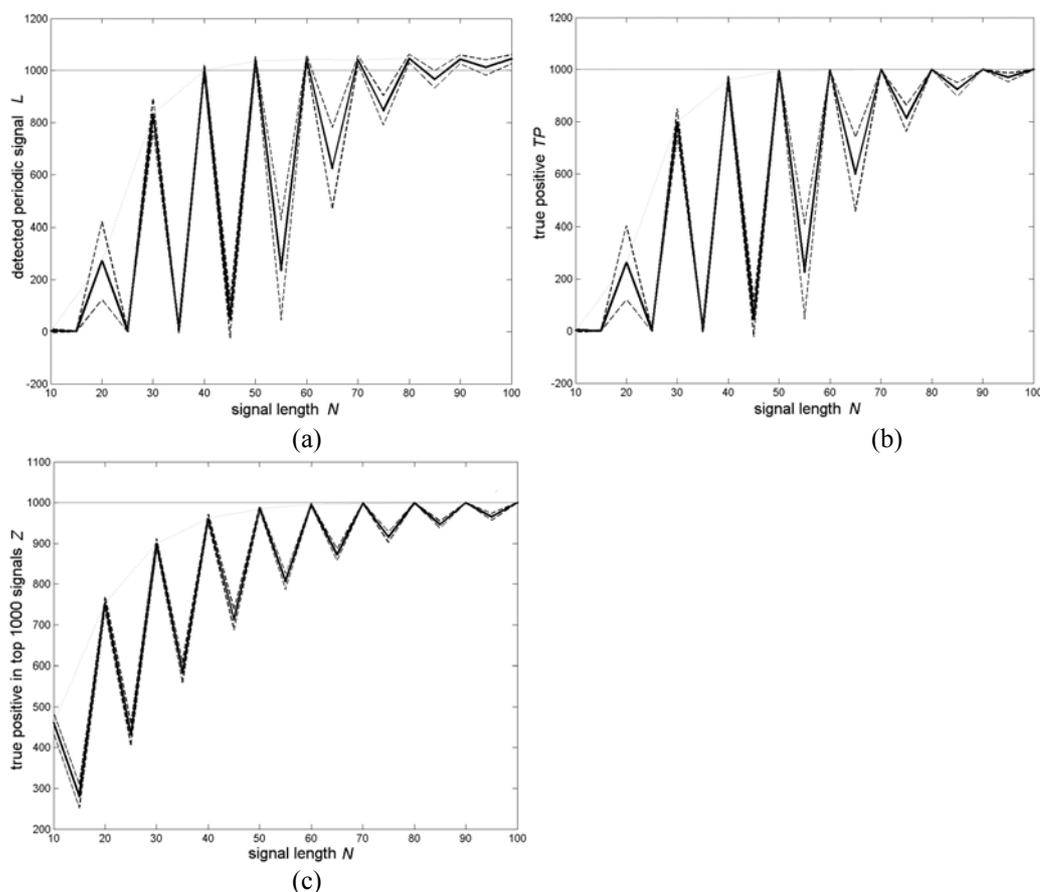


FIGURE 3. (a) Detected periodic signals as a function of signal length N . (b) The number of true positives in the detected periodic signals as a function of N . (c) The number of true positives in the top 1000 ranking as a function of N . In each plots, the dashed curves are $\pm 2\sigma$ around the mean μ (solid curve).

In Table 1, we tabulate the mean value of L , the mean value of TP , and the mean value of Z , for signal length of integer number of periods taken from the graphs in Fig. 3. It can be seen that the power of the test deteriorates significantly when signal length is short. A signal length of around 40 or longer is needed to produce good result when testing for periodicity using the Fisher test.

TABLE 1. The mean value of L , TP , and Z as a function of N taken from Fig. 3.

Signal Length	10	20	30	40	50	60	70	80	90	100
L	4	271	832	1001	1037	1043	1042	1046	1044	1044
TP	2	262	798	960	994	999	1000	1000	1000	1000
Z	461	754	899	961	985	995	998	999	1000	1000

We also investigate whether the power of the test is affected by the sampling rate of the underlying sinusoid. For this we set the signal length $N = 60$, but vary the period $T = \{10, 20, 30\}$. The other parameters are: $A = \sqrt{2}$, $FDR = 0.05$. We perform 30 simulation runs for each T , with each dataset consists of 4000 random signals and 1000 periodic signals embedded in Gaussian noise. The results are tabulated in Table 2. We see that higher sampling rate reduces the variance of the periodicity test.

TABLE 2. The $\mu \pm \sigma$ for L as a function of signal period T for a signal of length 60.

Signal Period	10	20	30
L	1043 ± 8.88	1043 ± 8.26	1043 ± 5.40

Power of Fisher Test versus SNR

We consider signals with different SNR ranging from -3dB to 3dB, with step of 1 dB. The amplitude A of the signals is set to give the required SNR by using the fact that the signal power of a sinusoid is given by $A^2/2$ and the noise power of zero mean unit variance iid Gaussian noise is 1. The other parameters of the simulated signals are: $T = 10$, $N = 50$. FDR is set to 0.05. For each SNR, we performed 30 simulation runs and calculated μ and σ . The results for L , TP , and Z versus SNR are plotted in Fig. 4. We see that the power of the test improves with higher SNR. For this experiment, a SNR of 0dB (i.e., $A = \sqrt{2}$) or higher gives very good result. However, even for very noisy signals (with SNR as low as -3dB), the power of the test is still quite high. For SNR = -3dB, the test was able to detect 748 periodic signals, of which 715 are true positives.

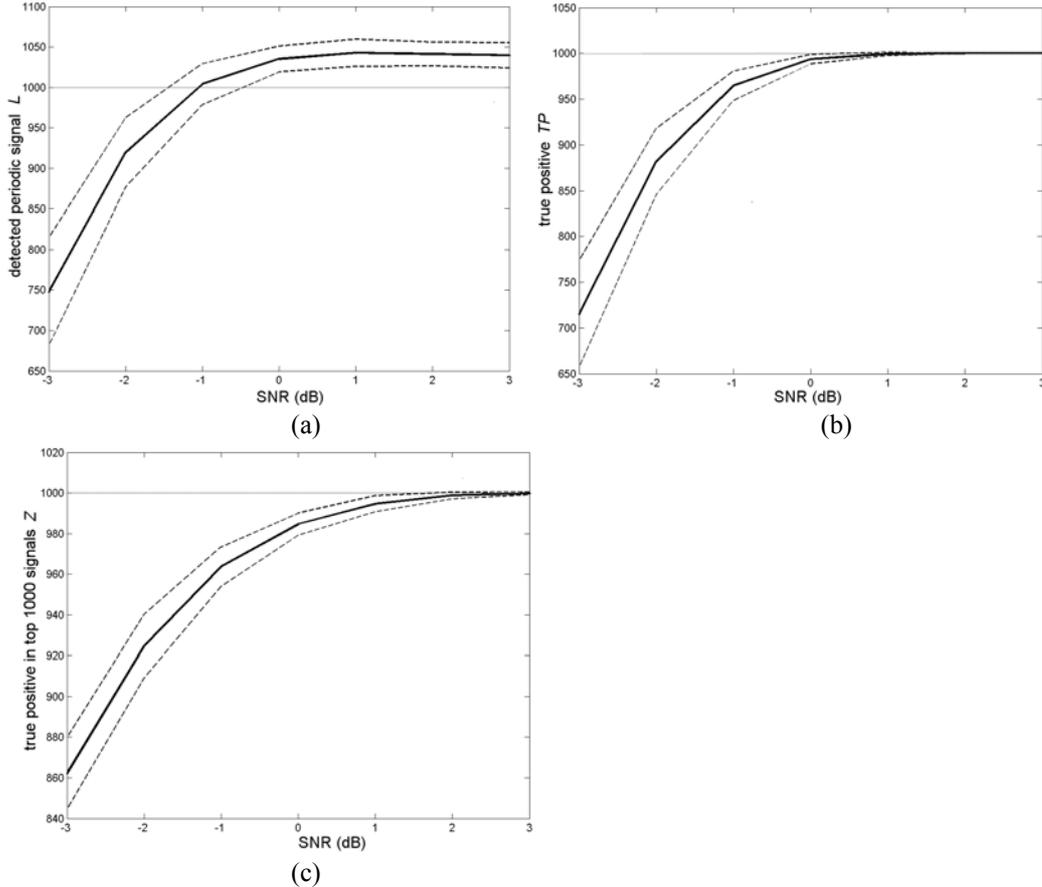


FIGURE 4. (a) Detected periodic signals as a function of SNR. (b) The number of true positives in the detected periodic signals as a function of SNR. (c) The number of true positives in the top 1000 ranking as a function of SNR. In each plots, the dashed curves are $\pm 2\sigma$ around the mean μ (solid curve).

Power of the Fisher Test versus FDR

In this experiment, we investigate the power of the test when different FDR are used. The signal parameters are: $T = 10$, $A = \sqrt{2}$, $N = 50$. For each FDR, we performed 30 simulation runs and calculated μ and σ . The results for L , TP , and Z versus FDR are plotted in Fig. 5. As FDR increases, the number of periodic signals detected (L) also increases as expected, since the number of false positives increases when FDR increases. However, looking at Figs. 5(a) and 5(b), we see that at FDR beyond 0.04~0.05, most of the increase in L is due to the increase in false positives. Note that since Z is independent of FDR, the plot in Fig.5(c) shows constant Z for different FDR.

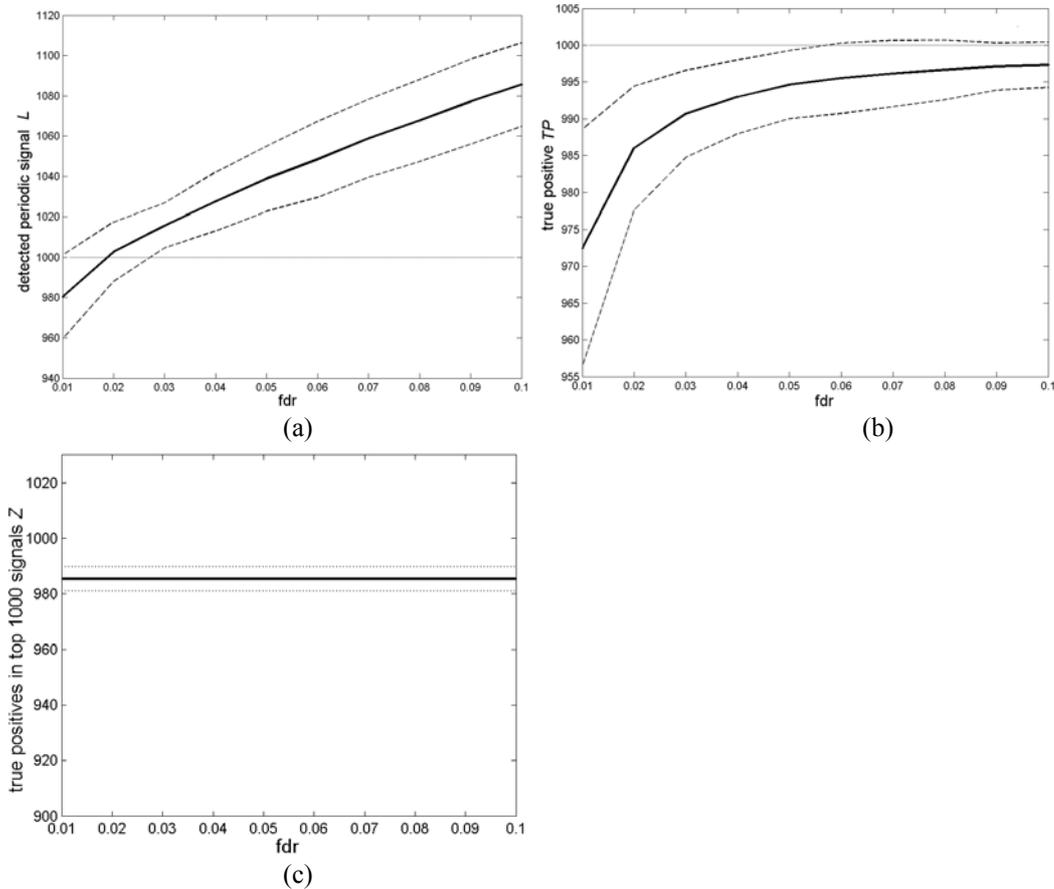


FIGURE 5. (a) Detected periodic signals as a function of FDR. (b) The number of true positives in the detected periodic signals as a function of FDR. (c) The number of true positives in the top 1000 ranking as a function of FDR. In each plots, the dashed curves are $\pm 2\sigma$ around the mean μ (solid curve).

CONCLUSIONS

In this paper, we have investigated the statistical power of the Fisher exact test for periodicity in a finite length signal by extensive simulation experiments. Although the theoretical null distribution was derived by Fisher for Gaussian noise, we found that deviation from it can be significant when signal length is short. Moreover, when the

signal does not cover an integer number of periods, significant drop in the statistical power of the test was observed. In this case, a much longer signal is needed for the test to return reliable result. We found that Fisher test is relatively robust to noise. We also investigate how the FDR multiple testing correction strategy affects the number of detected periodic signals. Although the Fisher test may be unreliable for short signal, the Fisher g -statistic has been observed to provide a useful ranking of periodic signals. All these findings have important implications for periodic gene expression profiles detection as these profiles are often noisy, of very short length, and often with unknown periodicity. In high likelihood, the number of periodic gene expression profiles can be severely underestimated for short length signal as is found with many of the publicly available gene expression datasets.

ACKNOWLEDGMENTS

This work is supported by a grant from the Hong Kong Research Grant Council (Project CityU122005).

REFERENCES

1. M.L. Whitfield, G. Sherlock, A.J. Saldanha, J.L. Murraray, C.A. Ball, K.E. Alexander, J.C. Matese, C.M. Perou, M.M. Hurt, P.O. Brown, and D. Botstein, *Mol. Biol. Cell* **13**, 1977–2000 (2002).
2. S. Wichert, K. Fokianos, and K. Strimmer, *Bioinformatics* **20**, 5-20 (2004).
3. J. Chen, *BMC Bioinformatics* **6**, 286-297 (2005).
4. M. Ahdesmäki, H. Lähdesmäki, R. Pearson, H. Huttunen, and O. Yli-Harja, *BMC Bioinformatics* **6**, 117 (2005).
5. A.W.C. Liew, J. Xian, S. Wu, D. Smith, and H. Yan, *BMC Bioinformatics* **8**,137 (2007).
6. T.S. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, B. Futcher, *Mol. Biol. Cell* **9**, 3273-3297 (1998).
7. R.A. Fisher, *Proc. R. Soc. A* **125**, 54–59 (1929).
8. Y. Benjamini, and Y. Hochberg, *JR Statist. Soc. B* **57**, 289–300 (1995).