A New Method to Identify Housekeeping Genes and Tissue Special Genes

Ji-xian RAO[1,*], Wei LIU[1] and Hong-wei XIE[1]

[1]Department of Automatic Control, College of Mechanical & Electronic Engineering and Automatization, National University of Defense Technology, Changsha 410073, China

*raojx_123@163.com

Keywords: Housekeeping genes (HK genes), Tissue special genes (TS genes),SVM (support vector machine), Machine Learning.

Abstract. Despite the detection methods of HK genes and TS genes were various, without the standard database, we could not judge which one was the best classifier. In this paper, we built a standard database of HK genes and TS genes, reviewed and evaluated the existed methods. By analyzing the features of HK genes and TS genes, we summarized 7 features: relative expression breadth (REB), signal to noise (S/N), tissue special index (TSI), Average, standard deviation (SD), Max/Average, dispersion measure (DPM). Based on the features, we used a new SVM (support vector machine) model to have identified more than 10000 HK genes and 6947 TS genes. It is the best classifier about HK genes and TS genes identification now.

Introduction

It is important to study HK genes and TS genes, which can explain the relationship of tissue special protein and disease or the protein-protein network [1,2,3,4]. In recent years, some high throughout experiments such as RNA-seq, Expressed Sequence Tag (EST), Microarray Chip and Mass Spectrum, produced more and more molecular expression data. By comparing small scale experiments, RT-PCR and Western blot, the high throughout experiments produced large volume of data and detected weak expression genes. Among these high throughout experiments, RNA-seq can detect weak expression genes, but it hard to distinguish noise or weakly expressed gene. And more and more researchers adopted RNA-seq method to reveal the relationship between HK genes and TS genes.

The definition of housekeeping genes is constantly evolving, and we have divided these definitions here into two major types. The early one represented by Watson et al. [5] and Warrington et al.[6] stated that the housekeeping genes need to be constitutively expressed in every tissue to maintain cellular functions. Due to measurement errors and stochastic noise, it is difficult to distinguish genes absent in the sample from those weekly expressed. The newer definition extends the first definition and emphasizes on a constant and stable expression, which was initially raised by Butte et al [7].Also the initial definition of TS genes meant that genes only expressed in one or several tissues. Another definition indicated that the expression level of tissue special genes in one or several tissues was higher than other tissues, which also emphasized a constant and stable expression.

It was existed many methods to distinguish HK genes and or TS genes, such as signal to noise (S/N) [8], relative expression breadth (REB), tissue special index (TSI) [9], HKera [10], etc.S/N identified genes as HK or TS based on the criterion of high or fairly constant expression, whereas REB did not focus on the magnitude of

expression, but instead, used a certain number as a threshold. TSI used a quantitative measure of variation in expression profiles in different tissues to evaluate the tendency of a gene to be HK gene (little tissue-wide variation) and TS gene (high variation). Different from all of the above, based on the tensor structure of tissue-wide gene expression profiles, HKera was a novel SVM classifier of designating a given human gene as HK or TS gene. Nowadays, we cannot judge which method is the best classifier. So we built a standard database to evaluate the existed methods. By analyzing the methods, we put forward a new method, using SVM to classify genes as HK or TS. And the results suggested our method was better than others methods.

Materials and Methods

Datasets

RT-PCR and Western blot were believed the best one to excavate HK genes and TS genes. By surveying papers, we built a standard database which could help us to evaluate the existed methods. We have found 1161 standard HK genes and 505 TS genes.

The GSE30611 RNA-seq data for Illumina Human Body Map 2.0 project transcription profiling of individual and mixture of 16 human tissues RNA was downloaded from GEO depositories and processed using previously described procedures [11]. This datasets contains gene expression profiles for 17419 genes in 16 tissues. Based on the standard database, we found standard HK genes and TS genes in tissue 16 database.

Methods

Performance Evaluation. To evaluate the HK prediction methods, we calculated the following performance measures:

$$\text{recall} = \frac{TP}{TP + FN}. \tag{1}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{2}$$

$$\text{specificity} = \frac{TN}{TN + FP}. \tag{3}$$

$$\text{accur}acy = \frac{TP + TN}{TP + FP + TN + FN}. \tag{4}$$

Where TP denotes true positive, FP false positive, TN true negative, and FN false negative. When computing the receiver operating characteristic (ROC) curves for HK prediction methods, TP was the number of correctly predicted HK genes in HK genes data sets, and TN was the number of correctly predicted TS genes in HK genes data sets, FP was the number of TS genes predicted to be HK and FN the number of benchmark HK genes predicted to be TS genes.

Methods Review. First of all, we evaluated the existed methods. REB meant how many tissues the gene expressed and it has been demonstrated in numerous studies [12,13,14]. S/N identified genes as HK or TS based on the criterion of high or fairly constant expression. TSI was defined as:

$$x_i = \frac{x_{\max} - x_i}{x_{\max} - x_{\min}}. \tag{5}$$

$$\tau = \frac{\sum_{i=1}^{N}(1 - x_i)}{N - 1}. \tag{6}$$

Where N is the number of tissue and x_i is the expression profile component normalized by the maximal component value. The DPM was a new method to identify HK gene or TS gene. To quantitatively estimate the relative expression specificity of a gene in a sample, the dispersion measure (DPM) [15,16] was introduced as following:

$$X = (x_1, x_2, x_3 \ldots x_i \ldots x_{n-1}, x_n). \tag{7}$$

$$X_i = (0,0,\ldots,x_i,\ldots 0,0). \tag{8}$$

$$SPM_i = \cos\theta = \frac{X_i \bullet X}{|X_i| \bullet |X|}. \tag{9}$$

$$DPM = \sqrt{\frac{\sum_{i=1}^{n}(SPM_i - \overline{\mathrm{SPM}})^2}{n-1}} \bullet \sqrt{n}. \tag{10}$$

Where X was the each gene expression profile. n was the number of samples in the profile. The $\overline{SPM}$ was the mean of SPMs in the gene expression profile. A vector X_i was created to represent the gene expression in a sample i. The SPM of a gene in a sample was then determined by calculating the cosine value of intersection angle θ between vector X_i and X in high-dimension feature space. Unlike conventional SD analysis, DPM was independent of gene expression level and sample number by scaling into a region of 0-1.0 as above. In this way, DPM made variability comparable between profiles or dataset. A value of DPM or TSI close to 0 suggested which gene can be defined as HK gene. And the value close to 1 suggested that gene could be TS genes.

Nowadays, Fourier analysis [17] transforms time-series gene expression data into Fourier spectra for a support vector machine (SVM, a machine learning method) to classify genes as HK or TS. And Austin et al.[10] found a novel classifier of HK and TS genes named HKera. HKera classified HK and TS based on the tensor structure of tissue-wide gene expression profiles. From all of the above, we summarized 7 features (REB, Average, SD, Max/Average, S/N, TSI, DPM) to training our SVM model. Our classifier was SVM model resulting from 5-fold cross validation on standard genes. We also used other machine learning algorithms (decision tree, BP neural network, naive Bayes) to build the model to classify HK and TS, but the SVM model was the best one to identify HK genes and TS genes.

Results and Discussion

By surveying the papers, we built a standard database about HK gene and TS genes. We have found 1161 HK genes and 505 TS genes. In the database, we summarized the geneID, symbol, annotation, source etc.

We evaluated four HK prediction methods (REB, S/N, TSI, DPM). The figure (Figure 1) are ROC curves of sensitivity vs. 1-specificity. Where sensitivity (i.e. Recall) and

specificity are defined respectively, by Equation (1) and Equation (3) in the methods. From the figure, the DPM and TSI were the best methods to identify HK gene or TS gene. While the method of S/N was better than REB. REB applied first HK definition (see Introduction) into HK genes classification, so some weak expressed gene or noise could influence the result of number of the HK genes. Because S/N accounted for the gene consistently expressed in all tissues, which was better than REB. DPM and TSI not only accounted for the gene consistently expressed in all tissues, but also avoided the noise and weak expressed.

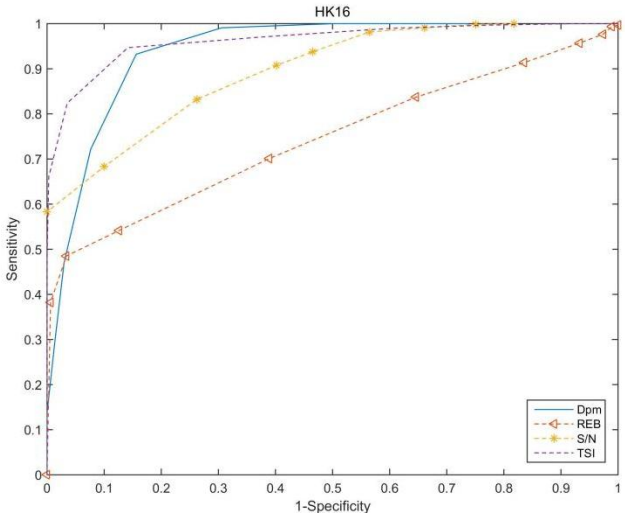

Figure1. Performance of four HK prediction methods.

Nowadays, machine learning is the most popular field of data science, one of which is SVM, the classification effect remarkable and easy to implement, lead to more and more researchers to use SVM to tissue special genes classification area. For example HKera and Fourier analysis are the SVM models which were shown to perform significantly better than other classifier that use different methodologies [10,17]. Table 1 summarized the performance of the four machine learning methods (1 decision tree, 2 BP neural network, 3 SVM in this study, 4 naive Bayes) to classify HK genes on training/test data. The results showed that our study was among the best performers.

Table 1. Performance of four machine learning models on training/test data

	Training(%)			Test(%)		
Model	Accuracy	Recall	Precision	Accuracy	Recall	Precision
1	88.3	81.7	91.2	85.7	80.3	90.5
2	67.3	87.7	92.0	72.3	85.6	89.3
3*	**94.2**	**93.7**	**92.5**	**93.6**	**91.2**	**90.3**
4	91.4	82.7	91.3	88.6	89.7	90.5

*model 3 was SVM in this study

Table 2. HK criterion and the resulting number of HK genes using different methods

Method	HK criterion	Number of HK genes	Reference
TSI	TSI<=0.5	2879	[9]
S/N	Consistently Expressed	2307	[8]
HKera	HKera score>=0.0	8070	[10]
DPM	DPM<=0.5	5389	[15,16]
Our method	SVM	10305	This work

Table 2 summarized the four different methods classifying our data (see Methods). TSI was bounded between 0 and 1, A lower TSI indicated a lower tendency for the gene to be TS (or a higher tendency for it to be HK). The 0.5 threshold was chosen following [18], and this method identified 2879 HK genes. While the HK criterion of the S/N identified 2307 HK genes. HKera was the SVM method with 5-fold cross validation which identified 8070 HK genes. The DPM was the new method to classify HK genes which could identify 5389 HK genes. In this study, we used 7 features to train our data and identify 10305 HK genes , which was the best one to identify HK genes among above methods.

Conclusions

The definition and detection methods of HK genes or TS genes are various.But without the standard data, we cannot evaluate which one is the best to classify genes as HK or TS. In this study:

(1) We built a standard database to test the method of DPM, S/N, TSI and REB classifying HK genes. The results showed the DPM and TSI were better than S/N and REB.

(2) We collected 7 features from a lots of papers, trained by SVM method to identify more than 10000 HK genes and 6947 tissue special genes. It is the best classifier about tissue special genes identification now.

But the effect of SVM method depends on the features, so it is import to summarized the features. While deep learning is one of popular algorithms in machine learning area, depends on big data, which can train the raw data without selecting features. Nowadays, deep learning algorithm has applied in visual recognition, image recognition etc. In the future, more and more studies with regard to deep learning will explain the relationship between HK genes and TS genes.

Acknowledgements

The authors would like to thank Liu Wei for many useful discussions. This work was supported by International Cooperation Project (2014DFB30010) and Chinese National Key Program of Basic Research (31171266).

References

[1] Uhlen M, Fagerberg L, Hallstrom B M, et al. Tissue-based map of the human proteome.Science, 2015, 347(6220) .

[2] Winter E E, Goodstadt L, Ponting C P. Elevated rates of protein secretion, evolution, anddisease among tissue-specific genes. Genome Res, 2004, 14: 54~61.

[3] Goh K I, Cusick M E, Valle D, et al. The human disease network. Proc Natl Acad Sci USA,2007, 104: 8685~8690.

[4] Chen M, Xiao J, Zhang Z, et al. Identification of human HK genes and gene expressionregulation study in cancer from transcriptomics data analysis. PLoS ONE, 2013, 8(1): e54082.

[5] Watson JD, Hopkins NH, Roberts JW, Steitz JA, Weiner AM. The functioning of higher eukaryoticgenes. Molecular Biology of the Gene. 1965;1.

[6] Warrington JA, Nair A, Mahadevappa M, Tsyganskaya M. Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. Physiological genomics. 2000; 2(3):143 – 7.

[7] Butte AJ, Dzau VJ, Glueck SB. Further defining housekeeping, or "maintenance,"genes Focus on"Acompendium of gene expression in normal human tissues". Physiological genomics. 2001; 7(2):95 – 6.

[8] Dezso Z, Nikolsky Y, Sviridov E, et al. A comprehensive functional analysis of tissuespecificity of human gene expression. BMC Biology, 2008, 6: 49.

[9] Yanai I, Benjamin H, Shmoish M, et al. Genome-wide midrange transcription profiles revealexpression level relationships in human tissue specification. Bioinformatics, 2005, 21:650~659.

[10]Chiang A, Shaw G, Hwang M. Partitioning the human transcriptome using HKera, a novelclassifier of housekeeping and tissue-specific genes. PLoS ONE, 2013, 8(12): e83040.

[11]Shaw GT, Shih ES, Chen CH, Hwang MJ (2011) Preservation of ranking order in the expression of human Housekeeping genes. PLoS One 6:e29314.

[12]Sémon M, Mouchiroud D, Duret L. Relationship between gene expression and GC-content in mammals: statistical significance and biological relevance. Human Molecular Genetics. 2005; 14(3):421 – 7.

[13]Lercher MJ, Urrutia AO, Hurst LD. Clustering of housekeeping genes provides a unified model of geneorder in the human genome. Nature genetics. 2002; 31(2):180 – 3.

[14]Vinogradov AE. Compactness of human housekeeping genes: selection for economy or genomic design? TRENDS in Genetics. 2004; 20(5):248 – 53.

[15]JPan J B, Hu S C, Wang H, et al. PaGeFinder: quantitative identification of spatiotemporalpattern genes. Bioinformatics, 2012, 28(11): 1544~1545.

[16]Pan J B, Hu S C, Shi D, et al. PaGenBase: a pattern gene database for the global and dynamicunderstanding of gene function. PloS ONE, 2013, 8 (12): e80747.

[17]Dong B, Zhang P, Chen X, et al. Predicting housekeeping genes based on Fourier analysis.PLoS One, 2011, 6: e21012.

[18]JChang CW, Cheng WC, Chen CR, Shu WY, Tsai ML, et al. (2011) Identification of human housekeeping genes and tissue-selective genes by microarray meta-analysis. PLoS One 6: e22859. doi: 10.1371/journal.pone.0022859.