2006－BIO－4（3）
2006／2／9

恣意的な判断基準を持たない時系列データの周期性判定法

富永大介 *、ポール・ホートン
独立行政法人産業技術総合研究所生命情報科学研究センター
〒 135-0064 東京都江東区青海 2-42, *tominaga@cbrc.jp

生物学全般において、たとえば概日周期変動や個体数の通年での変化の解析など、定量的観測値の時間変化が周期性を持つかどうかの判断を必要とすることは多い。しかしその判断基準はデータの特性に応じて経験的に、ケースバイケースで決められており、研究者や状況によって判断が変わることがありえるため、客観性を欠いた解析、解釈が行われやすい。したがって、恣意性を含まない判断基準が必要である。そこで我々は判断基準として情報量基準を導入し、恣意性を排除した判断を行うアルゴリズムを考案した。そのアルゴリズムを乱数データと遺伝子発現観測データに適用し、この手法の有効性を示した。

Periodicity judgment for time series data without arbitrary criterion

TOMINAGA Daisuke*, Paul HORTON
Computational Biology Research Center,
National Institute of Advanced Industrial Science and Technology
2-42 Aomi, Koto, Tokyo 135-0064, Japan, *tominaga@cbrc.jp

For biological time series data, such as circadian expression of genes, population of individuals, etc., judgment whether time series is periodic or not is often done widely. However, in most cases, criteria of judgments are defined by arbitrary parameters based on characteristics of data and analysts experiences, and there is no universal judgment criterion. Non-arbitrary judgment is important and needed. We developed an algorithm for non-arbitrary periodicity judgment by introducing Information Criterion. We applied the algorithm to randomly generated time series data and gene expression profile of mice to find circadian genes, and compared with widely used conventional judgment methods. Our algorithm shows both high sensitivity and specificity.

1　はじめに

生物学全般において、定量的観測値の時間変化が周期性を持つかどうかを判断する場面は多い。たとえば遺伝子発現を時系列で観測したデータから概日周期変動しているものを探し出すことなどがあるが [1]、その判断は、常に客観的に行われているとは言えない [3][14][15][16]。

観測値の示す時間変動が周期性を持つかどうかは、周期関数の当てはめ精度、自己相関係数の高さ、フーリエ変換によるスペクトルでの、注目する周期成分の大きさなどで判断される。たとえばコサイナー法 [9] では観測値時系列に単一の三角関数 ($y(t) = A\cos(Bx + C) + D$ など) を当てはめ、振幅 A、周期 B、位相 C、オフセット D を最適化し、近似精度が良ければ周期 B を持つと判断する。特定の周期、たとえば 24 時間周期を持つかどうかを判断したい場合は、B が 24(あるいは $\frac{2\pi}{24}$) に近いかどうかを判断基準に加えるか、B を 24 に固定してフィッティングを行う。

時系列の観測では一般に、離散的にサンプリングが行われる。従ってフーリエ変換 [10] を用いる場合は離散フーリエ変換 (DFT)[13] を適用する。DFT によりデータのサンプリング数と同じ個数の複素数からなるのフーリエ係数ベクトルが得られる。その中のたとえば 24h に対応するフーリエ係数の絶対値が他の成分よりも大きいか、既定のしきい値よりも大きいかなどでデータの周期性の有無を判断する。

自己回帰モデル (AR モデル [6]) でも同じよう

にモデル当てはめの精度と着目している周期成分の大きさから判断される。自己相関解析では相関の大きさで判断する。

コサイナー法、DFT、AR モデル、自己相関解析などで実際に解析を行うときに用いる判断基準は、データの性質や解析目的に応じて、経験に基づいて個々の研究者が各人の判断で決定するのが一般的である。したがって、解析を行う研究者によって結論が変わってしまうことは日常よく見られることであり、発表者にとって都合の良い結果を導くような判断基準が使われることで解析の信頼性が問題になることもあり得る。こういった問題を排除し、客観的で普遍的な解析を行うためには、恣意性を含まない判断基準が必要である。

我々はそういった恣意的な判断基準の代わりに情報量基準 [11] を使うことで、任意性をを排除した客観的な判断を行うアルゴリズムを考案した。このアルゴリズムでは、複数の周波数成分をパラメータとして持つモデル (フーリエ係数、AR モデルなど) の良さを情報量基準で判断し、もっとも良いモデルが持っている周波数成分が、元の信号の持つ周期性であると判断する。たとえば概日周期変動を示す遺伝子を探したいときは、各遺伝子について、その遺伝子の変動を表すもっとも良いモデルに 24 時間周期成分が含まれているかどうかを見ることで判断ができる。

我々が行ったアルゴリズムの実装では、モデルに DFT、情報量基準に BIC (Bayesian Information Criterion)[12] を使っている。モデルは複数のパラメータで定義され時系列を計算することのできるモデルならどんなものでも使うことができる。たとえば DFT の代わりに AR モデルを使っても全く同様に実装することができる。また情報量基準も AIC などのほかの情報量基準を用いることもできる。

考案したアルゴリズムの有効性を示すため、汎用のパソコン上にこれを実装し、正規分布乱数で生成したものと、web で公開されているマウスの遺伝子発現量の二種類の時系列データに対して適用、判定精度、計算機による実行速度を検証した。また広く用いられているコサイナー法、DFT を使った簡便な判定法との比較を行った。

2 アルゴリズム

定量的時系列データに対して、着目する周波数成分に対応するフーリエ係数が、BIC を最小とするモデルに含まれていれば、その時系列データはその周波数について周期的であると見なす。観測データを離散フーリエ変換したものを初期モデルとし、初期モデルからいくつかのフーリエ係数を選びだしたものをモデルとする。モデルを逆離散フーリエ変換した時系列データから、式 (1) によってそのモデルの BIC を計算する。ここではデータのノイズ、ばらつきは正規分布を仮定している。式 (1) は多項式の AIC[11] と BIC の定義 [12] を参考に導出した。式 (1) の値は、iDFT の結果と選び出す係数の個数により変化する。

$$\mathrm{BIC} = n\log 2\pi + n\log\hat{\sigma}^2 + n + (p+1)\log n \quad (1)$$

ここで n はデータのサンプリング点数、$\hat{\sigma}^2$ は観測データとモデルから計算した時系列データの残差二乗和 ($\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - X_i)^2$、$x_i$ はモデルの値、X_i は観測値)、p はモデルが持つパラメータ数、つまり選び出したフーリエ係数の個数である。フーリエ係数ベクトルの第一成分はオフセットであるため、モデルを生成する際に常に選ぶこととする。離散フーリエ変換の性質から、フーリエ係数を $F_i(1 \le i \le n)$ とすると、$F_i = F_{n-i+2}(2 \le i \le \frac{n}{2})$ である。第一成分をのぞいた $n-1$ 個のフーリエ係数から p 個の係数を選び出す組み合わせは ${}_{n-1}\mathrm{C}p$ 通りあるが、この対称性があるので選ぶ対象となる係数は $\frac{n}{2}$ 個でよく、ある p に対するモデルの総数は ${}_{\frac{n}{2}}\mathrm{C}_p$ になり、p は $0 \le p \le \frac{n}{2}$ の範囲内になる。BIC を計算する際に必要となる iDFT を行うときは、選び出されたフーリエ係数と対称位置にある係数も用いる (用いなければ iDFT の結果が複素数の時系列になり、式 (1) 右辺第二項が計算できない)。また情報量基準の性質から、p が大きくなると式 (1) における $\hat{\sigma}^2$ が小さくなりすぎて正しい評価ができなくなる。そのため、$0 \le p \le \min(\frac{n}{4}, 2\sqrt{\frac{n}{2}})$ とする。

従ってアルゴリズムは以下のようになる (図 1)。

1. 与えられた n 点からなる観測値時系列に DFT を適用して得られるフーリエ係数ベクトルを初期モデルとする。

2. $0 \leq p \leq \min(\frac{n}{4}, 2\sqrt{\frac{n}{2}})$ について、初期モデルから p 個のフーリエ係数を選び出すすべての組み合わせについて、以下を行う。

   (a) 選び出したフーリエ係数 (モデルと呼ぶ) から iDFT で時系列データを計算する。

   (b) 与えられている観測値時系列、モデルから得られる時系列データ、および p から式 (1) によって BIC の値を計算する。

3. BIC が最小となるモデルに、着目する特定の周期に対応するフーリエ係数が含まれていたら、与えられた時系列データはその周期での周期性を持つ、含まれていなければその周期での周期性はない、と判断する。

我々はこのアルゴリズムを piccolo 法と名付けた。これは周期性 (periodicity)、客観的判定 (clinical judgment)、情報量基準 (information criterion) などの単語と、簡潔でコンパクトなアルゴリズムであることからの連想である。

3 検証結果

平均 0、分散 1 の正規分布乱数で作った時系列と、web 上で公開されている DNA マイクロアレイによる時系列データから、24 時間周期で変動する遺伝子 (概日周期遺伝子) を探すことを目的として、piccolo 法とコサイナー法、DFT による判定法を適用、比較した。piccolo 法と DFT は GNU Octave(fftw[4] をリンクした) の、コサイナー法は gnuplot のスクリプトとして実装した。GNU Octave は version 2.1.71 (powerpc-apple-darwin7.9.0)、gnuplot は Version 4.1 patchlevel 0 であり、PowerPC G4 1.5GHz, 1.25GB ram で実行した。

検証に用いた公開データは、米国立衛生研究所 (NIH) の GEO データベース [2] に登録番号

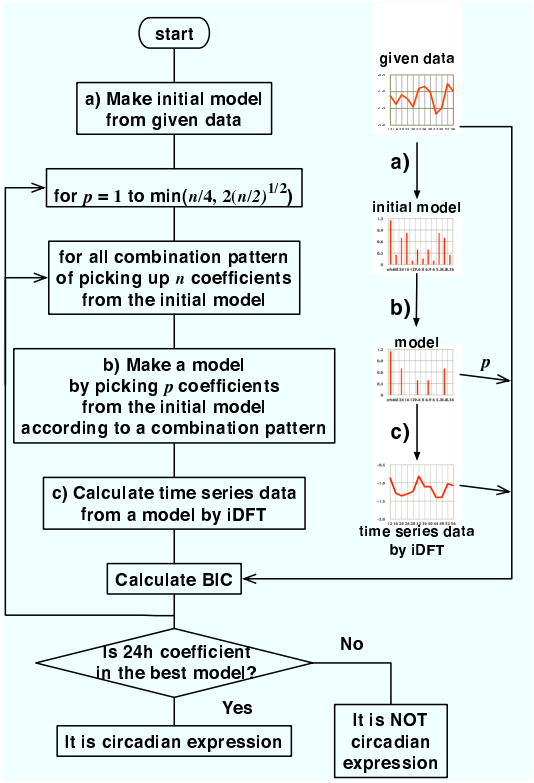

図 1: piccolo 法のフローチャート。

GDS404 として公開されているマウスのデータで、遺伝子数 12488、サンプリング時刻は 4 時間おきに 4h から 48h までの 12 点である。最初の 4h だけ 2 回サンプリングされているため、ここでは 4h データはその平均値を使った。このデータには、発現量が低いスポットなどの無効なデータが null として含まれている。一つの遺伝子がもつ 13 点のデータのうち、8 点以上が null のものは信頼性が低いものとして無視した。そうでないものでは null を 0.0 と置き換えた。これにより解析対象の遺伝子数は 10068 になる。乱数データは平均 0、分散 1 の正規分布乱数で作成した。GDS404 にあわせてサンプリング点数 12 の遺伝子が 10068 あることを想定し、合計 1006800 本の時系列データを作成して 100 セットに分割したデータを用意した。これを用いて 10068 遺伝子を想定した判定を 100 回行い、統計的な検証を行った。

コサイナー法では当てはめに用いるモデルを

$$y(t) = A\cos((2\pi/24)x + B) + C$$

と定義し、与えられた時系列データに対して、最小二乗法による当てはめを A, B, C をパラメータとして行った。これには gnuplot で実装されている Marquardt-Levenberg 法 [8](繰り返し計算により収束する最小二乗法) を用い、それぞれ初期値は、A はその時系列中のデータの最大値、B は最大値を取る時刻に $\frac{2\pi}{24}$ を乗じた値、C はデータの平均値とした。コサイナー法そのものは当てはめを行うだけであり、それ自体では概日周期性の判定基準を含まない。ここでは、他の方法と比較する際に当てはめの精度のみを判定基準に用いた。

DFT のみによる判定では、第一成分と、前半と対称な後半の係数を除いた $6(=\frac{n}{2})$ 個のフーリエ係数のうち、24h に対応するものの絶対値が最大の場合に概日周期と見なした。

piccolo 法では、与えられるデータのサンプリング点数が 12 であるため、式 (1) における p の範囲は $0 \leq p \leq 3$ となる。

概日周期と判定される遺伝子の数と、計算コストについて比較した結果を表 1 に示す。乱数データに対する判定ではそれぞれの判定法について、100 回の判定による平均と分散を示した。また GSE404 には概日周期遺伝子として既に知られているもの変動が含まれている。そのうちの 7 個の主な概日周期遺伝子について、各方法でどれが概日周期と判定されるかを表 2 に示す。

DFT 法、piccolo 法の二つでは、概日周期と判定される遺伝子の数が乱数データと発現データで約 1.3 倍違う。コサイナー法で piccolo 法と同じ数の遺伝子を概日周期と判定するためには、近似精度に RMSD(Root Mean Square Deviation) を使うとき、この値を乱数データでは 0.678(S.D. 0.0148) 以下、発現データでは 23.4 以下を周期的であると見なすようにすればよい。

コサイナー法は発現データだと収束しにくい傾向があるが、計算にかかる時間はコサイナー法がもっとも少なく、piccolo はコサイナー法の 21.9 倍 (乱数データ) および 16.6 倍 (GDS04) かかる。繰り返し計算の回数も piccolo の方が多く、11.1 倍 (乱数データ) および 6.86 倍 (GDS404) である。DFT 法は乱数データと発現データで計算時間はほぼ同じで、コサイナー法と同程度の時間である。

GDS404 に含まれている既知の概日周期遺伝子 7 個について、これらはいずれも、主観的には概日周期変動をしているように思われる (図 2)。コサイナー法で上述のように、RMSD $\leq$ 23.46 のものを概日周期とみなすとすると、7 つのうち 1 つだけを概日周期変動していると判定することになる。DFT 法では 5 個、piccolo 法では 6 個であり、もっとも人間に近い判断をすることができるのは piccolo 法である (表 2)。フーリエ変換を使う二つの方法 (DFT および piccolo) で周期性が判定できなかった CRY は、フーリエ係数のうち 24h よりも 12h の成分が大きく、それぞれ絶対値は 202.7 および 229.9(振幅はそれぞれ 33.8 と 38.3) であり、12h の方が大きいために判定されにくかったものと考えられる。この発現データでは、piccolo 法では直流成分のみのモデル (フーリエ係数を一つだけ含むモデル) がもっとも BIC の低い、良いモデルとなった。

DFT では拾えなかった Per1 のフーリエ係数を見ると、24h よりも 12h の方が絶対値が大きい (表 3, 24h での振幅は 60.0、12h での振幅は 61.8)。これも振幅の大きなもの上位二つを見るなどとすれば Per1 も 24h 周期性を持つと判定できるが、DFT では概日周期と判定された遺伝子数がすでに piccolo 法の 2 倍以上あり、基準を緩めると目標を絞り込むことができない。またこの Per1 の場合は、どの周波数成分を取り除いても近似精度が大きく悪くなるため、すべての周期成分を持ったモデルが、BIC 最小のもっとも良いモデルである。

コサイナー法で piccolo 法と同じ数の遺伝子を概日周期性と判定するためには、判断基準を RMSD le23.46 とすることになる。三角関数に近いように見える Arntl と DBP、Clock では RMSD はそれぞれ 146.5, 107.2, 221.4 である (表 3)。フィッティングを行うときにパラメータの初期値を対象にあわせて適切に選べば、もっと良い当てはめを行うことができると考えられるが、それを行うための一般的な方法は、興味深い研究対象ではあるが簡単ではない。またコサイナー法で 7 つすべての概日周期遺伝子を選び出すためには RMSD の

表 1: 概日周期変動の判定にかかる計算コストと判定されるものの割合。C はコサイナー法、D は DFT、P は piccolo 法である。time は秒である。iteration はコサイナー法では収束するまでの繰り返し計算の回数、piccolo 法では生成するモデルの個数であり、DFT は繰り返し計算を含まない。circadian は総数 10068 の遺伝子のうち、概日周期変動と判定されたものの数である。乱数データに対しては 100 回の判定による平均値を示した。SD はその際の標準偏差である。

	random data		
	time	iteration	circadian
C	41.6	5.70	-
(SD)	(0.742)	(0.0159)	-
D	56.8	1	2014
(SD)	(0.389)	-	(39.4)
P	816.	42	840.
(SD)	(19.2)	-	(26.2)
	GDS404		
	time	iteration	circadian
C	51.6	9.18	-
D	59.0	1	2730
P	856.	42	1110

しきい値を 234.5 にしなければならない。そのとき概日周期と判定される遺伝子数は 6325 になり、全遺伝子数 10068 の約 63%である。

4 考察

DFT、piccolo 法のいずれでも乱数データと遺伝子発現データでは判定結果が異なっており、乱数データの場合の標準偏差と比較すると、有意な違いがあると考える。

コサイナー法、DFT、piccolo 法のそれぞれの感度 (sensitivity、すべての正解のうち正しく判定できたものの割合) は 0.286、0.714、0.857 であり、piccolo 法が優れていることを表 2 に示した。

表 2: GDS404 における既知の概日周期遺伝子に対する、各判定法での判定結果。C はコサイナー法、D は DFT、P は piccolo 法である。'◯' は概日周期変動と判定されたもの、'×' は概日周期変動ではないとされたものである。

gene (clone ID)	C	D	P
Clock (92257_at)	×	◯	◯
Per1 (93619_at)	×	×	◯
Per2 (93694_at)	×	◯	◯
Arntl (102382_at)	×	◯	◯
CRY (101879_s_at)	◯	×	×
DBP (160841_at)	×	◯	◯
NFIL3 (101805_f_at)	×	◯	◯

表 3: 各手法での既知の概日周期遺伝子に対する判定。C はコサイナー法、D は DFT、P は piccolo 法である。RMSD はデータと当てはめた三角関数の平均分散値の平方根、dominant は DFT で得られるもっとも大きな周期成分、coef. は piccolo 法で BIC 最小のモデルに含まれるフーリエ係数の個数である。下線は、その遺伝子がその判定法で概日周期変動である判定されることを示す。

	C	D	P
gene	RMSD	dominant	coef.
Clock	221.36	<u>24h</u>	<u>3</u>
Per1	234.49	12h	<u>7</u>
Per2	145.92	<u>24h</u>	<u>3</u>
Arntl	146.45	<u>24h</u>	<u>3</u>
CRY	<u>13.873</u>	48h	1
DBP	107.18	<u>24h</u>	<u>5</u>
NFIL3	<u>5.0753</u>	<u>24h</u>	<u>3</u>

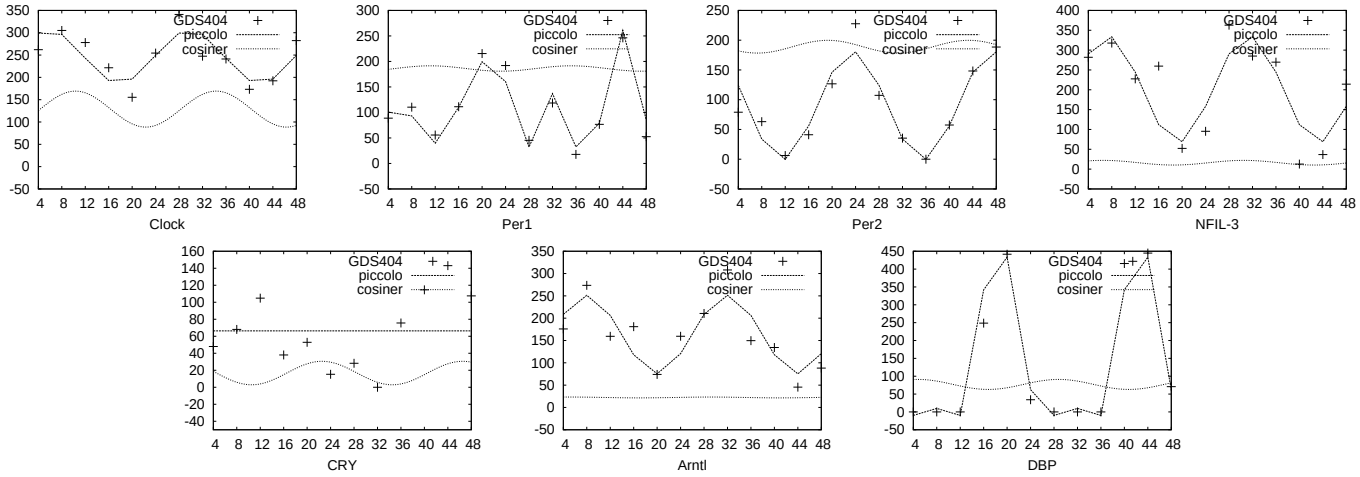

図 2: GDS404 データに含まれる、概日周期システムの要素として広く知られている遺伝子の発現量の変動時系列 (+) と、コサイナー法による最適近似関数 (破線) と、piccolo 法による最適モデル (実線) が示す変動時系列。

piccolo 法の欠点はその計算量である。観測値時系列のサンプリング点数を n とするとき、コサイナー法で使った Levenberg-Marquardt 法は連立一次方程式の解法を含むのでその計算量は $O(n^3)$、DFT の計算量は一般には $O(n^2)$ でどちらも多項式時間だが、piccolo 法には $n\mathrm{C}r$ の総探索が含まれるので、組み合わせ爆発を生じる。しかも r の値も走査対象である。データのサンプリング点数が増えると、計算時間の差は表 1 よりも急激に大きくなる。DNA マイクロアレイによるサンプリングは時間や労力のコストが高く時系列サンプリングには適さないため、現在公開されているデータにはサンプリング点数が多く含むものは少ないが、セル・アレイやトランスフェクション・アレイのような連続モニタリングの手法によるデータでは、数百点以上のサンプリングを容易に行うことができる。こういったデータに piccolo 法を適用するためには、データを間引いて点数を減らす、組み合わせ総探索の部分をたとえばシミュレーティド・アニーリングや遺伝的アルゴリズムのような発見的探索法にする、などの工夫が必要である。

コサイナー法には、MS-Excel や OpenOffice などの一般的な表計算ソフトウェア等で実装できる、高速な数値計算ソフトウェア等で大規模な解析にも適用できる、結果の解釈が容易であるなどの手軽な利点があるため、広く用いられている。しかし生命現象は、複数の要素が相互作用しあうことで全体の挙動が生じる複雑系としての特性が強いことを考えると、複数の周波数成分が大きな振幅を持つものに対しては RMSD が大きくなってしまうコサイナー法は、特に今回適用した遺伝子発現のような複数の要因が複雑に絡みあって変動を起こすような系を観測したデータには適切ではない。

同じことはここで比較実験を行った DFT を単純に用いる方法にも言える。表 3 に示すように、ここで使った DFT が CRY を拾えないのは、もっとも大きな成分のみに着目しているからである。しかしこの方法ではすでに piccolo 法の 2 倍以上の個数の遺伝子を概日周期変動であると判定しており、たとえば振幅の上位 2 成分が 24 時間に対応する成分を含んでいればよい、というように判断基準を甘くすると、大量データからの絞り込みができなくなる。

piccolo 法では、各周波数成分を持つモデル、持たないモデルを全探索で比較することにより、観測データにおける 24 時間成分の寄与の大きさが他の成分と比較して大きいかという相対的な判断基準に加えて、その成分が観測データを再現するのに重要かどうかという絶対的な基準も考慮することになり、複合周期に対応できるバランスのい

い判定基準であると言える。また piccolo 法は高い感度 (ここでは $\frac{6}{7} \approx 85.7\%$、表 2) を示しているが、これは絞り込み ($1110/10068 = 11\%$, 表 1) とバランスよく両立されていると言える。

参考文献

[1] Bar-Joseph, Z. (2004) Analyzing time series gene expression data, *Bioinformatics*, **20**, 16, 2493-2503.

[2] Barrett, T., Suzek, T. O., Troup, D. B., Wilhite, S. E., Ngau, W.-C., Ledoux, P., Rudnev, D., Lash, A. E., Fujibuchi, W., Edgar, R. (2005) NEBI GEO: mining millions of expression profiles - database and tools, *Nucleic Acids Research* **33**, Database issue, D562-D566.

[3] Chen. J. (2005) Identification of significant periodic genes in microarray gene expression data, *BMC Bioinformatics*, 6:286.

[4] Frigo, M., Johnson, S. G. (2005) The Design and Implementation of FFTW3, *Proceedings of the IEEE*, **93**, 2, 216-231.

[5] Harvery, A. C. (1993) Time Series Models, MIT Press, Cambridge, Massachusetts, USA.

[6] Kitagawa, G., Gersch, W. (1996) Smoothness Priors Analysis of Time Series (Lecture Notes in Statistics), Springer-Verlag, Heidelberg, Germany.

[7] Leloup, J.-C., Goldbeter, A. (2003) Toward a detailed computational model for the mammalian circadian clock, *PNAS*, **100**, 12, 7051-7056.

[8] Marquardt, D. (1963) An Algorithm for Least-Squares Estimation of Nonlinear Parameters, *SIAM Journal on Applied Mathematics,* **11**, 431-441.

[9] Nelson, W., Tong, Y. L., Lee, J. K., Halberg, F. (1979) Methods for cosinor-rhythmometry, *Chronobiologia*, 6, 305-323.

[10] Ronald, B. N. (1986) The Fourier Transform and its Applications, second edition, McGraw-Hill, New York, USA.

[11] Sakamoto Y., Ishiguro M., Kitagawa G. (1986) Akaike Information Criterion Statistics, D. Reidel Publishing Company, Tokyo, Japan.

[12] Schwarz, G. (1978) Estimating the dimension of a model, *Annals of Statistics*, **6**, 461-464.

[13] Smith, S. W. (1999) The Scientist and Engineer's Guide to Digital Signal Processing, 2nd edition *California Technical Publishing*, San Diego, USA.

[14] Storch, K-F., Lipan, O., Leykin, I., Viswanathan, N., Davis, F. C., Wong, W. H., Weitz, C. J. (2002) Extensive and divergent circadian gene expression in liver and heart, *Nature* **417**, 78-83.

[15] Wichert, S., Fokianos, K., Strimmer K. (2004) Identifying periodically expressed transcripts in microarray time series data, *Bioinformatics*, **20**, 1, 5-20.

[16] Ueda H. R, Chen, W., Adachi, A., Wakamatsu, H., Hayashi, S., Takasugi, T., Nagano, M., Nakahama, K., Suzuki, Y, Sugano, S., Iino, M., Shigeyoshi, Y., Hashimoto, S. (2002) A transcription factor response element for gene expression during circadian night, *Nature* **418**, 534-539.