CHEMICAL ENGINEERING TRANSACTIONS

VOL. 46, 2015

A publication of AIDIC

The Italian Association
of Chemical Engineering
Online at www.aidic.it/cet

Guest Editors: Peiyu Ren, Yancang Li, Huiping Song
Copyright © 2015, AIDIC Servizi S.r.l.,
ISBN 978-88-95608-37-2; **ISSN** 2283-9216

DOI: 10.3303/CET1546064

An Improved Method for Exon Array Data Analysis

Shenghua Jin

School of Computer engineering, Huaiyin Institute of Technology, Huaian, Jiangsu, China ;
Huaian key laboratory of the study and application of Internet of Things, Huaian, Jiangsu, China;
Jiangsu "Internet of Things" mobile Internet technology engineering laboratory, Huaian, Jiangsu, 223003;
zybjsh819@163.com

Alternative splicing of genes is usually induced by some severe diseases. Exon array manufactured by Affymetrix Company has been widely used to detect alternative splicing. Utilizing the mapping between the gray values of splice isoforms and the array probes, we proposed the calculation of gene expression level based on Kseq model. Aiming at the massive amount of array data, an algorithm for parallel computation using multi-core processor was presented. The algorithm was verified through experiments using real datasets and compared with the commonly used algorithm. It was found that parallel computation improved the computation efficiency of the model and the new model enhanced the computation accuracy in subsequent bioanalysis.

Keywords: Exon array, gene expression, alternative splicing Parallel Computing

1. Introductions

Human genome project (HGP) was completed in 2003. Since then, increasing concern has been given to gene functions and the mining of the correlations between genes and the diseases, which was confirmed in (Caceres and Kornblihtt, 2002). In the field of bioinformatics, exploring the mechanism of gene expression and transcriptional regulation represents new direction of research. DNA microarray technology makes monitoring simultaneously the gene expression level of thousands of genes under different samples possible, in gene expression data the type of alternative splicing plays an important role in gene expression and transcriptional regulation. Alternative splicing produces transcripts through various combinations of exons. Here we employed the exon array manufactured by Affymetrix Company.

PM probes are designed on the Affymetrix exon array, which was confirmed (Karin, 2010). The probes total about 6.5 million and constitute about 1.4 million probe sets. Santa Clara (2005) reported that four probes can cover about 90% of the exons and the array covers nearly 1 million exons. This high-integration exon array is able to detect the expressions of transcripts on the level of isomers, exons and genes, which was confirmed (KapurK, 2007). There are several methods for the calculation of isomer expression, including multi-mapping Bayesian gene expression (MMBGX), which was confirmed (Turro, 2010), and multiple exon array preprocessing (MEAP), which was confirmed (Chen, 2011). Based on the mapping between the gray values of splice isomers and probes, a data model conforming to chi-squared distribution was proposed to calculate the gene expression level (Yin and Liu.2015). Comparison between the proposed algorithm and other popular algorithms on real datasets reveals that Kseq model can effectively reduce the noises in the original experimental data and improved the accuracy and efficiency of subsequent bioanalysis.

2. Methods

2.1 Mapping between genes, isoforms and probes

Table 1 provides the relationships of 5 splice isomers to gene ENSG00000000457, which respectively are ENST00000367770, ENST00000367771, ENST00000367772, ENST00000423670, ENST00000470669 and ENST00000470238. In table 1 the second and third columns of pos-x and pos-y are the coordinates of the corresponding probes on the array, by which the probes corresponding to the isomers can be found ,and the fourth and fifth column of Probe-set is the name of probe set. The relationships among the gene, its splices and probes are presented as Fig 1.

Please cite this article as: Jin S.H., 2015, An improved method for exon array data analysis, Chemical Engineering Transactions, 46, 379-384 DOI:10.3303/CET1546064

Table 1: Mapping between genes and isoforms

Probe-set	pos-x	pos-y	Gene name	Isoform name
2443560	110	2055	ENSG00000000457	ENST00000367770
2443540	112	2516	ENSG00000000457	ENST00000367770
2443560	295	2543	ENSG00000000457	ENST00000367771
2443558	339	2192	ENSG00000000457	ENST00000367771
2443548	349	1782	ENSG00000000457	ENST00000367771
2443541	486	2531	ENSG00000000457	ENST00000367772
2443559	495	2125	ENSG00000000457	ENST00000423670
2443559	468	2324	ENSG00000000457	ENST00000423670
2443551	496	943	ENSG00000000457	ENST00000423669
2443551	523	1042	ENSG00000000457	ENST00000423669
2443552	652	1112	ENSG00000000457	ENST00000470238
2443552	661	956	ENSG00000000457	ENST00000470238

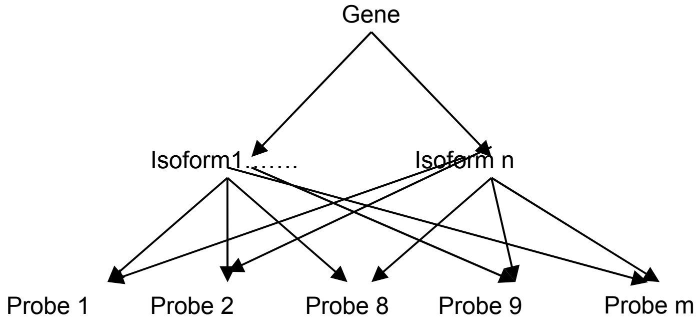

Fig 1: Illustration of the relationships among genes, splice isoforms and probes

Suppose that $y_{gjc} = \sum_k s_{gjkc}$, where y_{gjc} is the gray value of the j -th PM probe of gene g on array c . This gray value corresponds to several isomers of the gene. s_{gjkc} is the expression of the k -th isomer corresponding to this probe, which satisfies the chi-squared distribution of α_{gkc} and β_{gj} . β_{gj} is the random variable shared by several isomers corresponding to the probe. Given the properties of random variable conforming to chi-squared distribution, there is

$$y_{gjc} \sim Ga\left(\sum_k \alpha_{gkc}, \beta_{gj}\right) \tag{1}$$

Suppose β_{gj} conforms to the chi-squared distribution of c_g and d_g , then $\beta_{gj} \sim Ga\left(c_g, d_g\right)$, and each isomer satisfies the following:

$$p\left(s_{gjkc}\right) = \int p\left(s_{gjkc} \mid \alpha_{gkc}, \beta_{gj}\right) p\left(\beta_{gj} \mid c_{g,} d_g\right) d\beta_{gj} \tag{2}$$

Using (2), the joint distribution of each isomer is calculated, and the logarithm is taken for its likelihood function. The deduction is shown below:

$$\begin{aligned} L_g\left(\alpha_{gc}, c_g, d_g\right) &= \log P(Y_g) \\ &= \sum_j \log \int d\beta_{gj} P\left(\beta_{gj} \mid c_g, d_g\right) \prod_c P\left(y_{gjc} \mid \sum_k \alpha_{gkc}, \beta_{gj}\right) \\ &= \sum_j \log \left[\frac{d_g^{\,c_g} \Gamma(q)}{\Gamma(c_g)\omega^q} \prod_c \frac{y_{gjc}^{\sum_k \alpha_{gkc}-1}}{\Gamma(\sum_k \alpha_{gkc})} \right] \end{aligned} \tag{3}$$

where $\alpha_{gc} = \left[\alpha_{gkc}\right]$, $q = \sum_c \sum_k \alpha_{gkc} + c_g$, $\omega = \sum_c y_{gjc} + d_g$.

Estimates $\hat{\alpha}_{gkc}, \hat{c}_g, \hat{d}_g$ of the parameters α_{gkc}, c_g, d_g are obtained by maximum likelihood method. Then the probability density function of specific signal s_{gjkc} is solved using the estimated parameter values:

$$\begin{aligned} P(s_{gjkc} \mid \hat{a}_{gkc}, \hat{c}_g, \hat{d}_g) &= \int d\beta_{gj} P(s_{gjkc} \mid \hat{\alpha}_{gkc}, \beta_{gj}) P(\beta_{gj} \mid \hat{c}_g, \hat{d}_g) \\ &= \frac{\Gamma(\hat{c}_g + \alpha_{gkc}) \hat{d}_g^{\,\hat{c}_g} s_{gjkc}^{\hat{a}_{gkc}-1}}{\Gamma\left(\hat{\alpha}_{gkc}\right)\Gamma\left(\hat{c}_g\right)\left(\hat{d}_g + s_{gjkc}\right)^{\hat{c}_g + \hat{\alpha}_{gkc}}} \end{aligned} \tag{4}$$

The mean and variance of $\log\left(s_{gjkc}\right)$ are calculated:

$$\begin{aligned} \langle \log\left(s_{gjkc}\right) \rangle &= \log\left(\hat{d}_g\right) + \Psi(\hat{\alpha}_{gkc}) - \Psi(\hat{c}_g) \\ Var\left[\log(s_{gjkc})\right] &= \Psi'(\hat{\alpha}_{gkc}) + \Psi'(\hat{c}_g) \end{aligned} \tag{5}$$

where $\Psi(.)$ denotes the derivative of $\log\Gamma(.)$; $\Psi'(.)$ denotes the first-order derivative of $\Gamma(.)$. Thus the mean and variance of the corresponding gene are

$$\begin{aligned} \langle \log\left(\sum_k s_{gjkc}\right) \rangle &= \log\left(\hat{d}_g\right) + \Psi(\sum_k \hat{\alpha}_{gkc}) - \Psi(\hat{c}_g) \\ Var\left[\log(\sum_k s_{gjkc})\right] &= \Psi'(\sum_k \hat{\alpha}_{gkc}) + \Psi'(\hat{c}_g) \end{aligned} \tag{6}$$

Since the distribution of α_{gkc} is unimodal and α_{gkc} is larger than zero, the distribution of α_{gkc} is fitted by the Gaussian distribution truncated at point zero. The mean and variance of the isomer are

$$\begin{aligned} P\left(\alpha_{gc}\right) &\propto \exp\left(L'_{gc}(\hat{\alpha}_{gc})^T \left(\alpha_{gc} - \hat{a}_{gc}\right) + \frac{1}{2}\left(\alpha_{gc} - \hat{a}_{gc}\right)^T H^{gc}\left(\alpha_{gc} - \hat{a}_{gc}\right)\right) \\ &= N\left(\alpha_{gc}; \mu_{gc}, \textstyle\sum_{gc}\right) \end{aligned} \tag{7}$$

where Hessian matrix H^{gc} is written as the following:

$$H_{ij}^{gc} = \frac{\partial^2 L_{gc}\left(\alpha_{gc}\right)}{\partial\alpha_{gic}\partial\alpha_{gjc}} |_{\alpha_{gc}=\hat{\alpha}_{gc}}, \mu_{gc} = \sum_{gc} L'_{gc}\left(\hat{\alpha}_{gc}\right) + \hat{\alpha}_{gc}, \sum_{gc} = \left(-H^{gc}\right)^{-1} \tag{8}$$

$$\hat{\alpha} = C\left[\frac{\sigma}{\sqrt{2\pi}}\exp\left(-\frac{\mu}{2\sigma^2}\right) + \frac{\mu}{2} - \frac{\mu}{2}erf\left(-\frac{\mu}{\sqrt{2}\sigma}\right)\right] \tag{9}$$

$$\hat{\sigma}^2 = C\left[\frac{1}{2}\left(\sigma^2 + (\mu - \hat{\alpha})^2\right)\left(1 - erf\left(-\frac{\mu}{\sqrt{2}\sigma}\right)\right) + \frac{\sigma}{\sqrt{2}\sigma}(\mu - 2\hat{\alpha})\exp\left(-\frac{\mu^2}{2\sigma^2}\right)\right] \tag{10}$$

The mean and variance of the isomer are calculated by (9) and (10), respectively.

3. Experimental results and discussion

The real dataset GSE13072 was used to verify whether the proposed algorithm could reduce the noise in the original data. The dataset was derived from Gene Expression Omnibus Database, which is a source for alternative splice events and isomer expressions. Human Exon 1.0ST Array was used. Each array was a 2560×2560 matrix consisting of 6,553,600 probes. The dataset included two samples, which were brain and reference, respectively. VTbrain and VTreference came from Virginia Tech, and each sample had 5 replicates.

3.1 Comparison of computation accuracy

Comparison was made with RMA and iterPLER through the calculation of gene expressions on 5 replicates for each sample. The correlation between the replicates under each algorithm was calculated. The higher the correlation, the lower the noise was. It can be seen from the table that the correlation between the replicates for each sample using Kseq model was much higher than that using the other two methods. Therefore, Kseq model can more significantly reduce the noises in original data than the conventional RMA and iterPLER on a real dataset.

Table 2: Correlation of gene expression

Sample	5 replicates	RMA	Iterplier	Kseq
VTbrain	(replicate4, replicate 5)	0.9911	0.9935	0.9980
	(replicate1, replicate 2)	0.9915	0.9927	0.9973
	(replicate1, replicate 3)	0.9906	0.9925	0.9978
	(replicate1, replicate 4)	0.9883	0.9909	0.9966
	(replicate1, replicate 5)	0.9887	0.9912	0.9968
	(replicate2, replicate 3)	0.9919	0.9935	0.9979
	(replicate2, replicate 4)	0.9899	0.9922	0.9975
	(replicate2, replicate 5)	0.9893	0.9914	0.9972
	(replicate3, replicate 4)	0.9879	0.9900	0.9970
	(replicate3, replicate 5)	0.9882	0.9902	0.9970
	(replicate4, replicate 5)	0.9870	0.9912	0.9969
VTreference	(replicate1, replicate 2)	0.9903	0.9919	0.9973
	(replicate1, replicate 3)	0.9901	0.9908	0.9979
	(replicate1, replicate 4)	0.9887	0.9905	0.9968
	(replicate1, replicate 5)	0.9897	0.9913	0.9976
	(replicate2, replicate 3)	0.9866	0.9862	0.9968
	(replicate2, replicate 4)	0.9873	0.9894	0.9957
	(replicate2, replicate 5)	0.9866	0.9876	0.9963
	(replicate3, replicate 4)	0.9857	0.9850	0.9969
	(replicate3, replicate 5)	0.9895	0.9916	0.9976
	(replicate4, replicate 5)	0.9865	0.9877	0.9969
Mean±standard deviation		0.9881±0.0028	0.9896±0.0034	0.9968±0.0011

To further test the computation accuracy, MAQC QRT-PCR results were incorporated and the differential genes were identified through t-test, which was confirmed (Cui and Churchill, 2003). The larger the AUC, the better the performance of the algorithm is, of which was confirmed (Sing et al, 2005). Table 2 shows the test results in which the column of sample represents the sample type, the column of 5 replicates is given the number of the found replicate, the correlation coefficients of RMA, iterPLER and Kseq respectively. From the result of Table 2, the performance of Kseq is higher than of RMA and iterPLER, and the last line in Table 2 respectively presents the mean values and standard variations of RMA, iterPLER and Kseq which also show that the performance of Kseq is the best.

The AUC values of RMA, iterPLER and Kseq are given in fig2, it can be seen that Kseq model was superior to PLIER based on AUC, while RMA had the highest AUC. Since RMA cannot calculate the expression of isomers, Kseq model was chosen for the analysis of alternative splicing for improving the accuracy of bioanalysis.

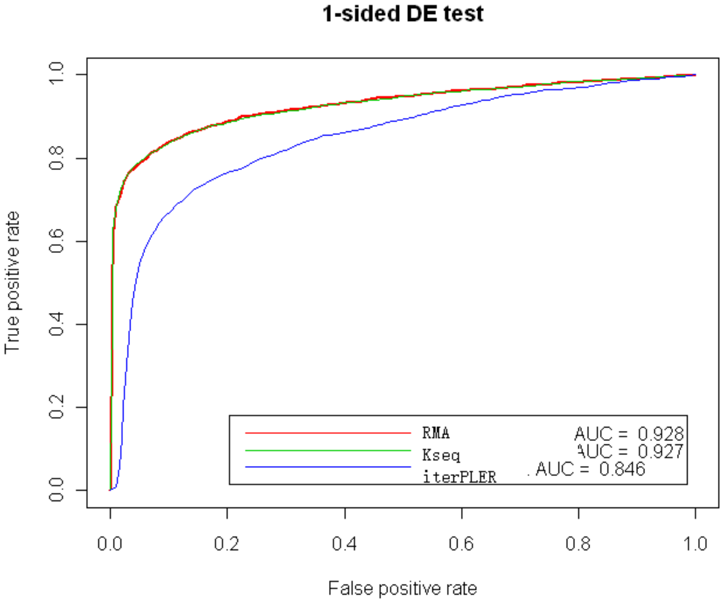

Fig 2: AUC values for the three way rule conductions

3.2 Comparison of computation speed

To compare the efficiency of different algorithms, Table 4 shows the computation time of the three methods on dataset VTbrain and VTreference with parallel computation. The experimental platform was the 64-bit Linux system with Intel Pentium 4 processor (3.4GHZ, 8G RAM). Because the computer had four cores, parallel computation with simultaneous use of 4 computers could be simulated with MapReduce on a single computer.

Table 3: Computation time of three different methods on different datasets (h)

	RMA	Iterrpler	Kseq
VTreference	5.4	5.4	3.5
VTbrain	7.9	8.1	6.2

It can be seen from Table 3 that Kseq had higher computation efficiency, because the mapping between genes, isomers and probes was fully utilized in the algorithm. The efficiency can be further improved by parallel computation using several computers simultaneously

3. Conclusions

This paper aims to solve the multi-source mapping of the gene isoform by using the relationship of the gene, the gene isoform and the gray value of the gene probe in Affymetrix extron chip. Firstly, the expression level of the splice isoform of the gene is calculated through using chi-squared distribution algorithm and the percentage of the gene which corresponds to the splice isoform is also computed. Secondly, the proposed model in this paper is verified by the real data of GSE13072, and the experiment results show that Kseq model can improve the computing accuracy and efficiency which supports the mechanism of alternative splicing. Finally, this model uses parallel computing to improve the computing methods and fully use multi-core processor, which makes the subsequent processing of large-scale data have better development prospects and provide the good reference such as differential analysis and cluster analysis.

Conflict of interest

The authors confirm that this article content has no conflicts of interest

Acknowledgment

This work is supported by the National natural fund project, China (No. 61402192), Jiangsu natural science fund project, China (No. 14KJB520006)

Reference

Caceres J. F., Kornblihtt A R. 2002. Alternative splicing: multiple control mechanisms and involvement in human disease [J], Trends in Genetics, 18; 186-193

Chen P, 2011. Comprehensive exon array data processing method for quantitative analysis of alternative spliced variants [J].Nucleic Acids Research, 39:e123.

Cui X., Churchill G., 2003, Statistical tests for differential expression in cDNA microarray experiments. Genome Biology, 4:210-235.

Kapur K. 2007, Exon arrays provide accurate assessments of gene expression [J], Genome Biology, 8(5); R82 Santa Clara. 2005. Affymetrix GeneChip exon array design [R]. www.affymetrix.com/support/technical/sample_data/exon_array_data.affx.

Karin Z. U. L., 2010, Analysis of Affymetrix Exon Arrays [R].

Sing T., 2005, ROCR: visualizing Classifier Performance in R [J]. Bioinformatics, 21(20):3940:3941.

Turro E., 2010. MMBGX: A method for estimating expression at the isoform level and detecting differential splicing using while-transcript Affymetrix arrays [J].Nucleic Acids Research. 38:e4.

Yin L., Liu Y.G. 2015, Biclustering of the Gene Expression Data by Coevolution Cuckoo Search. International Journal of Bioautomation, 19(2), 161-176.

Zhao Z., Liu X., Zhang L., 2011, A probabilistic model for the analysis of RNA-Seq data, CBME'2011, Wuhan, China.