GEsture: an online hand-drawing tool for gene expression pattern search

Chunyan Wang[1], Yiqing Xu[1], Xuelin Wang[1], Li Zhang[1], Suyun Wei[1], Qiaolin Ye[1], Youxiang Zhu[1], Hengfu Yin[2], Manoj Nainwal[3], Luis Tanon[4], Feng Cheng[5], Tongming Yin[6], Ning Ye[1*]

1. College of Information Science and Technology, Nanjing Forestry University, Nanjing, Jiangsu, China
2. Key Laboratory of Forest genetics and breeding，Research Institute of Subtropical Forestry, Chinese Academy of Forestry, Hangzhou, Zhejiang, China 311400
3. Department of Computer Science, Nantong University, Nantong, Jiangsu, China
4. Department of Cell Biology, Microbiology and Molecular Biology, University of South Florida, Tampa, FL, 33620, USA
5. Department of Pharmaceutical Science, College of Pharmacy, University of South Florida, Tampa, FL, 33612, USA
6. College of Forest Resources and Environment, Nanjing Forestry University, Nanjing, Jiangsu, China

* Corresponding author: Ning Ye, yening@njfu.edu.cn

Abstract

Gene expression profiling data provide useful information for the investigation of biological function and process. However, identifying a specific expression pattern from ~~big~~extensive time series gene expression data is not easy. Clustering is a popular method to classify similar expression genes~~. H~~; however, genes with a ‘desirable’ or ‘user-defined’ pattern cannot be efficiently detected by clustering methods. To address these limitations, we developed an online tool called GEsture. Users can draw, or graph, a curve using a mouse instead of inputting abstract parameters. GEsture explores genes showing similar, opposite, time-delay or ~~ahead~~predicted expression patterns with the input from time series data sets. We presented three examples that illustrate the capacity of GEsture in hunting for genes while conforming to users’ expectations. GEsture also provides visualization tools (such as expression pattern figure, heat map and correlation network) to display the searching results. The output results may provide useful information for researchers to understand the targets, function and biological processes of the input genes.

Keywords: bioinformatics, gene expression, microarrays, time series, hand-drawing

Introduction

Gene expression profiling (such as Microarray and RNA-seq) data provide important information for researchers to investigate biological function and process. Many public databases including gene expression omnibus (GEO) (Barrett & Edgar, 2006), gene signatures database (GeneSigDB) (Culhane et al., 2012), and molecular signatures database (MSigDB) (Liberzon, 2014) are available to identify the relationship between gene expression and biological functions/process. For many biological studies, researchers hope to find genes showing “anticipated” expression patterns. For example, biologists know that during a day, the expression levels of light rhythm genes change (increase or decrease) with the intensity of light, and change back with the darkness of night. However, it is hard for them to find the genes showing this pattern from large gene expression data studies without strong bioinformatics background.

Multiple approaches have been developed to find genes expressing similar~~ly~~ expression patterns across all samples (Androulakis, Yang, & Almon, 2007; Sharan & Shamir, 2000). Clustering is widely used to solve the problem.(Eisen et al., 1998; Wen et al., 1998; Jiang, Tang, & Zhang, 2004; Schliep et al., 2005). Clustering algorithms include hierarchical clustering (Jiang, Pei, & Zhang, 2003), self-organizing maps (Tamayo et al., 1999), K-means clustering (Tavazoie et al., 1999; Wu, 2008) and so on. Clustering approaches indeed performed well in grouping genes with similar expression patterns without any prior knowledge. Balasubramaniyan (Balasubramaniyan et al. , 2005) proposed a CLARITY algorithm using a local shape-based similarity measurement to dig for similarly expressed genes ~~(Balasubramaniyan et al. , 2005)~~. Qian et al. (2003) proposed a local clustering algorithm to identify time-delayed and inverted expression genes. (time-delayed means gene expression has a time difference but the expression trend is the same, and inverted refers to some genes that are high expression while the other genes are low expression at the same time)~~ (Qian et al., 2003)~~. Xia designed an eLSA package (Xia et al, 2013), which filters out insignificant results and constructs a partial and directed association network.

However, the clustering algorithm has some disadvantages. First, the computation complexity of clustering algorithms exponentially increases as the dataset becomes larger. Second, the issue of determining the optimum cluster number is not yet rigorously solved (Yeung, 2001). Third, during the data processing, as Ye et al. (2015) says, ‘expression data

3

vary greatly, clustering algorithms generally require pre-processing of the original data which may cause loss of useful information’~~(Ye et al., 2015)~~. Fourth, clustering algorithms will discard some clusters with smaller number of genes. Fifth, unrelated groups may be merged into one cluster. For example, classical K-means clustering extracts categories of a given number from the gene expression profile. However, it cannot always cluster a category ~~we~~that is expected. Because a time-delayed phenomenon often appears in gene expression~~.~~, K-means clustering cannot recognize it and mistakenly divides it into many categories instead of classifying only one category. Therefore, clustering methods can help to understand global profiles of gene expression, but not efficiently enough to detect genes with user-defined expression patterns.

In this paper, we presented an online tool, GEsture, short for Gene Expression gesture. The program searched specific gene expression pattern from time-series gene expression data using anticipated gene expression patterns drawn by users instead of using clustering algorithms. GEsture addresses the current shortcomings of the clustering algorithm and allows users to analyze time-series data from a different angle. This method not only can identify co-expression genes but also can detect opposite and time-delayed expression genes. Furthermore, it provides user-friendly interface for users to input and visualize the results. The output results may provide useful information for researchers to understand the targets, function and biological processes of the input genes.

Materials & Methods

The workflow of GEsture

The primary function of GEsture is to identify genes showing specific patterns from gene expression data. The workflow is illustrated in Figure 1. The first step is uploading time-series gene expression data. Two modes of operations, user-defined pattern and K-means clustering, are then provided for pattern searching. For user-defined pattern, users can either draw an expression curve by the mouse in the drawing board or select a pre-defined gene expression pattern in the system to search. Classical K-means clustering method extracts expression patterns from the gene expression profile based on the given number of data points (?).

There are three functions for gene expression pattern searching in GEsture: brush pattern search, contrast pattern search, and shift pattern search.

1. **Brush pattern search (co-expression pattern search).** It is the default pattern search function in GEsture. Users can draw a gene expression curve with mouse in the drawing board. GEsture will identify the genes showing similar pattern (co-expression genes) with the input curve. It is noted that users should include as many time points as possible in curve drawing to achieve accurate matching.

2. **Contrast pattern search.** This function searches the genes showing opposite expression pattern to the user's input. It aims at helping users to explore negative regulated genes. For example, target genes of a transcription factor that inhibits expression can be found~~ed~~ using this function.

3. **Shift pattern search.** This function is designed to find genes showing similar but with a time-delayed (or ahead) expression pattern. It will help users to identify possible downstream/upstream genes. The range of -4 to 4 can be chosen for the shifting of the gene expression pattern.

The expression levels of output genes are shown by the heat maps. The network map is generated to display the relationship of output genes.

Data analysis process

GEsture takes a curve as an input, and it allows the user to search genes with similar expression patterns. As a result, users can see the gene expression curves intuitively rather than in abstract parameters and data. ~~R~~A raw data set uploaded by user will be checked to filter low-quality (such as missing and low entropy) data in the uploaded file. The process of searching includes the drawing of an anticipated curve, followed by fitting the system in a line and sampling the data. Afterwards, genes are compared with each other in the gene expression file to calculate the similarity between them. Lastly, an assessment function is adopted using the Pearson correlation coefficient (Horyu & Hayashi, 2013; Wang, Mo, & Wang, 2015) to select closely-related genes. It may take a while when performing this kind of search on a large dataset, but it is significantly faster than clustering. GEsture only compares every gene expression curve in the file, while clustering~~,~~; on the other hand, it needs to determine the initial centers and iteration numbers of the algorithm, inevitably leading to higher computation and time complexity. The cutoff ~~of~~for the correlation coefficient for gene outputs can be adjusted by users if too many or too ~~less~~few genes are identified.

5

Output of GEsture

As shown in Figure 2, the output of GEsture includes a gene expression pattern figure and gene information table. To clearly show the expression patterns, a slider is provided for users to adjust the correlation coefficient cutoff value. At the same time, the information of the output genes in the figure is also shown in a table. In the output table, each row of data represents a gene. The information of gene name, p-value, correlation value and detailed time-series expression data is included. If a user clicks on the gene name in the table, the corresponding gene expression curve will be shown in the expression pattern figure. The gene information table can be exported as a CSV file.

Two visualization tools, a heat map and a comprehensive relationship network map were provided to visualize the search results. Figure 3 shows the co-expression genes of YNL309W. Each row in the heat map represents one gene and different colors to display the gene expression level. In GEsture, the maximum number of genes for a heat map is 500. The heat map can be exported as a PNG formated file. In addition, the comprehensive network map (shown in Figure 4) was also generated to shows the relationship of output genes. The center point is the searched gene or the most similar gene to the drawn-curve in the figure. ~~It~~The software was built on the Cytoscape.js program, and the size of the figure can be adjusted by users using the mouse interface.

Datasets used in GEsture

We used 3 examples to demonstrate the effectiveness of GEsture. In example 1, a yeast Cell-cycle data set was chosen to assess the performance of GEsture. The data set contains 6187 genes and 18-time points (Spellman et al., 1998). ~~It~~The data is available at http://genome-www.stanford.edu/cellcycle/data/rawdata/. The same yeast data set was also used in example 2 to identify the target genes of transcription factors. In example 3, the circadian rhythm genes of *Arabidopsis thaliana* were identified using GEsture. Columbia diurnal gene expression data of *Arabidopsis thaliana* (Mockler Lab) ~~which is~~ measured in the condition of growing with 12h-light 12h-dark/24h-hot (COL_LDHH) was chosen (Mockler et al., 2007).

Results

Example 1: identifying anticipated expression genes

Here, we used two methods, a user-defined pattern searching in GEsture and a K-means clustering method, to identify genes whose expression levels increased over time in the cell cycle. K-means clustering was firstly applied to cluster different gene expression categories. As shown in Figure 5, a variety of gene expression patterns were identified at k=16 and 25. However, the pattern of interest did not showpresent itself in the results.

GEsture was then applied to find the genes showing the increasing pattern over time. We drew the anticipated expression curve in GEsture (Figure 6A) and eleven genes were detected to express increasingly over time (Figure 6B). Among these genes, four genes (YOR010C, YDR534C, YOR382W, and YNL066W) are the cellular component: cell wall proteins. Three genes are transporters (YHL047C, YMR058W, and YBR102C) (Chervitz et al., 1999). This example showeddemonstrated that GEsture was more straightforward and efficient to identify genes wethat were expected. The results may also provide important information about the transcriptional mechanisms of cell cycle regulation.

Example 2: identifying target genes of transcriptional factors

As shown in the Saccharomyces Genome Database (SGD https://www.yeastgenome.org/), YNL309W(STB1) encodes a protein that contributes to the regulation of SBF and MBF target genes (Chervitz et al., 1999). 'SBF and MBF are sequence-specific transcription factors that activate gene expression during the G1/S transition of the cell cycle in yeast', as suggested by Iyer VR et al. (2001) says.(Iyer et al., 2001). We hypothesized that the genes showing similar, contarrast, or time-delay expression pattern may be controlled by a similar regulatory mechanism. In this example, we searched genes using GEsture and explored whether there arewere other genes regulated by the same transcription factor.

We have drawn a curve like the expression of *YNL309W* (shown in Figure 7) with the mouse and searched GEsture using three different patterns. Then we found 155 co-expression genes, 15 contrast expression genes and 44 one--interval--shift expression genes. More detail

7

information ~~of~~about these genes can be found in Table S1. YNL309W ~~were~~was detected in the co-expression gene list, which shows the accuracy of the tool.

We then used a database, YEASTRACT (Yeast Search for Transcriptional Regulators and Consensus Tracking http://www.yeastract.com) (Teixeira et al., 2006), to assess which identified genes ~~we acquired~~ were regulated by the same TFs as YNL309W. YEASTRACT provides the known TF-target genes association of yeast in the cell-cycle process. ~~We knew~~ *TEC1p* and *STE12p* ~~are~~were known TFs of *YNL309W* in the cell cycle (Madhani et al. , 1999), TEC1p is responsible for positive regulation, and STE12p is responsible for negative regulation. According to YEASTRACT, we found that 124 similar expression genes, 5 contrast expression genes and 27 shift expression genes (one interval) ~~are~~were regulated by TEC1p and STE12p (Table 1). Detailed gene information is listed in the Table S2. The example indicated that GEsture can efficiently identify ~~the~~ target genes ~~of the same~~associated with related transcription factors.

Example 3: identifying circadian rhythm genes of *Arabidopsis thaliana*

In higher plants, the circadian rhythm phenomenon is a universal, intrinsic and autonomous timing mechanism of approximately 24-hours. This mechanism allows organisms to adapt to daily external changes in the environment, such as light, temperature and so on (Bass & Takahashi, 2010; Bellpedersen et al., 2005; Hardin & Panda, 2013; Joska, Zaman, & Belden, 2014). The most noticeable characteristic of circadian rhythms is that the period of rhythm is close to 24 hours in the absence of environmental stimuli. The expression pattern of circadian rhythms genes in the period of rhythm almost does not vary (Hsu & Harmer, 2014; Wijnen & Young, 2006). As of now, some circadian rhythms-associated genes of the *Arabidopsis thaliana* have been identified and cataloged by *The Arabidopsis Information Resource* (TAIR https://www.arabidopsis.org) (David et al., 2008). Here, we input the expression pattern of known circadian rhythms genes and checked whether GEsture could efficiently identify genes related to circadian rhythm. As shown in Figure 8, we drew an expression curve ~~that~~approximating circadian rhythm genes ~~show~~ patterns. Three pattern searches were ~~applied~~attempted for gene identification. The TAIR database ~~were~~was finally used to check whether output genes were related to circadian rhythm.

8

GEsture found 40 circadian rhythm genes using three search patterns (Table 2). Detailed information was listed in the Table S3. In these 40 circadian rhythm genes 14, 11, and 15 genes ~~showing~~demonstrated similar, contrast and shift patterns with the input mouse-entered patterns. Among the co-expression circadian rhythm genes, we have found *AT5G25830*, *AT5G15850*, *AT5G56860* are TFs of *Arabidopsis*. We also found some genes with an expression pattern ~~like~~similar to circadian rhythm but ~~are not~~not classified as rhythm genes, such as *AT2G31990*, *AT1G32630*, *AT1G05320* and so on. The~~ir~~ expression curves of these genes are similar to circadian rhythm genes, but their biological process is still annotated as unknown in the TAIR database. The results may provide some hints for biologists to study biological functions and processes of these genes.

Discussion

Two data sets, ~~the~~one from a yeast dataset, and the other from Arabidopsis thaliana were selected to assess the performance of our program. Three examples demonstrated the effectiveness of GEsture in searching co-expression, contrast and time-delay expression genes. The biological meaning of the output genes ~~was~~were explored. For reference, the three example data sets are provided on the GEsture website.

GEsture is built for searching specific gene expression patterns from time-series gene expression data. ~~And it is~~Programming was written in PHP, JAVASCRIPT, HTML5, and Bootstrap. Also, two plugins of cytoscape.js (Franz et al., 2016) and Echart.js ~~are~~were utilized for graphical visualization. Here, we used GEsture to identify anticipated expression genes, target genes of transcriptional factors and circadian rhythm genes of *Arabidopsis thaliana*. The results of the first example indicate that clustering algorithms cannot efficiently dig out all gene expression patterns because the algorithm will discard some clusters with small number of genes. It is possible to identify the pattern by increasing the cluster number. However, it ~~need~~may require more time to ~~try~~attempt different cluster numbers and the process is not efficient. ~~It showed that~~ GEsture was shown more straightforward and efficient to identify genes ~~we~~which were expected and it would be good supplement to test options of other clustering methods.

In the second example, GEsture searched ~~the~~ similarly express~~ion~~ed genes by drawing ~~the~~a familiar gene expression curve ~~not according to concrete one~~usually associated with a more characterized gene, such as annotated gene names. ~~i~~It showed that GEsture was effective

9

and efficient ~~to~~in explor~~e~~ing other genes with the similar expression ~~genes~~patterns. And further, after checking, about 73% of similarly express~~ion~~ed genes GEsture ~~searched out are~~identified related genes controlled by the same TFs. The third example showed another function of G~~e~~Esture~~. It;~~ not only was it capable of seeking target genes of the TFs, but it also performed well in detecting genes with similar functions by curves.

In short, GEsture provides an interactive interface for pattern searching and is convenient and easy for users to edit the gene expression curve, then further explore the similar expression genes. In contrast to abstract parameters and data, it provides a visualization searching method to detect target genes and visualizes the result in heat map and network map furtherly. GEsture enriches the diversity of methods analyzing time-series expression data. It is available at http://bio.njfu.edu.cn/GEsture.

Conclusions

In conclusion, GEsture is a web-based and user-friendly tool, which can detect expression genes from time series gene expression data. It has some advantages. First, users can quickly identify genes showing three expression patterns (similar, opposite, and shift) with input gene expression pattern. Three examples showed that GEsture performed well. It can detect some expression patterns more efficiently than K-mean clustering. Therefore, GEsture will be an alternative method for users if the clustering methods failed. Second, GEsture provides an easy-to-use input interface. Users can draw a curve using mouse instead of inputting abstract parameters from defined algorithms. Thirdly, GEsture provides visualization tools (such as expression pattern figure, heatmap and correlation network) to display the searching results. The output results may provide useful information for users to understand the targets, function and biological processes of the input gene of choice.

Acknowledgements

The data of Examples 1 and 2 was collected at the Yeast Cell Cycle Analysis Project. The data of Example 3 was collected from Mockler Lab. We thank Anna Jiang for her advice for designing this tool.

Reference

Androulakis, I. P., Yang, E., & Almon, R. R. (2007). Analysis of Time-Series Gene Expression Data: Methods, Challenges, and Opportunities. *Annual Review of Biomedical Engineering, 9*(9), 205-228.

Balasubramaniyan, R., Hüllermeier, E., Weskamp, N., & Kämper, J. (2005). Clustering of gene expression data using a local shape-based similarity measure. *Bioinformatics, 21*(7), 1069-1077.

Barrett, T., & Edgar, R. (2006). Mining microarray data at NCBI's Gene Expression Omnibus (GEO)*. *Methods in Molecular Biology, 338*, 175.

Bass, J., & Takahashi, J. S. (2010). Circadian Integration of Metabolism and Energetics. *Science (New York, N.Y.), 330*(6009), 1349-1354.

Bellpedersen, D., Cassone, V. M., Earnest, D. J., Golden, S. S., Hardin, P. E., Thomas, T. L., & Zoran, M. J. (2005). CIRCADIAN RHYTHMS FROM MULTIPLE OSCILLATORS: LESSONS FROM DIVERSE ORGANISMS. *Nature Reviews Genetics, 6*(7), 544-556.

Chervitz, S. A., Hester, E. T., Ball, C. A., Dolinski, K., Dwight, S. S., Harris, M. A., . . . Roe, T. Y. (1999). Using the Saccharomyces Genome Database (SGD) for analysis of protein similarities and structure. *Nucleic Acids Research, 27*(1), 74-78.

Culhane, A. C., Schröder, M. S., Sultana, R., Picard, S. C., Martinelli, E. N., Kelly, C., . . . Flahive, W. (2012). GeneSigDB: a manually curated database and resource for analysis of gene expression signatures. *Nucleic Acids Research, 40*(Database issue), 1060-1066.

David, S., Christopher, W., Philippe, L., Berardini, T. Z., Margarita, G. H., Hartmut, F., . . . Larry, P. (2008). The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Research, 36*(Database issue), 1009-1014.

Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America, 95*(25), 14863-14868.

Franz, M., Lopes, C. T., Huck, G., Dong, Y., Sumer, O., & Bader, G. D. (2016). Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics, 32*(2), 309-311.

Hardin, P. E., & Panda, S. (2013). Circadian timekeeping and output mechanisms in animals. *Current Opinion in Neurobiology, 23*(5), 724-731.

Horyu, D., & Hayashi, T. (2013). Comparison between Pearson Correlation Coefficient and Mutual Information as a Similarity Measure of Gene Expression Profiles. *Japanese Journal of Biometrics, 33*(2), 125-143.

Hsu, P. Y., & Harmer, S. L. (2014). Wheels within wheels: the plant circadian system. *Trends in Plant Science, 19*(4), 240-249.

Iyer, V. R., Horak, C. E., Scafe, C. S., Botstein, D., Snyder, M., & Brown, P. O. (2001). Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature, 409*(6819), 533.

Jiang, D., Pei, J., & Zhang, A. (2003). *DHC: A Density-Based Hierarchical Clustering Method for Time Series Gene Expression Data.* Paper presented at the IEEE Symposium on Bioinformatics and Bioengineering.

Jiang, D., Tang, C., & Zhang, A. (2004). *Cluster Analysis for Gene Expression Data: A Survey*: IEEE Educational Activities Department.

Joska, T. M., Zaman, R., & Belden, W. J. (2014). Regulated DNA Methylation and the Circadian Clock: Implications in Cancer. *Biology, 3*(3), 560.

Liberzon, A. (2014). *A Description of the Molecular Signatures Database (MSigDB) Web Site*: Springer New York.

Madhani, H. D., Galitski, T., Lander, E. S., & Fink, G. R. (1999). Effectors of a Developmental Mitogen-Activated Protein Kinase Cascade Revealed by Expression Signatures of Signaling Mutants. *Proceedings of the National Academy of Sciences of the United States of America, 96*(22), 12530-12535.

Mockler, T. C., Michael, T. P., Priest, H. D., Shen, R., Sullivan, C. M., Givan, S. A., . . . Chory, J. (2007). The DIURNAL project: DIURNAL and circadian expression profiling, model-based pattern matching, and promoter analysis. *Cold Spring Harbor Symposia on Quantitative Biology, 72*(8), 353.

Qian, J., Lin, J., Luscombe, N. M., Yu, H., & Gerstein, M. (2003). Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. *Bioinformatics, 19*(15), 1917-1926.

Schliep, A., Costa, I. G., Steinhoff, C., & Schnhuth, A. (2005). Analyzing Gene Expression Time-Courses. *IEEE/ACM Transactions on Computational Biology & Bioinformatics, 2*(3), 179-193.

Sharan, R., & Shamir, R. (2000). *CLICK: a clustering algorithm with applications to gene expression analysis.* Paper presented at the Eighth International Conference on Intelligent Systems for Molecular Biology.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., . . . Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell, 9*(12), 3273-3297.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., . . . Golub, T. R. (1999). Interpreting Gene Expression with Self-Organizing Maps. *Proceedings of the National Academy of Sciences, 96*(6), 2907-2912.

Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., & Church, G. M. (1999). Systematic determination of genetic network architecture. *Nature Genetics, 22*(3), 281-285.

Teixeira, M. C., Monteiro, P., Jain, P., Tenreiro, S., Fernandes, A. R., Mira, N. P., . . . Sácorreia, I. (2006). The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae. *Nucleic Acids Research, 34*(Database issue), 446-451.

Wang, L., Mo, Q., & Wang, J. (2015). MIrExpress: A Database for Gene Coexpression Correlation in Immune Cells Based on Mutual Information and Pearson Correlation. *Journal of immunology research, 2015*(1), 140819.

Wen, X., Fuhrman, S., Michaels, G. S., Carr, D. B., Smith, S., Barker, J. L., & Somogyi, R. (1998). Large-scale temporal gene expression mapping of central nervous system. Proc Natl Acad Sci USA 95: 334-339. *Proceedings of the National Academy of Sciences, 95*(1), 334-339.

Wijnen, H., & Young, M. W. (2006). Interplay of circadian clocks and metabolic rhythms. *Annual Review of Genetics, 40*(40), 409.

Wu, F. X. (2008). Genetic weighted k-means algorithm for clustering large-scale gene expression data. *Bmc Bioinformatics, 9*(S6), S12.

Xia, L. C., Ai, D., Cram, J., Fuhrman, J. A., & Sun, F. (2013). Efficient statistical significance approximation for local similarity analysis of high-throughput time series data. *Bioinformatics, 29*(2), 230-237.

Ye, N., Yin, H., Liu, J., Dai, X., & Yin, T. (2015). GESearch: An Interactive GUI Tool for Identifying Gene Expression Signature. *BioMed research international, 2015*, 853734.

Yeung, K. Y. (2001). Model-based clustering and validation techniques for gene expression data. *Faculty.washington.edu*, *17*(10), 977--987.