

Gene Selection in Time-Series Gene Expression Data

Prem Raj Adhikari¹, Bimal Babu Upadhyaya¹,
Chen Meng², and Jaakko Hollmén¹

¹ Aalto University School of Science,
Department of Information and Computer Science,
P.O. Box 15400, FI-00076 Aalto, Espoo, Finland
`prem.adhikari@aalto.fi`

² Royal Institute of Technology,
School of Computer Science and Communication,
Department of Computational Biology,
P.O. Box S-100 44, Stockholm, Sweden
`chenm@kth.se`

Abstract. The dimensionality of biological data is often very high. Feature selection can be used to tackle the problem of high dimensionality. However, majority of the work in feature selection consists of supervised feature selection methods which require class labels. The problem further escalates when the data is time-series gene expression measurements that measure the effect of external stimuli on biological system. In this paper we propose an unsupervised method for gene selection from time-series gene expression data founded on statistical significance testing and swap randomization. We perform experiments with a publicly available mouse gene expression dataset and also a human gene expression dataset describing the exposure to asbestos. The results in both datasets show a considerable decrease in number of genes.

Keywords: Feature Selection, Statistical Significance, Time-series, Randomization.

1 Introduction

Biological datasets are often high-dimensional in terms of the number of measurements whereas the number of samples is often very low. However, performance of learning algorithms is poor when data is high-dimensional and sample size is small because of curse of dimensionality. The problem of high dimensionality has inspired several concentrated research in feature selection and feature extraction [1–3]. Furthermore, biological systems are very complex systems and measurements in biology are performed using high-resolution and high-throughput techniques thus producing high-dimensional data. Hence, bioinformatics community is also highly interested in feature selection and feature extraction methods to be used in their learning algorithms [3]. Moreover, data incorporating

temporal information, often referred to as time-series data in machine learning literature, increases complexity of feature selection. Each subsequent feature consists of temporal information in that time-stamp thus making feature selection a difficult task. Hence, several researchers have extensively investigated the problem of feature selection in time-series data [4]. In [5], authors study the problem of feature selection in a time-series prediction problem.

In biology, recent advancement in technology has also provided means to study the complex biological system. Micro-array technology, such as CGH (Comparative Genomic Hybridization) and aCGH (Array Comparative Genomic Hybridization) offer the facilities to study the genomes and the genes in high resolution [6]. Hence, in most biological experiments that measure genes of a sample (for e.g. a cancer patient), data generated would consist of dimensions from tens of thousands to hundreds of thousands. However, number of samples are severely restricted in biological experiments. For example, only small number of measurement samples are available from cancer patients. Hence, feature selection is inevitable in biological problems such as the analysis of genes. Considering its importance, research community has coined a new term called gene selection to describe the feature selection procedure in genes [3, 7]. Additionally, gene selection also helps in determining relevant genes for future detailed studies of gene expression by extensive and expensive methods such as qRT-PCR or further detailed experiments with specific focus on selected genes leading to bio-marker identification. In some cases, problem of bio-marker identification could be solved by feature selection alone.

Time-series data in biology usually originates from the effects of external stimuli such as drug treatment on certain biological entity such as genes in different time intervals. Difficulty of gene selection in time-series is further compounded by presence of few data points in the time-series because it is difficult to capture a trend from small number of data samples. Available gene selection methods in bioinformatics such as [7] and [8] would not be applicable in time-series data. Gene selection methods in time-series data has received considerable interest. Authors in [9] use average-periodogram to select the most discriminative genes. However, there is a graphical component in their pipeline that needs to be inspected to determine the periodic components in data thus involving human intuition. Additionally, Fisher's test is used only on the peaks but we are interested in overall changes in data not only the peaks. In [10], authors define candidate temporal profiles in terms of inequalities among mean expression levels at the time points and select the genes that meet a bootstrap-based criterion for statistical significance and assign each selected gene to the best fitting candidate profile. However, the algorithm requires specifying the candidate profile which is often an arduous task. Similarly, authors in [11] use ranking of genes, HMMs and cross-validation to select genes. However, their method uses backward selection which is computationally very inefficient for large number of genes. Moreover, all of these methods are supervised and require a class label to select the features. Similar to other application areas, biology also has difficulty in getting the class label.

In wrapper settings, gene selection requires a modelling technique before the genes are selected and if the model selection procedure is problematic, the gene selection results are not reliable. Moreover, traditional feature selection algorithms such as forward/backward selection requires modelling of each gene and their combination, which is computationally expensive and inefficient. Feature extraction methods such as Principal Component Analysis (PCA) and Independent Component Analysis (ICA) [12] are not suitable because these methods can not determine most discriminative set of genes instead they transform the entire data and extract new set of features. Determining set of most discriminative genes is important because often these analysis of micro-array data is used as a guide to further precise studies of gene expression by extensive and expensive methods such as qRT-PCR. Additionally, the available gene selection methods are complicated and difficult to implement. They also require considerable amount of prior knowledge and a number of parameters.

This paper proposes an unsupervised method founded on statistical significance of gene expression and swap randomization for gene selection. The proposed algorithm is based on existing methods but novelty in this contribution is the construction of the pipeline combining the available methods with application to gene selection. The selection of biologically motivated special test statistic of longest run of consecutive 1s improves the selection of discriminative genes which is also a novel contribution of this paper. We consider the problem of gene selection as a series of hypothesis testing problems; first in determining if the gene is expressed at any given time instance and second when we select the final set of genes using swap randomization. The methods used in the pipeline are easy to implement individually but the implementations are also freely available. We experiment our method on two time-series gene expression datasets and obtain considerable reduction in number of genes.

2 Time-Series Gene Expression Dataset

We performed experiments on two different time-series gene expression datasets. One of them on asbestos exposure was the data presented in [13] and the other dataset on T cell differentiation is publicly available from NCBI¹ database [14]. The T cell differentiation dataset is an Affymetrix mouse genome array and mouse T cell line CTLL-2 data. It describes the expression from Murine T cell in response to Interleukin-2 at different time points in 24 hours after Interleukin-2 treatment. This dataset is used in [14] to analyze the gene expression profiles of Interleukin-2 regulated genes over ten different time points. There are 45101 genes measured in ten different time points. Control for every time point in time-series is not available. Hence, a random sample not treated with Interleukin-2 is taken as the control group.

Asbestos exposure dataset [13] is also a micro-array data that describes the effect of drug crocidolite exposed through an asbestos on the gene expression of humans. It represents the evolution of gene expression levels over time for both

¹ <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE6085>

asbestos exposed and non-exposed samples of A549 Caucasian lung carcinoma. There were 27 micro-arrays and each array has 54675 probe sets; 38500 of which corresponds to 38500 well established genes. Hence, there are 54675 samples of data measured in 5 different time intervals. Unlike T cell differentiation dataset, control is available for every five data points in the time-series.

3 Methodology of Pipeline

We had access to two datasets as discussed in Section 2 which were already pre-processed using image processing techniques. Quality control was also already performed on the dataset. Our pipeline proceeds using the methodology as shown in Figure 1 starting with normalization of data and ending with the swap randomization to determine the statistically significant genes. These statistically significant genes are the final result of gene selection.

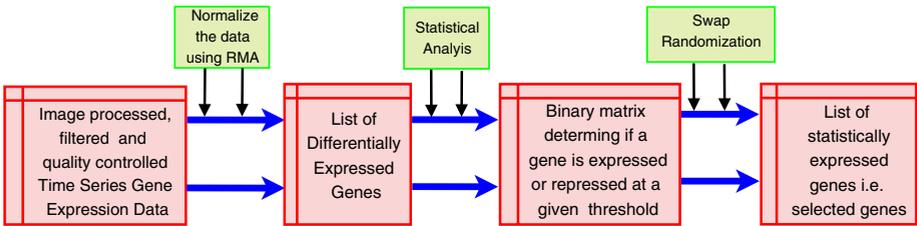


Fig. 1. The pipeline of the methodology. It starts with generic micro-array processing techniques and completes with swap randomization which enlists the statistically significant genes in the gene expression time-series. These statistically significant genes are the selected genes by our pipeline.

3.1 Normalization of Micro-Array Data

Normalization is the first transformation method applied to the image processed data from the micro-array experiments. Normalization adjusts the hybridization intensities such that they are suitable for biological comparisons [15]. It is important to adjust unequal amounts of initial RNA, to account for differences in labeling or detection efficiencies of fluorescence dyes used, and remove the bias of the system arising during the measurement [15]. There are different methods of normalization, such as \log_2 normalization, and intensity based filtering of array elements. The choice of normalization method also depends upon the source micro-array platform used. For example, methods such as MAS5.0, RMA or GC-RMA are suitable for Affymetrix micro-array data. For spotted arrays the normalization method depends on the experimental setup. For example, for within array normalization, \log_2 normalization is one of the most suitable option while methods such as median centering and scale normalization is suitable in other situations. In our experiments we used RMA (Robust Multi-chip

Average) normalization [15] method to normalize the time-series gene expression data. RMA normalization comprises of three distinct steps: background adjustment, quantile normalization, and summarization [15].

3.2 Gene Expression Analysis

Our test of statistical significance is based on simple comparison of two gene expression measurements of experimental and control patients [16]. Since we had access to two sets of measurements control and experimental in T cell differentiation dataset, we use the statistical significance to determine the differentially expressed genes. Asbestos exposure data at our disposal had no replicate copies required for determining the statistical significance. Hence, we generated the random samples from within the 95% confidence interval of the original gene expression measurements as the new replicate measurements. Our hypothesis is that “statistically significant genes differentially expressed whereas statistically insignificant genes are not differentially expressed” [16]. We calculate the statistical significance using paired t -test from the logarithm of the expression values to assess the significance of the measured relative expression or repression of the genes at different time instances. Multiple testing corrections were performed using bootstrapping with replacement. Confidence interval is chosen as discussed in Section 4 which is directly correlated with the number of genes required for the final output. If fewer number of genes are required; confidence interval is higher and if more number of genes are required confidence interval is lower.

3.3 Swap Randomization

After statistical significance testing is performed as discussed in Section 3.2, we can determine the list of genes that are differentially expressed at different time-points. It is important to note that a single gene can be expressed at certain time instance whereas it can be unexpressed at other time instances. Hence, we generate a binary matrix where 1 denotes that the gene is differentially expressed at that time instance where as 0 denotes the absence of differential expression of gene at that time instance. Now, the gene expression dataset is a 0-1 dataset where each row represents one gene at different time instances. Now the problem that arises is that the normal statistical significance testing can not be used to select the statistically significant genes because data belongs to a class of empirical distributions thus integrating over the PDF (Probability Density Function) is not feasible. Additionally, for any given data set \mathcal{D} , its PDF i.e. true generating model is often unknown. However, we can fix a null distribution and sample the data from the null distribution [17]. Swap randomization can be used to sample the data from null distribution [18].

Test Statistic. We introduce a new test statistic in this paper i.e. the longest run of consecutive 1s and 0s. It measures how often 1s and 0s comes together (adjacently) in each row i.e. for a gene in all time instances. Table 1 gives an example of longest run of consecutive 1s and consecutive 0s. To accommodate

this special test statistic, we only consider more extreme values while calculating p -values instead of the original definition which also considers the results that are as extreme as the original result i.e. equal to the original result.

Table 1 shows an example of calculation of longest run of consecutive 1s and 0s test statistic in a 5 dimensional dataset. In our experiments, if the longest run of consecutive 0s is statistically significant and occurs more than three times then Gene E is not selected.

Table 1. A toy data explaining the procedure of counting consecutive 1s and consecutive 0s

Data Points						Consecutive	
Gene	x_1	x_2	x_3	x_4	x_5	1s	0s
A	1	1	1	1	0	4	1
B	1	1	0	0	1	2	2
C	1	1	1	0	1	3	1
D	1	0	1	0	1	1	1
E	0	1	0	0	0	1	3

Null Distribution

The null distribution considered throughout this paper is similar to [18] and satisfies the following properties for a 0-1 dataset, \mathcal{D} :

1. The size of the random datasets, $\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_n$ is same as the original dataset, \mathcal{D} . If the original dataset, \mathcal{D} , has m rows and k columns, each of the sampled datasets $\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_n$ also has m rows and k columns.
2. The random datasets, $\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_n$ has the same column and row margins (sums) as the original dataset, \mathcal{D} . This constraint preserves the number of times the genes are expressed at any time instance and also the number of times a gene is expressed different time instance. It also automatically preserves the number of 1s in the dataset i.e. total number of genes differentially expressed in the data.

If we add more properties to this list, the randomization is more conservative. This is advantageous because we want to compare our results to the closely related datasets. If the randomization is unconstrained, then most of the results obtained by an algorithm will be statistically significant. For example, if we compare the height of grade 10 students with grade 5 students, almost everyone will be significantly tall. However, we want to state that a student in grade 10 is significantly tall than we need to compare his height with closely related students such as other students in grade 10. This constrained randomization is one of the main advantages offered by swap randomization.

Swap randomization [18] algorithm proceeds as follows: Let us consider a 0-1 dataset, \mathcal{D} and let $\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_n$ be the randomized datasets. Also consider an algorithm, \mathcal{A} , run on the dataset, \mathcal{D} , resulting in $\mathcal{A}(\mathcal{D})$ which determines the structural measure of the dataset \mathcal{D} . The structural measure used in our case is

the longest run of consecutive 1s and 0s in each row i.e. number of times genes are expressed in successive time intervals (1s) or not expressed in successive time intervals (0s). The algorithm \mathcal{A} is also applied to the randomized datasets producing results $\mathcal{A}(\mathcal{D})_1, \mathcal{A}(\mathcal{D})_2, \dots, \mathcal{A}(\mathcal{D})_n$. Now, p -values can be calculated from these results. The calculated p -values can be used to compare if the result on original data is significantly different from random data. The null hypothesis, \mathcal{H}_0 , considered in this paper is that for all datasets, \mathcal{D} , satisfying the above constraints, test statistic follows the same distribution.

p -Values

p -value is the probability of attaining a test statistic in randomized data as extreme as the original data, assuming that the null hypothesis, \mathcal{H}_0 , is true [17, 19]. For a dataset, \mathcal{D} , let us sample i.i.d. datasets $\hat{\mathcal{D}} = \{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_n\}$ from the null distribution, \mathcal{H}_0 . Now, one-tailed *empirical p -value* of $\mathcal{A}(\mathcal{D})$ for $\mathcal{A}(\mathcal{D})$ being as extreme can be defined by

$$\tilde{p} = \frac{1}{n+1} \left(\sum_{i=1}^n I(\mathcal{A}(\mathcal{D}_i) \geq \mathcal{A}(\mathcal{D})) + 1 \right), \quad (1)$$

where $i \in \{1, 2 \dots n\}$

Equation 1 defines the fraction of randomized dataset, $\hat{\mathcal{D}}$, whose structural measure $\mathcal{A}(\hat{\mathcal{D}})$ is as extreme as the original data $\mathcal{A}(\mathcal{D})$. Since our test statistic is special, we consider only those results that are more extreme than original results. Swap randomization [18] produces 10^4 samples of randomized dataset data using Markov Chain Monte Carlo (MCMC) using forward backward approach. Furthermore, Holm-Bonferroni correction was used to address the problem of multiple hypothesis testing [20].

4 Experiments and Results

We ran the proposed gene selection pipeline depicted in Figure 1 on two time-series gene expression datasets which were presented in Section 2. The CEL files was extracted and relevant preprocessing was performed necessary for data from Affymetrix micro-array data using Affy package in R. The dataset was also normalized to account for experimental variations using RMA [15] as discussed in Section 3.1. The normalized data was subjected to statistical significance analysis as discussed in Section 3.2. We implemented a paired t -test to the logarithm of the expression levels to assess the significance of the measured relative expression or repression of the genes at various time points in the time-series. After the application of statistical significance, the resulting matrix is a matrix of p -values where each elements, p_{ij} , take the value between 0 and 1 such that $0 \leq p_{ij} \leq 1$. The lesser the value, the more statistically significant the gene is at that time instance.

The p -values were obtained by bootstrapping strategy with replacement to correct for the multiple tests. We arbitrarily selected 90% as the confidence

interval to determine if a gene is statistically expressed ($\alpha = 0.1$). In 90% confidence interval more than two-thirds of the genes were significantly expressed at different time points. We created a 0-1 matrix where 1 represents that the gene at that time instance is differentially expressed while 0 denotes that the gene in that time instance is not differentially expressed at 90% confidence interval. A number of genes were not statistically expressed at any time instance so we excluded them from further experiments i.e. swap randomization as these genes are not interesting and their expression is not affected by external stimuli. Thus, achieving reduction in number of genes from 45101 to 21478 genes in T cell differentiation dataset and from 54675 to 38510 genes in asbestos exposure dataset.

4.1 Swap Randomization and Convergence Analysis

One of the important parameters to consider in the swap randomization process is number of swaps to be performed on the dataset i.e. the parameter \mathcal{J} . In our experiments and as suggested in [18], we consider the swap randomization procedure to converge when the distance, Frobenius norm, between the original data and the randomized data has converged.

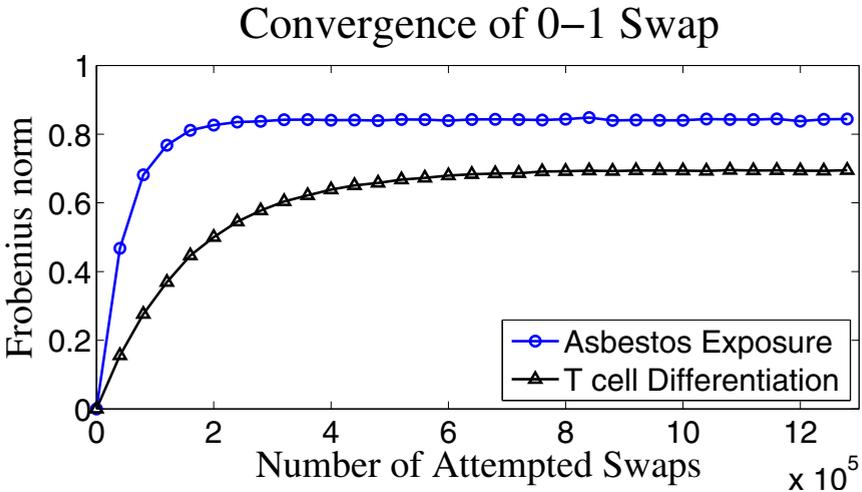


Fig. 2. Convergence analysis of 0-1 Swap. Y-axis shows the Frobenius norm between original data and swapped data after the number of performed swaps specified in the X-axis.

First we initialize the number of attempted swaps to 1 and increase it in the step-size of 1 which is different from the strategy used in [18] where the number of attempted swaps is fixed to \mathcal{K} equal to the number of 1s in the dataset. Figure 2 shows that Frobenius norm initially increases speedily as the number of attempted swaps increases but tends to reach steady state after certain number

of attempted swaps. Reaching the steady state is considered as the point of convergence. We repeat this experiment 10 times and report the mean of the results. For example, in the case of T cell differentiation dataset as shown in Figure 2, the convergence is reached when the number of swaps is approximately 6×10^5 . Similarly, for the Asbestos exposure dataset convergence is reached when the number of swaps is approximately 2×10^5 . However, taking a conservative approach we use 9×10^5 and 3×10^5 swaps for T cell differentiation and asbestos exposure datasets, respectively.

4.2 Selection of Genes

We generate 10^4 samples of random datasets after performing selected number of swaps for each random dataset. Now, we calculate the longest run of consecutive 1s in original dataset and also in each of the randomized datasets for each gene. As shown in the Figure 3, consecutive occurrence of 1s is mostly rare in the original data. However, among these rare occurrence of 1s; we find if these rare occurrence of 1s are statistically significant. Hence, we calculate the p -values comparing two p -values from original and randomized data as discussed in Section 3.3. The p -values were corrected for multiple corrections using Holm-Bonferroni [20]. The crux of the problem we are trying to solve is the selection of a set of genes. Finally, we select the genes that are significant at 99% confidence interval. In T cell differentiation dataset, number of statistically significant genes at 99% confidence interval were 12814. However, these 12814 sets of genes also includes those genes which have only one as the longest run of consecutive occurrence of 1s.

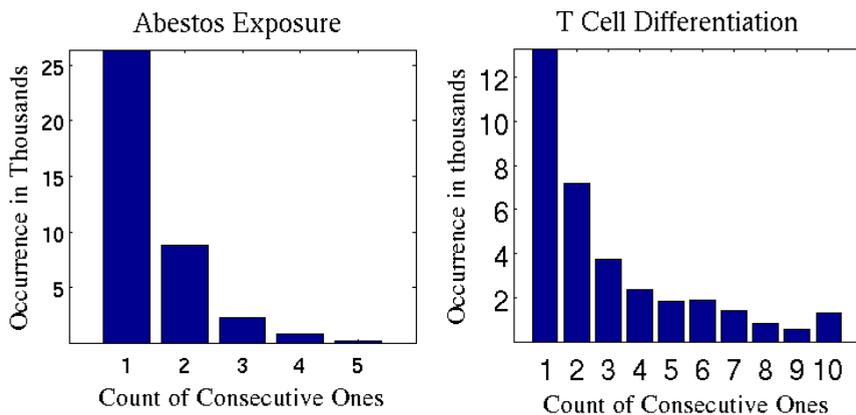


Fig. 3. Histogram showing number of longest run of consecutive 1s in different genes in original dataset. X-axis shows the number of longest run of consecutive 1s in both the figures. Consecutive occurrence of more than two 1s is very rare. We are interested in those genes that have significantly high number of consecutive occurrence of 1s and significantly less number of consecutive 0s.

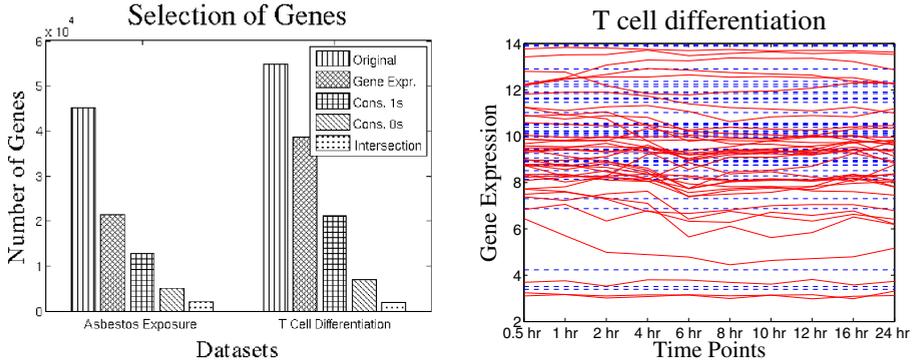


Fig. 4. Left panel shows the reduction in number of genes during different phases of analysis pipeline. Initially the number of genes are greater than 45000 in both datasets but finally the number of selected genes is around 2000. Right panel shows the gene expression profiles of first 40 selected genes at ten different time points. Solid line shows the expression profiles of treatment group and dashed lines represents the expression profiles of control group. Figure in the right panel shows that expression of selected genes are comparatively different from the control group over a period of time.

It is also to be noted that we are also not interested in genes having significantly high longest run of consecutive occurrence of 0s. Hence, we also calculated the genes that have significantly high occurrence of longest run of consecutive 0s with a threshold of 3; i.e. we calculate if a gene shows occurrence of 3 or more longest run of consecutive 0s. Unlike the longest run of consecutive 0s longest run of consecutive 1s were not subjected to a threshold. Since, we’ve not thresholded the longest run of consecutive 1s property, it allows for some incorrect assignment of differential expression at a time instance. Finally, as shown in the Figure 4, we take the intersection of genes that have significantly high longest run of consecutive 1s (12814 genes) and genes that do not have significantly high occurrence of 3 or more longest run of consecutive 0s (4980 genes). The intersection results in 2015 genes which computationally and even biologically seems appropriate number of genes for any experiment. Instead of longest run of consecutive 1s other thresholding methods would not produce useful results. Row sums are definitely one of them but we can not test for their statistical significance because row and column sums are preserved during randomization process. Removing rows having column sums below the specified threshold as with zero vectors would decrease the size of data significantly thus affecting the randomization process. Similarly, in asbestos exposure dataset, number of genes having significant number of consecutive 1s is 21174 and that of consecutive 0s with threshold of 3 is 7104. As shown in the Figure 4, the intersection of these two sets results in 1811 genes which is the final set of selected genes.

Figure 4 shows that gene expression profiles of selected genes differ from their control group over a longer timespan showing that selected genes are relevant set of genes and most responsive to the applied external stimuli. From the numbers

seen in the results, we can perceive that the number of genes selected from the total set of genes is small. These selected set of genes are statistically significant and are more likely to be affected by the external stimuli such as the application of drugs. Since each micro-array is a snapshot whereas the biological process is dynamic, the time-series selection helps us to trace the dynamics of biology from snapshots. These selected genes can be used for diagnostics using them as features for different classification and regression models.

5 Summary and Conclusions

In this paper, we propose a novel unsupervised method for gene selection in time-series gene expression data. The pipeline is based on the combination of statistical testing founded on a paired t -test and statistical significance testing founded on swap randomization [18]. Experiments are performed on two datasets: one of them on a publicly available mouse T cell differentiation and the other one on asbestos exposure. The experiments showed that number of genes are considerably reduced in both the experimented datasets. One of the major strength of the proposed methodology is its unsupervised nature thus it does not require expensive and difficult-to-get class labels and results in large reduction in number of genes.

References

1. Kira, K., Rendell, L.A.: A practical approach to feature selection. In: Proceedings of the ninth international workshop on Machine Learning, ML 1992, pp. 249–256. Morgan Kaufmann Publishers Inc., San Francisco (1992)
2. Blum, A.L., Langley, P.: Selection of relevant features and examples in machine learning. *Artificial Intelligence* 97(1-2), 245–271 (1997)
3. Saeys, Y., Inza, I., Larrañaga, P.: A review of feature selection techniques in bioinformatics. *Bioinformatics* 23(19), 2507–2517 (2007)
4. Mörchen, F.: Time series feature extraction for data mining using DWT and DFT. Technical Report 33, Department of Mathematics and Computer Science, University of Marburg, Germany (2003)
5. Tikka, J., Hollmén, J.: A Sequential Input Selection Algorithm for Long-term prediction of Time Series. *Neurocomputing* 71(13-15), 2604–2615 (2008)
6. Heller, M.J.: DNA microarray technology: Devices, systems, and applications. *Annual Review Of Biomedical Engineering* 4, 129–153 (2002)
7. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for Cancer Classification using Support Vector Machines. *Machine Learning* 46, 389–422 (2002)
8. Diaz-Uriarte, R., Alvarez de Andres, S.: Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* 7(1), 3 (2006)
9. Wichert, S., Fokianos, K., Strimmer, K.: Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics* 20, 5–20 (2004)
10. Peddada, S.D., Lobenhofer, E.K., Li, L., Afshari, C.A., Weinberg, C.R., Umbach, D.M.: Gene selection and clustering for time-course and dose-response microarray experiments using order-restricted inference. *Bioinformatics* 19(7), 834–841 (2003)

11. Lin, T., Kaminski, N., Bar-Joseph, Z.: Alignment and classification of time series gene expression in clinical studies. *Bioinformatics* 24(13), i147–i155 (2008)
12. Hyvärinen, A., Karhunen, J., Oja, E.: Independent component analysis. Adaptive and learning systems for signal processing, communications, and control. John Wiley and Sons (2001)
13. Nymark, N., Lindholm, P.M., Korpela, M.V., Lahti, L., Ruosaari, S., Kaski, S., Hollmén, J., Anttila, S., Kinnula, V.L., Knuutila, S.: Gene Expression Profiles in Asbestos-exposed Epithelial and Mesothelial Lung Cell Lines. *BMC Genomics* 8(1), 62 (2007)
14. Zhang, Z., Martino, A., Faulon, J.: Identification of expression patterns of IL-2-responsive genes in the murine T cell line CTLL-2. *Journal of Interferon & Cytokine Research* 27(12), 991–995 (2007)
15. Bolstad, B.M., Irizarry, R.A., Astrand, M., Speed, T.P.: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19(2), 185–193 (2003)
16. Parmigiani, G.: The analysis of gene expression data: methods and software. Springer, Heidelberg (2003)
17. Good, P.: Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses, 2nd edn. Springer, Heidelberg (2000)
18. Gionis, A., Mannila, H., Mielikäinen, T., Tsaparas, P.: Assessing data mining results via swap randomization. *ACM Transactions on Knowledge Discovery from Data* 1(3), 14 (2007)
19. Schervish, M.J.: P Values: What They Are and What They Are Not. *American Statistician* 50(3), 203–206 (1996)
20. Holm, S.: A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6, 65–70 (1979)