Vol.2019-BIO-57 No.2
2019/3/8

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

奥 牧人[1,a)]

概要：本稿では，前回のバイオ情報学研究会で発表した，遺伝子発現量データから同期性揺らぎ遺伝子（Synchronously Fluctuated Genes，SFGs）を抽出するための方法とは別のやり方について述べる．前回発表した 2 つの方法のうちの片方では主成分分析を用いていたが，本稿ではその前処理と後処理の部分を変えたものを提案する．具体的には，前処理部において中央絶対偏差と順位を用いたデータの変換を行い，後処理部ではカイ二乗分布による外れ値検出法を用いるよう変更した．人工データおよび実データを用いてこの手法の性能評価を行ったところ，前回提案した 2 つの方法と比べて良い点だけでなく悪い点もあることが分かった．今回新たに提案した手法は，同期性揺らぎ遺伝子抽出法の選択肢の一つにはなり得るのではないかと考えられる．

キーワード：遺伝子発現量データ，同期性揺らぎ遺伝子，主成分分析，外れ値検出

An alternative method for extracting synchronously fluctuated genes based on principal component analysis

Makito OKU[1,a)]

Abstract: In this paper, I propose an alternative method for extracting synchronously fluctuated genes (SFGs) from gene expression data, which is different from the methods I proposed at the previous seminar. Two methods were proposed last time, and one of them was based on principal component analysis. By modifying its pre- and post-processing steps, a new method is developed. At the pre-processing step, original data is transformed using median absolute deviations and ranks. At the post-processing step, outliers are detected using a chi-squared distribution. The performance of the proposed method was evaluated using artificial and real data, and its advantages as well as disadvantages against the two methods proposed last time were identified. The newly proposed method would be regarded as an optional method for extracting SFGs.

Keywords: gene expression data, synchronously fluctuated gene, principal component analysis, outlier detection

1. はじめに

同期性揺らぎ遺伝子（Synchronously Fluctuated Genes，SFGs）[1] とは，通常の遺伝子発現量データ解析において注目される発現変動遺伝子（Differentially Expressed Genes，DEGs）とは異なる性質を持った遺伝子である．発現変動遺伝子が図 1 に示すように異なる条件間の比較において発現量の平均値が顕著に増加または減少した遺伝子のことを指すのに対し，同期性揺らぎ遺伝子は図 2 に示すように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，かつ，互いの活動パターンが強く同期・相関した遺伝子集合のことを指す．揺らぎの異常な増加やそれに伴う隠れた相関性の顕在化は，生体の安定性や恒常性維持機能の低下と関わる可能性があるため，例えば病気に対する抵抗力が弱まっている疾病前状態 [2] の特徴付けに使えるのではないかと期待されている．

同期性揺らぎ遺伝子を遺伝子発現量データから取り出す

[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

© 2019 Information Processing Society of Japan

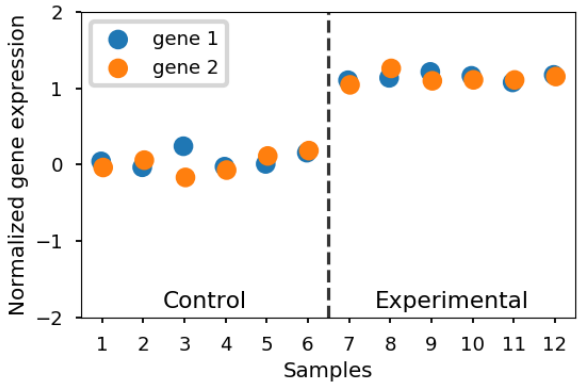

図 1　発現変動遺伝子の発現パターンの例
Fig. 1　An example of an expression pattern of DEGs

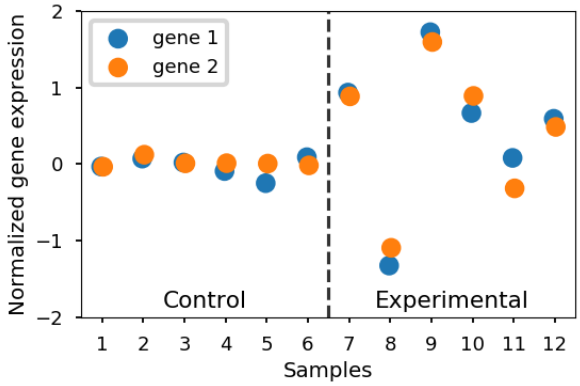

図 2　同期性揺らぎ遺伝子の発現パターンの例
Fig. 2　An example of an expression pattern of SFGs

ための手法がこれまでに幾つか提案されてきた [2–5]．既存の手法はいずれも外れ値に弱いという問題があったため，筆者は前回のバイオ情報学研究会において外れ値に強い同期性揺らぎ遺伝子抽出法を 2 つ提案した [1]．一つ目の手法は，まず最初に対照群と実験群を比較し中央絶対偏差（Median Absolute Deviation，MAD）が 2 倍より増えた遺伝子を選択した後，実験群データのスピアマンの相関係数に基づいて階層的クラスタリングをかけ，サイズの大きなクラスタを取り出すというものである．揺らぎと相関性を別々に扱っていることから，この手法のことを本稿でも引き続き二段階法と呼ぶ．

二つ目の手法では，対照群と実験群それぞれのデータに関して，前処理として外れ値を 0 に置換した後，主成分分析（Principal Component Analysis，PCA）をかけて各遺伝子の寄与度を算出する．そして，後処理として対照群と実験群それぞれにおける寄与度を比較し，大きく増加したものを外れ値検出法の一つである 3σ 則により取り出す．本稿ではこの手法のことを，後述の手法と区別するため，PCA 法その 1 と呼ぶ．

二つ目の手法では，主成分分析をかけた後で実験群と対照群を比較していた．しかし，図 3 に示すように，同一条件下で測定された少数サンプルのデータから計算された相関係数のほとんどは再現性が低いため，対照群の相関係数

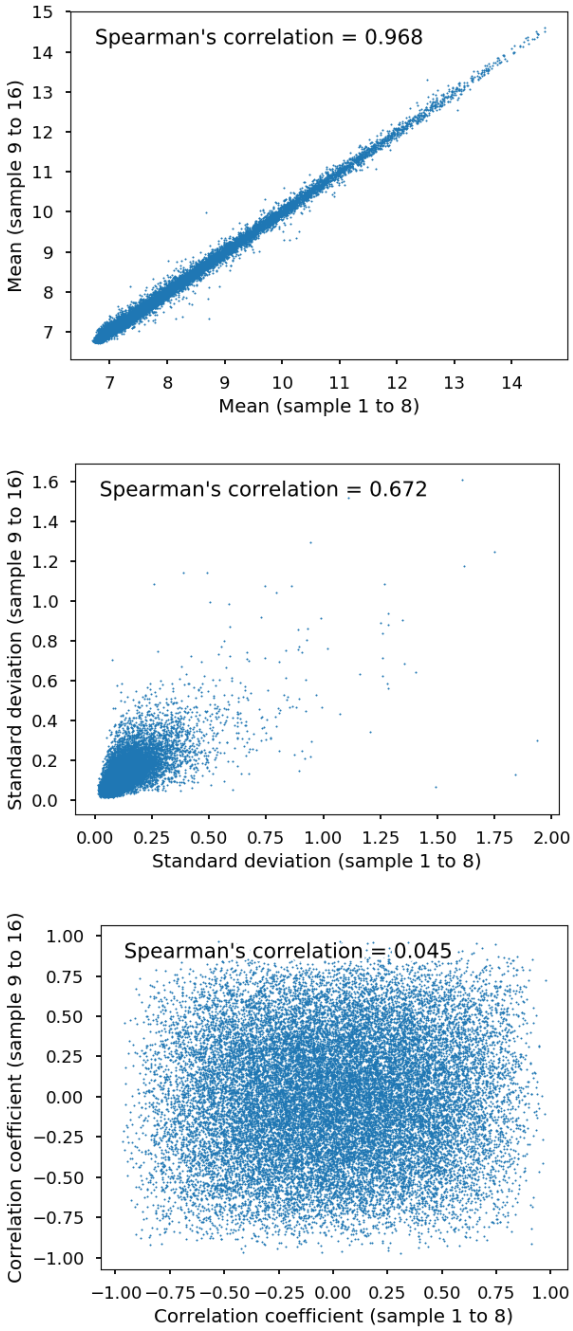

図 3　同一条件下における平均，標準偏差，相関係数の再現性の比較．GSE77578 で溶媒のみ投与された癲癇モデルマウスのデータ（$N = 17$）のうち最初の 8 サンプルと次の 8 サンプルを使用．下の図はサイズを減らしたデータを使用．
Fig. 3　Comparison of the reproducibility of means, standard deviations, and correlation coefficients under the same conditions. The first 8 samples and the next 8 samples of the data of epileptic mice administered with vehicle only ($N = 17$) in GSE77578 were used. The bottom plot used a reduced-size data.

© 2019 Information Processing Society of Japan

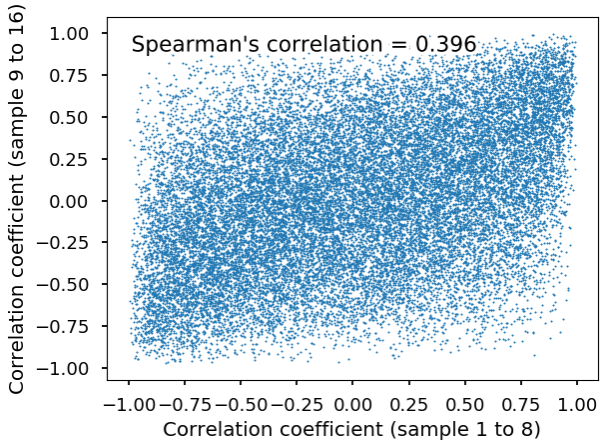

図 4　揺らぎの大きな遺伝子間の同一条件下における相関係数の再現性．最初の 8 サンプルで計算した中央絶対偏差の上位 1000 遺伝子を使用．

Fig. 4　Reproducibility of correlation coefficients between genes with large fluctuations under the same conditions. Top 1000 genes with the largest MADs calculated using the first 8 samples were used.

やそれと関連した共分散は比較対象として妥当でない可能性がある．ただし，少数サンプルで計算された全ての相関係数の再現性が低いという訳ではなく，例えば図 4 に示すように揺らぎの大きな遺伝子間に限ればある程度の再現性が保たれる．従って，実験群において揺らぎの増大した遺伝子間の相関性の情報を同期性揺らぎ遺伝子の抽出に用いること自体にはある程度の妥当性があるだろう．なお，二段階法では元々対照群の相関性は考慮していなかったため，この問題による影響はないと考えられる．

以上を踏まえ，本稿では二つ目の手法を基にしつつ，対照群の相関性を考慮しない新たな手法を提案する．この手法を本稿では PCA 法その 2 と呼ぶ．前処理では，外れ値を 0 で置き換える代わりに外れ値に強い中央絶対偏差と順位を用いてデータを変換するようにした．後処理では，寄与度の大きな遺伝子の抽出法として 3σ 則の代わりに χ^2 分布による外れ値検出法を使用した．

以降では，提案手法の詳細，性能評価法，結果，まとめと考察について順に述べる．

2. 提案手法: PCA 法その 2

入力として実験群のデータ行列 X と対照群のデータ行列 Y が与えられたとする．行方向には遺伝子や転写産物，列方向にはサンプルがそれぞれ並んでいる．行数は X と Y で一致している必要がある．列数は同じである必要は無い．前処理として，実験群データの i 行 k 列の値 $x(i,k)$ を，全ての $i,\ k$ に関して以下の値で置き換える:

$$\tilde{x}(i,k) = \frac{d_X(i)}{d_Y(i)+c}\, r_X(i,k), \qquad (1)$$

ここで $d_X(i)$ は実験群の遺伝子 i の中央絶対偏差，$d_Y(i)$ は対照群の遺伝子 i の中央絶対偏差を表す．分母の $c > 0$ は調節可能なパラメータであり，本稿では対照群の全遺伝子の中央絶対偏差の 1 %点の値とした．また，$r_X(i,k)$ は実験群の $x(i,k)$ の昇順の順位を表す．ただし，順位は各行毎に計算するものとする．

変換後のデータで遺伝子間のピアソンの相関係数を計算すると，元データにおけるスピアマンの相関係数の値になる．また，変換後のデータで各遺伝子の不偏分散を計算すると，データの列数を N として，元データにおける中央絶対偏差の比を c で補正したものに $N(N+1)/12$ を掛けた値になる．なお，予めその平方根で割っておけば，比の部分が 1 のとき不偏分散が 1 となるようにすることも出来る．

変換後のデータを主成分分析にかけ，第一主成分方向に対する各遺伝子の重みを計算する．その分布が比較的正規分布に近い形をしていることから，正規分布を当てはめて p 値を求め，外れ値判定に使用する．実装上は，標準正規分布に従う確率変数の 2 乗が自由度 1 の χ^2 分布に従うことから，その生存関数を使って p 値を計算することが出来る [6]．本手法では標準化手順として，中央値を差し引いた後，中央絶対偏差で割り，正規分布の 75 %点の値（約 0.6745）を掛けるようにした．有意水準は $\alpha = 0.05$ とし，多重比較に対応するため Bonferroni 補正を用いた．なお，偽発見率（False Discovery Rate，FDR）制御を用いた場合は後述の性能評価において適合率が比較的低くなってしまったため，本手法では Bonferroni 補正を選んだ．

3. 性能評価法

前回提案した 2 つの手法と今回新たに提案した手法の性能評価法について簡単に説明する．内容は前回とほぼ同一のため，詳細は文献 [1] および [7] を参照されたい．

まず，正解が分かっている人工データを用意し，各手法の抽出精度を評価した．人工データは実験群データ X と対照群データ Y からなり，それぞれ 10 000 行 N 列とする．ただし N は可変とする．X の上から 500 行を同期性揺らぎ遺伝子の正例，残り 9500 行を負例とし，正例では標準偏差の期待値が 5，互いの相関係数の期待値が 0.96，負例及び対照群データでは標準偏差の期待値が 1，互いの相関係数の期待値が 0 となるように人工データを生成した．性能評価には F1 スコア，適合率，再現率を用いた．

次に，実データを用いて各手法の再現性を評価した．実データは GSE77578 [8] 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．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

© 2019 Information Processing Society of Japan

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 [9] 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，Fisher の正確検定で p 値を求めて FDR 制御を適用した．

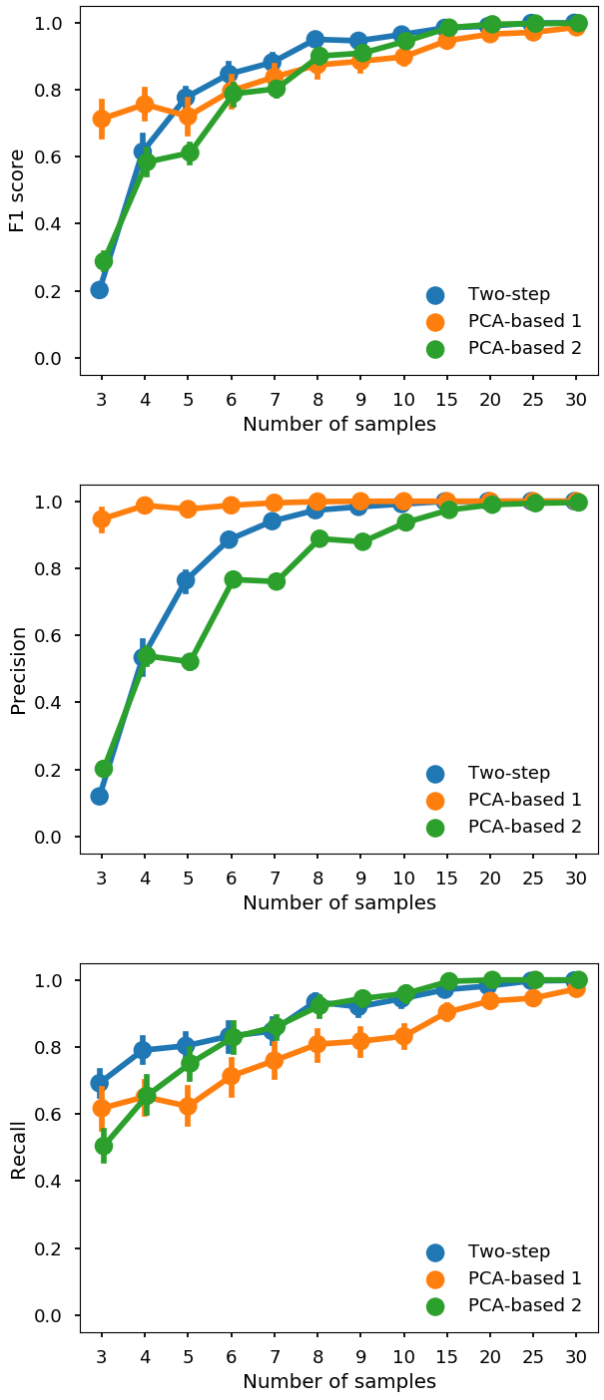

図 5　人工データに対する 3 つの手法の F1 スコア，適合率，再現率（100 回試行，エラーバーは 95 %信頼区間）

Fig. 5　F1 score, precision, and recall of the three methods for the artificial data (100 trials, error bars show 95 % confidence intervals)

4. 結果

図 5 に人工データに対する 3 つの手法の F1 スコア，適合率，再現率を示す．PCA 法その 2 は $N \geq 6$ で他の 2 つと同程度の F1 スコアを示した．同手法の適合率は他の 2 つより全体的に低かったが，再現率は $N \geq 6$ で二段階法と同程度に高かった．図 6 に実データに対する再現性評価の結果を示す．PCA 法その 2 は 1 サンプル除外時の Jaccard 指数が他の 2 つより高く，それ以降も比較的高い値を示した．図 7 に PCA 法その 2 によって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を示す．実験群のサンプル全体を通して分布幅や相関性が高まっていることが分かる．ここまでの結果より，PCA 法その 2 は人工データに対する適合率がやや低いものの，人工データに対する再

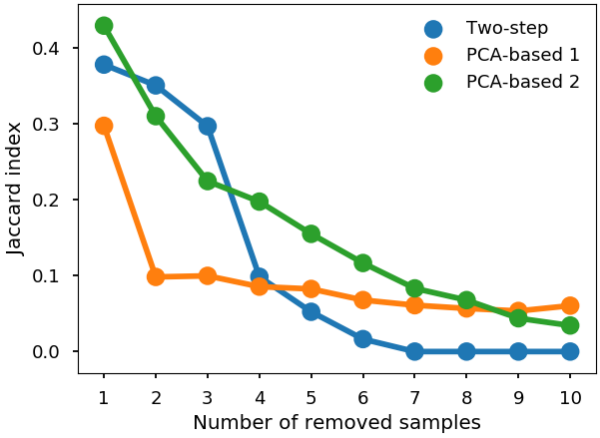

図 6　実データに対する 3 つの手法の Jaccard 指数

Fig. 6　Jaccard indices of the three methods for the real data

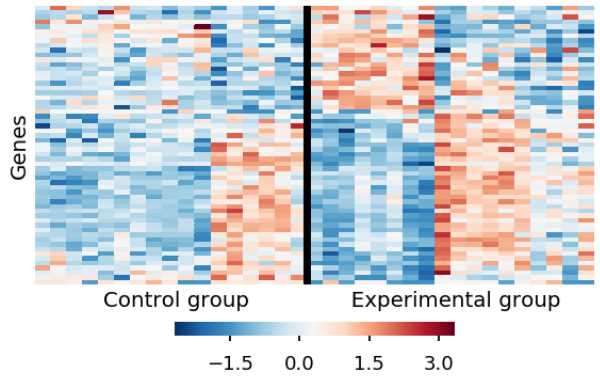

図 7　実データから PCA 法その 2 により抽出された同期性揺らぎ遺伝子のヒートマップ

Fig. 7　Heatmap of SFGs extracted from the real data by the PCA-based method 2

© 2019 Information Processing Society of Japan

表 1　実データから PCA 法その 2 により抽出された同期性揺らぎ遺伝子（$n = 42$）の GO エンリッチメント解析の結果（q 値 <0.05；重複数の多い順；26 個中上位 10 個を表示）．多重比較数は 446，総遺伝子数は 17911．

Table 1　Enriched GO annotations in the SFGs ($n = 42$) extracted from the real data by the PCA-based method 2 (q-value < 0.05; sorted by the overlap size; top 10 annotations among 26 are shown). The number of multiple comparisons was 446, and the total gene number was 17911.

GO 注釈	重複数	その注釈を持つ遺伝子数	p 値	q 値
cellular component biogenesis	15	2654	6.5E-04	2.6E-02
macromolecular complex subunit organization	14	2229	3.6E-04	2.6E-02
cellular component assembly	14	2418	8.3E-04	2.7E-02
macromolecular complex assembly	13	1544	3.3E-05	1.4E-02
protein complex subunit organization	12	1414	6.6E-05	1.4E-02
protein complex assembly	11	1277	1.3E-04	1.4E-02
protein complex biogenesis	11	1278	1.3E-04	1.4E-02
cellular macromolecular complex assembly	9	887	1.8E-04	1.6E-02
cellular macromolecule catabolic process	7	828	2.9E-03	4.9E-02
protein oligomerization	6	524	1.3E-03	3.4E-02

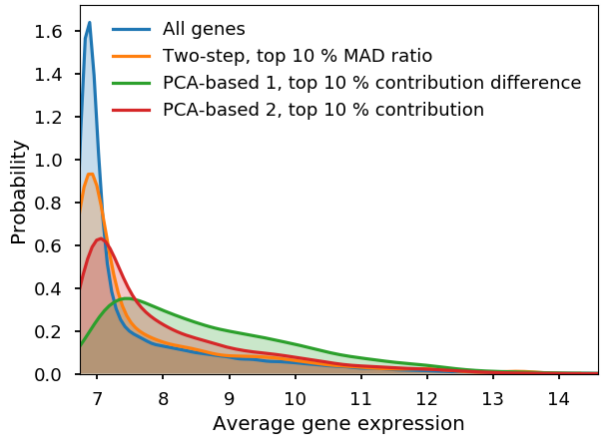

図 8　3 つの手法によるスコア付けで上位 10 %に含まれる遺伝子の平均発現量の分布

Fig. 8　Distributions of the average expressions of the top 10 % genes scored by the three methods

現率や実データに対する再現性が比較的良好であり，ヒートマップからも前回指摘した「一点問題」の発生は見られないため，同期性揺らぎ遺伝子抽出法として悪くない方法であることが示唆される．

図 8 に各手法によるスコア付けで上位 10 %に位置付けられた遺伝子の平均発現量の分布を示す．PCA 法その 1 では高発現遺伝子が優先的に選ばれる傾向があったが，PCA 法その 2 ではその傾向が弱まっていることが分かる．実際には高発現遺伝子に対する選好性はパラメータ c の値に依存し，値が小さければ二段階法の分布に，値が大きければ PCA 法その 1 の分布に近づく．しかし，c をどれだけ小さくしても二段階法と PCA 法その 2 の実データに対する結果はあまり一致しなかった．このことから，平均発現量以外の何らかの性質に関して，2 つの手法で抽出される遺伝子の傾向に違いがあることが示唆される．

表 1 に PCA 法その 2 で実データから抽出した同期性揺らぎ遺伝子のエンリッチメント解析の結果を示す．これらの注釈はいずれも，複数の部品を組み合わせて蛋白質複合体を作る過程に関わるもののようである．何故そのような機能に関する遺伝子群が薬物投与を受けた癲癇モデルマウスの海馬で同期性揺らぎを示したのかは詳しく調べないと分からない．いずれにせよ，この結果は本手法によって何らかの機能的まとまりを持った遺伝子群が抽出され得ることを示している．

5.　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新たな手法を一つ提案し，その性能評価結果について報告した．その結果，前回提案した 2 つの方法と比べて，実データに対する再現性が全体的に改善するなど良い面もあった（図 6）一方で，人工データに対する適合率がやや低いなどの悪い面も見つかった（図 5）．これらを総合的に判断すると，本手法は同期性揺らぎ遺伝子抽出法の選択肢の一つにはなり得るのではないかと考えられる．

二段階法と PCA 法その 2 の結果があまり一致しなかった理由について考察する．いずれの手法も実験群の各遺伝子の中央絶対偏差および順位と対照群の各遺伝子の中央絶対偏差に基づいて遺伝子選択を行う点は同じである．しかし，二段階法では遺伝子間の相関性は第一ステップで選択された遺伝子集合内のみで考慮されているのに対し，PCA 法その 2 では全遺伝子間の相関性が考慮される点が異なる．従って，おそらく二段階法では局所的な同期性揺らぎを示す遺伝子が選ばれやすく，一方で PCA 法その 2 ではより多数の遺伝子が関与する同期性揺らぎに関して寄与度の大きな遺伝子が選ばれやすいのではないかと予想される．詳

© 2019 Information Processing Society of Japan

細な検討は今後の課題とする．なお，パラメータ c の有無は本質的な違いではないと考えられる．その理由は，PCA 法その 2 で c をどれだけ小さくしても，その反対に二段階法に c を導入してどれだけ大きな値に設定しても，2 つの手法の結果をよく一致させることが出来なかったためである．

図 7 のヒートマップについて注意を述べる．この図の行の順番は見やすいように階層的クラスタリングを用いて並べ替えてあるが，列の順番は元データと同じである．それにも関わらず，実験群の左端の 8 サンプルとその隣の約 6 サンプルは，それぞれのグループ内でよく似た発現パターンを示している．さらに，対照群の右端の約 6 サンプルの発現パターンも互いにある程度似ている．このようなブロック状の発現パターンが偶然発生することも考えられるが，系統誤差などが影響した可能性も考慮すべきである．例えば，測定機器の調子，測定の時間帯，実験者など何らかの要因が遺伝子発現パターンのブロックの境目で変化したかもしれない．同様のブロック化現象は他のデータセットから抽出した同期性揺らぎ遺伝子に関しても観測されている．今後，系統誤差の影響が疑われた場合にそれが偶然によるものかどうかの良い判定法が見つかることを期待する．

複数クラスタがある場合の対応について述べる．PCA 法その 1 とその 2 はいずれも第一主成分のみを用いており，逆位相の場合を除いては 2 つ以上のクラスタを同時に取り出すことが出来ない．この問題は 2 番目以降の主成分も考慮すれば解決できる可能性がある．しかし，人工データに複数のクラスタを導入して調べたところ，必要以上に多数の主成分を用いた場合には誤検出の割合が大幅に増えることが分かった．従って，何番目までの主成分を用いるかを正確に見積もることが重要である．今後，同期性揺らぎ遺伝子抽出に適した主成分数決定法が見つかることを期待する．

謝辞 本研究は JSPS 科研費 JP15H05707 の助成を受けたものである．

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] Chen, L., Liu, R., Liu, Z.-P., Li, M. and Aihara, K.: Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers, *Sci. Rep.*, Vol. 2, p. 342 (online), DOI: https://doi.org/10.1038/srep00342 (2012).

[3] Li, Y., Jin, S., Lei, L., Pan, Z. and Zou, X.: Deciphering deterioration mechanisms of complex diseases based on the construction of dynamic networks and systems analysis, *Sci. Rep.*, Vol. 5, p. 9283 (online), DOI: https://doi.org/10.1038/srep09283 (2015).

[4] Vafaee, F.: Using multi-objective optimization to identify dynamical network biomarkers as early-warning signals of complex diseases, *Sci. Rep.*, Vol. 6, p. 22023 (online), DOI: https://doi.org/10.1038/srep22023 (2016).

[5] Richard, A., Boullu, L., Herbach, U., Bonnafoux, A., Morin, V., Vallin, E., Guillemin, A., Papili Gao, N., Gunawan, R., Cosette, J., Arnaud, O., Kupiec, J.-J., Espinasse, T., Gonin-Giraud, S. and Gandrillon, O.: Single-cell-based analysis highlights a surge in cell-to-cell molecular variability preceding irreversible commitment in a differentiation process, *PLOS Biol.*, Vol. 14, No. 12, p. e1002585 (online), DOI: https://doi.org/10.1371/journal.pbio.1002585 (2016).

[6] Taguchi, Y.-H.: Principal components analysis based unsupervised feature extraction applied to gene expression analysis of blood from dengue haemorrhagic fever patients., *Sci. Rep.*, Vol. 7, p. 44016 (online), available from ⟨https://doi.org/10.1038/srep44016⟩ (2017).

[7] Oku, M.: Two novel methods for extracting synchronously fluctuated genes, TBIO, to appear.

[8] 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).

[9] 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).

© 2019 Information Processing Society of Japan