Bioinformatics and Biology Insights

METHODOLOGY

Periodicity Detection Method for Small-Sample Time Series Datasets

Daisuke Tominaga

Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, Aomi 2-4-7, Koto, Tokyo, 135-0064, Japan. Corresponding author email: tominaga@cbrc.jp

Abstract: Time series of gene expression often exhibit periodic behavior under the influence of multiple signal pathways, and are represented by a model that incorporates multiple harmonics and noise. Most of these data, which are observed using DNA microarrays, consist of few sampling points in time, but most periodicity detection methods require a relatively large number of sampling points. We have previously developed a detection algorithm based on the discrete Fourier transform and Akaike's information criterion. Here we demonstrate the performance of the algorithm for small-sample time series data through a comparison with conventional and newly proposed periodicity detection methods based on a statistical analysis of the power of harmonics.
We show that this method has higher sensitivity for data consisting of multiple harmonics, and is more robust against noise than other methods. Although "combinatorial explosion" occurs for large datasets, the computational time is not a problem for small-sample datasets. The MATLAB/GNU Octave script of the algorithm is available on the author's web site: http://www.cbrc.jp/%7Etominaga/piccolo/.

Keywords: periodicity detection, gene expression time series, information criterion, discrete Fourier transform, circadian rhythm

Bioinformatics and Biology Insights 2010:4 127–136

doi: 10.4137/BBI.S5983

This article is available from http://www.la-press.com.

Introduction

Life phenomena are observed as changes in time, and many of these phenomena, such as circadian rhythm and the cell cycle, exhibit periodic behavior. These phenomena are common to many species and are thought to be expression of essential mechanisms of life. In addition, irregular periodicity is caused by abnormal stimuli or disorder of these mechanisms. Thus, a periodicity detection technique for time series observation data is important in many areas of biology and medicine.

Generally, the observation of life phenomena incurs certain costs and thus the number of sampling points in time is often small, as in the case of DNA microarray data.[1] For this reason, a reliable method of periodicity detection is needed for small datasets.

Time series data on life phenomena can be represented by a mathematical model consisting of noise and various simple formulae, such as polynomial functions or harmonics (sinusoidal functions).[2] A model of time series data of periodic phenomena should contain harmonics. If these harmonics are judged significantly large by a statistical test, the phenomena can be considered periodic.

Generally, life phenomena are the result of complex interactions of biological networks (gene regulatory networks, metabolic pathways, signal transduction networks, etc.); thus, time series data on constituents of these networks can contain multiple harmonics with different periods.

Periodicity detection techniques which are widely used can be classified into two categories: 1) model fitting in the time domain, and 2) statistical significance tests on power spectra.

The first category includes methods based on direct curve fitting to the observed data. When the data can be modeled by n harmonics, the number of parameters that are optimized by the fitting method is $3n + 1$,[3] which is too many parameters for small datasets.[4]

Methods in the second category are widely used in many area of science. The basic method is to calculate the power spectra by using the discrete Fourier transform (DFT) or the autocovariance matrix, and then to test the significance of each spectrum of a harmonic by outlier detection methods.[3] A simple method uses quantiles of spectra to detect outliers. Both parametric tests (such as Dixon's Q test or Fisher's G test) and non-parametric tests (eg, the quantile/box-plot) are used to detect outliers. These methods frequently do not detect the periodicity of interest if the significance of the period is close to that of other periods, even if the significance is high. Thus, these methods are inadequate for data with multiple periodicities. In addition, a certain number of spectrum elements are needed to make the outlier tests meaningful and robust to noise. Thus, these methods are not optimal for small sampled time series data.

Other advanced algorithms, such as wavelet-based methods[5,6] and model fitting using directional statistics[7] have been proposed, however, few applications of these methods have been reported to date; therefore, their utility for the analysis of small sampled biological datasets remains an open question.

An clustering method[1] and an AR (autoregression) model based periodicity detection method[8] are developed to be special for small sampled data. The first one classifies genes by expression time series but do not detect period or periodicity. The second one can find a period and its P-value for each time series data, but do not detect multiple harmonics, ie, do not find 'the second significant period'.

Our previously proposed method, called the 'piccolo',[9] consists of the DFT and Bayesian Information Criterion (BIC),[2] and is not based on an outlier detection. The algorithm is a exhaustive search to find the best combination of Fourier coefficients in terms of the information criterion.[4] The combinatorial search does not require a long computational time for most DNA microarray time series datasets found on the web, such as the datasets in the Gene Expression Omnibus[10] and ArrayExpress.[11]

We improve the peridicity detection performance of the piccolo method by introducing Akaike's Information Criterion (AIC)[4] instead of BIC, and demonstrate its performance through a comparison with two conventional methods, one newly developed method and the old version of our method (BIC version of the piccolo) on two simulation datasets and twelve microarray datasets. The piccolo algorithm (new AIC version) is shown to be highly sensitive and robust against noise on simulated short time series data which consist of multiple (two) harmonic signals and noise. In addition, the present method can achieve high detection rates of

a period of interest for DNA microarray datasets, thus satisfying the expectations for biological data.

Methods

We choose two widely used conventional methods, one recently proposed methods and the older version of the piccolo method for comparison with the improved new piccolo method. The methods selected for the comparison except the old version of the piccolo are based on statistical tests on the logarithms of power spectra.

The two conventional methods are the quantile method and Dixon's Q test. The other method is a non-parametric test for significance of the logarithms of power spectra, recently proposed by Ahdesmäki et al.[12]

Dixon's Q test requires the assumption that the distribution of samples is normal. The other four methods, namely, the quantile method, Ahdesmäki's method, and the old and new piccolo method, do not require this assumption. In the piccolo method (both old and new), the error distribution at each data point (sampling points at various times in the time series data) is assumed to be normal and its variance is assumed to be the same as that of the data.[4]

According to results of statistical tests for normality of powers and logarithms of powers of each time series in all datasets (Tables 1 and 2), no conclusion can be reached regarding the distribution of powers and logarithms of powers. Note that the power of a harmonic is calculated as the product of its Fourier coefficient, which is calculated by DFT, and the complex conjugate of the coefficient; therefore, the power of a harmonic is a real value. Logarithms of powers are perhaps more suitable than the values of powers themselves for the quantile method and Dixon's Q test, considering histograms of logarithms of powers for each dataset (Fig. 1). Accordingly, the quantile method and Dixon's Q test method are applied to logarithms of powers.

Quantile method

Outlier detection using an inter quantile range (IQR) is a basic and widely used technique in many scientific fields because it has been found empirically to be useful for outlier elimination.[13,14]

In the quantile method, the DFT is applied to the data, and the power of each harmonic is calculated as the product of its Fourier coefficient and its complex conjugate. Then, quantile points of the logarithm of the powers and IQR are calculated. All logarithms of powers are compared with the outlier bound, which is the sum of the third quartile point (75 percentile point) and the IQR multiplied by 1.5 (for normally distributed samples, this is same as that critical value is 0.9541 one-sided). If a logarithm of a power is larger than the outlier bound, the harmonic corresponding to the power is significant, and thus the given time series data is considered periodic. The periodicity of the time series is the same as the significant harmonics.

The quantile method requires a sample size (data length) of 8 or more for power spectra. Power spectra (real numbers) are calculated from Fourier coefficients (complex numbers), which have symmetry; thus, the number of unique samples of the spectra is half the data length. The unique samples size is $(n-1)/2$ for an odd data length n. If the number of samples is less than 4, the quantile method cannot detect any outliers because the bound is larger than the largest sample. Thus, this method cannot be used for time series data with data length of 7 or less.

Dixon's Q test

Dixon's Q test[15,16] is a widely used outlier detection algorithm, in which the sample distribution is assumed

Table 1. *P*-values and standard deviations (sd) of the normality test for the distribution of powers and logarithms of powers of time series data in simulation datasets for the one-harmonic and two-harmonic conditions.

	N	T	Int.	Power	Log of power
One harmonics	500	12	4	0.755 (sd: 0.204)	0.862 (sd: 0.175)
Two harmonics	500	12	4	0.771 (sd: 0.218)	0.855 (sd: 0.165)

Notes: Signal to noise ratio (R_{SN}) is 0.1. *P*-values are calculated using the Kolmogorov-Smirnov test.
Abbreviations: N, number of time series data in each dataset; T, length of each time series in the dataset; Int., interval between each two samplings (h) in each time series.

Table 2. *P*-values and their standard deviations (sd) for the normality test of the distribution of powers and logarithms of powers of time series data in datasets taken from the Gene Expression Omnibus database.

	N	T	Int.	Power	Log of power
GDS1629	6346	8	6	0.898 (sd: 0.127)	0.911 (sd: 0.112)
GDS2110	14904	6	4	0.936 (sd: 0.0712)	0.936 (sd: 0.0712)
GDS2232	29109	12	4	0.845 (sd: 0.172)	0.828 (sd: 0.182)
GSE3424	22759	6	4	0.922 (sd: 0.0751)	0.936 (sd: 0.0719)
GDS404	6484	12	4	0.871 (sd: 0.153)	0.865 (sd: 0.159)
GSE6542-1	11699	6	4	0.936 (sd: 0.0712)	0.936 (sd: 0.0712)
GSE6542-2	11699	6	4	0.936 (sd: 0.0720)	0.936 (sd: 0.0718)
GSE6542-3	11699	6	4	0.945 (sd: 0.0682)	0.944 (sd: 0.0681)
GSE6542-4	11699	6	4	0.931 (sd: 0.0735)	0.932 (sd: 0.0735)
GSE6542-5	11699	12	4	0.877 (sd: 0.145)	0.875 (sd: 0.147)
GSE6542-6	11699	6	4	0.937 (sd: 0.0714)	0.936 (sd: 0.0715)

Note: *P*-values are calculated using the Kolmogorov-Smirnov test.
Abbreviations: N, number of time series data in each dataset; T, length of each time series in the dataset; Int., interval between each two samplings (h) in each time series.

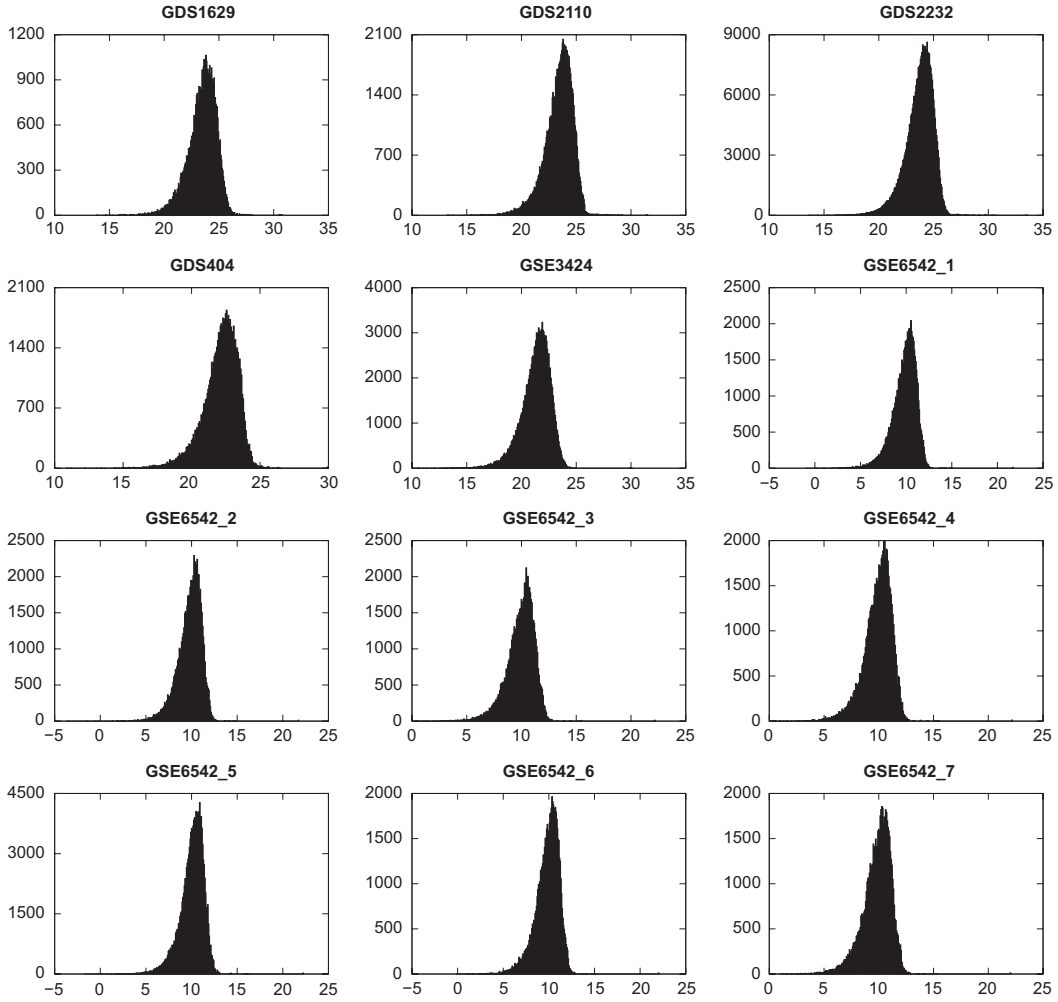

Figure 1. Histograms of logarithms of powers for all twelve DNA microarray datasets for the performance comparison in the result section. The x axes are bins of histograms. Each bin is a range of natural logarithms of powers of each time series in each dataset. The y axes are frequencies of logarithms of powers in each range. The sum of all frequencies are same as the number of probes in each dataset.

to be normal. This test, which ignores redundant information from half of a two-sided power spectrum, is used to detect outliers from a set of logarithms of power spectra. The criterion of the test is a critical value of 0.95, one-sided.[17]

Ahdesmäki's method

The Ahdesmäki's method[12] uses the kernel density estimation[18] of the distribution of the square root of the targeted harmonic's power (proportional to the logarithm of power). The distribution is approximated by shuffling the order of samples in the time series data and calculating the power of the targeted harmonic by least-square fitting. We use Yi Cao's 'gkde' kernel density estimation method*[1] to calculate the approximate probability density function (PDF) of powers of the harmonics, and we use the built-in function 'ols' in GNU octave version 3.2.3*[2] to calculate the power of the harmonic.

The criterion of the test is a critical value of 0.95, one-sided.

The piccolo method

The 'piccolo' algorithm[9] is an exhaustive search for the optimal combination of Fourier coefficients calculated by DFT from a given time series data. The algorithm searches for all possible subsets of conjugate pairs of Fourier coefficient, but the search range for a size of subsets is limited to keep the information criterion value (AIC, BIC, etc.) reliable.[4]

Our previously presented version of the method incorporates BIC (Bayesian Information Criterion) as the information criterion. Here we introduce AIC (Akaike's Information Criteriron) instead of BIC to improve detection performance. The previous version is called 'piccolo/B' in this paper. The 'piccolo' implies new AIC version.

The optimal subset is defined such that the AIC value calculated from the subset and given data is minimal. AIC is used as the information criterion under the assumptions that the error distribution of the datum at each time point is normal and that its variance is the same as the variance among the time series data.[4]

*[1]http://www.mathworks.com/matlabcentral/fileexchange/19160
*[2]http://www.gnu.org/software/octave/doc/interpreter/Linear-Least-Squares.html

The model is a subset of the set of the Fourier coefficients obtained by DFT from given time series data. The number of Fourier coefficients is n when n is the number of samples in the time series; however, half of these coefficients are complex conjugates of the other half. A Fourier coefficient must always be selected with its conjugate. This allows the inverse DFT of the model to be real numbers, which is necessary to calculate the AIC. Thus, the number of model parameters is the number of the coefficient pairs. When the data length is even, a coefficient corresponds to the Nyquist frequency is a pure real number and its complex conjugate do not appear in the set of Fourier coefficients in the model. This coefficient does not form a pair when it is chosen.

The AIC value is calculated using the following equation:[4]

$$\mathrm{AIC} = n \log(2\pi) + n \log(\sigma^2) + n + 2p, \qquad (1)$$

where n is the number of samples (data length of the time series), σ is the variance of errors between the given time series data and the time series calculated from the model by inverse DFT, and p is the number of parameters (pairs of Fourier coefficients) in the model.

In the piccolo method, the Fourier coefficients in the subset that minimize the AIC value are taken to be significant constituents to represent the given data. Accordingly, periods corresponding to these Fourier coefficients are considered significant, and the given time series data judged to be periodic with periods corresponding to these Fourier coefficients. Thus, multiple periods can be found simultaneously even if their powers are close each other.

Result

Fourteen datasets are used to compare the five methods for periodicity detection, comprising two simulated datasets and twelve DNA microarray datasets taken from an online database.

Robustness against noise

Data

We tested the robustness against noise of the five periodicity detection methods, namely the quantile method, Dixon's Q test, Ahdesmäki's method, the piccolo/B and the piccolo method, using simulation

data consisting of one or two harmonic signals and log-normal noise.

Considering that the distribution of DNA microarray data is log-normal,[19] each datum is created as a sum of a log-normally distributed random number and the value of one harmonic (the one-harmonic condition), or two harmonics (the two-harmonic condition). For both conditions, 15 datasets are generated by changing the signal-to-noise ratio as follows:

$$\mathrm{logN}(0,1) + A_i \cos\left(\frac{2\pi}{24}t + C_i\right) + B_i \cos\left(\frac{2\pi}{16}t + D_i\right)$$

where logN(0,1) is log-normaly distributed random noise whose mean is 0 and variance is 1, i (i = 1, …, 500) is the suffix for time series, A_i and B_i are amplitude of harmonic signals whose period are 24-hour and 16-hour respectively (B_i = 0 for the one-harmonic condition), and C_i and D_i are phase of each signal. Values of A_i and B_i are detemined by log-normal random numbers according to the signal-noise ratio (described later). Values of C_i and D_i are detemined by uniformly distributed random numbers within the range of [0,48]. t is the time. Values of time are discrete and its intervals are fixed to 4 hours. The number of time points is 12. Each dataset consists of 500 time series data (each time series data consists of twelve sampling points).

The signal-to-noise ratio (R_{SN}), which is defined as a ratio of the variance of signal and noise, is set at various values of R_{SN} = (0.001, 0.002, 0.005, …, 50.0) under each condition. Thus, all generated time series data in all datasets contain a circadian rhythm.

To consider whether or not Dixon's Q test is appropriate, the normality of distribution of the spectra and the logarithms of spectra is tested. The P-values calculated by the Kolmogorov-Smirnov test for datasets of R_{SN} = 0.1 under both conditions are shown in Table 1. For both spectra and logarithms of spectra, the null hypothesis (the distribution is normal) cannot be rejected at the 90% confidence level. Thus, Dixon's Q test cannot be considered inappropriate.

Detection performance

The numbers of detected time series from the datasets are compared to evaluate the robustness of the methods against noise.

In the two-harmonic condition, simulation data consist of two signals (16 and 24 hour harmonics) and noise, however, the three detection method except piccolo and piccolo/B can hardly detect plural signals simultaneously in principle. Therefore we tested the five methods on detection of a 24-hour siginal.

Plots of the number of detected time series data on each dataset are shown in Figure 2. For both the one-harmonic and two-harmonic conditions, the piccolo method achieved a high detection rate, especially for noisy (low R_{SN}) data.

The detection performance was relatively lower at R_{SN} = 1.0 in the one-harmonic condition except for the piccolo method. In this dataset, the variance of the signal and noise is the same; thus, the signal and

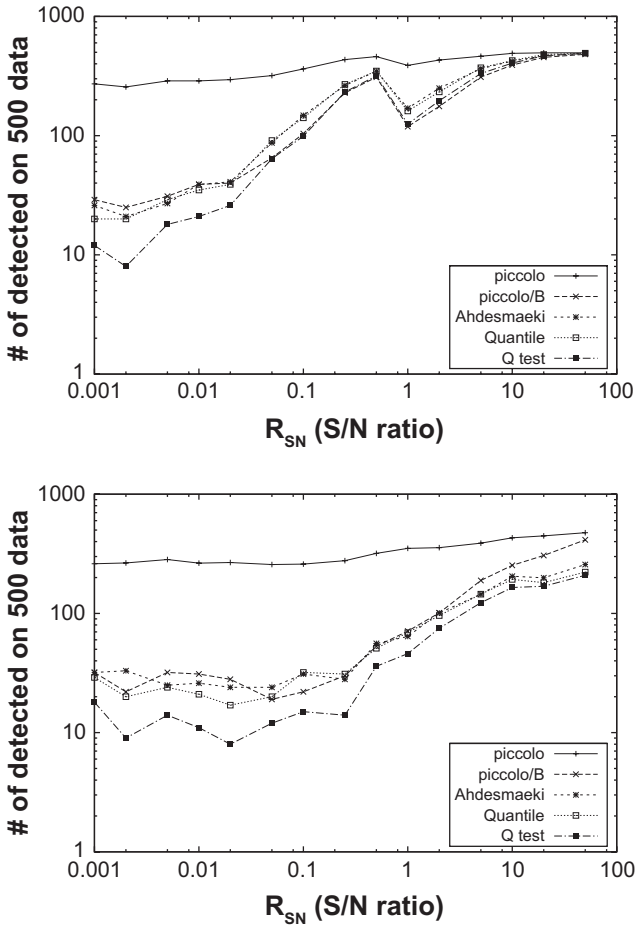

Figure 2. Log/log plots of the signal-to-noise ratio versus the number of detected time series data out of 500. The time series data consist of log-normal random noise and a harmonic (above), and log-normal random noise and two harmonics (below). Since all simulated data contains periodic signal to be detected, the possible maximum number of the detection is 500. The R_{SN} is defined as a division of the variance of the signal by the variance of the noise. Therefore smaller R_{SN} value of the simulated time series means that it is noisy data.

noise are difficult to distinguish, especially for small sampled time series data.

The number of detected time series in the two-harmonics condition is lower than that in the one-harmonic condition for $R_{SN} > 0.01$. The difference between the one-harmonic condition and two-harmonic condition is smaller for the piccolo method than for the other methods.

Detection of circadian rhythm

Data

The five detection methods are applied to experimentally observed DNA microarray data taken from the Gene Expression Omnibus online database by NCBI, NIH,[10] to detect genes (probes) which have 24-hour periodicity, or 'circadian rhythm'.

The *P*-values obtained by the Kolmogorov-Smirnov test for the normality of the distribution of powers and logarithms of powers are shown in Table 2. *P*-values are calculated for time series data in datasets, and means and standard deviations of the *P*-values are calculated and listed in the table. For both powers and logarithms of powers, the null hypothesis (the distribution is normal) cannot be rejected at the 95% confidence level. Although the samples sizes are small (6 to 12), it can be said that Dixon's Q test cannot be considered inappropriate.

It is not defined whether or not the time series in the datasets are circadian; however, some of them are labeled with the GO term[20] 'circadian rhythm'. Here, detection performance is evaluated in terms of the total number of detected probes and the number of detected probes labeled 'circadian rhythm' for each dataset. The quantile method cannot be used on datasets in which the data length of each time series is 7 or less.

Biological description of datasets

All twelve DNA microarray datasets are time series observations intending to analyze circadian rhythm. GDS1629 is a set of fourty five samples of a immortalized suprachiasmatic nucleus cell line of normal rat for 42 hours, every 6 hours (eight time points). The dataset contains five or six samples for each time point. We only use one of them, whose sample ID is the largest. GDS2110 is a set of six samples of normal Macaca mulatta adult females adrenal glands for 20 hours, every 4 hours (six time points). GDS2232 is a set of twenty four samples of normal mouse adrenal glands for 44 hours, every 4 hours (twelve time points). The dataset contains two samples for each time point. We only use one of them, which appears earlier in the published data file. GDS404 is a set of thirteen samples of normal mouse aortae for 44 hours, every 4 hours (twelve time points). The dataset contains two samples for the first time point. We only use one of them, which appears earlier in the published data file. GSE3424 is a set of eight samples of normal Arabidopsis thaliana for 20 hours, every 4 hours (six time points). The dataset contains two samples for two time points (0-hour and 12-hour). We only use one of them, which appears earlier in the published data file. GSE6542 is a set of fourty eight samples of three mutants of Drosophila melanogaster in two experimental conditions (seven conditions in total). We divide it into seven sub-datasets here. Six sub-datasets consist of six time points and one consists of twelve points. All these datasets are normalized by publishers for further analysis.

Data of duplicate probes for same gene and data of probes which contain a numerically invalid value are ignored for this performance comparison.

Detection performance

The detection results are shown in Table 3. For both the total number of detected probes and the number of detected probes labeled circadian, the piccolo method is superior to the other four methods, including previous version of the piccolo (piccolo/B), for all datasets. Ratios of *S* in Table 3, which is the number of probes detected by the piccolo method but not by other four methods, to the number of total probes in each dataset are 0.333 (GDS1629) to 0.776 (GSE6542_3). This means that using the piccolo method we find that 77.6% of all probes in GSE6542_3 are under the influence of circadian oscillation mechanisms but other four methods cannot detect these probes.

On the other hand, ratios of the numbers of probes detected by one or more of the other four methods but not detected by the piccolo method to the number of total probes in each dataset are in the range of 0.0 (GSE6542_2, GSE6542_4, GSE6542_6) to 0.0418 (GDS404), or less than 5% (data not shown).

Table 3. Results of detection of circadian oscillation on the twelve DNA microarray datasets. Numbers before and after a slash are the number of detected probes and detected circadian annotated probes respectively. The annotated probes are labeled with the GO term 'circadian rhythm' in the chip definition files of the microarrays.

	C	Quantile	Q test	Ahdesmäki	Piccolo/B	Piccolo	S
GDS1629	22	146 / 1	60 / 0	121 / 0	163 / 1	2231 / 7	1981
GDS2110	26	–	667 / 0	457 / 2	0 / 0	10658 / 16	9745
GDS2232	37	4005 / 1	3053 / 3	5343 / 4	9118 / 11	23057 / 28	10892
GSE3424	33	–	2853 / 1	1436 / 2	0 / 0	19233 / 29	15837
GDS404	12	714 / 2	497 / 2	655 / 3	752 / 3	4044 / 6	2670
GSE6542-1	28	–	529 / 0	401 / 1	0 / 0	8554 / 23	7829
GSE6542-2	28	–	413 / 0	339 / 2	0 / 0	7706 / 23	7118
GSE6542-3	28	–	720 / 0	340 / 1	0 / 0	9939 / 23	9099
GSE6542-4	28	–	495 / 0	361 / 0	0 / 0	8924 / 23	8235
GSE6542-5	28	799 / 3	623 / 0	656 / 5	1046 / 4	6038 / 19	4335
GSE6542-6	28	–	534 / 0	378 / 1	0 / 0	8513 / 20	7800
GSE6542-7	28	–	501 / 0	403 / 0	0 / 0	8236 /19	7517

Notes: C is the number of circadian probes in the chip used for each dataset (duplicate probes for each gene and probes containing invalid numerical data are omitted). S is the number of probes detected only by the piccolo method but not by other four methods.

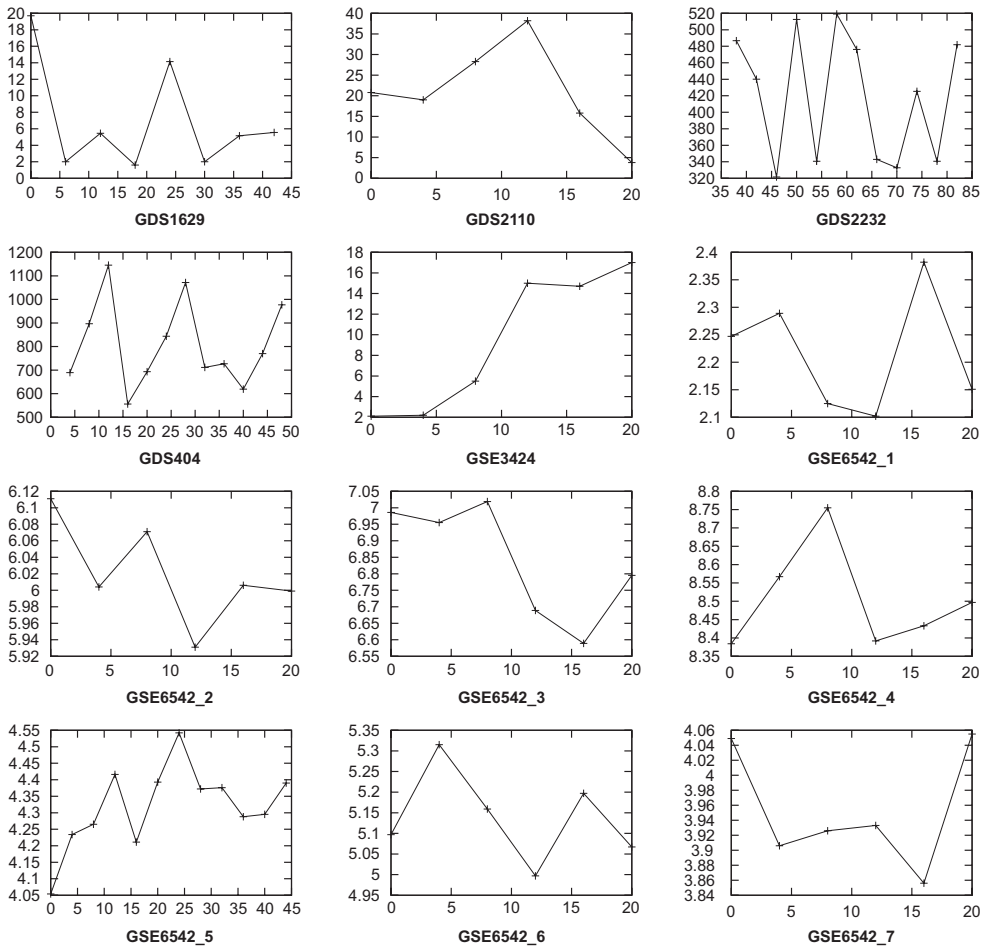

Figure 3. Plots of time series data which are detected only by the piccolo method and not by the other four methods. For each dataset, the time series data of the probes with the largest ratio between the maximum power and the second largest power is plotted. Ranges of sampling time points are different by datasets. Datasets and its time ranges are: Top (left to right)—GDS1629 (44 h), GDS2110 (20 h), GDS2232 (44 h), Second (left to right)—GDS404 (44 h), GSE3424 (20 h), GSE6542_1 (20 h), Third (left to right)—GSE6542_2 (20 h), GSE6542_3 (20 h), GSE6542_4 (20 h), Bottom (left to right)—GSE6542_5 (44 h), GSE6542_6 (20 h), GSE6542_7 (20 h).

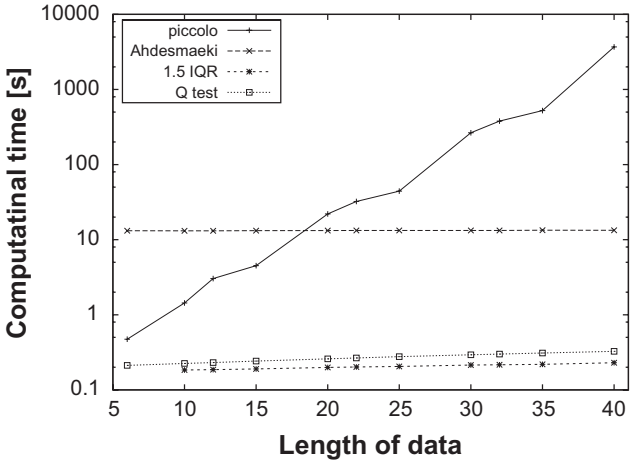

Figure 4. Plot of computational time which is needed to perform detection on 500 time series data. The x axis is the length of time series data (the number of time points). The y axis is elapse CPU time in second to perform detection in a logarithmic scale. 500 time series data are generated by normally distributed random numbers. The CPU time of piccolo/B method (previous version of the piccolo method, not shown here) is very similar to the piccolo method which incorporates AIC.

The time series of a probe detected by only the piccolo method is plotted in each panel in Figure 3 (one probe is chosen for each dataset).

Computational cost

We measured the increase in computational time required to perform detection on 500 time series when the data length of each time series is increased from 6 to 40. The dataset consist of normally distributed random numbers with a mean of 0 and variance of 1. The results are shown in Figure 4. In the performance evaluation, all detection programs are run on GNU octave version 3.2.3[*3] on Mac OS X 10.6.3, and the computer is equipped with two 3 GHz Dual-Core Intel Xeon and 8 GB of 667-MHz DDR2 core memory.

The computational time of the quantile method, Dixon's Q test and Ahdesmäki's method increase linearly with increasing data length. This increase is exponential in the case of the piccolo method. The CPU time of the piccolo/B is almost same to the piccolo and not shown here.

The curves fit to data, $ax + b$ for Ahdesmäki's method and $\exp(ax + b)$ for piccolo method, intersect at $x = 18.8$ (x is data length). The piccolo method is faster than Ahdesmäki's method for small datasets with a data length of less than 19.

[*3]http://octave.sourceforge.net/

Discussion

Five methods for periodicity detection, namely, two simple methods (the quantile method and Dixon's Q test), one recently proposed method (Ahdesmäki's method) and two methods by the authors (piccolo and piccolo/B) are compared for small sampled (short length) time series of two simulated datasets which consist of twelve time points and twelve sets of experimentally observed DNA microarray data, which consist of 6, 8, 12 time points for observation of the circadian rhythm.

Dixon's Q test requires the assumption that the distribution of samples is normal. *P*-values of the normality of the distribution of the spectra and logarithm of spectra of each time series in the given datasets were calculated by the Kolmogorov-Smirnov test. The null hypothesis (the distribution is normal) was not rejected for the logarithms of spectra of all datasets.

The piccolo method selects significant harmonics to model the data. Harmonics included in the best model that minimizes the AIC are significant. A harmonic whose power is not a maximum can be detected as significant more frequently by using the piccolo method compared with other outlier based methods. These smaller power harmonics are selected according to the AIC and therefore are considered to be significant statistically. The high detection sensitivity of the piccolo method is shown by results of analyses using both simulations and experimentally observed data. These results satisfies the expectations that most genes in a living cell are involved in one or more gene regulatory networks and that these networks are interconnected. The oscillation of the core circadian clock genes are expected to spread over whole gene networks.

S in Table 3 shows that many genes exhibiting periodicity in the form of circadian rhythm can be detected only by the piccolo method and not the other four methods. This finding can be attributed to the magnitude of circadian periodicity, which is thought to depend on the 'distance' in the whole interconnected gene regulatory networks from central circadian clock systems. Many genes further from the central clock systems could have lower magnitude circadian periodicity and can not be detected by other methods than the piccolo.

A comparison of the five methods using simulation data shows that the piccolo method is most robust against noise. The detection performance of the methods, except the piccolo method, was worse for the two-harmonic data than for the one-harmonic data. The piccolo method exhibited more consistent performance between datasets than the other methods. This suggests that the piccolo method has high detection performance for data with multiple periodicity.

The computational cost of the piccolo method represents a potential problem for large datasets. In future work, we will attempt to reduce the computational cost by introducing the branch and bound method to the exhaustive search for the combination of Fourier coefficients.

Acknowledgement

We wish to thank Drs. Wataru Fujibuchi and Sachiyo Aburatani of the CBRC, AIST, for fruitful discussions.

Disclosure

This manuscript has been read and approved by the author. This paper is unique and is not under consideration by any other publication and has not been published elsewhere. The author and peer reviewers of this paper report no conflicts of interest. The author confirms that they have permission to reproduce any copyrighted material.

References

1. Ernst J, Bar-Joseph Z. STEM: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics*. 2006;7:191.
2. McQuarrie ADR, Tsai CL. *Regression and Time Series Model Selection*. World Scientific; 1998.
3. Artis M, Hoffmann M, Nachane D, Toro J. *The detection of hidden periodicities: A comparison of alternative methods*. EUI Working Paper ECO. 2004;10.
4. Sakamoto Y, Ishiguro K, Kitagawa G. *Akaike Information Criterion Statistics*. Springer verlag; 1986.
5. Benedetto JJ, Pfander GE. Periodic wavelet transforms and periodicity detection. *SIAM Journal of Applied Mathematics*. 2002;62(4):1329–68.
6. Janer L, Bonet JB, Lleida-Solano E. *Pitch detection and voiced/unvoiced decision algorithm based on wavelet transform*. Proceedings of The Fourth International Conference on Spoken Language Processing. 1996;2(FrP2P1): 1209–12.
7. Okamura H, Semba Y. A novel statistical method for validating the periodicity of vertebral growth band formation in elasmobranch fishes. *Canadian Journal of Fisheries and Aquatic Sciences*. 2009;66(5):771–80.
8. Yang R, Su Z. Analyzing circadian expression data by harmonic regression based on autoregressive spectral estimation. *Bioinformatics*. 2010;26: i168–74.
9. Tominaga D, Horimoto K. Judgment algorithm for periodicity of time series data based on bayesian information criterion. *Journal of Bioinformatics and Computational Biology*. 2008;6(4):747–57.
10. Barrett T, Suzek TO, Troup DB, et al. NCBI GEO: mining millions of expression profiles-database and tools. *Nucleic Acids Research*. 2005;33:D562–6.
11. Parkinson H, Kapushesky M, Kolesnikov N, et al. ArrayExpress update-from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Research*. 2009;37:D868–72.
12. Ahdesmäki M, Lähdesmäki H, Pearson R, et al. Robust detection of periodic time series measured from biological systems. *BMC Bioinformatics*. 2005;6:117.
13. Hogg RV, McKean JW, Craig AT. *Introduction to Mathematical Statistics*. 6th ed. Peason Prentice Hall; 2005.
14. Rousseeuw PJ, Leroy AM. *Robust Regression and Outlier Detection*. Wiley-Interscience; 2003.
15. Dixon WJ. Analysis of extreme values. *Annals of Mathematical Statistics*. 1950;21:488–506.
16. Dixon WJ. Ratios involving extreme values. *Annals of Mathematical Statistics*. 1951;22:68–78.
17. Rorabacher DB. Statistical treatment for rejection of deviant values: critical values of Dixon's "Q" parameter and related subrange ratios at the 95% confidential level. *Analytical Chemistry*. 1991;63(2):139–46.
18. Silverman BW. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall/CRC; 1986.
19. Konishi T. Three-parameter lognormal distribution ubiquitously found in cDNA microarray data and its application to parametric data treatment. *BMC Bioinformatics*. 2004;5:5.
20. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nature Genetics*. 2000;25(1):25–9.