Vol.2021-BIO-65 No.5
2021/3/11

同期性揺らぎ遺伝子の二段階抽出におけるパラメータ調整法

奥 牧人[1,a)]

概要：同期性揺らぎ遺伝子とは特定の条件下で発現量が同期的に大きく揺らぐ遺伝子集合のことである．それらは生体の恒常性維持機能の低下と関係している可能性がある．同期性揺らぎ遺伝子の二段階抽出では，まず揺らぎの大きな遺伝子を選択し，次にその中から同期性遺伝子クラスタを抽出する．この二段階法は 2 つの重要なパラメータを持つが，それらの調整の仕方は確立されていなかった．この問題に対処するために，本稿では，同期性揺らぎ遺伝子の二段階抽出におけるパラメータ調整法を提案する．提案手法は，パラメータを先行研究で使用していた値に固定した場合と比べて，様々な条件下における試行の 77 % で人工データに対する抽出性能を改善した．また，提案手法はパラメータ固定の場合と同様に実データに対する中程度の再現性を示した．

キーワード：トランスクリプトーム，同期性揺らぎ遺伝子，二段階法，パラメータ調整

Parameter adjustment method in two-step extraction of synchronously fluctuated genes

Makito OKU[1,a)]

Abstract: Synchronously fluctuated genes (SFGs) are a set of genes whose expressions fluctuate largely and synchronously under certain conditions. They may be associated with reduced homeostasis of living organisms. In the two-step extraction of SFGs, the genes with large fluctuations are selected first, and then synchronous gene clusters are extracted from them. The two-step method has two important parameters, but no method has been established to adjust them. In order to address this issue, in this paper, I propose a parameter adjustment method in two-step extraction of SFGs. The proposed method improved extraction performance for artificial data in 77 % of trials under various conditions, as compared to the case where the parameters were fixed at the values used in the previous study. In addition, the proposed method showed a moderate level of reproducibility for the real data as in the case of fixed parameters.

Keywords: transcriptome, synchronously fluctuated genes, two-step method, parameter adjustment

1. はじめに

トランスクリプトームデータとは，試料中に含まれる様々な mRNA の量を網羅的に計測したものであり，その解析手法開発は生命情報科学における重要な研究テーマの一つである．本稿では，トランスクリプトームデータから同期性揺らぎ遺伝子（Synchronously Fluctuated Genes, SFGs）[1, 2] を抽出する手法を扱う．同期性揺らぎ遺伝子とは「異なる条件間の比較において発現量の分布幅が顕著に増加したもののうち，互いの発現パターンが強く同期・相関した遺伝子集合のこと」[1] である（図 1）．

同期性揺らぎ遺伝子は生体の恒常性維持機能の低下と関係している可能性がある．ここでの仮説は，生体の恒常性維持機能が低下することにより，一部の遺伝子の発現量が一定値を保てなくなり，それらが相互作用により同期して大きく上下するだろう，というものである．この仮説が実際の生体に当てはまるとは限らないが，同期性揺らぎ遺伝子は，一般的に着目される発現変動遺伝子と共に，疾病の

[1] 富山大学 和漢医薬学総合研究所
Institute of Natural Medicine, University of Toyama, Toyama 930–0194, Japan
[a)] oku@inm.u-toyama.ac.jp

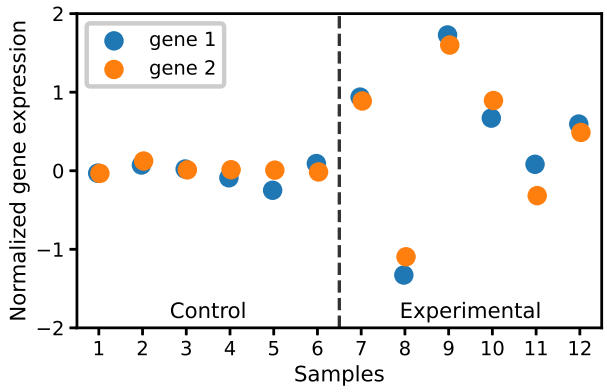

図 1 同期性揺らぎ遺伝子の発現パターンの例
Fig. 1 An example of an expression pattern of SFGs

発症メカニズムの解明や予防・治療に役立つ新たな知見をもたらすと期待されている [3,4].

同期性揺らぎ遺伝子の抽出法として，筆者はこれまでに二段階法，主成分分析に基づく方法その 1，主成分分析に基づく方法その 2 を開発した [1,2,5]．これら 3 つの手法のうちどれが最も良いかは分かっていない．本稿では二段階法を扱う．二段階法は 2 つの重要なパラメータを持つが，それらの調整の仕方は確立されていなかった．そこで本稿では，同期性揺らぎ遺伝子の二段階抽出におけるパラメータ調整法を提案する．

本稿の以降の構成について述べる．2 節では二段階法について，3 節では性能評価用の人工データについて，4 節では二段階法のパラメータ調整法について，5 節では再現性評価用の実データについて説明する．6 節で結果を示し，7 節でまとめと考察をする．

2. 二段階法

二段階法は，揺らぎの大きな遺伝子を選択する第一段階と，同期性遺伝子クラスタを抽出する第二段階から成る．第一段階では，各遺伝子について実験群データと対照群データそれぞれで中央絶対偏差を計算し，前者の値が後者の値の θ 倍より大きい遺伝子を選択する．パラメータ θ は調節可能であり，先行研究 [1,2] では 2 としていた．

第二段階では，第一段階で選択された遺伝子に関する実験群データに階層的クラスタリングを適用する．類似度にはスピアマンの相関係数，連結法には平均連結法，分割基準には類似度に対する閾値 ϕ を用いる．パラメータ ϕ は調節可能であり，先行研究 [1,2] では 0.75 としていた．

クラスタ分割の後，どのクラスタを取得するかを決める．先行研究 [1,2] では，最大クラスタ及びその半分より大きな二番目以降のクラスタを取得し，それらに属する遺伝子を同期性揺らぎ遺伝子として抽出していた．しかし，この方法には，ある程度大きなクラスタが存在しない場合も同期性揺らぎ遺伝子を出力してしまうという問題がある．

そこで，本研究ではクラスタ取得条件を変更し，対照群データにも実験群データと同様に階層的クラスタリングを適用し，対照群の最大クラスタの 1.2 倍より大きな実験群のクラスタを取得することにした．閾値 1.2 は状況に応じて変更しても問題ない．

3. 人工データ

性能評価用の人工データの生成法について説明する．まず，サンプル数 N，正例数 K，揺らぎ強度 γ を決める．次に，各要素が標準正規分布に独立に従う 10 000 行 N 列のデータ行列を 2 つ用意する．そして，平均が 0 で分散が $\gamma^2 - 1$ の正規分布に各要素が独立に従う長さ N の横ベクトルを用意し，それを片方のデータ行列の K 行分に繰り返し加算する．このデータ行列を実験群データ，他方を対照群データとする．実験群データのうち値を加算された K 行を同期性揺らぎ遺伝子の正例，残りを負例とする．

正例は，標準偏差の期待値が γ，互いの相関係数の期待値が $1 - \gamma^{-2}$ である．負例は，標準偏差の期待値が 1，互いの相関係数の期待値が 0 である．

4. パラメータ調整法

二段階法のパラメータ θ と ϕ の調整法について説明する．最初に，予備実験について説明する．様々な N, K, γ の人工データを生成し，それらに対して様々な θ と ϕ の組み合わせを試し，どうすれば F1 スコアが高くなるかを調べた．その結果，ϕ は N が大きいほど小さくした方が良いことと，最適な θ はデータ毎に異なり，N, K, γ にあまり強くは依存しないことが分かった．また，先行研究と同じ $\theta = 2, \phi = 0.75$ という設定は，様々な条件下で比較的良い性能を示し，固定値としては適切であることも分かった．

そこで，ϕ は以下のように調整することにした（図 2）：

$$\phi = \tanh \frac{3}{\sqrt{N-3}}, \tag{1}$$

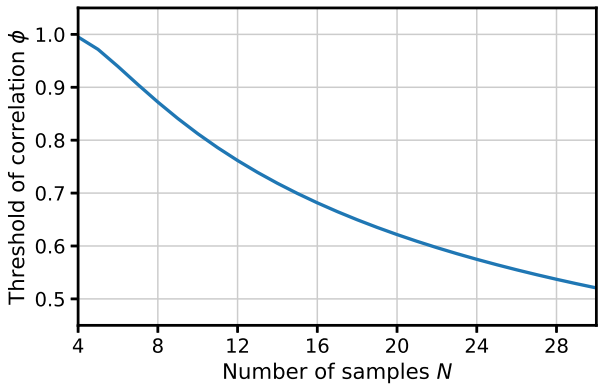

図 2 サンプル数 N と相関係数の閾値 ϕ の関係
Fig. 2 Relation between number of samples N and threshold of correlation ϕ

ただし $N \geq 4$ とする．この ϕ の値はピアソンの相関係数をフィッシャー変換した際の標準偏差の 3 倍に相当する．

続いて，パラメータ θ の調整法について説明する．まず，$\theta = 2$ として同期性揺らぎ遺伝子を仮抽出し，それらの中央絶対偏差比の平均 z を計算する：

$$z = \frac{1}{|S|} \sum_{i \in S,\ d_Y(i) \neq 0} \frac{d_X(i)}{d_Y(i)}, \tag{2}$$

ここで，S は同期性揺らぎ遺伝子の番号集合，$d_X(i)$ と $d_Y(i)$ は遺伝子 i の実験群および対照群データの中央絶対偏差をそれぞれ表す．この z は，$\theta \in \{1.5,\ 1.75,\ 2,\ 2.25,\ 2.5\}$ の中で F1 スコアに関して最適な θ の値と正の相関を持つ（図 3）．そこで，最適な θ の値毎の z の中央値を用いて線形回帰を行い，以下のように θ を調節することにした（図 4）：

$$\theta = \begin{cases} \theta_{\min} & \text{if} \quad az + b \leq \theta_{\min}, \\ az + b & \text{if} \quad \theta_{\min} < az + b \leq \theta_{\max}, \\ \theta_{\max} & \text{if} \quad \theta_{\max} < az + b, \end{cases} \tag{3}$$

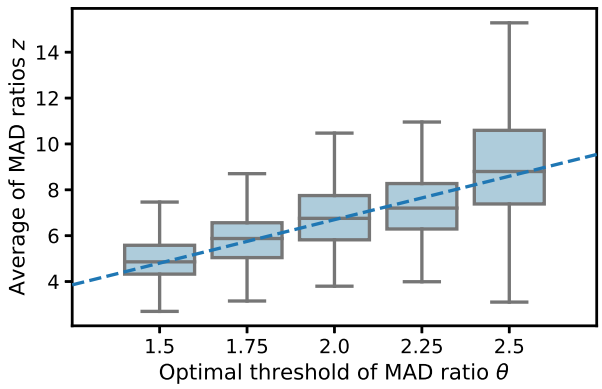

図 3 最適な閾値 θ と中央絶対偏差比の平均 z の関係
Fig. 3 Relation between optimal threshold of median absolute deviation (MAD) ratio θ and average of MAD ratios z

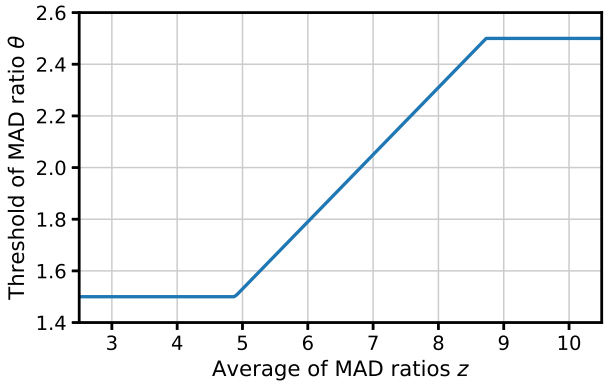

図 4 中央絶対偏差比の平均 z と中央絶対偏差比の閾値 θ の関係
Fig. 4 Relation between average of MAD ratios z and threshold of MAD ratio θ

ここで，$\theta_{\min} = 1.5$, $\theta_{\max} = 2.5$, $a = 0.26$, $b = 0.23$ である．閾値 θ に上限と下限を設けた理由は，設けない場合と比べて性能が良くなるからである．

5. 実データ

再現性評価用の実データについて説明する．実データは GSE77578 [6] の 4 条件のうち，溶媒のみ投与された癲癇モデルマウス（$N = 17$）と PLX3397 を 3 mg/kg 投与された癲癇モデルマウス（$N = 18$）のデータをそれぞれ対照群，実験群データとして用いた．

先行研究 [5] において，このデータにはサンプル番号と関連したブロックパターン（図 5）が含まれ，その該当遺伝子の一部が同期性揺らぎ遺伝子として抽出される場合があることが分かった．しかし，このようなパターンは生体揺らぎとは考え難い．そこで，本研究では遺伝子毎にラグ 1 の自己相関を計算し，フィッシャー変換した際に標準偏差の 1 倍範囲の外側だった場合，該当する遺伝子データを予め除外した．

再現性評価法は先行研究 [1, 2, 5] と同じやり方にした．最初に全データを用いて同期性揺らぎ遺伝子を抽出し，それ以降は実験群データから 1 サンプルずつ順に取り除きながら同様に同期性揺らぎ遺伝子を抽出した．取り除くサンプルは，最初の結果との重複が最も少なくなるような最悪ケースのものとし，重複度は Jaccard 指数で評価した．

最後に，実データから抽出された同期性揺らぎ遺伝子のエンリッチメント解析を行った．解析には DAVID（Database for Annotation, Visualization and Integrated Discovery）データベース [7] バージョン 6.8 を用いた．GO（Gene Ontology）の生物学的プロセスに関する注釈のうち入力した遺伝子リストとの重複数が 2 以上のものを取り出し，フィッシャーの正確検定で p 値を求めて偽発見率制御を適用した．

6. 結果

図 6 に，先行研究 [1, 2] と同じ $K = 500$, $\gamma = 5$ の人工

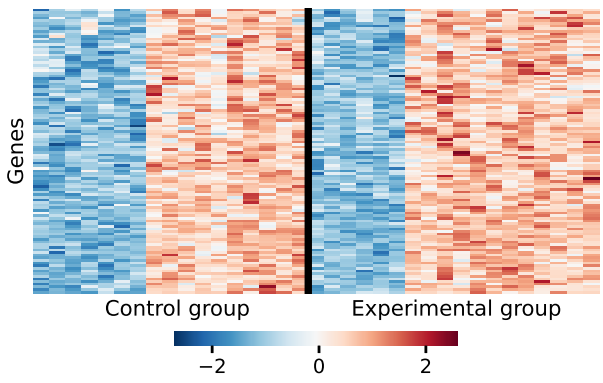

図 5 実データに含まれるブロックパターンの例
Fig. 5 An example block pattern in the real data

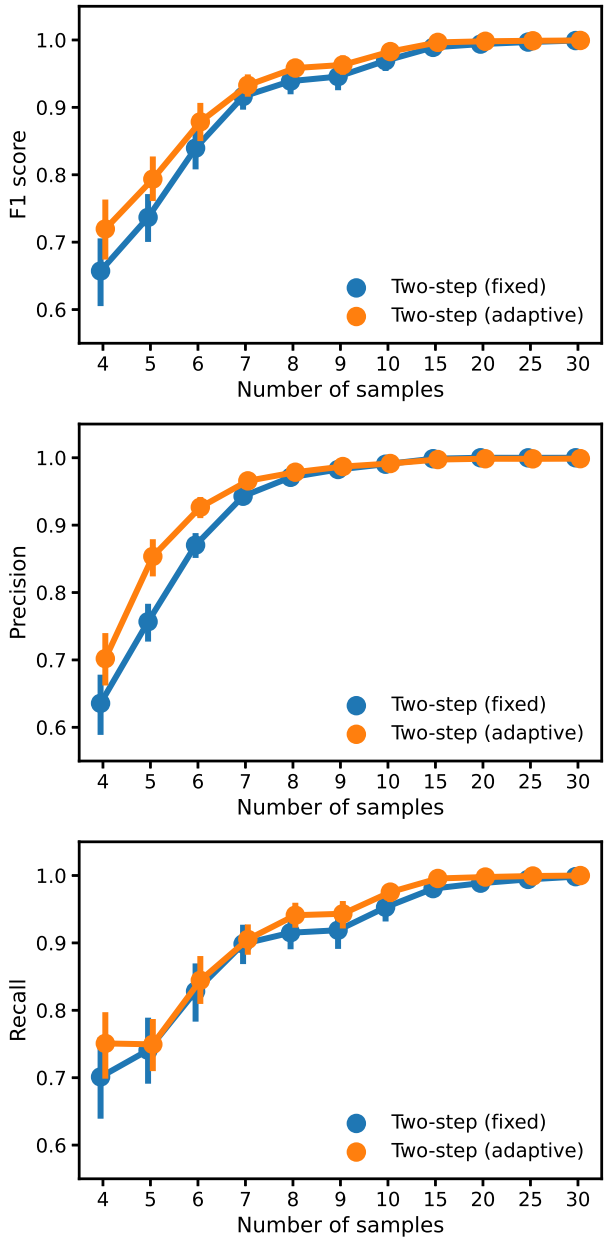

図 6 固定パラメータと適応的パラメータを用いた二段階法の人工データに対する F1 スコア，適合率，再現率（$K = 500$, $\gamma = 5$, 100 回試行，エラーバーは 95 %信頼区間）

Fig. 6 F1 score, precision, and recall of the two-step method with fixed and adaptive parameters for the artificial data ($K = 500$, $\gamma = 5$, 100 trials, error bars show 95 % confidence intervals

データを用いた抽出性能評価の結果を示す．固定パラメータでは先行研究と同じ $\theta = 2$, $\phi = 0.75$ とし，適応的パラメータでは式 1 から式 3 を用いて θ と ϕ を調節した．この条件下では，F1 スコア，適合率，再現率の全てに関して，適応的パラメータは固定パラメータと同程度またはそれを上回る性能を示した．

図 7 に様々な条件の人工データを用いた抽出性能評価の結果を示す．人工データの各パラメータの範囲は $N \in$

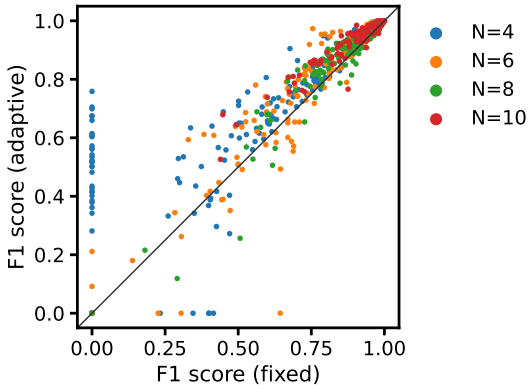

図 7 様々な条件下における固定パラメータと適応的パラメータを用いた二段階法の人工データに対する F1 スコア

Fig. 7 F1 score of the two-step method with fixed and adaptive parameters for the artificial data under various conditions

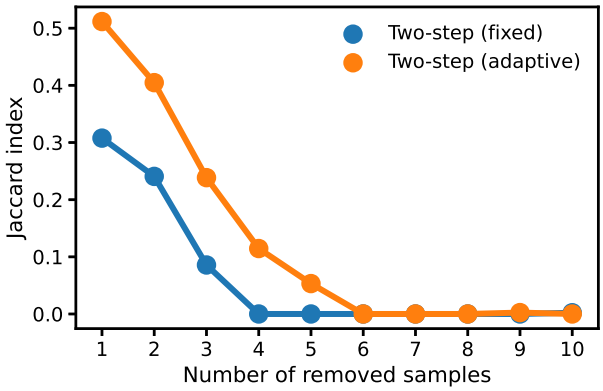

図 8 固定パラメータと適応的パラメータを用いた二段階法の実データに対する Jaccard 指数

Fig. 8 Jaccard index of the two-step method with fixed and adaptive parameters for the real data

$\{4, 6, 8, 10\}$, $K \in \{100, 200, 500, 1000\}$, $\gamma \in \{3, 4, 5, 6\}$ とし，各条件毎に 10 回ずつ試した．斜線より上側の点は，適応的パラメータの方が固定パラメータよりも F1 スコアが高かった場合を示す．全体の 77 %の点が斜線より上側にあり，18 %の点が斜線より下側にあった．残りは斜線上だった．また，θ を 2 に固定し ϕ のみを式 1 により調節した場合は，斜線より上が 53 %，下が 33 %だった．

図 8 に実データを用いた再現性評価の結果を示す．適応的パラメータは固定パラメータより高い再現性を示した．図 9 に適応的パラメータを用いて抽出された同期性揺らぎ遺伝子のヒートマップを示す．実験群と対照群を比べると，これらの遺伝子は発現量の分布幅が増加し，互いに強く同期していた．表 1 にそれらのエンリッチメント解析の結果を示す．いずれの GO 注釈も重複数が少ないものの，細胞接着，細胞遊走，脳の発生などと関わる遺伝子が含まれることが分かった．

表 1 適応的パラメータを用いた二段階法により実データから抽出された同期性揺らぎ遺伝子（$n = 33$）の GO エンリッチメント解析の結果（q 値 <0.05；14 個中上位 10 個を表示）．多重比較数は 38，総遺伝子数は 18082.

Table 1 Enriched GO annotations in the SFGs ($n = 33$) extracted from the real data by the two-step method with adaptive parameters (q-value < 0.05; top 10 annotations among 14 are shown). The number of multiple comparisons was 38, and the total gene number was 18082.

GO 注釈	重複数	その注釈を持つ遺伝子数	p 値	q 値
calcium-independent cell-cell adhesion via plasma membrane cell-adhesion molecules	2	24	8.7E-04	2.3E-02
acute-phase response	2	35	1.9E-03	2.3E-02
positive regulation of epithelial cell migration	2	36	2.0E-03	2.3E-02
midbrain development	2	40	2.4E-03	2.3E-02
insulin receptor signaling pathway	2	51	3.9E-03	3.0E-02
odontogenesis of dentin-containing tooth	2	63	5.9E-03	3.7E-02
chloride transport	2	80	9.3E-03	4.8E-02
forebrain development	2	86	1.1E-02	4.8E-02
protein heterooligomerization	2	91	1.2E-02	4.8E-02
wound healing	2	94	1.3E-02	4.8E-02

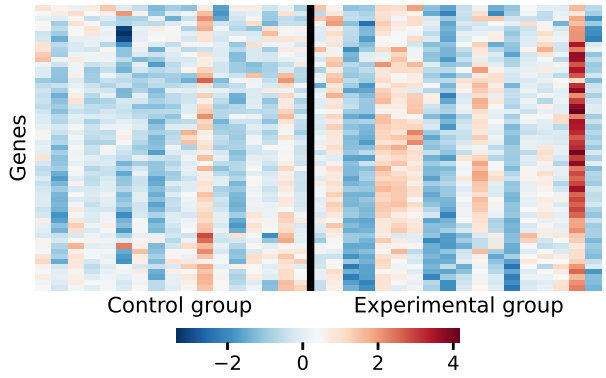

図 9 適応的パラメータを用いた二段階法により実データから抽出された同期性揺らぎ遺伝子のヒートマップ

Fig. 9 Heatmap of SFGs extracted from the real data by the two-step method with adaptive parameters

7. まとめと考察

本稿では，同期性揺らぎ遺伝子抽出のための二段階法のパラメータ θ と ϕ の調整法を提案した（式 1 から式 3）．提案手法は，先行研究と同じ $\theta = 2, \phi = 0.75$ に固定した場合と比べて，様々な条件下における試行の 77 %で人工データに対する抽出性能を改善した（図 7）．また，提案手法はパラメータ固定の場合と同様に実データに対する中程度の再現性を示した（図 8）．

提案手法の限界と問題点について述べる．まず，式 1 はピアソンの相関係数に関するものだが，実施に使用するのはスピアマンの相関係数であり，より適切な式があるかもしれない．そもそもフィッシャー変換自体が近似式である．次に，式 2 の z は最適な θ と良く相関する統計量だが（図 3），これより良い統計量があるかもしれない．出来れば同期性揺らぎ遺伝子の仮抽出を必要としないものの方が計算時間の観点からも良い．さらに，式 3 のパラメータは人工データから求めたものだが，それらが実データに対しても良いとは限らない．最後に，提案手法を用いることで固定パラメータより性能が悪くなる場合があった（図 8）．サンプル数 N が小さいときに多いが，その対策を考えることは今後の課題である．

実データを用いた再現性評価の固定パラメータの結果（図 8）が先行研究 [1,2] より悪化した理由について考察する．解析手順の変更箇所は，クラスタ取得とブロックパターン除去の 2 つである．いずれも結果の再現性を必ずしも悪化させるとは限らず，データによっては結果を改善する場合も考えられる．問題は，データの前処理や後処理を少し変えただけで，出力される遺伝子リストがしばしば大きく変わることである．結果をさらに安定させることが今後の課題である．

参考文献

[1] 奥　牧人：同期性揺らぎ遺伝子の二つの新規抽出法，情処研報，Vol. 2018-BIO-56, No. 1, pp. 1–6（オンライン），入手先 ⟨http://id.nii.ac.jp/1001/00192709/⟩ (2018).
[2] Oku, M.: Two novel methods for extracting synchronously fluctuated genes, *TBIO*, Vol. 12, pp. 9–16 (online), DOI: https://doi.org/10.2197/ipsjtbio.12.9 (2019).
[3] Koizumi, K., Oku, M., Hayashi, S., Inujima, A., Shibahara, N., Chen, L., Igarashi, Y., Tobe, K., Saito, S., Kadowaki, M. and Aihara, K.: Identifying pre-disease signals before metabolic syndrome in mice by dynamical network biomarkers, *Sci. Rep.*, Vol. 9, p. 8767 (online), DOI: https://doi.org/10.1038/s41598-019-45119-w (2019).
[4] Koizumi, K., Oku, M., Hayashi, S., Inujima, A., Shibahara, N., Chen, L., Igarashi, Y., Tobe, K., Saito,

S., Kadowaki, M. and Aihara, K.: Suppression of dynamical network biomarker signals at the pre-disease state (Mibyou) before metabolic syndrome in mice by a traditional Japanese medicine (Kampo formula) bofutsushosan, *eCAM*, Vol. 2020, p. 9129134 (online), DOI: https://doi.org/10.1155/2020/9129134 (2020).

[5] 奥 牧人：もう一つの主成分分析に基づく同期性揺らぎ遺伝子抽出法，情処研報，Vol. 2020-BIO-57, No. 2, pp. 1–6（オンライン），入手先 ⟨http://id.nii.ac.jp/1001/00194900/⟩ (2019).

[6] Srivastava, P. K., van Eyll, J., Godard, P., Mazzuferi, M., Delahaye-Duriez, A., Steenwinckel, J. V., Gressens, P., Danis, B., Vandenplas, C., Foerch, P., Leclercq, K., Mairet-Coello, G., Cardenas, A., Vanclef, F., Laaniste, L., Niespodziany, I., Keaney, J., Gasser, J., Gillet, G., Shkura, K., Chong, S.-A., Behmoaras, J., Kadiu, I., Petretto, E., Kaminski, R. M. and Johnson, M. R.: A systems-level framework for drug discovery identifies Csf1R as an anti-epileptic drug target., *Nat. Comm.*, Vol. 9, No. 1, p. 3561 (online), DOI: https://doi.org/10.1038/s41467-018-06008-4 (2018).

[7] Huang, D. W., Sherman, B. T. and Lempicki, R. A.: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, *Nat. Protoc.*, Vol. 4, No. 1, pp. 44–57 (online), DOI: https://doi.org/10.1038/nprot.2008.211 (2009).