RESEARCH ARTICLE

# Application of SVM classifiers in information design: case study of a new method of classifying exons and introns based on 10-dimensional vectors

Chuan Yang[1], Ronghao Tang[2], Huimin Feng[3, *]

[1]Sanjiang University, Nanjing, Jiangsu, China. [2]Jiangsu Provincial Academy of Social Sciences, Nanjing, Jiangsu, China. [3]Nanjing Medical University, Nanjing, Jiangsu, China

**Received:** September 7, 2023; **accepted:** October 10, 2023.

**Support vector machine (SVM) is a supervised machine learning algorithm based on statistical learning frameworks. This relatively simple algorithm can be used for classification and/or regression tasks in a wide range of fields including facial, speech, handwriting recognitions, image filtering, video caption extraction, image classification and retrieval. This study focused on the information design of classifying exons and introns by using an SVM classifier developed through scikit-learn (version 0.23.2) and pandas (version 1.0.5). In the initial stage, relevant data regarding information behaviors and signal-to-noise ratio (SNR) were collected, and a maximum margin hyperplane (MMH) was built in a vector space over the real numbers. The classifiers were then designed in 10-dimensional vectors by using a kernel trick which included (1) setting an SNR threshold by using the Parzen window estimation (PWE) method, (2) providing 10-dimensional vectors to characterize each DNA sequence based on the SNR threshold and Z-curve features, (3) distinguishing exons from introns by using Fisher and SVM classifiers. This newly developed method for classifying exons and introns on the genomic DNA sequences of *C. elegans* is currently available online at www.utoledo.edu/med/depts/bioinfo/database.html. The classification results indicated that this new method could achieve the accuracies of 94.4%, 89.0%, and 79.6% on average in long, middle, and short DNA sequences, respectively, if combining 10-dimensional feature vectors with an SVM classifier, which were superior to other traditional methods.**

**Keywords:** information design; SVM; exons; introns; vector.

[*]**Corresponding author:** Huimin Feng, Nanjing Medical University, Nanjing 210029, Jiangsu, China. Email: fenghuimin_1989@126.com.

## Introduction

A DNA sequence is a long sequence consisting of four nucleotides of adenine (A), guanine (G), thymine (T), and cytosine (C). Recent rapid expansion of genomic sequences has made it much more complicated than ever before to study how to distinguish exons from introns accurately. In the last 2 decades, several computational methods had been proposed to find protein-coding regions, in which some algorithms were based on signal processing. Tiwari *et al*. and Anastassiou applied Fourier spectrum analysis to recognize protein-coding regions based on the 3-base periodicity behavior, by which the Fourier power spectrum of an exonic sequence with length N had a prominent peak at frequency N/3 while an intronic sequence had no such feature [1, 2]. Fickett and Yin *et al*. proved that the 3-base periodicity behavior in an exonic sequence was partly caused by the unbalanced nucleotide distributions in the three coding positions [3, 4]. Furthermore, on the strength of 3-base periodicity behavior, Lorenzo-

Ginori *et al*. used digital filters for predicting protein-coding genes and showed that digital filtering could clearly identify coding regions at a very low computational cost [5]. In addition, by employing entropy measure, Román-Roldán proposed a new complexity measure based on the entropic segmentation of DNA sequences into compositionally homogeneous domains [6]. By reviewing these signal processing approaches, the method of using the 3-base periodicity behavior with fixed value of signal-to-noise ratio (SNR) threshold is a simple and practical way in distinguishing exons from introns, but the fixed value of SNR threshold is not suitable for diverse DNA sequences.

Differing from signal processing approaches, another type of method is based on machine learning, which requires the use of Support Vector Machine (SVM). SVM belongs to and is believed to be an exception to Tikhonov regularization. SVM classifiers map vectors into a higher dimensional space where a maximum margin hyperplane (MMH) will be built. SVM is also known as a maximal margin classifier for its effectiveness in empirical error minimization and geometric margin maximization. In statistics, an expectation-maximization (EM) algorithm is used to discover the maximum likelihood estimates of parameters in a probabilistic model dependent on unobserved latent variables. The EM algorithm is often applied to data clustering in machine learning and computer vision fields. Thanks to their outstanding performance, SVM classifiers have opened a new door for information classification. Law *et al*. treated the Z-curve feature as a feature vector in the first place and then utilized Fisher classifier to make a distinction between exons and introns [7].

Although the above machine learning and signal processing techniques are from different perspectives and both prove effective in the field of gene recognition, fewer researchers have considered integrating both techniques together. This study focused on integrating the strengths of both techniques and generating a synthetic feature vector for improved classification accuracy at different DNA length levels.

## Materials and methods

### The strategy of the study

The strategy of this study was to extract some biological features from DNA sequences followed by the construction of a proper classifier to recognize coding regions from non-coding regions. In particular, the Parzen Window Estimation (PWE) method was employed first to set an SNR threshold, and then, 10-dimensional vectors were provided to characterize each DNA sequence based on the SNR threshold and Z-curve features. Finally, the Fisher classifier and SVM classifier were used to distinguish exons from introns.

### Data resource

The genomic DNA sequences of *C. elegans* were obtained from The Bioinformatics Program, The University of Toledo, Toledo, OH, USA. The sequences have already been separated into two sets and each set is marked with a label exon or intron. The first 2,000 sequences from both sets were employed for this study. In order to identify a method that performed optimally in distinguishing exons from introns at varied length levels, the sample DNA sequence data were divided into 3 groups including long sequences (> 500 bp), middle sequences (200 – 500 bp), and short sequences (< 200 bp) with each group containing additional 2 subgroups of exons and introns, which made a total of six groups for analysis. The frequency distributions among the 6 groups were listed in Table 1.

**Table 1.** The frequency distributions among the 6 groups of *C. elegans* DNA sequence data.

| Type<br>Group | Long | Middle | Short | Overall |
|---|---|---|---|---|
| Exon | 149 | 627 | 1,224 | 2,000 |
| Intron | 871 | 262 | 867 | 2,000 |
| Total | 1,020 | 889 | 2,091 | 4,000 |

**DNA sequence feature extraction**

**(1) 3-base periodicity behavior and determination of SNR threshold**

The DNA sequence $f(n)(n=1,2,...,N)$ is consisted of A, T, G, and C, and can be represented by 4 indicator sequences as $u_\alpha(n)\,(n=1,2,...,N)$, $\alpha \in \{A,T,C,G\}$.

$$u_\alpha(n)=\begin{cases}1 & \alpha\, appears\, at\, location\, n,\\ 0 & otherwise,\end{cases}$$

The Fourier power spectrum of a DNA sequence $PS(k), k=1,2,...,N$ is the sum of the power spectrum of its four binary indicator sequences [8], which is defined as follows:

$$PS(k)=\sum_\alpha P_\alpha\left(\mathrm{k}\right)=\sum_\alpha\sum_{n=1}^{N}u_\alpha(n)e^{-2\pi kn/N} \quad (1)$$

$$k=1,2...,N,\ \alpha\in\{A,T,C,G\}.$$

It is a well-known fact that an exonic sequence at length *N* has a 3-base periodicity behavior, which means a prominent peak can be found in its Fourier power spectrum at frequency *N*/3, while an intronic sequence has no such feature. The ratio of the Fourier power spectrum at the frequency *N*/3, *PS*(*N*/3), to the average Fourier power spectrum over all the frequencies, *E*, denoted by:

$$R=PS(N/3)/E. \quad (2)$$

The equation (2) is considered as the SNR, which has a higher value in exons rather than introns. According to Yin's work, 2 was assigned to the SNR threshold, meaning that a DNA sequence was treated as an exon when its SNR was larger than 2 [4]. Obviously, this was a brief way of classifying exons and introns but assigning 2 to the SNR threshold did not take into account the discrepancy among different organisms, which would lead to a higher error probability for the classification in some organisms. Thus, the threshold should be set by using some accurate tools. The Parzen Window Estimation (PWE) method was utilized in this study to calculate the SNR threshold which could minimize the error probability. PWE is a traditional and effective non-parametric estimation method [9]. Given an instance of the random sample, $x_i\,(i=1,2,3...,n)$, Parzen windowing was adopted to estimate a probability distribution function (PDF), *P*(*x*), from samples derived. To estimate the value of the PDF at point *x*, it was essential to place window functions at each observation $x_i\,(i=1,2,3...,n)$ to depict their contributions to this point. Then, the PDF value *P*(*x*) was assigned as the total sum of contributions acquired from these observations as follows:

$$P(x)=1/n\sum_{i=1}^{n}1/h_n^d K((x-x_i)/h_n). \quad (3)$$

where *K*(*x*) was the window function in d-dimensional space, and $h_n$ was a smoothing parameter called bandwidth. If a SNR threshold $\overline{R}$ was determined, the classification error probability $e$ could be described as:

$$e=\int_0^{\overline{R}}f_1(x)dx+\int_{\overline{R}}^{+\infty}f_2(x)dx. \quad (4)$$

Based on the Bayes decision theory, this study would get the minimized classification error probability if selecting the intersection point of the two distribution curves, $R_0$, as the SNR threshold.

**(2) 10-D vector**

Z-curve feature is a 9-dimensional feature for analyzing DNA sequences and recognizing coding sequences in the human genome [10]. Let the frequencies of bases A, C, G, and T at the positions 0, 3, 6; 1, 4, 7; and 2, 5, 8, respectively, be $A_i$, $C_i$, $G_i$, and $T_i$, ($i$ = 0, 1, 2). Z-curve feature was then defined as follows:

$$F_{3i} = (A_i + G_i) - (C_i + T_i) \qquad i = 0,1,2.$$

$$F_{3i+1} = (A_i + C_i) - (G_i + T_i) \qquad i = 0,1,2. \qquad (5)$$

$$F_{3i+2} = (A_i + T_i) - (C_i + G_i) \qquad i = 0,1,2.$$

The biological interpretation of the above three measures was as follows. Component $F_{3i}$ was the distribution of purine bases (A or G) and pyrimidine bases (C or T) along the sequence. Component $F_{3i+1}$ was the distribution of the bases in amino form (A or C) and Keto form (G or T). Component $F_{3i+2}$ was the distribution of the bases of the weak hydrogen bond (A or T) and strong hydrogen bond (G or C) [11]. Although the 3-base periodicity behavior was one of the most significant properties in coding region recognition, such periodicity was not obviously observed in short exons which were very common in human genome sequences [12]. This study also verified the fact that it was very easy to detect the 3-base periodicity behavior if an exon was longer than 500 bp, but such periodicity was hardly detectable when an exon was shorter than 200 bp. Meanwhile, the machine learning technique based on Z-curve feature vectors could classify short DNA sequences much more effectively than depending on 3-base periodicity behavior, considering its biological properties [13]. So, to find a way that could precisely distinguish coding regions from non-coding regions without the limitation of length levels, the spectral property was integrated with the properties of physics and chemistry to exert their advantages at varied length levels, respectively. This study combined the 3-base periodicity property with the Z-curve feature to represent DNA sequences in forms of 10-dimensional vectors, which were defined as:

$$V = [F_1, F_2 ... F_9, P]^T. \qquad (6)$$

where $F_i (i = 0,1,2...8)$ was gained from Z-curve feature and $P$ was a new measurement of 3-base periodicity relying on SNR threshold, which could be denoted as $R / R_0$, where $R$ was the SNR of a DNA fragment, and $R_0$ denoted the SNR threshold derived from the PWE method.

### Design of the classifier

#### (1) Fisher classifier

Fisher classifier is a linear classifier. It searches for directional vector *W* in the data that can maximize the Fisher discriminant function [14].

$$J_F(W) = \frac{W^T S_b W}{W^T S_\omega W}, \qquad (7)$$

where $S_b$ was "between class scatter" while $S_\omega$ was "within class scatter", and both of them were defined in formulas (8) and (9) below.

$$S_b = (m_A - m_B)(m_A - m_B)^T. \qquad (8)$$

$$S_\omega = \sum_{\alpha \in \{A,B\}} (x - m_\alpha)(x - m_\alpha)^T. \qquad (9)$$

where $m_A$ was the mean of the training samples in class A, while $m_B$ was the mean of the training samples in class B. The method of Lagrange multipliers was employed to get the optimal directions vector $W^*$ from [15].

$$W^* = S_\omega^{-1}(m_A - m_B) . \qquad (10)$$

The major steps in adopting Fisher classifier were list below:

1. Computing the means of the training samples, $m_A$, $m_B$, in both two classes.
2. After acquiring $S_\omega$ (=within class scatter), calculating the optimal directions vector $W^*$ according to formula (10).
3. Mapping the training sample spaces into 1-D projection space on the direction of $W^*$ and computing the means of the two classes, $\bar{m}_A$, $\bar{m}_B$, in the 1-D projection space.

4. Computing the threshold $y_0$ in the 1-D projection space by using formula (11).

$$y_0 = \frac{N_A \ \bar{m}_A + N_B \ \bar{m}_B}{N_A + N_B}. \tag{11}$$

where, $N_A$ and $N_B$ were the sample capacity in classes A and B.

5. For every testing sample *X*, firstly calculating the projection alue *y* on the optimal direction $W^*$ presented in formula (10). Then, as per a rule depicted in formula (12), the class that every sample belonged to could be acquired.

$$\begin{cases} y \ge y_0 \Rightarrow X \in A \\ y < y_0 \Rightarrow X \in B \end{cases}. \tag{12}$$

**(2) SVM classifiers based on kernel trick**

SVM is a powerful method used for classification, regression, and other tasks. Given some training data ***D***, a set of ***n*** points of the form:

$$D = \{(X_i, y_i) \mid X_i \in R^P, y_i \in \{-1,1\}\}_{i=1}^{n}. \tag{13}$$

where $y_i$ was either 1 or -1, indicating the class to which the point $X_i$ belonged. The goal of applying a linear classifier was to find a hyperplane that divided the points having $y_i = 1$ from those having $y_i = -1$. If the training data were linearly separable, any hyperplane could be written as:

$$W \cdot X - b = 0. \tag{14}$$

In contrast to a linear classifier, the core idea of SVM method was to look for an optimal direction vector $W$ that could maximize the distance between two hyperplanes, and thus no points were between them. These hyperplanes could be expressed as equations 15 and 16.

$$W \cdot X - b = 1, \tag{15}$$

$$W \cdot X - b = -1. \tag{16}$$

Based on geometry, it could be inferred that the maximum distance between these two hyperplanes was $\frac{2}{\|W\|}$, which equaled to seeking for the minimum $\frac{1}{2}\|W\|^2$. For each $i \ (i = 1, 2, \ldots, n)$, the following constraint was added:

$$y_i (W \cdot X_i - b) \ge 1. \tag{17}$$

So, the above problem could be transformed into a quadratic programming optimization problem, and a common strategy to solve it was the method of Lagrange multipliers.

This study applied the kernel trick to maximize the gap between hyperplanes, which was one of the most attractive SVM properties. This powerful tool made the design of a linear classifier into high dimensional space. The kernel trick in this study was defined as linear kernel function and the LIBSVM, a popular machine learning library, was also applied to support classification [16].

## Results and discussion

The SNR distribution of exons and introns in *C. elegans*, $f_1(x)$, $f_2(x)$, which were calculated by using the PWE method was shown in Figure 1, where standard Gaussian PDF was used as window function and the bandwidth was 1. The results showed that the short sequences occupied a large proportion in the genomic DNA sequences of *C. elegans*, especially for exons. The overall classification accuracy was largely subjected to the classification accuracy of short sequences. The classification of *C. elegans* based on the threshold fixed as 2 and the PWE method was shown in Table 2.

**Table 2.** The classification of *C. elegans* based on the threshold fixed as 2 and the PWE method.

| | Long | | Middle | | Short | | Overall | |
|---|---|---|---|---|---|---|---|---|
| | $R_0 = 2$ | PWE method | $R_0 = 2$ | PWE method | $R_0 = 2$ | PWE method | $R_0 = 2$ | PWE method |
| Th | 2.000 | 1.820 | 2.000 | 1.500 | 2.000 | 1.164 | 2.000 | 1.259 |
| CA | 0.919 | 0.927 | 0.747 | 0.820 | 0.592 | 0.716 | 0.715 | 0.773 |
| SN | 0.913 | 0.926 | 0.669 | 0.789 | 0.324 | 0.641 | 0.476 | 0.613 |
| SP | 0.925 | 0.928 | 0.825 | 0.851 | 0.860 | 0.791 | 0.954 | 0.933 |

**Notes:** Th: threshold. CA: classification accuracy. SN: sensitivity. SP: specificity.

**Table 3.** The classification of *C. elegans* based on SVM and Fisher classifiers. The ratio of training sample capacity to the whole capacity was fixed as 0.3.

| | Long | | Middle | | Short | | Overall | |
|---|---|---|---|---|---|---|---|---|
| | Fisher | SVM | Fisher | SVM | Fisher | SVM | Fisher | SVM |
| CA | 0.602 | 0.920 | 0.852 | 0.833 | 0.749 | 0.739 | 0.714 | 0.770 |
| SN | 0.653 | 0.883 | 0.884 | 0.964 | 0.736 | 0.791 | 0.894 | 0.876 |
| SP | 0.550 | 0.957 | 0.820 | 0.702 | 0.762 | 0.687 | 0.534 | 0.664 |

**Notes:** CA: classification accuracy. SN: sensitivity. SP: specificity.

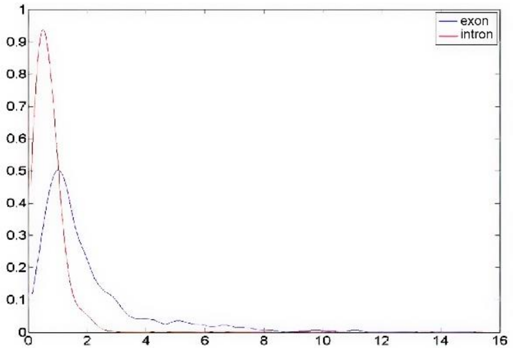

**Figure 1.** The signal-to-noise distribution of exons and introns in *C. elegans*. The minimized classification error probability could be obtained if selecting the intersection point of the 2 distribution curves as SNR threshold.

As indicated in Table 2, no matter how long the sequence was, the value of CA and SN became higher when PWE was used to calculate SNR rather than just fixing it as 2, which verified the worth of PWE in classification. Therefore, PWE method was applied to compute the SNR threshold when utilizing the 3-base periodicity behavior to process the classification. The results listed in Table 2 also manifested that the sharp increment of classification accuracy (typically for exons) came with the increase of DNA length level. Concretely, with the SNR determined by the PWE method, only 64.1% of 1,224 short exonic sequences could be detected with the 3-base periodicity property. The 3-base periodicity behavior expressed more evidently when exonic sequences increased in length, which significantly impacted classification accuracy. Moreover, when exonic sequences were longer than 500 bp, 92.6% exons held the 3-base periodicity behavior, which enabled the classification accuracy to reach 92.7%. However, this method could hardly find the 3-base periodicity behavior in an intronic sequence no matter how long the sequence was. These observations revealed that the 3-base periodicity behavior was not a universal property in exons, and the PWE method could only apply this feature to make the classification if the data set was mainly composed with long DNA sequences. Therefore, in order to improve overall classification accuracy, especially for short sequences, the machine learning technique should offer further support for this study. For the purpose to find a classifier that could express better performances when classifying short DNA sequences, Z-curve feature vectors were employed to calculate classification accuracy (CA), sensitivity (SN), and specificity (SP) values relying on Fisher classifier and SVM classifier. The training data were randomly selected and the

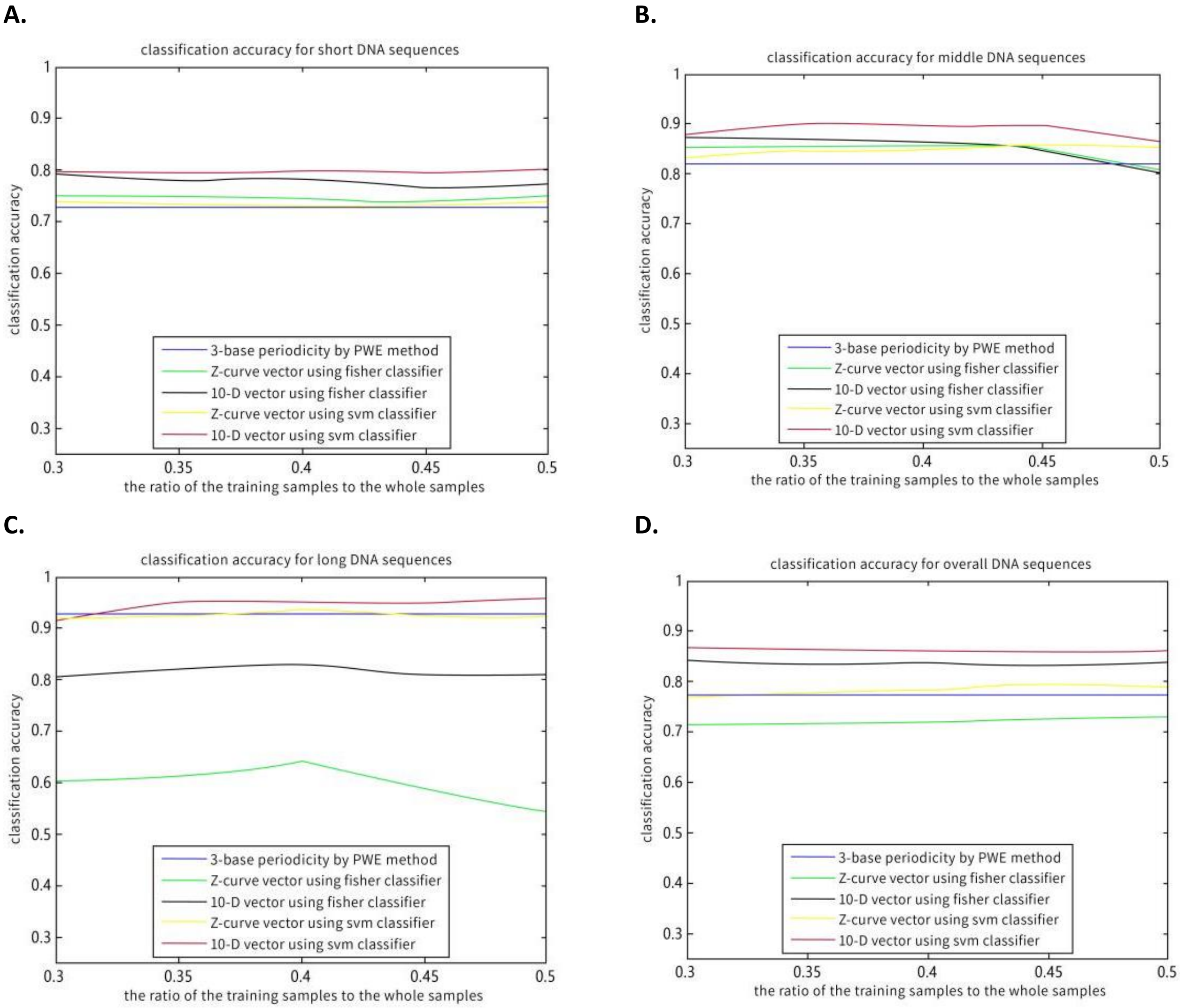

**Figure 2.** Classification accuracy for short sequences (A), middle sequences (B), long sequences (C), and overall sequences (D).

ratio of training sample capacity to the whole capacity was fixed as 0.3. Table 3 demonstrated the average classification accuracy of 10 independent experiments. By comparing the results of Table 3 with that of Table 2, it suggested that, with the notable increment of SN on short sequences, the approach of Z-curve features combined with Fisher and SVM classifiers raised classification accuracy from 71.6% to 74.9% and 73.9%, respectively. However, the classification accuracy of long sequences had declined from 92.7% to 60.2% and 92.0%, respectively due to the rapid decrease of SP values, which indicated that the machine learning technique could promote the strength of classifying short DNA sequences, despite difficulty in classifying long DNA sequences to some extent. Such a character in Z-curve feature vectors had just made up the disadvantages in PWE method for the classification of short sequences. Therefore, the 10-D vector was generated by combining the Z-curve features with the 3-base periodicity property, with a view to utilizing the strengths of both techniques. The fraction of sequences which were randomly selected as training samples were from 0.3 to 0.5. The average classification accuracy using varied collocations of different types of feature vectors

**Table 4.** The classification of *C. elegans* using five different approaches.

| Classification Accuracy for short sequences | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0.3 | | | 0.35 | | | 0.40 | | | 0.45 | | | 0.50 | | | Average | | |
| | AC | SN | SP | AC | SN | SP | AC | SN | SP | AC | SN | SP | AC | SN | SP | AC | SN | SP |
| PWE | 0.716 | 0.641 | 0.791 | 0.716 | 0.641 | 0.791 | 0.716 | 0.641 | 0.791 | 0.716 | 0.641 | 0.791 | 0.716 | 0.641 | 0.791 | 0.716 | 0.641 | 0.791 |
| Z+F | 0.749 | 0.736 | 0.762 | 0.749 | 0.720 | 0.778 | 0.744 | 0.707 | 0.781 | 0.737 | 0.679 | 0.795 | 0.751 | 0.687 | 0.815 | 0.746 | 0.706 | 0.786 |
| 10D+F | 0.792 | 0.723 | 0.861 | 0.781 | 0.706 | 0.856 | 0.784 | 0.699 | 0.869 | 0.765 | 0.664 | 0.866 | 0.772 | 0.673 | 0.871 | 0.778 | 0.693 | 0.865 |
| Z+SVM | 0.739 | 0.791 | 0.687 | 0.732 | 0.786 | 0.678 | 0.729 | 0.773 | 0.685 | 0.735 | 0.759 | 0.710 | 0.738 | 0.761 | 0.714 | 0.734 | 0.774 | 0.694 |
| 10D+SVM | 0.795 | 0.824 | 0.766 | 0.794 | 0.810 | 0.778 | 0.797 | 0.801 | 0.792 | 0.790 | 0.776 | 0.805 | 0.801 | 0.784 | 0.818 | 0.795 | 0.799 | 0.792 |
| Classification Accuracy for middle sequences | | | | | | | | | | | | | | | | | | |
| | 0.3 | | | 0.35 | | | 0.40 | | | 0.45 | | | 0.50 | | | Average | | |
| | AC | SN | SP | AC | SN | SP | AC | SN | SP | AC | SN | SP | AC | SN | SP | AC | SN | SP |
| PWE | 0.819 | 0.789 | 0.851 | 0.819 | 0.789 | 0.851 | 0.819 | 0.789 | 0.851 | 0.819 | 0.789 | 0.851 | 0.819 | 0.789 | 0.851 | 0.819 | 0.789 | 0.851 |
| Z+F | 0.852 | 0.884 | 0.820 | 0.854 | 0.877 | 0.832 | 0.856 | 0.856 | 0.855 | 0.855 | 0.840 | 0.869 | 0.809 | 0.764 | 0.854 | 0.845 | 0.844 | 0.846 |
| 10D+F | 0.871 | 0.900 | 0.843 | 0.872 | 0.885 | 0.859 | 0.865 | 0.895 | 0.871 | 0.853 | 0.829 | 0.875 | 0.802 | 0.760 | 0.843 | 0.852 | 0.846 | 0.858 |
| Z+SVM | 0.833 | 0.964 | 0.702 | 0.846 | 0.961 | 0.731 | 0.842 | 0.950 | 0.735 | 0.859 | 0.930 | 0.778 | 0.854 | 0.923 | 0.784 | 0.847 | 0.947 | 0.746 |
| 10D+SVM | 0.884 | 0.968 | 0.800 | 0.908 | 0.951 | 0.865 | 0.897 | 0.934 | 0.860 | 0.898 | 0.927 | 0.868 | 0.863 | 0.927 | 0.799 | 0.890 | 0.941 | 0.838 |
| Classification Accuracy for long sequences | | | | | | | | | | | | | | | | | | |
| | 0.3 | | | 0.35 | | | 0.40 | | | 0.45 | | | 0.50 | | | Average | | |
| | AC | SN | SP | AC | SN | SP | AC | SN | SP | AC | SN | SP | AC | SN | SP | AC | SN | SP |
| PWE | 0.927 | 0.926 | 0.928 | 0.927 | 0.926 | 0.928 | 0.927 | 0.926 | 0.928 | 0.927 | 0.926 | 0.928 | 0.927 | 0.926 | 0.928 | 0.927 | 0.926 | 0.928 |
| Z+F | 0.602 | 0.653 | 0.550 | 0.612 | 0.665 | 0.558 | 0.640 | 0.690 | 0.590 | 0.588 | 0.655 | 0.520 | 0.542 | 0.613 | 0.471 | 0.597 | 0.655 | 0.538 |
| 10D+F | 0.805 | 0.835 | 0.775 | 0.822 | 0.850 | 0.793 | 0.828 | 0.854 | 0.823 | 0.807 | 0.838 | 0.777 | 0.808 | 0.836 | 0.779 | 0.814 | 0.842 | 0.785 |
| Z+SVM | 0.920 | 0.883 | 0.957 | 0.922 | 0.885 | 0.958 | 0.935 | 0.898 | 0.971 | 0.925 | 0.890 | 0.960 | 0.923 | 0.888 | 0.959 | 0.925 | 0.889 | 0.961 |
| 10D+SVM | 0.913 | 0.878 | 0.948 | 0.950 | 0.918 | 0.982 | 0.948 | 0.911 | 0.985 | 0.950 | 0.917 | 0.983 | 0.957 | 0.927 | 0.986 | 0.944 | 0.910 | 0.977 |
| Classification Accuracy for overall sequences | | | | | | | | | | | | | | | | | | |
| | 0.3 | | | 0.35 | | | 0.40 | | | 0.45 | | | 0.50 | | | Average | | |
| | AC | SN | SP | AC | SN | SP | AC | SN | SP | AC | SN | SP | AC | SN | SP | AC | SN | SP |
| PWE | 0.773 | 0.613 | 0.933 | 0.773 | 0.613 | 0.933 | 0.773 | 0.613 | 0.933 | 0.773 | 0.613 | 0.933 | 0.773 | 0.613 | 0.933 | 0.773 | 0.613 | 0.933 |
| Z+F | 0.714 | 0.894 | 0.534 | 0.717 | 0.899 | 0.535 | 0.715 | 0.881 | 0.549 | 0.725 | 0.890 | 0.559 | 0.730 | 0.917 | 0.543 | 0.720 | 0.896 | 0.544 |
| 10-D+F | 0.841 | 0.814 | 0.869 | 0.833 | 0.814 | 0.852 | 0.840 | 0.863 | 0.817 | 0.831 | 0.867 | 0.795 | 0.837 | 0.889 | 0.785 | 0.836 | 0.849 | 0.823 |
| Z+SVM | 0.770 | 0.876 | 0.664 | 0.779 | 0.869 | 0.690 | 0.783 | 0.848 | 0.717 | 0.793 | 0.831 | 0.755 | 0.788 | 0.828 | 0.748 | 0.783 | 0.850 | 0.715 |
| 10D+SVM | 0.867 | 0.856 | 0.878 | 0.862 | 0.838 | 0.885 | 0.860 | 0.818 | 0.902 | 0.861 | 0.824 | 0.898 | 0.862 | 0.826 | 0.897 | 0.862 | 0.832 | 0.892 |

**Notes:** PWE: Parzen Window Estimation. Z: Z-curve feature vector. F: Fisher classifier. 10-D: 10-D vector. SVM: SVM classifier.

and classifiers were demonstrated in Figure 2 with the detailed information listed in Table 4, in which the values of SN and SP were also added. The results indicated that when a classifier was selected, classification accuracy could be increased if the Z-curve feature vector or 3-base periodicity property was replaced with 10-D vector at different DNA sequence length levels. Classification accuracy increased to 94.4%, 89.0%, 79.6% on average if 10-D feature vector was combined with SVM classifier in long, middle and short sequences, respectively. The overall classification accuracy was up to 86.2% on average if 10-D vector was integrated with SVM classifier, which was higher than that of other collocations. In addition, the results found that the values of SP and SN were both beyond 80% (except for short sequences, which were close to 80%) if SVM classifier combined with 10-D vector, which suggested that this combination mode could be effective in both classification methods.

Classifying a DNA sequence into 2 categories (exon and intron) is a tough task. In this study, two commonly used properties, 3-base periodicity property and Z-curve were combined to generate a synthetic 10-D vector for classification. Thanks to a fusion of the spectral property and the biological property, recognition accuracy had improved whatever classifier was chosen. Moreover, the results showed that the SVM classifier offered significant advantages over linear Fisher classifier in DNA sequences classification. In conclusion, the 10-D vector and SVM classifier could be combined to produce the best recognition ratio.

## References

1. Tiwari S, Ramachandran S, Bhattacharya A, Bhattacharya S, Ramaswamy R. 1997. Prediction of probable genes by Fourier analysis of genomic sequences. Comput Appl Biosci. 13(3):263-270.
2. Anastassiou D. 2000. Frequency-domain analysis of biomolecular sequences. Bioinformatics. 16(12):1073-1081.
3. Fickett JW. 1996. The gene identification problem: an overview for developers. Comput Chem. 20(1):103-118.
4. Yin C, Yau SS. 2005. A Fourier characteristic of coding sequences: origins and a non-Fourier approximation. J Comput Biol. 12(9):1153-1165.
5. Lorenzo-Ginori JV, Rodriguez-Fuentes A, Abalo RG, Rodriguez RS. 2009. Digital signal processing in the analysis of genomic sequences. Curr Bioinform. 4(1):28-40.
6. Román-Roldán R, Bernaola-Galván P, Oliver JL. 1998. Sequence compositional complexity of DNA through an entropic segmentation method. Phys Rev Lett. 80(6):1344-1347.
7. Law NF, Cheng KO, Siu WC. 2006. On relationship of Z-curve and Fourier approaches for DNA coding sequence classification. Bioinformation. 1(7):242-246.
8. Li C, Marzani F, Yang F. 2018. Demodulation of chaos phase modulation spread spectrum signals using machine learning methods and its evaluation for underwater acoustic communication. Sensors (Basel). 18(12):4217.
9. Xiong F, Zhang Z, Ling Y, Zhang J. 2022. Image thresholding segmentation based on weighted Parzen-window and linear programming techniques. Sci Rep. 12(1):13635.
10. Zhang R. 2011. A rebuttal to the comments on the genome order index and the Z-curve. Biol Direct. 6:10.
11. Cui Y, Xu Z, Li J. 2019. ZCMM: A novel method using Z-curve theory-based and position weight matrix for predicting nucleosome positioning. Genes (Basel). 10(10):765.
12. Yin C, Yau SS. 2007. Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence. J Theor Biol. 247(4):687-694.
13. Sanders WS, Johnston CI, Bridges SM, Burgess SC, Willeford KO. 2011. Prediction of cell penetrating peptides by support vector machines. PLoS Comput Biol. 7(7):e1002101.
14. Misaki M, Kim Y, Bandettini PA, Kriegeskorte N. 2010. Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. Neuroimage. 53(1):103-118.
15. Ruiz-Gonzalez R, Gomez-Gil J, Gomez-Gil FJ, Martínez-Martínez V. 2014. An SVM-based classifier for estimating the state of various rotating components in agro-industrial machinery with a vibration signal acquired from a single point on the machine chassis. Sensors (Basel). 14(11):20713-20735.
16. Chang CC, Lin CJ. 2007. Libsvm: a library for support vector machines. ACM Trans Intell Syst Technol. 2(3):1-27.