Two Novel Methods for Extracting Synchronously Fluctuated Genes

Makito Oku1,a)

Received: October 30, 2018, Accepted: December 17, 2018

Abstract: In this paper, I propose two novel methods for extracting synchronously fluctuated genes (SFGs) from a transcriptome data. Variability and synchrony in biological signals are generally considered to be associated with the system’s stability in some sense. However, a standard method for extracting SFGs from a transcriptome data with high reproducibility has not been established. Here, I propose two novel methods for extracting SFGs. The first method has two steps: selection of remarkably fluctuated genes and extraction of synchronized gene clusters. The other method is based on principal component analysis. It has been confirmed that the two methods have high extraction performance for artificial data and a moderate level of reproducibility for real data. The proposed methods will help to extract candidate genes related to the stability and homeostasis in living organisms.

Keywords: transcriptome, synchronously fluctuated gene, clustering, principal component analysis

1. Introduction

The development of methods for transcriptome data analysis is one of the important research topics in the field of bioinformatics. In this study, I focus on the extraction of synchronously fluctuated genes (SFGs), which were also called dynamical network biomarkers in previous studies [1], [2], [3], [4], [5], [6], [7], [8], [9]. SFGs are similar to but different from differentially expressed genes (DEGs). DEGs are genes whose average expressions considerably increase or decrease under a certain condition as compared to the control group (Fig. 1). In contrast, SFGs are genes whose fluctuation levels considerably increase under a certain condition as compared to the control group, and whose expressions fluctuate in a synchronous or correlated manner (Fig. 2). DEGs and SFGs are both important for understanding various biological phenomena.

SFGs have been investigated by several research groups [1], [2], [3], [4], [5], [6], [7], [8], [9] based on an expectation that they are useful for characterizing the predisease state [1], that is, a state before the disease onset with lowered stability and abnormal homeostasis. The characterization of the predisease state may lead to rational and efficient disease prevention. In addition, SFGs are also expected to be useful for the diagnosis of the predisease state [1], although the cost of DNA microarray or RNA-seq measurements seems to be too high at present for the purpose of health monitoring. Putative predisease states before several diseases have been claimed to be identified and characterized using SFGs: acute respiratory distress syndrome [1], [5], [6], hepatocellular carcinoma [8], [9], diabetes mellitus [2], [3], [5], and influenza [5]. In addition, preceding states before cell differentiation in in-vitro experiments have also been characterized using SFGs [4], [7].

Here, I explain why SFGs are supposed to be associated with the predisease state [1] characterized by lowered stability and abnormal homeostasis. It is because if the inherent homeostatic functions of a living organism is impaired, it would become dif-

---

1 Institute of Natural Medicine, University of Toyama, Toyama 930-0194, Japan

a) oku@inm.u-toyama.ac.jp
ficult for some genes to keep constant expressions, and those genes’ expressions may begin to fluctuate synchronously due to interactions [1]. Although this hypothesis has not been proved by rigorous experiments, it is gaining attention, and several studies have been conducted based on it [1], [2], [3], [4], [5], [6], [7], [8], [9].

The prediction that the decrease of the stability of a living organism may cause the increase of fluctuation levels and enhancement of synchrony or correlations for some gene expressions comes from theoretical analyses [1], [10], [11]. The theories are based on stationary, unbounded, linear, and monostable dynamical systems’ limiting behavior. Under these conditions, the variances of the state variables go to infinity [10], the correlation coefficients between the state variables go to either 1 or −1 if the dominant eigenvalue of the Jacobian matrix is real-valued [1], and the stationary distribution of the state variables expands along a two-dimensional plane determined by the dominant eigenvectors if the dominant eigenvalues of the Jacobian matrix are complex-valued [11]. On the other hand, real living organisms are nonstationary, bounded, nonlinear, and multistable, and state transitions can occur well before the limit state. Therefore, it is important to notice the huge gap between the theory and reality. Criticisms are especially needed when we (a generic third person in the field of bioinformatics) encounter claims emphasizing theoretical underpinnings.

The term synchronously fluctuated genes is used for the first time in this study. They were called dynamical network biomarkers in previous studies [1], [2], [3], [4], [5], [6], [7], [8], [9]. There are three reasons for the renaming. First, the adjective dynamical is indirect to tell that we are focusing on fluctuations. Second, although we assume that the synchronization of SFGs occurs due to interactions between genes, the network structure is not necessarily considered. Third, although SFGs were expected to serve as biomarkers, it should be better to clarify the contrast to DEGs. In fact, the superiority of SFGs against DEGs as biomarkers for diagnosing predisease states, in which SFGs were supposed to be preferable to DEGs [1], has been neither proved nor disproved.

Previous studies used various methods for extracting SFGs [1], [2], [3], [4], [5], [6], [7], [8], [9]. A standard method has not been established, and almost all studies used mutually distinct procedures. To the best of the author’s knowledge, all of the previous methods were sensitive to outliers. A typical case is that genes showing particularly high or low expressions at only one sample are preferentially selected. Such genes may have some biologically important meanings in that they are differentially expressed for some reason at only one sample even though the condition is the same to that of the other samples. However, it is difficult to interpret that their activity patterns reflect the increase of fluctuations due to the lowered stability of the organism. This problem is referred to as the one-point problem henceforth.

In this study, I propose two novel methods for extracting SFGs. The first method has two steps: selection of remarkably fluctuated genes and extraction of synchronized gene clusters. The other method is based on principal component analysis (PCA). The main advantage of the proposed methods against the previous methods are that both of them are insensitive to outliers. In addition, the two methods are easier to implement than most of the previous methods.

The rest of the manuscript is organized as follows. In Section 2, an overview of previous methods is given. In Sections 3 and 4, the two-step method and PCA-based method are introduced, respectively. In Section 5, non-robust methods are introduced for comparison. In Sections 6 and 7, the procedures for performance evaluation using artificial data and real data are described, respectively. Their results are shown in Section 8. Finally, discussion and conclusions are given in Section 9.

2. An Overview of Previous Methods

The first extraction method for SFGs was proposed by Chen et al. [1]. Precisely, this method was originally proposed for selecting SFGs from DEGs extracted using the Student’s t-test with the false discovery rate (FDR) control. Therefore, the final results were in fact SFGs as well as DEGs, although differences between SFGs and DEGs were emphasized in the paper. However, the same method can be used for selecting SFGs from all genes. Since the original procedure was complicated including many steps and parameters, I here explain its major parts.

At first, each gene’s standard deviation was calculated for each of the experimental and control groups. Then, genes that had experimental standard deviation more than twice as large as that of the control group were selected. The reason for the threshold value 2 was not explained in the paper. Next, hierarchical clustering was applied to the selected genes. The similarity was measured by the Pearson’s correlation coefficient, and the single linkage method was used. It is important to note that the usage of the single linkage method should be avoided in transcriptome data analysis due to the chaining phenomenon [12], [13]. Anyway, the genes were then separated into 40 clusters, and clusters consisting more than or equal to 10 genes were selected. Finally, the obtained clusters were compared based on a few statistics such as the average standard deviation and average of the absolute values of correlation coefficients, and only one cluster was chosen as SFGs. However, from my experience of data analysis of several real datasets, when two or more large clusters exist, all of them are usually worth considering because they often contain genes that are presumably relevant to the investigated biological phenomenon.

The first method was modified in various ways in subsequent studies. In Ref. [2], the filtering step based on the standard deviation was modified so that genes with the experimental standard deviation more than the average of them were selected because the standard deviations of the control group were not available. In Ref. [3], the clustering step was modified so that a cutoff value 0.75 for the correlation similarity was used. The authors of the paper also argued that increasing the cutoff value reduced the size of SFGs. In Ref. [8], the clustering step was modified in two ways. First, the absolute values of correlation coefficients were used as the similarity measure. This means that two clusters with anti-phase behavior are regarded as one cluster in this setting. Second, the cutoff value was set to the 95th percentile of all correlation coefficients. However, it is important to note that
the calculation of the correlation coefficients of all gene pairs has large computational cost. In Ref. [9], the similarity measure was the same to that of the first method, and the cutoff value for it was set to the 99th percentile of all correlation coefficients.

Other methods that did not use hierarchical clustering were also proposed. In Ref. [5], the network structure of genes for each experimental condition was initially inferred by integrating gene expression data and existing knowledge on protein-protein interactions. The procedure was actually not easy to implement. Then, a graph clustering algorithm called ClusterONE was used to extract clusters, which are also known as communities in network science. Finally, one cluster was selected as SFGs based on a few criteria in a similar manner to the first method.

In Ref. [6], an evolutionary algorithm for multi-objective optimization called the non-dominated sorting genetic algorithm-II was used for extracting SFGs from 304 DEGs. It is unknown whether this method is also applicable to all genes instead of DEGs because the size of the solution space becomes quite huge (> 22000).

In Ref. [7], the authors of the paper first filtered genes based on the Pearson’s correlation coefficient, and then the fluctuation level was considered. Precisely, gene pairs whose correlation increased by 0.5 or more between the experimental and control groups were selected at first, and the genes participated in at least one pair were taken. Next, genes whose coefficient of variations (CVs) increased between the experimental and control groups were extracted from those obtained at the first step. The CV is defined by the standard deviation divided by the mean.

3. Proposed Method 1: Two-step Method

The first proposed method has two steps: selection of remarkably fluctuated genes and extraction of synchronization gene clusters. At the first step, each gene’s median absolute deviation (MAD) is calculated for each of the experimental and control groups. Then, we select genes that have experimental MADs more than twice as large as that of the control group. The MAD is one of the statistics for describing the scale of a distribution, and it is less sensitive to outliers than the standard deviation. The MAD is defined by the median of the absolute values of the differences between the variable’s values and median. The threshold value 2 against the MAD ratio, which was chosen empirically, may be modified according to the situation. For example, when almost all genes are separated into distinct clusters, a smaller threshold value would be better for obtaining clusters of reasonable size. Keeping the maximum cluster size above 10 is recommended in most cases.

After the clustering, we select genes belonging to the largest cluster as well as the second and subsequent clusters whose sizes are more than a half of the largest one. Therefore, when two or more large clusters exist, all of them are included in order to avoid missing relevant genes as well as to enhance the reproducibility. The threshold value 0.5 for the cluster size ratio, which was chosen empirically, may be modified if necessary.

4. Proposed Method 2: PCA-based Method

Another proposed method is based on PCA. Here, I briefly explain the reason why PCA can be used for extracting SFGs. When the stability of a living organism decreases, the stationary distribution of the state variables becomes wider if the assumptions used in the theory are applicable [11]. The expansion rate is largest in the direction with the largest vulnerability to disturbances. Therefore, by obtaining the most expanded direction of the empirical distribution of the state variables using PCA, we may be able to extract SFGs that are the most responsive to the decrease of the system’s stability.

The procedure of the PCA-based method is as follows. At first, each gene’s interquartile range (IQR) is calculated for each of the experimental and control groups. Then, we define data points more than 1.5 times the IQR below the first quartile or more than 1.5 times the IQR above the third quartile as outliers. This criterion for outlier detection is commonly used in box plots. The conventional parameter value 1.5 for the IQR multiplier may be modified if necessary. We then subtract the mean of non-outlier values for each gene, and replace the outlier values with 0. Subsequently, for each of the experimental and control groups, we apply PCA to the mean-centered data, and multiply the absolute values of the first principal component (PC1) weights by the standard deviation of the PC1 scores, which are referred to as contributions henceforth.

We then subtract each gene’s contribution under the control condition from that under the experimental condition, and consider the distribution of the differences. Finally, we select genes whose contribution difference is more than 3 times the standard deviation of the distribution above the mean of the distribution. This extraction criterion is used because the contribution differences are approximately normally distributed. The conventional parameter value 3 for the multiplier of the standard deviation may be modified if necessary.

5. Non-robust Methods

I introduce non-robust counterparts of the proposed methods for comparison. The non-robust two-step method uses the standard deviation instead of the MAD, and the Pearson’s correlation coefficient instead of the Spearman’s correlation coefficient. The non-robust PCA-based method skips the outlier detection procedure.
The non-robust two-step method is used as a representative method of several previous methods that used the filtering step based on the standard deviation and clustering step based on the Pearson's correlation coefficient [1], [2], [3], [8], [9]. Although the implementation details differed among these previous methods, their core ideas for extracting SFGs are retained in the non-robust two-step method. However, this method does not represent other previous methods that did not use hierarchical clustering [5], [6], [7]. They were not investigated in this paper due to the difficulty of implementation or the computational cost.

It is important to note that if we follow the original implementations exactly, the methods used in Refs. [1] and [6] cannot extract SFGs from the artificial data explained in the next section. This is because the two methods are for extracting SFGs from DEGs, and thus are useless if no DEGs exist.

6. Performance Evaluation Procedure Using Artificial Data

In this study, artificial and real data were used in order to evaluate the performance of the proposed methods as well as the non-robust methods for extracting SFGs. The artificial data was generated as follows. First, two 10,000 × N data matrices whose elements independently followed the standard normal distribution were generated. The sample number N was adjustable. Next, a row vector of length N whose elements independently followed a normal distribution with mean 0 and variance 24 was generated. This vector was then repeatedly added to 500 rows of the first data matrix. It is important to note that adding two independent random variables with variances 1 and 24 results in an expected variance 25. The first and second matrices correspond to the experimental and control data, respectively. The modified 500 rows had the expected standard deviation 5 (= \sqrt{25}) and expected within-group correlation coefficient 0.96 (≈ 24/5²), which can also occur in many real data. The other rows of the experimental data and all rows of the control data had the expected standard deviation 1 and expected within-group correlation coefficient 0. The modified 500 rows were regarded as positive cases of SFGs and the remaining 9,500 rows were regarded as negative cases.

In order to evaluate the classification performance, precision, recall, and F1 score were used. Let TP, FP, and FN denote the numbers of true positives, false positives, and false negatives, respectively. The precision is defined by TP/(TP+FP). The recall is defined by TP/(TP+FN). The F1 score is defines by the harmonic mean of the precision and recall.

7. Performance Evaluation Procedure Using Real Data

The details of the real data used in this study are as follows. It was used in order to evaluate the reproducibility of the proposed methods even though the actual SFGs are unknown. Since no benchmark problem existed for evaluating the reproducibility of SFG extraction methods, the author searched appropriate public data having many samples in the Gene Expression Omnibus (GEO) database. Human's gene expression datasets were excluded because inter-individual variability is hard to distinguish from inherent fluctuations. For similar reasons, presumably heterogeneous datasets such as those from tumor samples or single-cells were also excluded. Only one dataset GSE77578 [14] was found under the condition that both the experimental and control groups consist of more than or equal to 15 samples. When the required sample number was lowered to 10, several other public datasets were also found.

The GSE77578 is a dataset of DNA microarray measurements of gene expressions of mice. According to the dataset's descriptions, the gene expression values have been base-2 log-transformed and quantile normalized. I have also confirmed them by checking a box-plot of all samples and descriptive statistics such as means, standard deviations, and medians. Therefore, the original data was used without any additional normalization in this study. In addition, no missing values were found in the data matrix. The overview of the dataset is shown in Table 1. PLX3397 is the name of a drug, and vehicle only means that mice were administered the vehicle that did not contain the drug as a control condition. In this study, only two classes in Table 1 were used: the first one (N = 17) as the control data and the second one (N = 18) as the experimental data.

The evaluation procedure for reproducibility was as follows. First, SFGs were extracted using the control and experimental data. Next, one sample was temporarily removed from the experimental data while the control data was kept unchanged, and SFGs were extracted again. The overlap between this result and the original one was measured using the Jaccard index. The Jaccard index between two sets is defined by the size of the intersection divided by the size of the union. The same procedure was repeated through all experimental samples, and the worst-case sample whose removal resulted in the largest reduction of reproducibility was permanently removed from the experimental data. Similarly, samples were removed one by one from the experimental data in order to investigate how fast the reproducibility decreases against the removed sample numbers. The reason why the worst-cases were used rather than focusing on the average reduction of reproducibility was to investigate the influence of the abovementioned one-point problem.

Enrichment analysis, which is widely used in transcriptome data analysis [13], [15], was also performed in order to reveal what kind of genes are mainly included in the SFGs extracted from the real data. The DAVID (Database for Annotation, Visualization and Integrated Discovery) database [16] version 6.8 was used for retrieving biological process terms of gene ontology (GO) that had two or more overlaps with each set of SFGs. The raw p-values were calculated based on the one-tailed Fisher's exact test. It is important to note that they were different from the p-values provided by the DAVID database that were based on a modified Fisher's exact test. Then, the q-values were calculated in order to keep the expected value of the FDR below 0.05. When no

| Condition                                      | Sample number |
|------------------------------------------------|---------------|
| epileptic mice administered vehicle only      | 17            |
| epileptic mice administered PLX3397 3 mg/kg | 18            |
| epileptic mice administered PLX3397 30 mg/kg | 15            |
| normal mice administered vehicle only         | 6             |

Table 1 Overview of GSE77578 data.
result was obtained, the acceptable FDR was relaxed to 0.2. The same procedure was also performed for DEGs extracted from the real data by a fold-change cutoff $|x_i - y_i| > 0.5$, where $x_i$ and $y_i$ are the $i$th gene's average log-expressions of the experimental and control groups, respectively. The cutoff value 0.5 was relaxed from the conventional value 1 corresponding to a two-fold change because no DEGs were obtained using it.

8. Results

All the data analyses were performed using Anaconda 5.0.0 for Python 3.6, which includes various Python packages such as NumPy, SciPy, pandas, Matplotlib, and scikit-learn. Source codes used in the present study will be available in the GitHub repository under MIT license (https://github.com/okumakito/sfg-tbio2018).

Figure 3 shows the results of the extraction performance evaluation of the proposed and non-robust methods using the artificial data. All of the F1 scores, precision, and recall showed increasing tendency as the sample number increased except for the precision of the PCA-based and non-robust PCA-based methods, for which the precision was almost 1 even at $N = 3$. Let us focus on the results of the two proposed methods at first. The F1 score was higher for the PCA-based method at $N = 3$, and it was comparable between the two methods for $4 \leq N \leq 8$. It was a little higher for the two-step method when $N \geq 9$. The required numbers of samples for achieving F1 scores above 0.9 were $N \geq 7$ for the two-step method and $N \geq 15$ for the PCA-based method. The precision was higher for the PCA-based method, whereas the recall was higher for the two-step method.

Next, let us focus on the results of the non-robust methods. The F1 score was comparable between the two non-robust methods for $N \geq 4$, and it increased monotonically as the sample number increased. These results suggest that two phenomena observed in the results of the proposed methods—sample number-dependent alteration of preferable methods and non-monotonic relationship between the F1 score and sample number—were caused by the countermeasures against outliers. It is important to note that the overall performance of the non-robust methods was better than that of the proposed methods because the artificial data completely followed normal distributions and did not contain outliers.

Figure 4 shows the evaluation results of the reproducibility of the proposed and non-robust methods using the real data. The two-step method showed much higher reproducibility than the non-robust two-step method. The PCA-based method showed higher reproducibility than the non-robust PCA-based method when three or more samples were removed. These results suggest that the two proposed methods prevented themselves from largely responding to a small number of samples and successfully extracted genes showing synchronous fluctuations throughout the experimental samples. The reproducibility of the two-step method was higher than that of the PCA-based method when one to three samples were removed, and the two were comparable when four samples were removed. The PCA-based method showed higher reproducibility than the two-step method when five or more samples were removed.

Figures 5 and 6 show the heatmaps of the SFGs extracted from the real data by the two-step method and non-robust two-step method, respectively. The color scale indicates the Z-score calculated in each row. In Fig. 6, only two samples showed notably high expressions whereas the rest of the experimental samples did not show remarkable increase in fluctuations and correlations as compared to the control group. This is a typical case of the abovementioned one-point problem. On the other hand, in Fig. 5, many more samples took values distant from the mean values, and the extent of synchronization was also strengthened in the experimental group as compared to the control group. Similar things can be said for the PCA-based and non-PCA based method, as
The SFGs extracted from the real data by the two proposed methods had little overlap. In order to investigate the reason, distributions of the average gene expressions of the top 10% genes scored by the proposed methods were obtained, as shown in Fig. 9. Here, the top 10% genes scored in the first step were used for the two-step method because scoring of genes in the second step was difficult. It can be seen that the distribution of high-ranked genes in the two-step method was similar to that of all genes, whereas the PCA-based method preferentially selected genes that were highly expressed on average. These results suggest that the two methods tend to extract different types of SFGs with respect to biological functions because there exist several types of genes such as constitutive genes, which are continuously transcribed in order to maintain basic cellular functions, and facultative genes, which are activated when needed in order to respond to stimuli or environmental changes.

Tables 2 and 3 show the enriched GO terms in the SFGs extracted from the real data by the two-step and PCA-based methods, respectively. An overlap means the number of common genes between two gene sets, SFGs and a set of genes having each GO term. It can be seen that the SFGs extracted by the two-step method included three genes associated with midbrain development, which is a tissue-specific function, whereas the SFGs
extracted by the PCA-based method included tens of genes involved in apparently basic cellular functions such as regulation of signal transduction and metabolic process.

Table 2 shows the enriched GO terms in the DEGs extracted from the real data by a fold-change cutoff (q-value < 0.05; sorted by the overlap size; all enriched terms are shown). The number of multiple comparisons was 2115, and the total gene number was 17911.

| Term                                      | Overlap | #Genes with the term | p-value | q-value |
|-------------------------------------------|---------|----------------------|---------|---------|
| midbrain development                      | 3       | 87                   | 3.0E-04 | 1.0E-01 |
| positive regulation of lamellipodium organization | 2       | 27                   | 7.5E-04 | 1.3E-01 |
| regulation of lamellipodium organization  | 2       | 40                   | 1.6E-03 | 1.9E-01 |

Table 3 shows the enriched GO terms in the SFGs extracted from the real data by the PCA-based method (q-value < 0.05; sorted by the overlap size; top 10 terms among 479 are shown). The number of multiple comparisons was 2196, and the total gene number was 17911.

| Term                                      | Overlap | #Genes with the term | p-value | q-value |
|-------------------------------------------|---------|----------------------|---------|---------|
| regulation of cell communication          | 53      | 2785                 | 2.8E-06 | 1.0E-03 |
| regulation of signaling                   | 53      | 2808                 | 3.6E-06 | 1.1E-03 |
| response to organic substance             | 48      | 2972                 | 6.0E-04 | 1.1E-02 |
| regulation of cellular component organization | 47      | 2399                 | 5.9E-06 | 1.4E-03 |
| cellular component biogenesis              | 46      | 2654                 | 1.7E-04 | 5.1E-03 |
| positive regulation of metabolic process   | 46      | 3004                 | 2.5E-03 | 2.2E-02 |
| regulation of signal transduction         | 45      | 2468                 | 6.0E-05 | 4.6E-03 |
| cellular response to chemical stimulus     | 45      | 2607                 | 2.2E-04 | 5.7E-03 |
| positive regulation of macromolecule metabolic process | 45 | 2794 | 1.0E-03 | 1.4E-02 |
| phosphate-containing compound metabolic process | 45 | 2869 | 1.7E-03 | 1.8E-02 |

Moreover, as shown in Figs. 5 and 7, the two proposed methods successfully resolved the one-point problem, which was one of the serious issues of the previous methods. More detailed comparisons between the proposed methods and previous methods [1], [2], [3], [4], [5], [6], [7], [8], [9] are important future works.

The required number of samples is discussed. Most of the previous studies used approximately five samples for each of the experimental and control groups. On the other hand, the result shown in Fig. 3 suggests that the F1 score did not saturate at N = 5, and more samples were required for better extraction performance. It is also important to note that the positive cases of the artificial data used in the present study had five times larger expected standard deviations than those of the negative cases and control group. Since the difficulty of the classification task depends on the strength of the common noise, much more samples will be needed to achieve a sufficient statistical power when the hypothesized effect size is smaller.

The strategy to select between the two proposed methods is discussed. The results shown in Figs. 3 and 4 seem to suggest that the PCA-based method is better when the sample size is small and the two-step method is preferable otherwise. However, since the degrees of fluctuations and correlations differ among real data, it is recommended to carefully select an appropriate method according to the situation. It is also important to note that the two methods tend to extract different types of genes, as shown...
in Fig. 9, Table 2, and Table 3. Therefore, they could be helpful for us to understand living organisms from biologically different aspects. The establishment of a standard protocol for extracting SFGs, which would be based on the proposed methods, is an important future work.

The reproducibility of the proposed methods is discussed. Figure 4 shows that the values of the Jaccard index of the proposed methods were not high, although they were relatively higher than those of the non-robust methods. The worst-case Jaccard indices for one sample removal were 0.38 for the two-step method and 0.30 for the PCA-based method. In contrast, DEGs extracted using a fold-change cutoff of 0.5 resulted in a Jaccard index value of 0.77 for one sample removal. Therefore, the reproducibility of the extraction of SFGs is currently much lower than that of DEGs. The improvement of extraction methods for better reproducibility is an important future work.

The way to deal with missing values is discussed. Although the GSE77578 dataset did not contain missing values, many gene expression datasets have them, which are represented by various ways such as ‘NA’, ‘null’, a blank cell, ‘-999’, ‘0’ without the fractional part, and so on. The two-step method is applicable to data with missing values by neglecting them in the calculations of the MADs and Spearman’s correlation coefficients. The PCA-based method is also applicable to such data by dealing with the missing values similarly to the outliers.

Acknowledgments The author would like to thank the anonymous referees for their careful reading of an earlier draft of this paper and giving valuable comments. This work was supported by JSPS KAKENHI Grant Number JP15H05707.

References

[1] 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 (2012). DOI: https://doi.org/10.1038/srep00342 (2012).
[2] Liu, X., Liu, R., Zhao, X.-M. and Chen, L.: Detecting early-warning signals of type 1 diabetes and its leading biomolecular networks by dynamical network biomarkers, BMC Med. Genomics, Vol.6, Suppl. 2, p.S8 (2013).
[3] Li, M., Zeng, T., Liu, R. and Chen, L.: Detecting tissue-specific early warning signals for complex diseases based on dynamical network biomarkers: Study of type 2 diabetes by cross-tissue analysis, Brief. Bioinform., Vol.15, No.2, pp.229–243 (2014). DOI: https://doi.org/10.1093/bib/bbt027 (2014).
[4] Chen, P., Liu, R., Chen, L. and Aihara, K.: Identifying critical differentiation state of MCF-7 cells for breast cancer by dynamical network biomarkers, Front. Genet., Vol.6, p.252 (2015). DOI: https://doi.org/10.3389/fgene.2015.00252 (2015).
[5] Li, Y., Jin, S., Liu, L., Pan, Z. and Zou, X.: Deciphering deterioration mechanisms of complex diseases based on the construction of the dynamic network and systems analysis, Sci. Rep., Vol.5, p.9283 (online). DOI: https://doi.org/10.1038/srep09283 (2015).
[6] 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).
[7] Richard, A., Boullé, L., Herbach, U., Bonnafoix, A., Morin, V., Vallin, E., Guillenmin, A., Papii Gao, N., Gunawan, R., Cosette, J., Arnaud, O., Kupiec, J.-J., Espinasse, T., Gomin-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).
[8] Lu, L., Jiang, Z., Dai, Y. and Chen, L.: Low-grade dysplastic nodules revealed as the tipping point during multistep hepatocarcinogenesis by dynamic network biomarkers, Genes, Vol.8, No.10, p.268 (online). DOI: https://doi.org/10.3390/genes8100268 (2017).

[9] Yang, B., Li, M., Tang, W., Liu, W., Zhang, S., Chen, L. and Xia, J.: Dynamic network biomarker indicates pulmonary metasta-
sis at the tipping point of hepatocellular carcinoma, Nat. Commun., Vol.9, No.1, p.678 (online). DOI: https://doi.org/10.1038/s41467-018-03024-2 (2018).
[10] Biggs, R., Carpenter, S.R. and Brock, W.A.: Turning back from the brink: Detecting an impending regime shift in time to avert it, Proc. Natl. Acad. Sci. U.S.A., Vol.106, No.3, pp.826–831 (online). DOI: https://doi.org/10.1073/pnas.0811729106 (2009).
[11] Oka, M. and Aihara, K.: On the covariance matrix of the station-
ary distribution of a noisy dynamical system, NOLTA, Vol.9, No.2, pp.166–184 (online). DOI: https://doi.org/10.1587/nolta.9.166 (2018).
[12] Kamishima, T.: A survey of recent clustering methods for data min-
ing (part 1): Try clustering! (in Japanese), J. Jpn. Soc. Artif. Intell., Vol.18, No.1, pp.59–65 (2003).
[13] Oka, M.: Fundamentals of gene expression data analysis, Annual Report, Institute of Natural Medicine, University of Toyama (in Japanese), Vol.44, pp.2–11 (2018).
[14] Srivastava, P.K., van Eyll, J., Godard, P., Mazzuferi, M., Delahaye-
Duriez, A., Steenwinckel, J.V., Gressens, P., Danis, B., Vandenplas, C., F. Laaant, L., Niespodzany, I., Kneun, J., Gasser, J., Gillel, G., Shika, K., Chong, S.-A., Behmousas, J., Kadi, I., Petrotto, E., Kaminski, R.M. and Johnson, M.R.: A systems-level frame-
work for drug discovery identifies Csf1R as an anti-epileptic drug target, Nat. Commun., Vol.9, No.1, p.3561 (online). DOI: https://doi.org/10.1038/s41467-018-06008-4 (2018).
[15] Nakaoka, S.: Introduction of gene expression analysis (in Japanese), http://coop-math.ism.ac.jp/report/show/2016S01 (2016).
[16] Huang, D.W., Sherman, B.T. and Lempicki, R.A.: Systematic and integrative analysis of large gene lists using DAVID bioinformat-
ics resources, Nat. Protoc., Vol.4, No.1, pp.44–57 (online). DOI: https://doi.org/10.1038/nprot.2008.211 (2009).

Makito Oku received his B.E. degree in systems innovation engineering and his Ph.D. degree in information science and technology from the University of Tokyo, Tokyo, Japan, in 2005 and 2011, respectively. He joined the Information Processing Society of Japan in 2018. He is currently a specially appointed associate professor with the Institute of Natural Medicine, University of Toyama. His current research interests include transcriptome data analysis. His ORCID ID is 0000-0002-0282-2361.

(Communicated by Takeshi Obayashi)