Analyzing Circadian Expression Data by Harmonic Regression Based on Autoregressive Spectral Estimation

Rendong Yang Zhen Su

China Agricultural University

ISMB, 2010

Outline

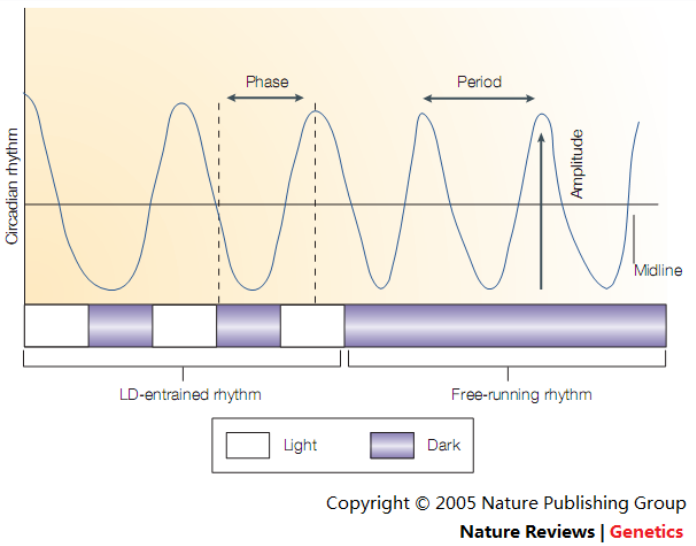
Outline

Circadian Rhythm: Concept and Description

Definition

A biological rhythm with a periodicity of approximate 24 hours that persists in constant conditions.

Description



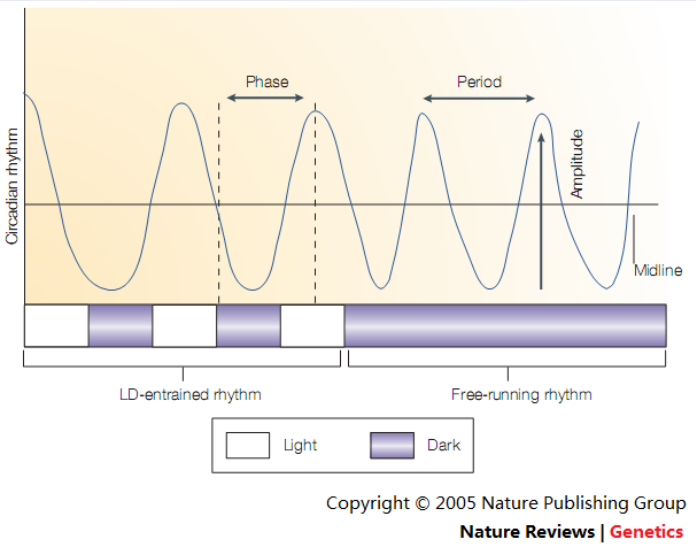
- Period - $[20h, 28h]$
- Phase - timing of peak expression
- Amplitude - half the range of oscillation
- Mean level - average value

Circadian Rhythm: Concept and Description

Definition

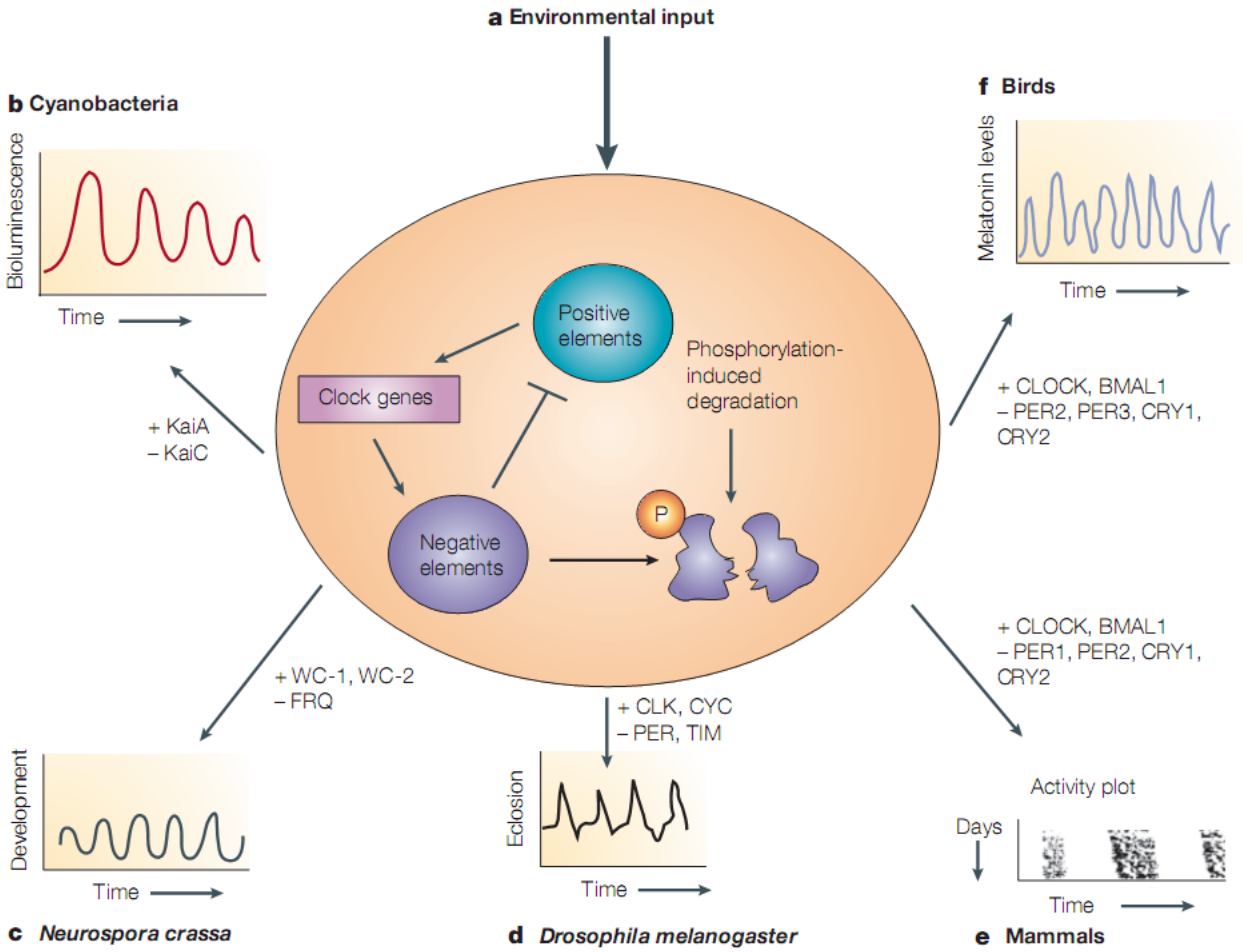
A biological rhythm with a periodicity of approximate 24 hours that persists in constant conditions.

Description



- Period - $[20h, 28h]$
- Phase - timing of peak expression
- Amplitude - half the range of oscillation
- Mean level - average value

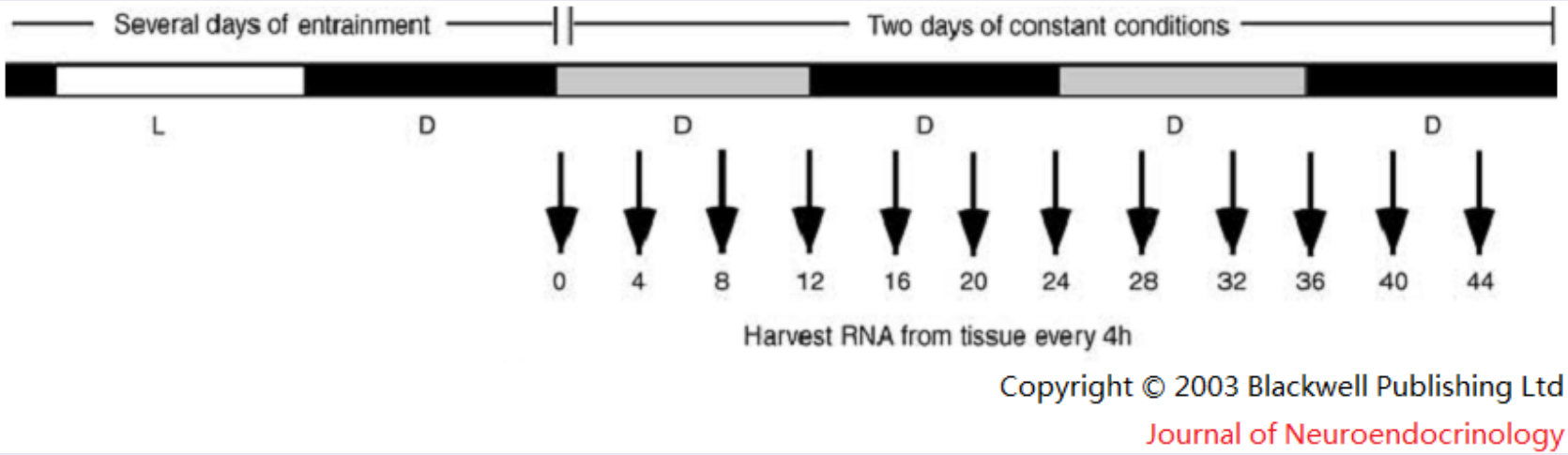
Circadian Clock: How Many Genes Involved?



Copyright © 2005 Nature Publishing Group
Nature Reviews | Genetics

Microarray and Circadian Rhythm

Design of Circadian Microarray Experiment



Copyright © 2003 Blackwell Publishing Ltd

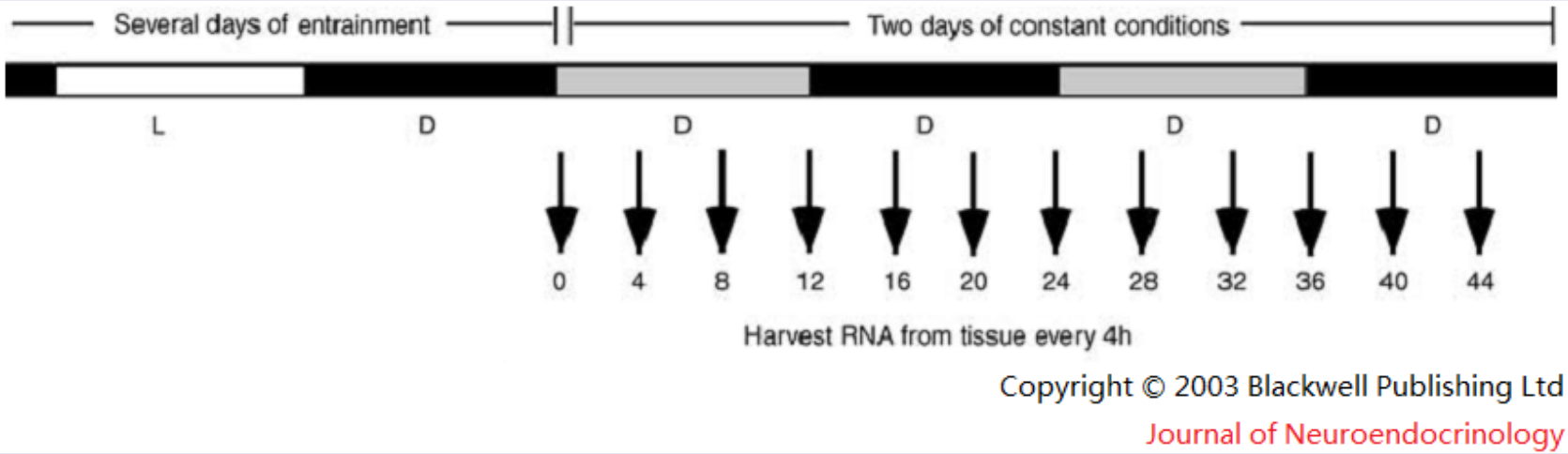
Journal of Neuroendocrinology

Computational Challenges

- Extremely sparse determination
- Extremely high dimensionality
- Low replicate numbers

Microarray and Circadian Rhythm

Design of Circadian Microarray Experiment



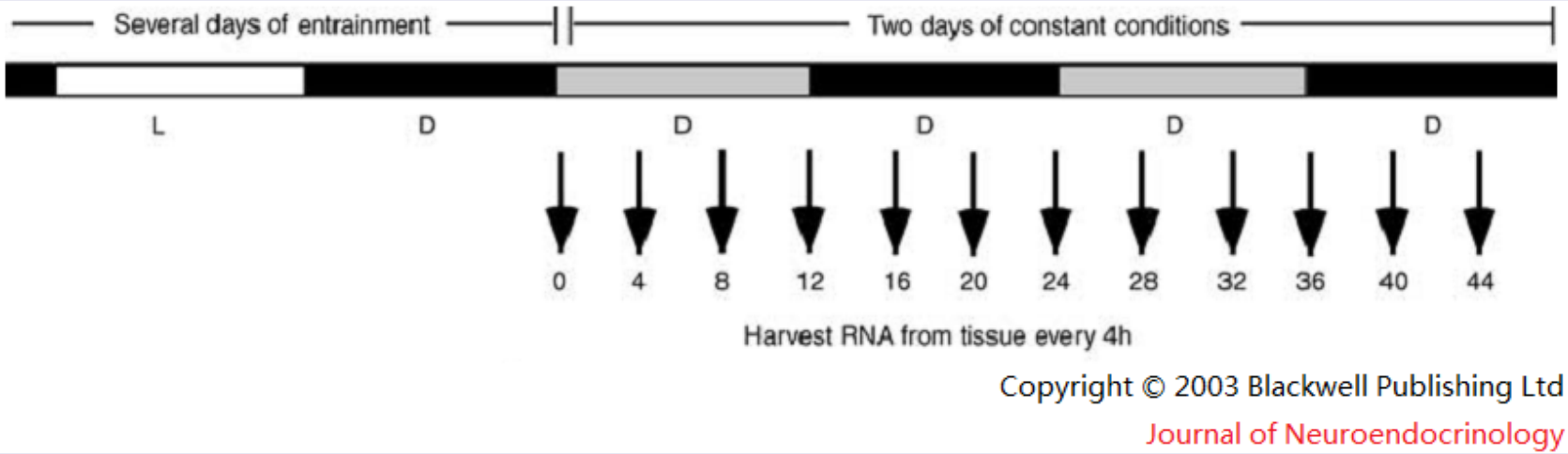
Copyright © 2003 Blackwell Publishing Ltd
Journal of Neuroendocrinology

Computational Challenges

- Extremely sparse determination
- Extremely high dimensionality
- Low replicate numbers

Microarray and Circadian Rhythm

Design of Circadian Microarray Experiment

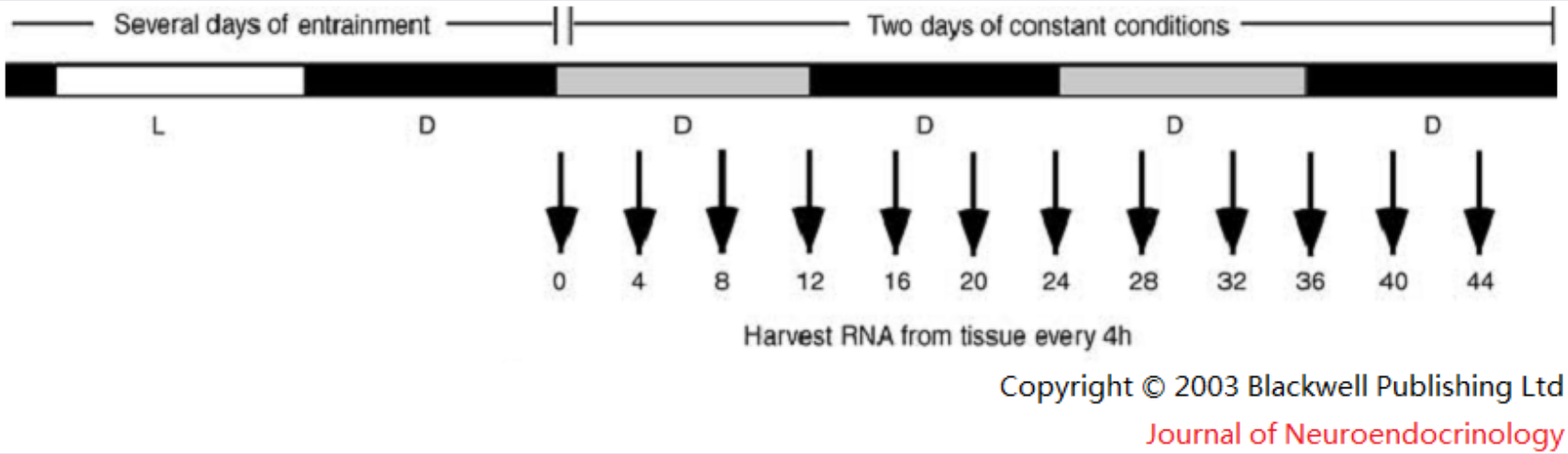


Computational Challenges

- Extremely sparse determination
- Extremely high dimensionality
- Low replicate numbers

Microarray and Circadian Rhythm

Design of Circadian Microarray Experiment



Copyright © 2003 Blackwell Publishing Ltd
Journal of Neuroendocrinology

Computational Challenges

- Extremely sparse determination
- Extremely high dimensionality
- Low replicate numbers

Outline

Statistical Assessment of Circadian Rhythms in Microarray Data

Categories of Prior methods

1. Time-domain algorithms (Such as *COSOPT*)
   - Pros: Efficient for short time-series
   - Cons: Model dependent, Predefine wavelength
2. Frequency-domain methods (Such as *Fisher's G-test*)
   - Pros: Model independent
   - Cons: Low resolution for short time-series

Our Solution

Combine both time-domain and frequency-domain analyses.

Statistical Assessment of Circadian Rhythms in Microarray Data

Categories of Prior methods

1. Time-domain algorithms (Such as *COSOPT*)
   - Pros: Efficient for short time-series
   - Cons: Model dependent, Predefine wavelength
2. Frequency-domain methods (Such as *Fisher's G-test*)
   - Pros: Model independent
   - Cons: Low resolution for short time-series

Our Solution

Combine both time-domain and frequency-domain analyses.

Outline

Frequency Domain: Period Detection

Autoregressive Spectral Estimation

Input

An evenly spaced time-series $\{x_t : t = 1, \cdots, n\}$

Algorithm

1. Apply the autoregressive(AR) model of order p, noted as AR(p), to fit $\{x_t\}$ by:

$$x_t = \sum_{i=1}^{p} \alpha_i x_{t-i} + \varepsilon_t \tag{1}$$

2. Estimate the frequency spectrum from AR coefficients α_i by:

$$p_x(f) = \frac{\sigma_\varepsilon^2}{|1 + \sum_{k=1}^{p} \alpha_k e^{-ifk}|^2} \quad 0 \le f < 0.5 \tag{2}$$

Time Domain: Rhythm Modeling

Harmonic Regression

Input

$\{x_t\}$ and frequency f_i (derived in Eq. (2))

Algorithm

Harmonic Regression models the rhythmic components of $\{x_t\}$ by

$$x_t = \mu + \sum_{i=1}^{n} \beta_i \cos(2\pi f_i t + \phi_i) + \varepsilon_t \tag{3}$$

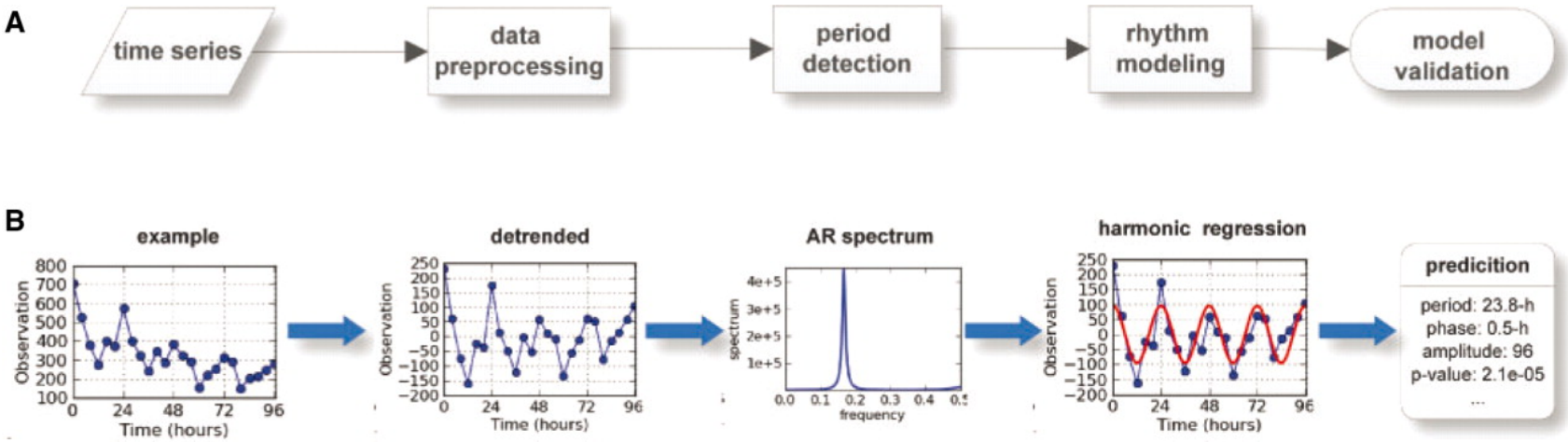
then Eq. (3) can be reduced to a simple linear regression form:

$$x_t = \mu + \sum_{i=1}^{n} \{p_i \cos(2\pi f_i t) + q_i \sin(2\pi f_i t)\} + \varepsilon_t \tag{4}$$

ARSER in Action

a synthetic time-series

Generated by $f(t) = 500e^{-0.01 \cdot t} + 140e^{-0.01 \cdot t} \cdot \cos(\frac{2\pi}{24} t) + \varepsilon$,
where $t \in [0, 96]$ with 4h intervals



The diagram of ARSER and a case study

Outline

Multiple Testing Corrections

ARSER

False discovery rate q-value (Storey et al. 2003)

Prior methods

- COSOPT
  - pMMC-β value to correct for multiple comparisons
- Fisher's G-test
  - false discovery rate method

Testing Datasets in Our Study

Dataset 1: Our generated synthetic data

- periodic time-series
  - Stationary cosine wave with *constant* amplitude and mean level
  - Nonstationary cosine wave with *exponentially damped* amplitude and mean level
- random time-series
  - white noise following ($\mu = 0, \sigma = 1$) normal distribution
  - AR(1) process
- time-series sampled every 4h over 48hrs

Dataset 2: Public synthetic data

- 120 time-series containing five circadian rhythmic patterns (Michael *et al.* 2008)
- time-series sampled every 4h over 48hrs

Dataset 3: Public microarray data

- *Arabidopsis* circadian expression data (Edwards *et al.* 2006)
- time-series sampled every 4h over 48hrs

Testing Datasets in Our Study

Dataset 1: Our generated synthetic data

- periodic time-series
  - Stationary cosine wave with *constant* amplitude and mean level
  - Nonstationary cosine wave with *exponentially damped* amplitude and mean level
- random time-series
  - white noise following ($\mu = 0, \sigma = 1$) normal distribution
  - AR(1) process
- time-series sampled every 4h over 48hrs

Dataset 2: Public synthetic data

- 120 time-series containing five circadian rhythmic patterns (Michael *et al.* 2008)
- time-series sampled every 4h over 48hrs

Dataset 3: Public microarray data

- *Arabidopsis* circadian expression data (Edwards *et al.* 2006)
- time-series sampled every 4h over 48hrs

Testing Datasets in Our Study

Dataset 1: Our generated synthetic data

- periodic time-series
  - Stationary cosine wave with *constant* amplitude and mean level
  - Nonstationary cosine wave with *exponentially damped* amplitude and mean level
- random time-series
  - white noise following ($\mu = 0, \sigma = 1$) normal distribution
  - AR(1) process
- time-series sampled every 4h over 48hrs

Dataset 2: Public synthetic data

- 120 time-series containing five circadian rhythmic patterns (Michael *et al.* 2008)
- time-series sampled every 4h over 48hrs

Dataset 3: Public microarray data

- *Arabidopsis* circadian expression data (Edwards *et al.* 2006)
- time-series sampled every 4h over 48hrs

Testing Datasets in Our Study

Dataset 1: Our generated synthetic data

- periodic time-series
  - Stationary cosine wave with *constant* amplitude and mean level
  - Nonstationary cosine wave with *exponentially damped* amplitude and mean level
- random time-series
  - white noise following $(\mu = 0, \sigma = 1)$ normal distribution
  - AR(1) process
- time-series sampled every 4h over 48hrs

Dataset 2: Public synthetic data

- 120 time-series containing five circadian rhythmic patterns (Michael *et al.* 2008)
- time-series sampled every 4h over 48hrs

Dataset 3: Public microarray data

- *Arabidopsis* circadian expression data (Edwards *et al.* 2006)
- time-series sampled every 4h over 48hrs

Testing Datasets in Our Study

Dataset 1: Our generated synthetic data

- periodic time-series
  - Stationary cosine wave with *constant* amplitude and mean level
  - Nonstationary cosine wave with *exponentially damped* amplitude and mean level
- random time-series
  - white noise following $(\mu = 0, \sigma = 1)$ normal distribution
  - AR(1) process
- time-series sampled every 4h over 48hrs

Dataset 2: Public synthetic data

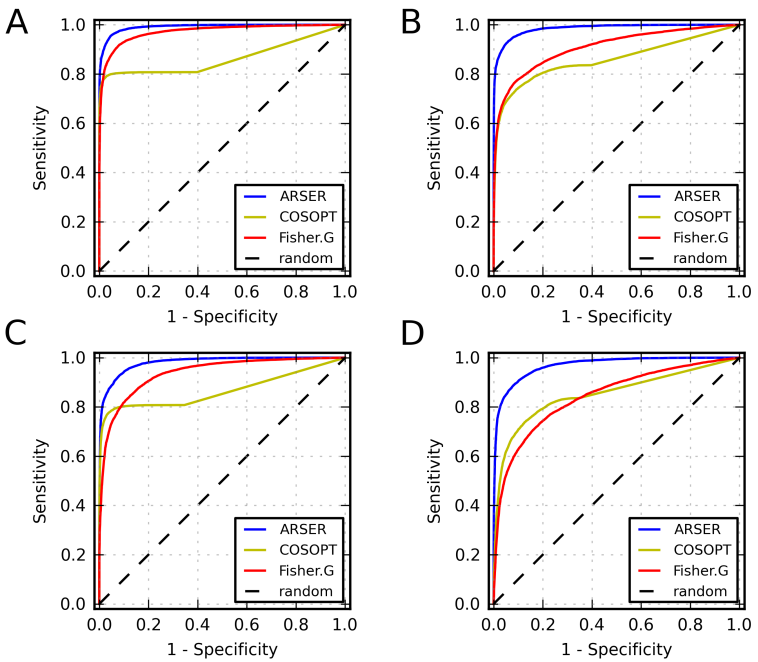
- 120 time-series containing five circadian rhythmic patterns (Michael *et al.* 2008)
- time-series sampled every 4h over 48hrs

Dataset 3: Public microarray data

- *Arabidopsis* circadian expression data (Edwards *et al.* 2006)
- time-series sampled every 4h over 48hrs

Outline

Periodicity Detection with Random Background Models

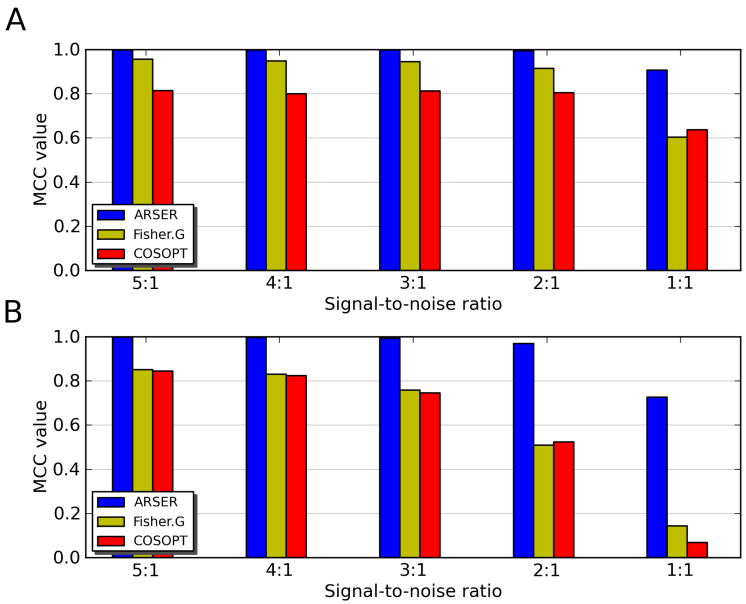


- A 10 000 *stationary* periodic signals and 10 000 *white noise* signals
- B 10 000 *non-stationary* periodic signals and 10 000 *white noise* signals
- C 10 000 *stationary* periodic signals and 10 000 *AR(1)* signals
- D 10 000 *non-stationary* periodic signals and 10 000 *AR(1)* signals

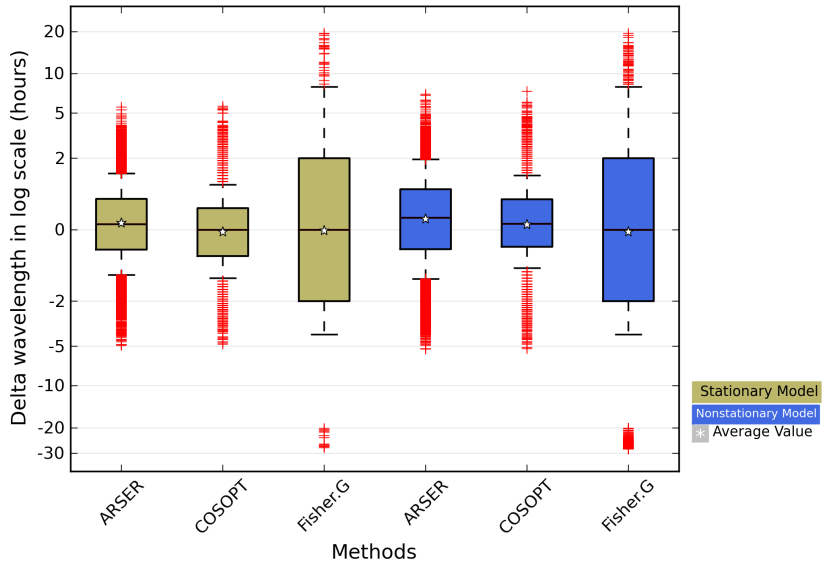
Performance measurement

- binary classification: periodic and non-periodic
- ROC curve

Robustness to Noise and Wavelength



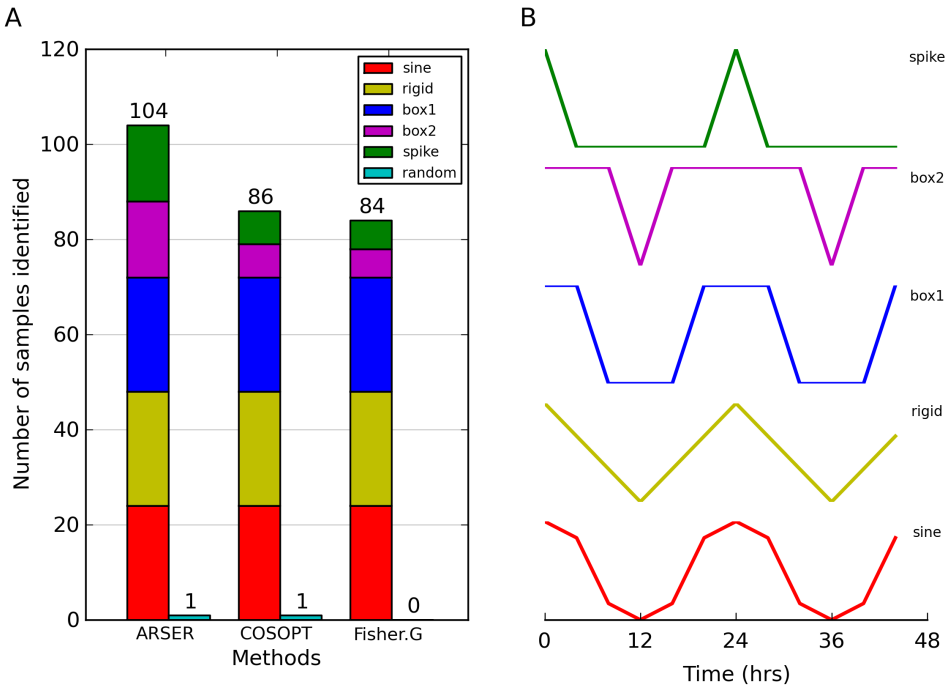
Identifying (A) stationary and (B) non-stationary periodic signals under decreasing **signal-to-noise ratio (SNR)**



Distribution of differences between predicted **wavelength** and the actual wavelength for each periodic signals (wavelength $\in$ [20h, 28h) with 0.1-h spaced)

Outline

Detection of Non-sinusoidal Periodic Waveforms

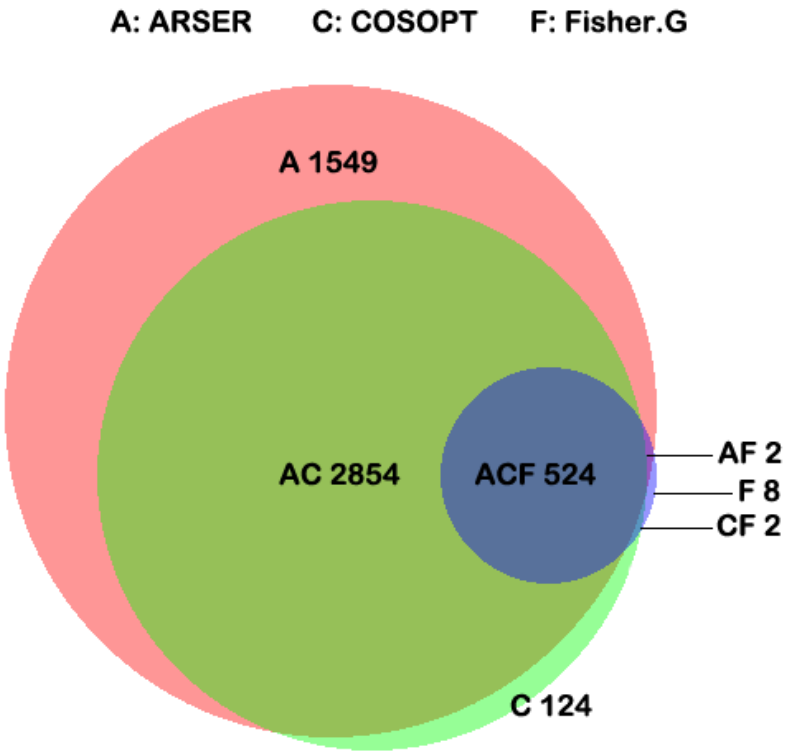


- 120 time-series
- 5 circadian rhythmic patterns, 24 samples for each
- ARSER identified 87% (104/120) periodic signals

Data downloaded from http://haystack.cgrb.oregonstate.edu/

Outline

Analysis of *Arabidopsis* Circadian Expression Data



Comparison of three algorithms for identifying *Arabidopsis* circadian-regulated genes

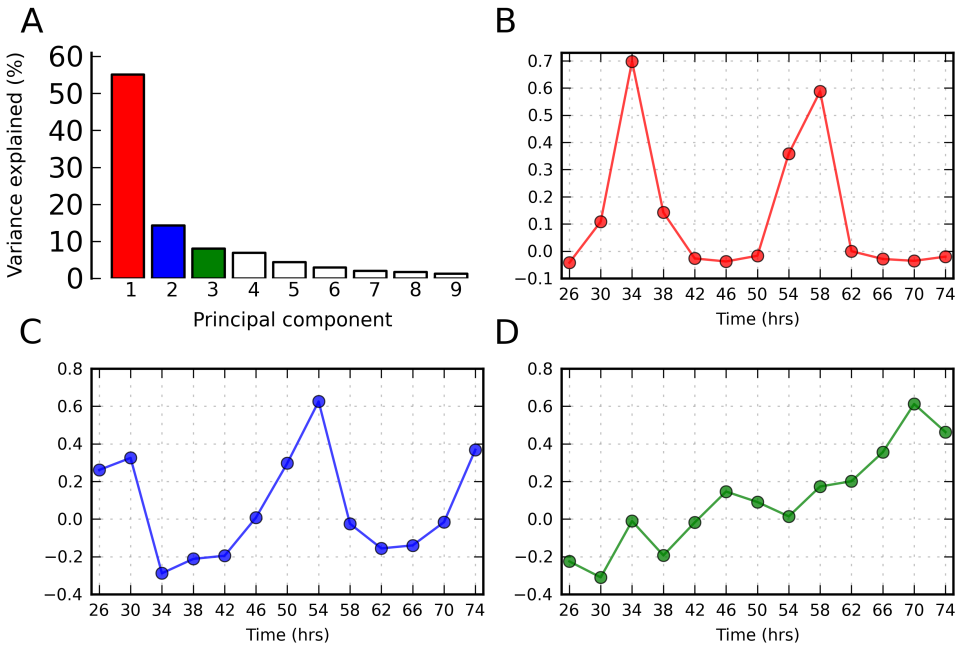
Original Report

- COSOPT algorithm by setting pMMC-$\beta < 0.05$
- 3505 genes, 16% of Arabidopsis whole genome are rhythmically expressed

ARSER identified

- 4929 genes rhythmically expressed (q-value<0.05)
- covering 96% of genes identified by COSOPT

Newly identified *Arabidopsis* Rhythmic Transcripts by ARSER



Principal component analysis of the 1549 newly-found rhythmic transcripts in *Arabidopsis* identified by ARSER.

Computational Validation

- PCA analysis
- first two components show rhythmic pattern

Biological Validation

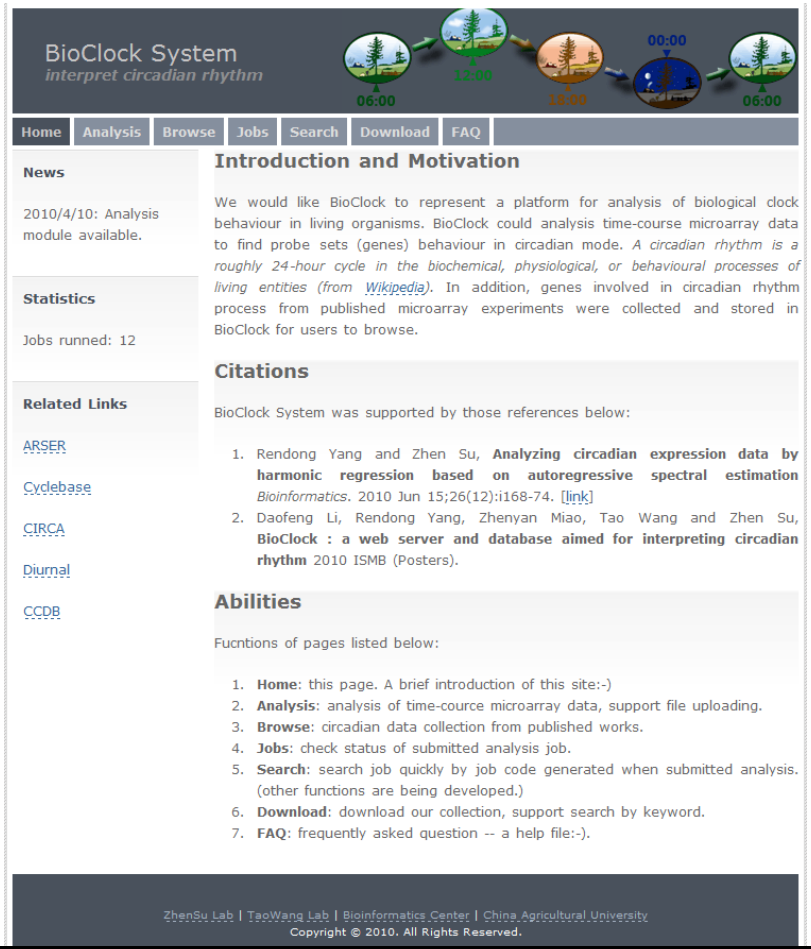
- find 2 core clock genes from 27 known *Arabidopsis* clock genes.
- CRY1 and PRR9

Summary

- ARSER combines the time-domain and frequency-domain analyses to efficiently identify sinusoidal and non-sinusoidal periodic patterns in short, noisy and non-stationary time-series.
- Tested on well defined simulation data, ARSER is superior to two former methods, COSOPT and Fisher's G-test.
- Analysis of *Arabidopsis* microarray data using ARSER led to identification of a novel set of periodic transcripts

- Outlook
  - ARSER can only used to analyze evenly spaced time-series. We are developing an algorithm for irregularly spaced samples

BioClock: a platform for analyzing circadian expression data



Website

http://bioinfo.cau.edu.cn/BioClock

Poster

Poster Section E34

Acknowledgement

- Prof. John Hogenesch for sharing COSOPT software
- Daofeng Li for developing the interface of BioClock
- Ms. Wenying Xu and our colleagues in the Lab for helpful discussions

- Funding
  - Ministry of Science and Technology of China (2008AA02Z312, 2006CB100105)
  - ISCB, DOE and NSF for travel fellowship award to support my presentation

Thank you!

Rendong Yang
Email: cauyrd@gmail.com