Classification of short human exons and introns based on statistical features

Yonghui Wu* and Alan Wee-Chung Liew†
Department of Computer Engineering and Information Technology, 83 Tat Chee Avenue, City University of Hong Kong, Kowloon, Hong Kong

Hong Yan‡
Department of Computer Engineering and Information Technology, 83 Tat Chee Avenue, City University of Hong Kong, Kowloon, Hong Kong
and School of Electrical and Information Engineering, University of Sydney, NSW 2006, Australia

Mengsu Yang§
Department of Chemistry and Biology, 83 Tat Chee Avenue, City University of Hong Kong, Kowloon, Hong Kong
(Received 7 October 2002; published 27 June 2003)

The classification of human gene sequences into exons and introns is a difficult problem in DNA sequence analysis. In this paper, we define a set of features, called the simple *Z* (SZ) features, which is derived from the *Z*-curve features for the recognition of human exons and introns. The classification results show that SZ features, while fewer in numbers (three in total), can preserve the high recognition rate of the original nine *Z*-curve features. Since the size of SZ features is one-third of the *Z*-curve features, the dimensionality of the feature space is much smaller, and better recognition efficiency is achieved. If the stop codon feature is used together with the three SZ features, a recognition rate of up to 92% for short sequences of length <140 bp can be obtained.

DOI: 10.1103/PhysRevE.67.061916 PACS number(s): 87.14.Gg

I. INTRODUCTION

The prediction of genes and the classification of coding and noncoding DNA sequences are popular research areas. In the past twenty years, numerous advanced statistical gene-finding algorithms have been developed. These algorithms operate on a basic assumption that every exon in a genome should have some distinct sequence features or properties that can distinguish it from the surrounding regions, such as introns or intergenic regions. Several review papers about these algorithms have been published by Fickett [1,2] and Guigo [3]. Some of the sequence features that have been used are compositional bias [4], position weight matrix [5], codon usage measure [6], dicodon usage measure [7] and three-base periodicity [8]. These features have been used either singly or in combinations with different algorithms such as MZEF [9], GLIMMER [10], MORGAN [11], GENEMARK.HMM, [12] and GENESCAN [13]. Although good results have been obtained in the recognition of coding and noncoding regions of prokaryotes gene, the strengths of the statistical features are not sufficient to identify exons in humans because of their limited average length. So the classification of coding and noncoding sequences in humans is still a difficult problem in bioinformatics.

Good recognition rates (i.e., 95–98 %) for the coding and noncoding sequences of yeast and vibrio cholerae, and recently of other bacterial and archaeal genomes, can be obtained by using the *Z*-curve features [14–17]. In this paper, we investigate whether these features are still as effective in recognizing the exons and introns of humans, which is a much more challenging problem since the human exons and introns are much shorter in length (137 bp in average) [10–12,18–20]. We then propose a set of more efficient statistical features, called the SZ features. We show that these features can be combined with other features to achieve a significant improvement in the recognition accuracy.

The *Z*-curve based method was suggested by Zhang and co-workers [21–23]. It is based on the differences of single nucleotide frequencies at three codon positions between the protein coding open reading frame (ORFs) and the noncoding ones. Assume that the frequencies of the bases A, C, G, and T occurring in an ORF or a fragment of DNA sequence with bases at positions 1,4,7, . . . ;2,5,8, . . . ; and 3,6,9, . . . , are $a_1,c_1,g_1,t_1;a_2,c_2,g_2,t_2;a_3,c_3,g_3,t_3$, respectively. In the *Z*-curve method, the variables $x_1,y_1,z_1;x_2,y_2,z_2;x_3,y_3,z_3$ are defined as

$$\left.\begin{aligned} x_i&=(a_i+g_i)-(c_i+t_i)\\ y_i&=(a_i+c_i)-(g_i+t_i)\\ z_i&=(a_i+t_i)-(g_i+c_i)\end{aligned}\right\}(i=1,2,3), \tag{1}$$

and the nine features are denoted by f_1 to f_9 as follows:

$$\begin{aligned} &f_1=x_1,\ f_2=y_1,\ f_3=z_1;\\ &f_4=x_2,\ f_5=y_2,\ f_6=z_2;\\ &f_7=x_3,\ f_8=y_3,\ f_9=z_3. \end{aligned} \tag{2}$$

*Email address: itwyh@cityu.edu.hk
†Email address itwcliew@cityu.edu.hk
‡Email address: ityan@cityu.edu.hk
§Email address: bhmyang@cityu.edu.hk

For a DNA sequence with N bases, the N-length Z curve is generated by computing the quantities $f_1, f_2, \ldots, f_9$ for the DNA segment from the first base position up to the base index n. Thus, the last position of the Z curve denotes the frequency differences of single nucleotides in this entire sequence, and can be used as features for the classification of a DNA sequence into an exon or an intron [14–16].

In this paper, we analyze the characteristics of a DNA sequence which are captured by the Z-curve features. A set of features, called the simple Z (SZ) features, is proposed for the recognition of short human exons and introns. Then, Z-curve features and SZ features are compared using an information-theoretic method, and the recognition rate for human exons and introns, using the SZ features, is evaluated using the K-nearest-neighbor (KNN) classifier.

II. DATABASES

We use the human exon and intron datasets (refer to Ref. [27]). We extracted 1500 human exons and 1500 human introns. Their lengths are all less than 140 bp, where bp stands for base pairs, and the exons are not frame specific. Although introns in humans can be potentially very long, short introns were selected since they are more easily confused with exons and also to avoid introducing any bias in recognition due to length. The exons are used as positive samples and introns as negative samples.

III. THE PROPOSED SZ FEATURES

In the nine Z-curve features, $(a_i+g_i)-(c_i+t_i)(i=1,2,3)$ displays the number of bases of the purines or pyrimidines types in frames 1, 2, and 3, respectively. Among the three frames, only one is at the correct coding position. Since the predominant bases at the first codon position are purines, this feature has a large positive value at the correct coding position [3]. Likewise, the feature $(a_i+c_i)-(g_i+t_i)$ displays the number of bases of the amino or keto ($M=$ A, C or $K=$ G, T) types in frames 1, 2, and 3, respectively; and the feature $(a_i+t_i)-(g_i+c_i)$ displays the number of bases of the hydrogen bonds types, i.e., bases of strong H bonds ($S=$ G, C) or weak H bonds ($W=$ A, T), in frames 1, 2, and 3, respectively [23].

In order to improve the recognition efficiency, we propose a set of features, called the SZ features, to replace the nine Z-curve features:

$$\left\{\begin{array}{l}\max\limits_i[(a_i+g_i)-(c_i+t_i)]\\ \max\limits_i[(a_i+c_i)-(g_i+t_i)]\\ \max\limits_i[(a_i+t_i)-(g_i+c_i)]\end{array}\right\}(i=1,2,3). \tag{3}$$

Since we expect the single nucleotide differences to be far different from a random residual when the start position corresponds to the correct reading frame, the max operation over the three positions will ensure that the SZ features capture information at the correct reading frame.

IV. THE MUTUAL INFORMATION CONTENT OF HUMAN EXONS AND INTRONS FEATURES

To measure the discrimination ability of the nine Z-curves features, $f_1, f_2, \ldots, f_9$, and the three SZ features f_{10}, f_{11}, f_{12}, we use an information-theoretic approach. Specifically, we want to measure how much information a particular feature f_j tells us about the class label ω, where the class label consists of exon or intron.

The mutual information [24,25] of the jth feature, f_j, with respect to the class labels ω is given by

$$G_j=\sum_{i=1}^{m}\sum_{k=1}^{v_j} p(\omega_i, f_j(k))\ln\frac{p(\omega_i|f_j(k))}{p(\omega_i)}, \tag{4}$$

where $p(A,B)$ denotes the joint probability of observing both events A and B, and $p(B|A)$ denotes the conditional probability of observing event B after event A has occurred. In Eq. (4), each feature f_j has v_j discrete values which are obtained by creating histograms. The mutual information G_j measures the information that feature f_j tells us about the class label. Since

$$p(\omega_i, f_j(k))=p(\omega_i)p(f_j(k)|\omega_i)=p(\omega_i|f_j(k))p(f_j(k)), \tag{5}$$

Eq. (4) can be rewritten as

$$G_j=\sum_{i=1}^{m}\sum_{k=1}^{v_j} p(\omega_i)p(f_j(k)|\omega_i)\ln\frac{p(f_j(k)|\omega_i)}{p(f_j(k))}. \tag{6}$$

For exon and intron classification, we have two classes, so $m=2$. The prior probabilities for coding (exon) and noncoding regions (intron) for human genome is roughly $p(\text{coding})=0.05$, $p(\text{noncoding})=0.95$.

In order to compare the effect of dataset size on mutual information, we randomly select 500 exons and 500 introns as training dataset 1, 750 exons and 750 introns as training dataset 2, and 1000 exons and 1000 introns as training dataset 3, all from the database of 1500 exons and 1500 introns of humans. In order to provide a baseline comparison about the mutual information of each feature, we construct training dataset 4 by randomly selecting 1000 exons twice and training dataset 5 by randomly selecting 1000 introns twice. The mutual information of five training datasets is computed and averaged over three experiments (Table I and Fig. 1). In Fig. 1, the markers square, circle, point, diamond, and asterisk represent datasets 1 to 5, respectively.

The following results can be summarized from Table I and Fig. 1.

(1) The mutual information of the training datasets 1, 2, and 3 is similar. It shows that if the training dataset represents the population well, the information of each feature does not increase with the size of the dataset.

(2) The mutual information of each feature in the training datasets 1, 2, and 3 is much larger than that of the corresponding features in the training datasets 4 and 5. It shows that the Z-curve features and the SZ features are both effective for recognizing human short exons and introns.

TABLE I. Mutual Information of different features for human exons and introns.

		500 Samples	750 Samples	1000 Samples	Training Set 4	Training Set 5
	$(a_1+g_1)-(c_1+t_1)$	0.0753	0.0726	0.0629	0.0143	0.0108
	$(a_1+c_1)-(g_1+t_1)$	0.0635	0.0610	0.0559	0.0076	0.0091
	$(a_1+t_1)-(c_1+g_1)$	0.1106	0.1182	0.1208	0.0122	0.0113
Z	$(a_2+g_2)-(c_2+t_2)$	0.0929	0.0757	0.0775	0.0107	0.0099
	$(a_2+c_2)-(g_2+t_2)$	0.1120	0.0974	0.0920	0.0078	0.0070
	$(a_2+t_2)-(c_2+g_2)$	0.1375	0.1491	0.1483	0.0132	0.0119
	$(a_3+g_3)-(c_3+t_3)$	0.0797	0.0639	0.0628	0.0071	0.0089
	$(a_3+c_3)-(g_3+t_3)$	0.0400	0.0331	0.0323	0.0055	0.0112
	$(a_3+t_3)-(c_3+g_3)$	0.0937	0.1092	0.1040	0.0121	0.0152
	$\max[(a_i+g_i)-(c_i+t_i)]$	0.1624	0.1476	0.1418	0.0093	0.0081
SZ	$\max[(a_i+c_i)-(g_i+t_i)]$	0.1099	0.1034	0.0960	0.0097	0.0091
	$\max[(a_i+t_i)-(c_i+g_i)]$	0.2726	0.2932	0.3096	0.0107	0.0128

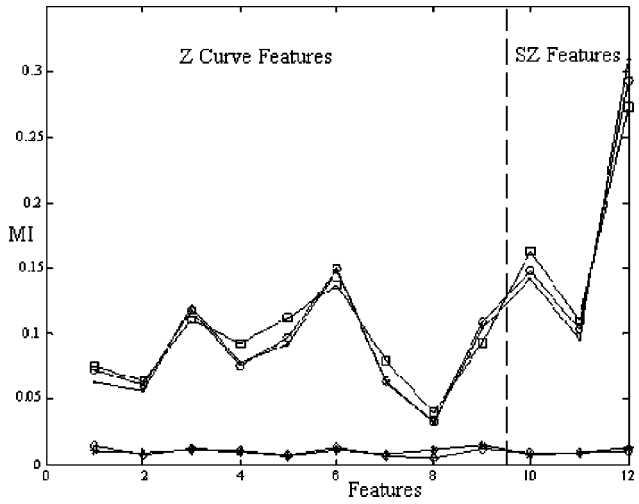

FIG. 1. The mutual information of different features in different training datasets.

(3) For the Z-curve features, the mutual information varies greatly among different features. The mutual information of f_3 and f_6 is larger, and the mutual information of f_8 is smaller.

(4) The order of Z-curve features and SZ features arranged by their mutual information is as follows: $f_{12}>f_{10}>f_6>f_3>f_{11}>f_9>f_5>f_4>f_1>f_7>f_2>f_8$. The mutual information of the SZ features is generally larger than the mutual information of the Z-curve features. Among them, the mutual information of f_{12} is the largest.

(5) The mutual information of the SZ features is generally larger than the maximum mutual information of the three corresponding Z-curves features for training datasets 1, 2, and 3.

V. CORRELATIONS BETWEEN FEATURES

The correlation coefficients of the Z-curve features and the SZ features are computed using Eq. (7) and are listed in Table II. In Eq. (7), x_i and y_i are the two features to be correlated, $\bar{x}$ and $\bar{y}$ are their mean values computed over the samples, and n is the number of samples in the dataset. In

TABLE II. The correlation coefficients of the Z-curve features and the SZ features of human exons and introns.

	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Z_9	SZ_1	SZ_2	SZ_3
Z_1	1	−0.12	0.01	0.33	0.03	0.16	0.29	−0.16	−0.02	0.70	−0.08	0.14
Z_2	−0.12	1	0.05	−0.06	0.23	0.11	0.00	0.23	−0.11	−0.07	0.62	0.08
Z_3	0.01	0.05	1	−0.03	−0.08	0.30	0.11	0.06	0.32	0.06	0.01	0.58
Z_4	0.33	−0.06	−0.03	1	−0.08	0.08	0.31	−0.07	0.03	0.60	−0.04	0.13
Z_5	0.03	0.23	−0.08	−0.08	1	0.10	−0.09	0.24	0.11	0.00	0.60	0.16
Z_6	0.16	0.11	0.30	0.08	0.10	1	−0.01	−0.12	0.25	0.11	0.04	0.65
Z_7	0.29	0.00	0.11	0.31	−0.09	−0.01	1	−0.17	0.05	0.60	−0.07	0.12
Z_8	−0.16	0.23	0.06	−0.07	0.24	−0.12	−0.17	1	0.00	−0.14	0.59	0.00
Z_9	−0.02	−0.11	0.32	0.03	0.11	0.25	0.05	0.00	1	0.05	−0.01	0.56
SZ_1	0.70	−0.07	0.06	0.60	0.00	0.11	0.60	−0.14	0.05	1	−0.02	0.23
SZ_2	−0.08	0.62	0.01	−0.04	0.60	0.04	−0.07	0.59	−0.01	−0.02	1	0.16
SZ_3	0.14	0.08	0.58	0.13	0.16	0.65	0.12	0.00	0.56	0.23	0.16	1

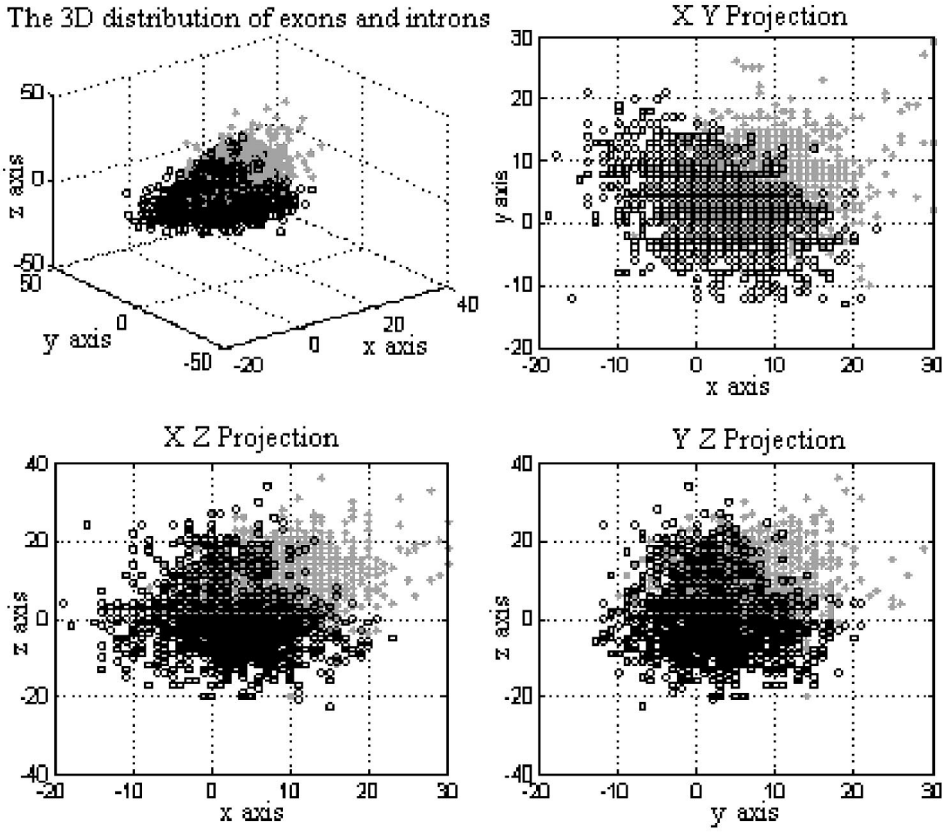

FIG. 2. The distribution of the three SZ features of exons and introns in 3D space and 2D projections in *X*-*Y*, *X*-*Z*, and *Y*-*Z* planes. Gray points represent exons and black circles represent introns.

Table II, $Z_i(i=1$ to 9) are the *Z*-curve features, and $\mathrm{SZ}_i(i=1$ to 3) are the SZ features. The correlation values in Table II show that the SZ features have less redundancy than the *Z*-curve features, as we would have expected:

$$f_{xy}=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2\sum_{i=1}^{n}(y_i-\overline{y})^2}}. \tag{7}$$

The following results can be summarized from Table II.

(a) The correlations between the *Z*-curve features are generally small, except as noted below.

(b) The correlations between the SZ features are generally small.

(c) The correlations between SZ_i and Z_i, SZ_i and Z_{i+3}, SZ_i and Z_{i+6} $(i=1,2,3)$, are large since they are closely related by Eq. (3). In addition, the correlations between Z_i, Z_{i+3}, and Z_{i+6} $(i=1,2,3)$, are larger than the ones between other *Z*-curve features, and also larger than the smallest correlation value among the SZ features.

VI. RECOGNITION OF SHORT HUMAN EXONS AND INTRONS

The exons and introns can be visualized in three-dimensional (3D) space and 2D projections using the SZ features (Fig. 2). As can be seen, although the two classes are somewhat separated, there is a considerable overlap between the two groups, and a linear decision plane dividing the two classes cannot be easily obtained. To evaluate the recognition rate using the nine *Z*-curve features and the three SZ features, the same datasets, with the two types of feature sets, are classified using the KNN classifier. Unlike the Fisher discriminant algorithm, KNN algorithm does not require the decision surface to be linear. The KNN algorithm uses $K=10$ (experimentation shown in Fig. 3 indicated that $K=10$ is a reasonable value to use) and the Euclidean distance metric. A sixfold cross-validation test is adopted.

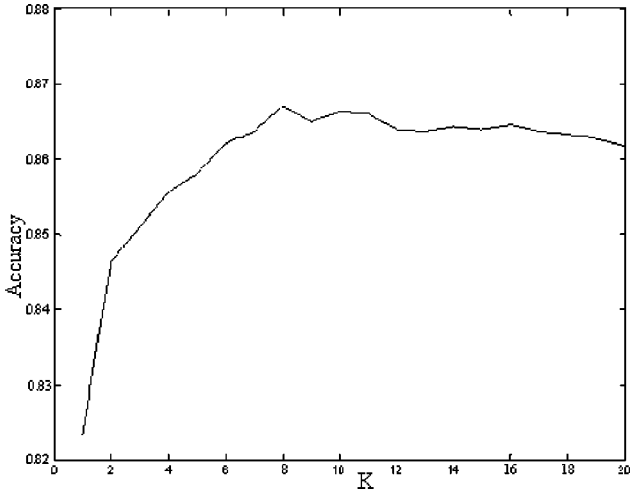

FIG. 3. Effect of *K* on the accuracy of the KNN algorithm, testing on 750 training samples and 750 testing samples use *Z*-curve features.

TABLE III. The average sensitivity, specificity, and accuracy for human exon and intron recognition based on the nine *Z*-curve features using the KNN classifier.

Number of training samples	Number of testing samples	Sensitivity	Specificity	Accuracy
250	1250	0.7768	0.9056	0.8412
500	1000	0.8010	0.916	0.8585
750	750	0.8267	0.9060	0.8663
1000	500	0.843	0.9140	0.8785
1250	250	0.8520	0.9160	0.8840

TABLE IV. The average sensitivity, specificity, and accuracy for exon and intron recognition based on the three SZ features using the KNN classifier.

Number of training samples	Number of testing samples	Sensitivity	Specificity	Accuracy
250	1250	0.8592	0.8148	0.8370
500	1000	0.8560	0.8310	0.8435
750	750	0.8540	0.8387	0.8463
1000	500	0.8540	0.8270	0.8405
1250	250	0.8680	0.8580	0.8630

TABLE V. The average sensitivity, specificity, and accuracy for exon and intron recognition based on the three SZ features using the Fisher algorithm.

Number of training samples	Number of testing samples	Sensitivity	Specificity	Accuracy
250	1250	0.7668	0.8420	0.8044
500	1000	0.7600	0.8555	0.8077
750	750	0.7680	0.8460	0.8070
1000	500	0.7710	0.8360	0.8035
1250	250	0.7800	0.8680	0.8240

TABLE VI. The average sensitivity, specificity, and accuracy for exon and intron recognition based on the three SZ features and the stop codon feature.

Number of training samples	Number of testing samples	Sensitivity	Specificity	Accuracy
250	1250	0.9012	0.8824	0.8918
500	1000	0.8970	0.8900	0.8935
750	750	0.9013	0.8800	0.8906
1000	500	0.9140	0.891	0.9025
1250	250	0.9160	0.9220	0.9190

In the dataset, 1500 exons and 1500 introns are randomly divided into two parts. Part 1 is taken as the training set and part 2 as the testing set. The sensitivity, specificity, and accuracy of the algorithm based on part 2 are calculated. Then, the procedure is applied again by reversing the roles of the two parts, i.e., part 2 is now taken as the training set and part 1 as the testing set. The above procedure is repeated three times. For the first time, part 1 contains 750 exons and 750 introns, and part 2 contains 750 exons and 750 introns. For the second and third times, the partition of the two parts becomes $500+1000$ and $250+1250$, respectively. The average sensitivity, specificity, and accuracy over the sixfold cross-validation test are calculated and listed in Tables III and IV, respectively. The sensitivity (S_n) and specificity (S_p) are as defined in Ref. [26], where S_n is the proportion of coding nucleotides that have been correctly classified as coding and S_p is the proportion of noncoding nucleotides that have been correctly classified as noncoding. The sixfold cross-validation test indicated that the accuracy of the *Z*-curve features is better than 84%, and the accuracy of the SZ features is better than 83%.

As a comparison, the recognition result using the Fisher discriminant algorithm is shown in Table V. The sixfold cross-validation test indicated that the accuracy of the SZ features based on the Fisher algorithm is about 80%. This result agrees with that suggested in Fig. 2, that is, a simple decision plane adequately separating the two classes cannot be obtained.

The recognition tests show that SZ features are able to maintain the good recognition rate of the *Z*-curve features, while having better recognition efficiency (i.e., similar recognition rate but using fewer features). This is due to the fact that the max operation in the SZ features is able to capture the information at the correct reading frame. If SZ features and other type of features are used together, the recognition rate can be improved considerably. For example, the three SZ features and the stop codon feature are used together for recognizing short human exons and introns, and the recognition results are list in Table VI. The stop codon feature has been used by Wang [18]. In the coding frame of a gene, at least one of the triplets TAA, TAG, and TGA is uniquely used as the last codon of the gene. On the average, the triplets TAA, TAG, and TGA occur about every 20 bases in the DNA sequences. The distribution of the triplets in the coding regions is apparently different from those in the noncoding and intergenic regions. In deriving the stop codon feature, the number of triplets TAA, TAG, and TGA occurring in each of the three frames of the sequence is counted. Let the total number of the triplets contained in all the three frames in a sequence be denoted by n. The number of frames containing the three triplets is a sequence denoted by K, i.e., $K=0,1,2,3$. The stop codon feature is then defined by $f_{\mathrm{SC}}=(1+K^2)n$ [18]. We can see that the recognition accuracy, specificity, and sensitivity have all improved when the SZ features are augmented with the stop codon features. The sixfold cross-validation test demonstrated that the accuracy of the SZ features with the stop codon added is better than 89%.

It is interesting to compare our recognition results with the results reported in Ref. [18]. In Table VII of Ref. [18], they compared the average recognition accuracy of their algorithm with two other algorithms, i.e., the lengthen-shuffling fast Fourier transform and the Markov chain model for short human coding and noncoding sequences. They reported the accuracies of 87.2%, 78.1%, and 89.5%, respectively, for sequences with length of 129 bp. For sequences of length 162 bp, the accuracies of 90.8%, 80.7%, and 90.1%, respectively, were obtained. Our results with sequences of length 140 bp indicated that the SZ features together with the stop codon feature can perform on par or better than the three algorithms on the recognition of short human coding and noncoding sequences.

VII. CONCLUSIONS

A set of features, called the SZ features, is proposed for the classification of short human exons and introns. Due to their ability to capture information at the correct reading frame, the SZ features were able to preserve the good recognition rate of *Z*-curve features while using much fewer features. If the SZ features and the other additional features, such as the stop codon feature, are used together, the recognition rate can be improved significantly. Experiments on recognizing the short human exons and introns (sequence length 140 bp) using the three SZ features and the stop codon feature were able to give a recognition accuracy of 89–92 % based on the sixfold cross-validation test.

ACKNOWLEDGMENT

This work was supported by a CityU interdisciplinary research grant (Grant No. 9010003).

[1] J.W. Fickett and C. Tung, Nucleic Acids Res. **20**, 6641 (1992).
[2] J.W. Fickett, Trends Genet. **12**, 316 (1996).
[3] R. Guigo, in *Genetic Databases*, edited by M. Bishop (Academic Press, New York, 1999), p. 53.
[4] J.W. Fickett, Nucleic Acids Res. **10**, 5303 (1982).
[5] R. Staden, Nucleic Acids Res. **12**, 505 (1984).
[6] R. Staden and A. McLachlan, Nucleic Acids Res. **10**, 141 (1982).
[7] R. Farber, A. Lapedes, and K. Sirottkin, J. Mol. Biol. **226**, 471 (1992).
[8] S.R.S. Tiwari, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy, CABIOS, Comput. Appl. Biosci. **13**, 263 (1997).
[9] M. Zhang, Proc. Natl. Acad. Sci. U.S.A. **94**, 565 (1997).
[10] S.L. Salzberg, A.L. Delcher, S. Kasif, and O. White, Nucleic Acids Res. **26**, 544 (1998).
[11] S.L. Salzberg, A.L. Delcher, K.H. Fasman, and J. Henderson, J. Comput. Biol. **5**, 667 (1998).
[12] A.V. Lukashin and M. Borodovsky, Nucleic Acids Res. **26**, 1107 (1998).
[13] C. Burge and S. Karlin, J. Mol. Biol. **268**, 78 (1997).

[14] J. Wang and C.T. Zhang, Eur. J. Biochem. **268**, 4261 (2001).
[15] C.T. Zhang and J. Wang, Nucleic Acids Res. **28**, 2804 (2000).
[16] C.T. Zhang, J. Wang, and R. Zhang, Comput. Chem. **26**, 195 (2002).
[17] F. Guo, H. Ou, and C.T. Zhang, Nucleic Acids Res. **31**, 1780 (2003).
[18] Y. Wang, C.T. Zhang, and P. Dong, Biopolymers **63**, 207 (2002).
[19] J.D. Hawkins, Nucleic Acids Res. **16**, 9893 (1988).
[20] T.A. Thanaraj, Nucleic Acids Res. **28**, 744 (2000).
[21] R. Zhang and C.T. Zhang, J. Biomol. Struct. Dyn. **11**, 767 (1994).
[22] C.T. Zhang and K.C. Chou, J. Mol. Biol. **238**, 1 (1994).
[23] C.T. Zhang, J. Theor. Biol. **187**, 297 (1997).
[24] S. Haykin, *Digital Communications* (Wiley, New York, 1988).
[25] C.E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948).
[26] M. Burset and R. Guigo, Genomics **34**, 353 (1996).
[27] See, http://bit.uq.edu.au/altExtron/ for human exon and intron datasets.