Communications and Network, **2013, 5, 601-605**
http://dx.doi.org/10.4236/cn.2013.53B2108 Published Online September 2013 (http://www.scirp.org/journal/cn)

Threshold Selection Study on Fisher Discriminant Analysis Used in Exon Prediction for Unbalanced Data Sets

Yutao Ma[1], Yanbing Fang[2], Ping Liu[1], Jianfu Teng[3]
[1]School of Physics and Electrical Information Engineering, Ningxia University, Yinchuan, China
[2]School of Mathematics and Computer Science, Ningxia University, Yinchuan, China
[3]School of Electronic Information Engineering, Tianjin University, Tianjin, China
Email: yutao_ma@163.com, nxfangzi@nxu.edu.cn, liuping@nxu.edu.cn

Received May 2013

ABSTRACT

In gene prediction, the Fisher discriminant analysis (FDA) is used to separate protein coding region (exon) from non-coding regions (intron). Usually, the positive data set and the negative data set are of the same size if the number of the data is big enough. But for some situations the data are not sufficient or not equal, the threshold used in FDA may have important influence on prediction results. This paper presents a study on the selection of the threshold. The eigen value of each exon/intron sequence is computed using the Z-curve method with 69 variables. The experiments results suggest that the size and the standard deviation of the data sets and the threshold are the three key elements to be taken into consideration to improve the prediction results.

Keywords: Fisher Discriminant Analysis; Threshold Selection; Gene Prediction; Z-Curve; Size of Data Set

1. Introduction

Protein coding region and non-coding region of a DNA sequence are also called exon and intron respectively. Gene prediction of eukaryotes is still one of the most important research domains of bioinformatics for the prediction accuracy is needed to be improved [1]. Fisher Discriminant analysis/algorithm (FDA) is widely used in solving binary classification problems like fault classification [2], gene expression data classification [3], image categorization [4], integrating heterogeneous data sets [5] and DNA sequence classification [6,7], and it has attracted more and more attention. Kernel FDA (KFDA) may present out performance in both simulation time consume and classification precision than support vector machine (SVM) for it does not need to solve any quadratic problem [8,9]. For maximizing the uniformity of class-pair separabilities and class separability in kernel space simultaneously, a novel kernel FDA kernel parameters optimization criterion is presented [10]. A novel dimensionality reduction algorithm based on FDA is proposed for ranking applications such as visual search re-ranking [11].

Usually, the size of the training and test data sets are selected equal and the threshold is determined by making the false negative rate and the false positive rate equal [6]; but for some real situations, the threshold could not be obtained by the method because the sizes of the positive and negative data sets are unbalanced. In the latter situation, there are five possible thresholds for making the coding/non-coding decision. Using the exon and intron data sets downloaded from the website:
http://www.fruitfly.org/seq_tools/datasets/Human/, the FDA based experiments show that the size and the standard deviation of the data sets are the two key elements to be taken into consideration to improve the prediction results.

2. Data Sets and Methods

2.1. The Data Sets

The Data sets used in this paper were downloaded from the above net address. The Data were separated into 7 parts in the website, and each part consists of an exon sub-set and an intron sub-set. There are 301 DNA fragments at least and 448 at most in each sub-set as detailed in **Table 1**.

The difference values between the numbers of exons and introns in each part are all 66 from the **Table 1**.

Table 1. The details of each part of the data sets.

	Part0	Part1	Part2	Part3	Part4	Part5	Part6
Exon	367	448	447	389	399	402	391
Intron	301	382	381	323	333	336	325
Total	668	830	828	712	732	738	716

2.2. The Z Curve Methods

The Z curve method is a powerful tool to visualize and analyze DNA sequences. It is also one of the most widely used mapping methods which maps a DNA sequence into three digital sequences. According to [6], the prediction results of the Z curve method with 69 variables are almost the same with that of 169 variables ones. So, the Z curve method with 69 variables is applied in this paper. The 69 variables are composed with 9 Z curve parameters for frequencies of phase-specific mononucleotides, 12 Z curve parameters for frequencies of phase independent di-nucleotides and 48 Z curve parameters for frequencies of phase independent tri-nucleotides. These Z curve parameters are defined below.

Let A, T, C and G represent base adenine, thymine, cytosine and guanine respectively. The bases A, T, C, G are occurring in a DNA fragment at positions 1, 4, 7, …; 2, 5, 8, …; and 3, 6, 9, …, with frequencies a_1, t_1, c_1, g_1; a_2, t_2, c_2, g_2; a_3, t_3, c_3, g_3 respectively. Then by using the Z-transform defined by Equation (1), a fragment of DNA sequence is transformed into the 9 Z curve parameters for frequencies of phase-specific mononucleotides.

$$\begin{cases} x_i = (a_i + g_i) - (c_i + t_i) \\ y_i = (a_i + c_i) - (g_i + t_i) \\ z_i = (a_i + t_i) - (g_i + c_i) \end{cases} . \qquad (1)$$

$$x_i,\ y_i,\ z_i \in [-1, 1],\ i = 1, 2, 3$$

Let $p(XY)$ be the frequency of the di-nucleotides XY ($X, Y = A$, C, G, T), the 12 variables for frequencies of phase independent di-nucleotides are given by

$$\begin{cases} x_X = [p(XA) + p(XG)] - [p(XC) + P(XT)] \\ y_X = [p(XA) + p(XC)] - [p(XG) + P(XT)] \\ z_X = [p(XA) + p(XT)] - [p(XG) + P(XC)] \end{cases} . \qquad (2)$$

$$X = A, C, G, T$$

The last 48 Z curve parameters for frequencies of phase independent tri-nucleotides can be obtained using the similar notations by

$$\begin{cases} x_{XY} = [p(XYA) + p(XYG)] - [p(XYC) + P(XYT)] \\ y_{XY} = [p(XYA) + p(XYC)] - [p(XYG) + P(XYT)] \\ z_{XY} = [p(XYA) + p(XYT)] - [p(XYG) + P(XYC)] \end{cases} . \qquad (3)$$

$$X = A, C, G, T, Y = A, C, G, T$$

Next, the 69 variables of each exon/intron are $u_1^9, u_2^9, \cdots, u_9^9, u_1^{12}, u_2^{12}, \cdots, u_{12}^{12}, u_1^{48}, u_2^{48}, \cdots, u_{48}^{48}$. They are defined by Equations (4) to (8). These

$$\begin{aligned} &u_1^9 = x_1, u_2^9 = x_2, u_3^9 = x_3, \\ &u_4^9 = y_1, u_5^9 = y_2, u_6^9 = y_3, . \\ &u_7^9 = z_1, u_8^9 = z_2, u_9^9 = z_3 \end{aligned} \qquad (4)$$

$$\begin{aligned} &u_1^{12} = x_A, u_2^{12} = x_C, u_3^{12} = x_G, u_4^{12} = x_T, \\ &u_5^{12} = y_A, u_6^{12} = y_C, u_7^{12} = y_G, u_8^{12} = y_T, . \\ &u_9^{12} = z_A, u_{10}^{12} = z_C, u_{11}^{12} = z_G, u_{12}^{12} = z_T \end{aligned} \qquad (5)$$

$$\begin{aligned} &u_1^{48} = x_{AA}, u_2^{48} = x_{AC}, u_3^{48} = x_{AG}, u_4^{48} = x_{AT}, \\ &u_5^{48} = x_{CA}, u_6^{48} = x_{CC}, u_7^{48} = x_{CG}, u_8^{48} = x_{CT}, \\ &u_9^{48} = x_{GA}, u_{10}^{48} = x_{GC}, u_{11}^{48} = x_{GG}, u_{12}^{48} = x_{GT}, \\ &u_{13}^{48} = x_{TA}, u_{14}^{48} = x_{TC}, u_{15}^{48} = x_{TG}, u_{16}^{48} = x_{TT}, \end{aligned} . \qquad (6)$$

$$\begin{aligned} &u_{17}^{48} = y_{AA}, u_{18}^{48} = y_{AC}, u_{19}^{48} = y_{AG}, u_{20}^{48} = y_{AT}, \\ &u_{21}^{48} = y_{CA}, u_{22}^{48} = y_{CC}, u_{23}^{48} = y_{CG}, u_{24}^{48} = y_{CT}, \\ &u_{25}^{48} = y_{GA}, u_{26}^{48} = y_{GC}, u_{27}^{48} = y_{GG}, u_{28}^{48} = y_{GT}, \\ &u_{29}^{48} = y_{TA}, u_{30}^{48} = y_{TC}, u_{31}^{48} = y_{TG}, u_{32}^{48} = y_{TT}, \end{aligned} . \qquad (7)$$

$$\begin{aligned} &u_{33}^{48} = z_{AA}, u_{34}^{48} = z_{AC}, u_{35}^{48} = z_{AG}, u_{36}^{48} = z_{AT}, \\ &u_{37}^{48} = z_{CA}, u_{38}^{48} = z_{CC}, u_{39}^{48} = z_{CG}, u_{40}^{48} = z_{CT}, \\ &u_{41}^{48} = z_{GA}, u_{42}^{48} = z_{GC}, u_{43}^{48} = z_{GG}, u_{44}^{48} = z_{GT}, \\ &u_{45}^{48} = z_{TA}, u_{46}^{48} = z_{TC}, u_{47}^{48} = z_{TG}, u_{48}^{48} = z_{TT}. \end{aligned} . \qquad (8)$$

2.3. The Fisher Discriminant Analysis/Algorithm

Let $\mathbf{X} \in R^{N \times M}$ be the data matrix, where *N* and *M* are the number of data/samples and the dimension of each data respectively. The *M* is also the number of the best FDA coefficients, which are $u_1, u_2, \cdots, u_M$. In this paper the dimension of the data is 69, that is $M = 69$. Each exon/Intron is described by a point or row vector $\mathbf{x}$ in a 69-dimensional (69-D) space spanned by

$$\begin{aligned} \mathbf{x} &= (u_1^9, u_2^9, \cdots, u_9^9, u_1^{12}, u_2^{12}, \cdots, u_{12}^{12}, u_1^{48}, u_2^{48}, \cdots, u_{48}^{48})^T \\ &= (x_1, x_2, \cdots, x_{69})^T \end{aligned} , \qquad (9)$$

There are two groups of samples in the training data set used in FDA. One contains positive samples which belong to sample space $\mathbf{w_P}$, and another contains the negative samples which belong to sample space $\mathbf{w}_\mathrm{n}$. The positive samples are DNA fragments from real exons and the negative samples are from the real introns. All the sequences longer than 2000 bases were cut down to 2000 bases. The numbers of positive and negative samples are N_P and N_n respectively, and $N = N_P + N_n$. The mean vector of the positive/negative samples in input space is defined by

$$\overline{\mathbf{u}}_P = \frac{1}{N_P} \sum_{\mathbf{x} \in \mathbf{w_p}} \mathbf{x}, \ \overline{\mathbf{u}}_n = \frac{1}{N_n} \sum_{\mathbf{x} \in \mathbf{w_n}} \mathbf{x}, \ \overline{\mathbf{u}}_P \in R^M, \overline{\mathbf{u}}_n \in R^M . \qquad (10)$$

The best FDA coefficients consists a column vector $\mathbf{u}$. Using the divergence matrix $\mathbf{S}_W$, the vector $\mathbf{u}$ is defined as the best projecting direction (BPD) and is given by

$$\mathbf{u} = (u_1, u_2, \cdots, u_M)^T = (\mathbf{S}_W)^{-1} \left(\overline{\mathbf{u}}_P - \overline{\mathbf{u}}_n \right), \quad \mathbf{u} \in R^M , \qquad (11)$$

where "*T*" indicates the transpose of a matrix, and the divergence matrix $\mathbf{S}_W$ is defined as

$$\mathbf{S}_W = \sum_{j=p,n}^{2} \sum_{i \in N_j} \left(\mathbf{x}_i - \mathbf{u}_j\right)\left(\mathbf{x}_i - \mathbf{u}_j\right)^T . \tag{12}$$

Once the $\mathbf{u}$ and the threshold c are obtained, the discrimination of exon/intron for each DNA fragment in the test set is carried out by $\mathbf{u}^T \bullet \mathbf{x} - c > 0 / \mathbf{u}^T \bullet \mathbf{x} - c < 0$.

2.4. The Threshold Selection of FDA

For situations where numbers of positive and negative samples are not equal, in other words $N_p \neq N_n$, there are five different thresholds (c_1 , $\cdots$, c_5) to be selected, which are represented by Equations (13) to (17).

$$c_1 = \frac{\mathbf{u}^T\left(\overline{\mathbf{u}_p} + \overline{\mathbf{u}_n}\right)}{2}, \tag{13}$$

$$c_2 = \frac{\mathbf{u}^T\left(N_p\overline{\mathbf{u}_p} + N_n\overline{\mathbf{u}_n}\right)}{N}, \tag{14}$$

$$c_3 = \frac{\mathbf{u}^T\left(N_n\overline{\mathbf{u}_p} + N_p\overline{\mathbf{u}_n}\right)}{N}, \tag{15}$$

$$c_4 = \frac{\mathbf{u}^T\left(\tilde{\sigma}_p\overline{\mathbf{u}_p} + \tilde{\sigma}_n\overline{\mathbf{u}_n}\right)}{\tilde{\sigma}_p + \tilde{\sigma}_n}, \tag{16}$$

where $\tilde{\sigma}_p$ and $\tilde{\sigma}_n$ are the average variances of the positive and negative samples respectively.

$$c_5 = \frac{\mathbf{u}^T\left(\tilde{\sigma}_n\overline{\mathbf{u}_p} + \tilde{\sigma}_p\overline{\mathbf{u}_n}\right)}{\tilde{\sigma}_p + \tilde{\sigma}_n}. \tag{17}$$

For situation where $N_p = N_n$, the threshold is uniquely determined by letting the false negative rate (FNR) be equal to the false positive rate (FPR) [6,7]. The details will be given in the next section of this paper.

2.5. The Z Curve Mark of the DNA Sequence

The best projecting direction $\mathbf{u}$ can be interpreted as the weight vector corresponds to the 69 Z curve values of each test sample. The Z curve score (ZCS) of each DNA sequence is defined as $ZCM = \mathbf{u} \times \mathbf{x}$, where the "×" is the multiplication cross symbol, "$\mathbf{u}$"and "$\mathbf{x}$" are the best project direction and row vector of a DNA sequence. The ZCS is too small to be display directly, so a transform named remainder and multiple (RM) is carried out. The RM transform is defined as

$$\begin{aligned} R(ZCS) &= MOD\ (ZCS \times 1000, M) \\ M\ (ZCS) &= floor(ZCS \times 1000 / M) \end{aligned}, \tag{18}$$

where "MOD" and "floor" are two MATLAB function names. Function "mod(x, y)" returns the modulus of "x" after it is devided by "y", and function "floor(x)" rounds the elements of "x" to the nearest integers. In this work, the "M" in the RM transform is set to 3. Thus, the ZCS of a DNA sequence can be displayed in a two dimensional surface with the rounds ($M\ (ZCS)$) and the modulus ($R(ZCS)$) are the two coordinate axis.

3. Results and Discussion

3.1. The Prediction Accuracy Measures

To measure the prediction accuracy, the sensitivity and the specificity are applied. They are defined by Equations (19) and (20) respectively.

$$S_n = TP / (TP + FN) \times 100\% \tag{19}$$

$$S_p = TN / (TN + FP) \times 100\% \tag{20}$$

where TP/TN is the number of true exons/introns which were predicted as exons/introns, and the FP/FN is the number of true introns/exons which were predicted as exons/introns [12]. The prediction average accuracy is defined as $a = (S_n + S_p)$. The FPR and the FNR are defined by Equations (20) and (21) respectively.

$$FPR = 1 - S_p = FP / (TN + FP) \tag{21}$$

$$FNR = 1 - S_n = FN / (TP + FN) \tag{22}$$

Let $FPR = FNR$, we have $FP/(TN + FP) = FN/(TP + FN)$. For the situation N_p equals to N_n , the expressions $N_n = TN + FP$ and $N_p = TP + FN$ are satisfied. Then we have $FP = FN$, and the threshold can be expressed as

$$c_0 = \mathbf{u}^T\left(N_p\overline{\mathbf{u}_p} + N_n\overline{\mathbf{u}_n}\right) / N = \mathbf{u}^T\left(\overline{\mathbf{u}_p} + \overline{\mathbf{u}_n}\right) / 2 . \tag{23}$$

The Statistics of the ZCS of Data Sets.

To clear the relationship between the data sets and the prediction results, the statistics nature of the 7 data parts are presented in **Tables 2** and **3**. These statistics nature include mean and standard deviation of the Z curve value. **Figure 1** gives the ZCS scatter diagram of the Part0 to Part6 with the RM transform is applied to the ZCS of each sequence.

Table 2. The ZCS statistics nature of the exon sequences.

Data sets	Part0	Part1	Part2	Part3	Part4	Part5	Part6
mean	0.0053	0.0039	0.0043	0.0035	0.0043	0.0053	0.0043
SD[a]	0.0053	0.0043	0.0039	0.0043	0.0044	0.0045	0.0041

Table 3. The ZCS statistics nature of the intron sequences.

Data sets	Part0	Part1	Part2	Part3	Part4	Part5	Part6
mean	−0.0103	−0.0086	−0.0055	−0.0078	−0.0075	−0.0065	−0.0064
SD[a]	0.0041	0.0034	0.0028	0.0036	0.0034	0.0033	0.0035

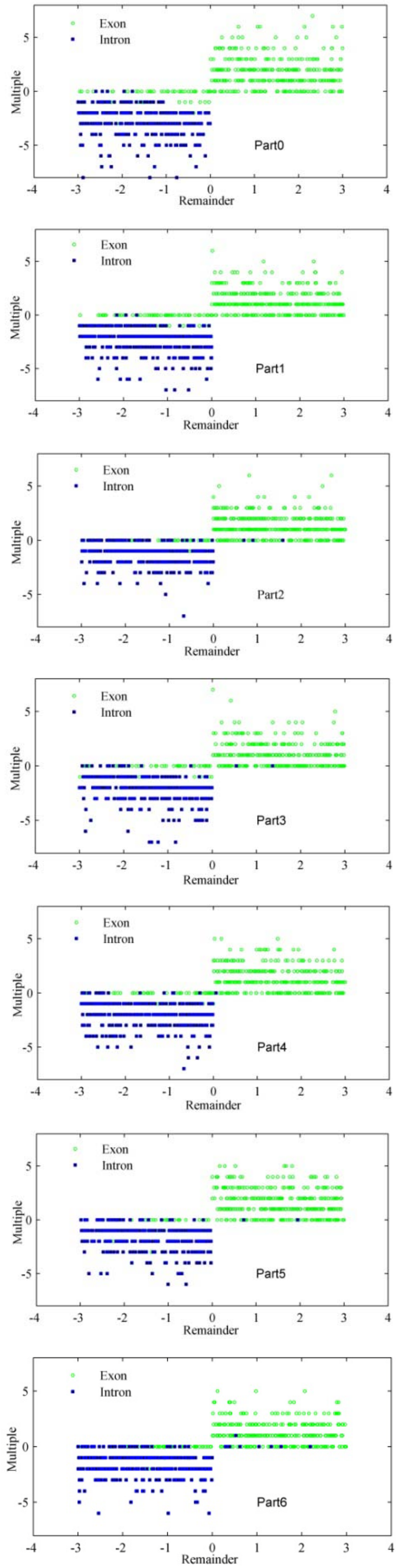

Figure 1. The ZCS scatter diagram of the data set.

3.2. Test with Different Training/Testing Sets

To study the relationship between size of the training samples and the prediction results, two experiments were carried out. In the first experiment, let part0 be the training data set (TRDS) and the rest 6 parts together (NPart0) be the test data set (TEDS) at first. Then let part1 be the training set and the rest 6 parts together (NPart1) be the test set, and so on. **Table 4** gives the prediction results. **Table 5** shows the mean prediction results obtained with the different training and testing sets (according to **Table 4**) for the five thresholds. It is clear that the fifth threshold (c_5) is the best choice in this experiment situation, c_3 is the next choice except c_5, and the c_1 is the third choice.

From the mean prediction accuracies listed in the **Table 5**, the third and the fifth thresholds present a little better prediction than the three others.

In the second experiment, let the part0 be the training set (1Parts) at first, and then let the part0 and the part1 (2Parts)be the training set, and so on. In this experiment, the testing set includes all the seven parts. **Table 6** gives the prediction result using the c_5 as the threshold. **Figure 2** shows the BPD obtained with training sets as mentioned like 1Parts, 2Parts, etc. The BPDs changed greatly as the size of training data set increased, which is confirmed by the prediction accuracy listed in **Table 6**. The mean and standard deviation of the Z curve value in **Tables 2** and **3** also show the same correlation.

Table 4. Prediction results with different training sets.

	TRDS	Part0	Part1	Part2	Part3	Part4	Part5	Part6
Threshold	TEDS	NPart0	NPart1	NPart2	NPart3	NPart4	NPart5	NPart6
	NS	4556	4394	4396	4512	4492	4486	4508
	S_n (%)	86.19	85.51	87.90	89.28	88.22	85.70	89.07
c_1	S_p (%)	97.89	97.94	97.95	95.93	97.82	97.89	96.51
	a (%)	91.99	91.68	92.85	92.38	92.94	91.75	92.61
	S_n (%)	83.00	82.38	85.60	87.49	85.88	83.08	86.87
c_2	S_p (%)	98.61	98.40	98.61	96.93	98.73	98.69	97.48
	a (%)	90.80	90.39	92.08	92.09	92.28	90.88	92.10
	S_n (%)	88.65	87.68	89.40	91.2	90.38	87.67	90.62
c_3	S_p (%)	96.74	97.13	97.19	94.35	96.89	96.75	95.12
	a (%)	92.55	92.29	93.15	92.34	93.46	92.07	92.54
	S_n (%)	82.27	81.88	83.01	87.61	84.98	80.58	87.07
c_4	S_p (%)	98.69	98.54	98.76	96.76	98.81	99.14	97.44
	a (%)	90.49	90.21	90.88	92.06	91.88	89.88	92.17
	S_n (%)	89.22	88.68	90.82	91.04	90.96	89.27	90.54
c_5	S_p (%)	96.25	96.81	96.33	94.42	96.61	95.61	95.24
	a (%)	92.54	92.59	93.33	92.31	93.57	92.19	92.57

Table 5. The mean prediction results over the different training sets.

Thresholds		c_1	c_2	c_3	c_4	c_5
mean	S_{nm} (%)	87.41	84.90	89.37	83.91	90.08
	S_{pm} (%)	97.42	98.21	96.31	98.31	95.90
	a_m (%)	92.31	91.52	92.63	91.08	92.73

Table 6. Prediction result with increased training set.

TRDS	Part0	Part0-Part1	Part0-Part2	Part0-Part3	Part0-Part4	Part0-Part5	Part0-Part6
S_n (%)	90.08	66.87	72.99	65.04	72.92	58.85	59.80
S_p (%)	96.57	75.98	85.43	74.02	91.85	71.96	74.73
a (%)	93.13	70.81	79.06	68.89	82.59	65.73	67.82

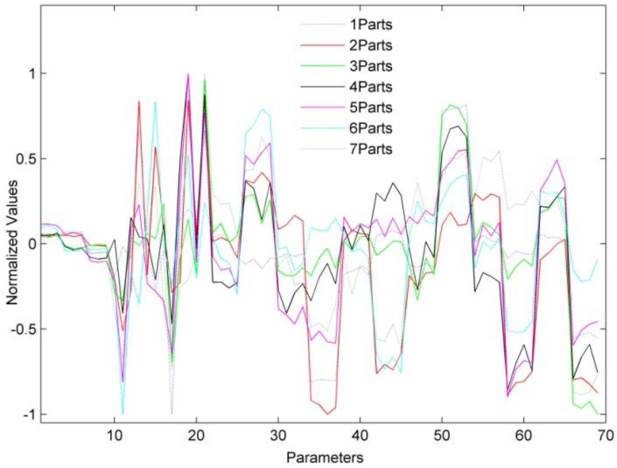

Figure 2. The best projecting directions obtained in the second experiment.

4. Concluding Remarks

In this work, some studies on the relationships between the prediction accuracy and some parameters alike have been carried out. The experiments based on FDA show that the mean, the standard deviation of the testing/training data sets and the threshold are the three key elements to improve the classification accuracy.

5. Acknowledgements

The work is partially supported by the Nature Science Foundation of China (11261041) and the Nature Foundation of Ningxia Hui Nationality Autonomous Region (NZ13xx), China.

REFERENCES

[1] J. P. Mena-Chalco, H. Carrer, Y. Zana, *et al.*, "Identification of Protein Coding Regions Using the Modified Gabor-Wavelet Transform," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 5, No. 2, 2008, pp. 198-206. http://dx.doi.org/10.1109/TCBB.2007.70259

[2] W. Y. Wang; X. B. Ma and R. Kang, "Fisher Discriminant Analysis for fault classification," *2012 IEEE Conference on Prognostics and System Health Management* (*PHM*), 2012, pp. 23-25.

[3] H. Huang, J. W. Li and J. M. Liu, "Gene Expression Data Classification Based on Improved Semi-Supervised Local Fisher Discriminant Analysis," *Expert Systems with Applications*, Vol. 39, No. 3, 2012, pp. 2314-2320.

[4] F. Yan, J. Kittler, K. Mikolajczyk and A. Tahir, "Non-Sparse Multiple Kernel Fisher Discriminant Analysis," *The Journal of Machine Learning Research*, Vol. 13, 2012, pp. 607-642.

[5] J. S. Hamid, C. M. T. Greenwood and J. Beyene, "Weighted Kernel Fisher Discriminant Analysis for Integrating Heterogeneous Data," *Computational Statistics & Data Analysis*, Vol. 56, No. 6, 2012, pp. 2031-2040.

[6] F. Gao and C.-T. Zhang, "Comparison of Various Algorithms for Recognizing Short Coding Sequences of Human Genes," *Bioinformatics*, Vol. 20, No. 5, 2004, pp. 673-681.http://dx.doi.org/10.1093/bioinformatics/btg467

[7] C.-T. Zhang and J. Wang, "Recognition of Protein Coding Genes in the Yeast Genome at Better Than 95% Accuracy Based on the Z Curve," *Nucleic Acids Research*, Vol. 28. No. 14, 2000, pp. 2804-2814. http://dx.doi.org/10.1093/nar/28.14.2804

[8] Y. Li and L. Jiao, "Target Recognition Based on Kernel Fisher Discriminant (In Chinese)," *Journal of Xidian University*, Vol. 30, No. 2, 2003, pp. 179-182.

[9] C. Zhao, W. Chen and C. Guo. "Research and Analysis of Methods for Multiclass Support Vector Machines (In Chinese)," *CAAI Transactions on Intelligent Systems*, Vol. 2, No. 2, 2007, pp. 11-17.

[10] J. Liu, F. Zhao and Y. Liu, "Learning Kernel Parameters for Kernel Fisher Discriminant Analysis," *Pattern Recognition Letters*, Vol. 34, No. 9, 1 2013, pp. 1026-1031.

[11] Z. Ji, P. G. Jing, T. S. Yu, Y. T. Su and C. S. Liu, "Ranking Fisher Discriminant Analysis," *Neurocomputing*, 2013. http://www.sciencedirect.com/science/article/pii/S0925231213002877)

[12] M. Burset and R. Guigo, "Evaluation of Gene Structure Prediction Programs," *Genomics*, Vol. 34, 1996, pp. 353-367. http://dx.doi.org/10.1006/geno.1996.0298