

Yonghong Wang<sup>1,2</sup>

Chun-Ting Zhang<sup>1</sup>

Puxuan Dong<sup>1</sup>

<sup>1</sup> Department of Physics,  
Tianjin University,  
Tianjin, 300072, China

<sup>2</sup> Department of Physics,  
Hebei University of  
Technology,  
Tianjin, China

Received 12 December 2000;  
accepted 20 September 2001

---

## Recognizing Shorter Coding Regions of Human Genes Based on the Statistics of Stop Codons

**Abstract:** With the quick progress of the Human Genome Project, a great amount of uncharacterized DNA sequences needs to be annotated copiously by better algorithms. Recognizing shorter coding sequences of human genes is one of the most important problems in gene recognition, which is not yet completely solved. This paper is devoted to solving the issue using a new method. The distributions of the three stop codons, i.e., TAA, TAG and TGA, in three phases along coding, noncoding, and intergenic sequences are studied in detail. Using the obtained distributions and other coding measures, a new algorithm for the recognition of shorter coding sequences of human genes is developed. The accuracy of the algorithm is tested based on a larger database of human genes. It is found that the average accuracy achieved is as high as 92.1% for the sequences with length of 192 base pairs, which is confirmed by sixfold cross-validation tests. It is hoped that by incorporating the present method with some existing algorithms, the accuracy for identifying human genes from unannotated sequences would be increased. © 2002 John Wiley & Sons, Inc. *Biopolymers* 63: 207–216, 2002; DOI 10.1002/bip.10054

**Keywords:** human genome; gene; gene recognition; stop codons; statistics of stop codons

---

### INTRODUCTION

Databases of human and model organism DNA sequences have been increasing quickly since the beginning of the Human Genome Program (HGP) in 1990. Currently the HGP is evolving into the large-scale sequencing phase. It is now indispensable to use computational methods for identifying human genes.

A large amount of uncharacterized DNA sequences needs to be annotated copiously by better algorithms. There exist two basic problems in gene recognition: recognition of protein coding regions and recognition of the functional sites of genes. They are not yet satisfactorily solved, especially in recognizing shorter coding regions of human genes. The present paper is devoted to studying this important problem.

---

Correspondence to: Chun-Ting Zhang; email: ctzhang@tju.edu.cn

Contract grant sponsor: 973 Project of China

Contract grant number: G1999075606

*Biopolymers*, Vol. 63, 207–216 (2002)

© 2002 John Wiley & Sons, Inc.

In the past twenty years or so, many algorithms of gene recognition have been developed. Among them are the algorithms based on compositional bias,<sup>1</sup> position weight matrix,<sup>2</sup> codon usage measure,<sup>3</sup> dicodon usage measure,<sup>4</sup> and 3-base periodicity.<sup>5, 6</sup> Fickett listed seven algorithms in a review paper published in 1996<sup>7</sup>; one of them was used to recognize protein coding regions<sup>8</sup> and the others were used to recognize complete genes. In 1996, Burset and Guigo constructed a dataset of vertebrate gene sequences, and evaluated many gene recognition algorithms by the dataset.<sup>9</sup> It turned out that the accuracy of the algorithms was less than 50% for recognizing internal exons. Several new algorithms have been proposed since then, such as MZEF,<sup>10</sup> GLIMMER,<sup>11</sup> MORGAN,<sup>12</sup> GeneMark.hmm,<sup>13</sup> and others.<sup>14,15</sup> An up-to-date list of references is maintained by Wentian Li at the website [linkage.rockefeller.edu/wli/gene/list.html](http://linkage.rockefeller.edu/wli/gene/list.html). The accuracy of these algorithms for complete gene recognition is generally high when tested using Guigo's dataset, but not so good for newly sequencing human genome sequences. Those algorithms, which used both coding information and splicing signals, performed better than those using only splicing signals.<sup>16</sup> There is still the need of new methods for gene prediction that utilize features of gene structure that have so far not been incorporated into programs already available.<sup>7</sup>

In this paper, the emphasis is placed on the recognition of shorter coding regions of human genes. To find properties of these regions, a larger set of protein coding regions and noncoding regions of human genes and intergenic regions is analyzed. The statistic properties to be analyzed include the distributions of bases at three codon positions in coding regions and the distributions of the triplets TAA, TAG, and TGA in three frames of shorter coding, noncoding and intergenic regions, respectively. The features of these two statistic distributions, together with the length-shuffling Fast Fourier Transform (FFT) technique,<sup>17</sup> constitute the basis of the present algorithm. The goal of this work is to find a new method suitable for recognizing shorter coding regions of human genes.

## DATABASE

The database consists of three sets, composed of coding, noncoding and intergenic fragments of human DNA, respectively. Coding sequences of the database are extracted from the file 4813\_Hum\_CDS.fa at the website [ftp://genome.lbl.gov/pub/genesets/Human/Release June 5, 1999](ftp://genome.lbl.gov/pub/genesets/Human/Release%20June%205,%201999). The file contains 4813 complete

coding sequences of human genes, beginning from the first codon and ending with one of the stop codons. Noncoding sequences are extracted from the files in the directory `intron_v105` at the above website, including complete intron sequences of 462 human genes. The intergenic sequences are partially extracted from the entries longer than 30,000 base pairs (bp) at the above website and partially from the human chromosome 4, respectively. Removing the regions of mRNA in the entries annotated, the remaining sequences are used. Each set of the database includes 4000 fragments with length of 200 bp. The coding fragments are used as positive samples. The noncoding and intergenic fragments are used as negative samples. The detailed procedure for the construction of the database is as follows. The first 4000 genes in the file 4813\_Hum\_CDS.fa with lengths longer than 210 bp are used to construct the positive sample set. For each positive sample, a fragment of 200 bp is used. To avoid using coding sequences with definite phase (see the discussion later), the first base of each fragment is selected randomly from the first 6 bases of the original coding sequence in the file. The negative samples consist of two parts: noncoding and intergenic fragments. The negative sample set 1 contains 4000 noncoding sequences, in which the length of each sequence is 200 bp. These 4000 noncoding sequences are randomly selected from the sequences in the intron files with length longer than 200 bp. There is no overlapping between any two adjacent noncoding sequences. Similarly, the negative sample set 2 is composed of 4000 intergenic fragments.

## ALGORITHM

### Recognition Variables

At the core of gene recognition algorithm, a coding measure is defined to make the coding/noncoding or coding/intergenic decision. In this paper such a coding measure will be derived from four variables defined as follows:

*The Asymmetric Variable  $x_i$ .* Many authors have found that the distribution of bases at three codon positions is asymmetric.<sup>18, 19</sup> This fact is used here to distinguish coding regions from noncoding and intergenic regions. The distribution of bases at three codon positions, shown in Table I, is obtained from the complete human genes in the file 4813\_Hum\_CDS.fa. It can be seen that the contents of T, G, and A are poor at the first, second, and third codon positions, respec-

**Table I** Content of Base A, C, G, and T at Three Codon Positions

Codon Position	A	C	G	T
1st	0.259243	0.251522	0.328768	<b>0.160467<sup>a</sup></b>
2nd	0.288732	0.241069	<b>0.197685<sup>a</sup></b>	0.272514
3rd	<b>0.167272<sup>a</sup></b>	0.312449	0.326228	0.194051

<sup>a</sup> The three bold figures are used to define the asymmetry variable. See the text for more detailed explanation.

tively. In coding, noncoding and intergenic sequences, each of the three phases is considered. The subsequence with bases at positions 1, 4, 7, ... is called the phase-1 sequence. Those with bases at the positions 2, 5, 8, ... and 3, 6, 9, ... are called the phase 2 and phase 3 sequences, respectively. As only one of the three phases in coding sequences is in-frame and as it is difficult to identify which is the in-frame phase at this stage, all three possibilities need to be analyzed. At first it is assumed that the bases at the first codon position are associated with the phase-1 sequence. Let  $y_1(1)$ ,  $y_2(1)$ , and  $y_3(1)$  represent the contents of T, G, and A at the first, second, and third codon positions, respectively. Let  $R_1$  be the product of  $y_1(1)$ ,  $y_2(1)$ , and  $y_3(1)$ , i.e.,  $R_1 = y_1(1) \times y_2(1) \times y_3(1)$ . Then the bases at the first codon position are assumed associated with the phase 2 sequence. Then  $y_1(2)$ ,  $y_2(2)$ , and  $y_3(2)$  are defined similarly and  $R_2 = y_1(2) \times y_2(2) \times y_3(2)$ . Finally, the bases at the first codon position are assumed to be associated with the phase 3 sequence, then  $R_3 = y_1(3) \times y_2(3) \times y_3(3)$ . The asymmetric variable  $x_1$  is defined as the minimum of  $R_1$ ,  $R_2$ , and  $R_3$ .

**The 3-Periodicity Variable  $x_2$ .** It is well known that there exists an imperfect 3-periodicity in coding regions.<sup>20–22</sup> For a longer coding sequence, say, 1024 bp, it is relatively easier to detect the 3-base periodicity by the traditional FFT algorithm. But for shorter coding sequences, say, shorter than 150 bp, the 3-periodicity cannot be easily detected by the traditional FFT algorithm. As the lengths of most eukaryotic exons are relatively short,<sup>23, 24</sup> to detect the 3-periodicity for such sequences, a lengthen-shuffling FFT technique was proposed.<sup>17</sup> The ratio of 3-periodicity signal/noise is increased by lengthening the shorter sequences first and then shuffling the lengthened sequence. Then the FFT is performed to calculate the power spectrum at the position of  $N/3$ , where  $N$  is the length of the lengthen-shuffling sequence.<sup>17</sup> Denoting

the power spectrum by  $P$ , the 3-periodicity variable is defined by  $x_2 = \ln P$ .

**The Purine Variable  $x_3$ .** It is well known that the predominant bases at the first codon position are purines and this fact is independent of species.<sup>18, 19</sup> Compared with coding sequences, the bases in both noncoding and intergenic regions tend to be randomly distributed. In coding, noncoding, and intergenic sequences of the database, the three phases of each sequence are considered. The occurrence frequencies of purines in the three phases are denoted by  $P_1$ ,  $P_2$ , and  $P_3$ , respectively. The purine variable is defined by  $x_3 = \max(P_1, P_2, P_3)$ .

**The Stop-Codon Variable  $x_4$ .** The stop codons are very strong signals in DNA sequences. In the coding frame of a gene, at least one of them 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 DNA sequences. The distribution of the triplets in coding regions is apparently different from those in noncoding and intergenic regions. To analyze the distributions of the triplets in three frames of coding, noncoding, and intergenic sequences, respectively, sequences with various lengths are studied (see the text below). The number of the triplets TAA, TAG, and TGA occurring in each frame of the sequence is counted. According to the number of frames containing the triplets, the sequences in each set of the database are classified into four groups: fstop 0, fstop 1, fstop 2, and fstop 3. The sequences in fstop 0 do not contain the triplets in any frames. The sequences in fstop 1 contain the triplets only in one frame. Similarly, the sequences in fstop 2 contain the triplets in any two frames, and those in fstop 3 contain the triplets in each of the three frames. Refer to Table II. For the sake of reliability, analysis is applied to sequences with various lengths by chunking the sequence with 200 bases into shorter fragments with, for instance, 42, 63, 87, 108, 129, 162, and 192 bases, respectively. Figure 1 shows that the number of noncoding and intergenic sequences containing the triplets TAA, TAG, and TGA in all three frames increases rapidly when the sequences get longer. To utilize the distribution of the triplets in three frames, the fourth, or the stop-codon variable, is defined as follows. The total number of the triplets contained in all three frames in a sequence is denoted by  $n$ . The number of the frames containing the three triplets in a sequence is denoted by  $K$ , i.e.,  $K = 0, 1, 2, 3$ . The stop-codon variable is defined by  $x_4 = (1 + K^2) \times n$ .

In summary, the first three variables are based on

**Table II The Distribution of the Stop Codons TAA, TAG, and TGA in all Three Frames of Coding, Noncoding, and Intergenic Sequences, Respectively, with Various Sequence Lengths<sup>a</sup>**

Sequence Type	Number of Sequences	Length of Sequences (bp)	fstop 0 (%)	fstop 1 (%)	fstop 2 (%)	fstop 3 (%)
Coding	16000	42	34.70	47.06	18.24	0.00
Noncoding	16000	42	13.42	39.38	36.83	10.37
Intergenic	16000	42	9.87	34.69	40.7	14.83
Coding	12000	63	21.80	46.83	31.37	0.00
Noncoding	12000	63	6.50	26.39	42.93	24.16
Intergenic	12000	63	4.42	21.47	43.58	30.53
Coding	8000	87	14.44	41.64	43.93	0.00
Noncoding	8000	87	3.31	15.84	40.39	40.35
Intergenic	8000	87	2.20	12.58	38.72	46.50
Coding	4000	108	10.33	38.38	51.30	0.00
Noncoding	4000	108	1.98	10.95	34.93	52.15
Intergenic	4000	108	1.18	7.87	30.95	60.00
Coding	4000	129	7.88	32.98	59.15	0.00
Noncoding	4000	129	1.33	7.68	30.13	60.88
Intergenic	4000	129	0.78	5.32	24.80	69.10
Coding	4000	162	4.45	25.25	70.30	0.00
Noncoding	4000	162	0.63	4.15	21.85	73.38
Intergenic	4000	162	0.35	2.80	17.65	79.20
Coding	4000	192	2.65	20.10	77.25	0.00
Noncoding	4000	192	0.32	2.82	16.33	80.53
Intergenic	4000	192	0.20	1.88	12.27	85.65

<sup>a</sup> The fstop 0 denotes the fraction of sequences that do not contain any triplets TAA, TAG, and TGA. Similarly, fstop 1, fstop 2, and fstop 3 denote the fractions of sequences in which any of the three triplets appears only in one frame, in two frames, and in each of the three frames, respectively.

well-known facts. They are not new ones. We just use them in the present study. So far as we know that the stop-codon variable might be a new one. It is introduced into the gene-finding field for the first time probably.

### The Fisher Linear Discriminant Analysis

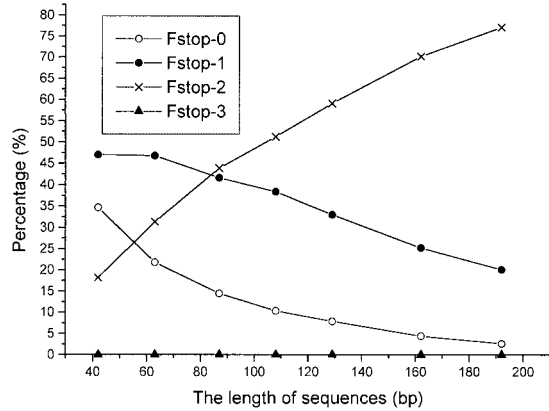
By introducing the four variables above, each sequence can be represented by a point or a vector in a four-dimensional space. There are two kinds of points, represent coding/noncoding or coding/intergenic sequences, respectively. The Fisher linear discriminant algorithm is adopted to distinguish between coding and noncoding sequences or coding and intergenic sequences. The Fisher discriminant function is actually a four-dimensional super plane, described by a vector  $\mathbf{C}$ , which has four components, i.e.,  $C_1$ ,  $C_2$ ,  $C_3$ , and  $C_4$ . The procedure to determine the vector  $\mathbf{C}$  can be found in any textbook of multivariable statis-

tics.<sup>25</sup> For a recognition task, say, coding/noncoding or coding/intergenic sequences, an appropriate threshold  $U_0$  is needed. The vector  $\mathbf{C}$  can be calculated based on the sequences in the training database. The decision of coding/noncoding or coding/intergenic is determined by the criterion of  $\mathbf{C} \cdot \mathbf{X} > U_0 / \mathbf{C} \cdot \mathbf{X} < U_0$ , where  $\mathbf{X} = (x_1, x_2, x_3, x_4)\mathbf{T}$ , and “ $\mathbf{T}$ ” is a transpose operator for a matrix.  $U_0$  is uniquely determined by letting the false positive and false negative rates identical, using the sequences in the training database.

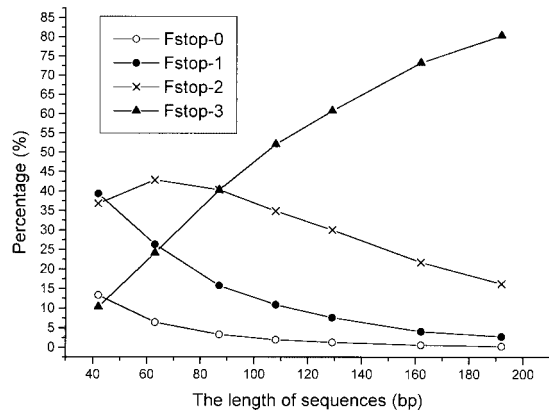
## RESULTS AND DISCUSSION

### The sensitivity, Specificity, and Accuracy of the Algorithm

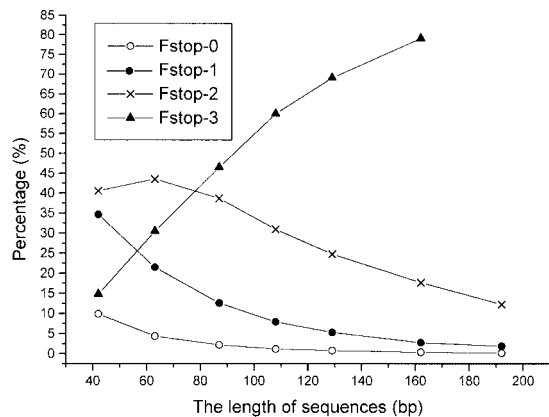
To evaluate the algorithm, sixfold cross-validation tests are adopted.<sup>15</sup> First, coding sequences are distinguished from noncoding sequences in the database.



(a)



(b)



(c)

**FIGURE 1** The distribution of the triplets TAA, TAG, and TGA in three frames along a DNA sequence. The  $x$  axis represents the sequence length, whereas the  $y$  axis indicates the percentage of sequences of the four groups, respectively. The fstop 0 denotes the fraction of sequences that do not contain any triplets TAA, TAG, and TGA. Similarly, fstop 1, fstop 2, and fstop 3 denote the fractions of sequences in which any of the three triplets appears only in one frame, in

The 4000 coding and 4000 noncoding sequences in the database are randomly divided into two parts: part 1 and part 2. Part 1 is taken as a training set and part 2 as a test set. The linear discriminative coefficients  $C_i$  and the threshold  $U_0$  can be determined based on the training set. Then the sensitivity, specificity<sup>15</sup> and accuracy of the algorithm based on part 2 are calculated. Next, the procedure is applied by reversing the roles of the two parts, i.e., part 2 is now taken as a training set and part 1 as a test set. The procedure repeats for three times. In the first time, part 1 contains 500 coding sequences and 500 noncoding sequences, and part 2 contains 3500 coding and 3500 noncoding sequences. In the second and third times, the partition becomes 1000 + 3000 and 2000 + 2000, respectively. Therefore, we have performed sixfold cross-validation tests, each of which is repeated for the sequences with various window lengths, 42, 63, 87, 108, 108, 129, 162, and 192 bp, respectively. The average sensitivity, specificity and accuracy over the six-fold cross-validation tests are calculated and listed in Table III. Similarly, coding sequences are distinguished from intergenic sequences in the database. The procedure is quite similar. The average sensitivity, specificity and accuracy over the sixfold cross-validation tests are listed in Table IV.

### The Sequence-Length Dependence of the Recognition Accuracy

As can be seen from Table III and IV, the accuracy for shorter sequences is lower than that for longer sequences. This is true for various algorithms. When sequences are shortened, the difference of statistical characteristics between coding and noncoding sequences tends to be small. For example, the ratio of the signals over noises of the 3-base periodicity gets low as the sequence involved gets shorter. Although the lengthen-shuffling technique enhances the ratio of signal/noise,<sup>17</sup> the signal at  $N/3$  is still dependent on the length of sequences. So it is for other statistical characteristics. To study the length dependence of the algorithm accuracy clearly, each individual recognition variable is used separately. In other words, only one variable is applied each time. The procedure is quite similar to that for the case of four variables. Totally seven different lengths are studied, i.e., 42, 63,

two frames, and in each of the three frames, respectively. The curves associated with the fstop 0, fstop 1, fstop 2, and fstop 3 are denoted by  $\circ$ ,  $\bullet$ ,  $\times$ , and  $\blacktriangle$ , respectively. (a) Coding, (b) noncoding, and (c) intergenic.

**Table III The Average Sensitivity, Specificity, and Accuracy<sup>a</sup> for Coding/Noncoding Sequence Recognition, Averaged Over Sixfold Cross-Validation Tests for Various Fragment Lengths**

Fragment length (bp)	192	162	129	108	87	63	42
Sensitivity (training)	92.7	90.7	88.0	85.7	82.9	77.6	73.4
Specificity (training)	92.7	90.7	88.0	85.7	82.9	77.6	73.4
Accuracy (training)	92.7	90.7	88.0	85.7	82.9	77.6	73.4
Sensitivity (test)	92.3	90.4	87.7	86.1	82.2	75.8	72.6
Specificity (test)	91.9	90.0	87.3	84.9	82.0	74.2	71.2
Accuracy (test)	92.1	90.2	87.5	85.5	82.1	75.0	71.9

<sup>a</sup> Accuracy is defined as the average of the sensitivity and specificity.

87, 108, 129, 162, and 192 bp. For each fragment length, 2000 coding and 2000 noncoding sequences are used as the training set, and the remaining 2000 coding and 2000 noncoding sequences as the test set. The accuracy is calculated based on the sequences in the test set. The resulting accuracy for various fragment lengths is listed in Table V. The result is also shown clearly in Figure 2. It is seen that the accuracy of an individual variable is dependent on sequence lengths obviously. When the length of sequences is shortened, the accuracy for the variable  $x_4$  reduces drastically, whereas the accuracy of the other variables reduces slowly.

### The Rank of Importance of the Four Variables

To study the rank of importance of the four variables, one variable is removed in turn each time, and we would like to see what happens for the accuracy. In other words, only three variables are applied to recognize coding sequences each time. The coding and noncoding sequences of the database are divided into two equal parts: part 1 and part 2. Each includes 2000 coding and 2000 noncoding sequences, respectively. Repeat the recognition procedure mentioned above for sequences with various lengths. As usual, part 1 is taken as a training set and part 2 as a test set. The

above procedure is then repeated by reversing the roles of the two parts, that is, part 2 is taken as the training set and part 1 as the test set. The average accuracy is listed in Table VI and intuitively shown in Figure 3. It can be seen that the accuracy after deleting the variable  $x_4$  apparently becomes much lower than that deleting other variables, i.e., the distribution of stop codons in three frames plays an important role in recognizing coding sequences. The same conclusion can be obtained for recognizing coding sequences from intergenic sequences. The results in Table V and VI indicate that  $x_4$  is the most important variable for recognizing shorter sequences. The variables  $x_1$  and  $x_2$  are the next. The variable  $x_3$  is the least important. Consequently, the rank of importance of the four variables may be arranged as follows:  $x_4$ ,  $x_1$ ,  $x_2$ , and  $x_3$ .

### Frame Independence of the Algorithm

It is well known that knowing the correct reading frame of coding regions would make the recognition of coding sequences easier. We should emphasize that the variables defined in the present algorithm are independent of the reading frames of sequences. The variable  $x_4$  contains the information on the distributions of the three stop codons along the sequence in all three phases. It is obvious that the variable  $x_1$  is

**Table IV The Average Sensitivity, Specificity, and Accuracy<sup>a</sup> for Coding/Intergenic Sequence Recognition, Averaged Over Sixfold Cross-Validation Tests for Various Fragment Lengths**

Fragment length (bp)	192	162	129	108	87	63	42
Sensitivity (training)	92.6	90.6	88.0	85.9	83.4	80.0	74.3
Specificity (training)	92.6	90.6	88.0	85.9	83.4	80.0	74.3
Accuracy (training)	92.6	90.6	88.0	85.9	83.4	80.0	74.3
Sensitivity (test)	92.9	90.9	88.2	86.9	83.5	79.8	74.1
Specificity (test)	91.9	90.1	87.6	84.7	82.1	78.6	73.5
Accuracy (test)	92.4	90.5	87.9	85.8	82.8	79.2	73.8

<sup>a</sup> Accuracy is defined as the average of the sensitivity and specificity.

**Table V The Accuracy<sup>a</sup> of the Algorithm Using Only One Individual Variable<sup>a</sup>**

Fragment length (bp)	192	162	129	108	87	63	42
Accuracy (%) $x_1$	80.75	79.29	77.50	75.98	74.31	71.39	67.18
Accuracy (%) $x_2$	79.60	79.00	78.06	75.88	75.16	70.64	67.33
Accuracy (%) $x_3$	63.60	62.09	60.90	59.55	57.59	55.53	55.16
Accuracy (%) $x_4$	86.67	83.15	78.53	75.96	72.34	71.60	66.35

<sup>a</sup> The accuracy is obtained based on a test set including 2000 coding and 2000 noncoding sequences. See the definition of the four variables  $x_1$ ,  $x_2$ ,  $x_3$ , and  $x_4$  in the Algorithm Section.

defined without knowing the phase in advance. Similarly, the variable  $x_3$  is defined as the maximum of purine contents of all three phases. It is also phase independent. According to the work,<sup>17</sup> the variable  $x_2$  is independent of the phases. Therefore, all the four variables and the whole algorithm presented here are phase or frame independent.

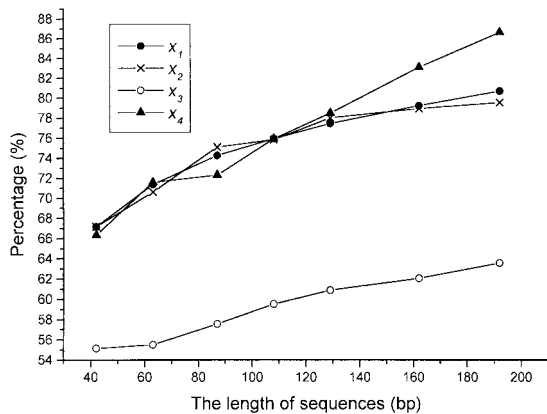
### The Comparison Between the Recognition of Coding/Noncoding and Coding/Intergenic Sequences

Table III and IV show that the algorithm can well distinguish coding fragments from noncoding or intergenic fragments. On the other hand, the accuracy for the recognition of coding/intergenic fragments is slightly higher than that of coding/noncoding fragments with lengths shorter than 87 bp. For longer fragments, the accuracy between the two recognition cases is roughly identical. This seems to come from the stop-codon variable. The A+T content in inter-

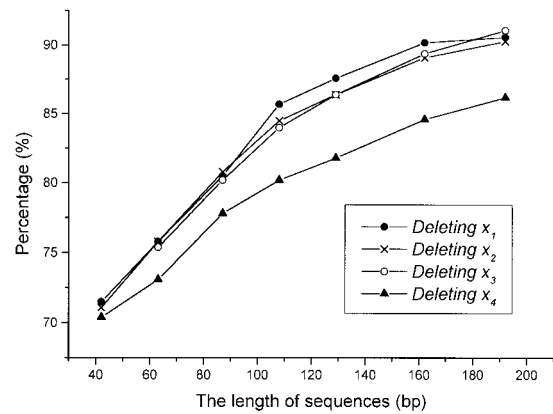
genic sequences is generally higher than that in gene regions.<sup>26</sup> On average, the content of stop codons in intergenic regions is relatively higher. Table II shows that the number of intergenic sequences containing the stop codons is obviously higher than that of non-coding sequences containing the stop codons in all three frames. On the other hand, it is well known that there are many pseudogenes in intergenic regions of prokaryotic and eukaryotic DNA sequences.<sup>27-29</sup> Harrison et al.<sup>28</sup> have analyzed pseudogenes in the worm and found that there is about one pseudogene for every eight genes. Pseudogenes also exist in each human chromosome. In general, pseudogenes have low G+C content, premature stop codons or frame-shift. The stop codons are easily found in all three frames. The variables of the algorithm are sensitive to the properties of pseudogenes. The algorithm would be helpful for analyzing pseudogenes.

### Comparison with Other Algorithms

It is of interest to compare the present algorithm with others. Two well-known algorithms currently avail-



**FIGURE 2** Relation between the accuracy of the algorithm and sequence length, when each individual variable is used separately. The x axis represents the sequence length, the y axis the accuracy. The ●, ×, ○, and ▲ denote the accuracy when the variable  $x_1$ ,  $x_2$ ,  $x_3$ , and  $x_4$  are used separately.



**FIGURE 3** Relation between the accuracy of the algorithm and sequence length, when each variable is removed separately. The x axis represents the sequence length, the y axis the accuracy. The ●, ×, ○, and ▲ denote the accuracy when the variable  $x_1$ ,  $x_2$ ,  $x_3$ , and  $x_4$  are in turn removed, respectively.

**Table VI The Accuracy of the Algorithm Using Three Variables Each Time<sup>a</sup>**

Fragment length (bp)	192	162	129	108	87	63	42
Accuracy (%) $x_1, x_2, x_3$	86.2	84.6	81.8	80.2	77.8	73.1	70.4
Accuracy (%) $x_1, x_2, x_4$	91.1	89.4	86.4	84.0	80.2	75.4	71.3
Accuracy (%) $x_1, x_3, x_4$	90.3	89.1	86.4	84.5	80.8	75.8	71.1
Accuracy (%) $x_2, x_3, x_4$	90.6	90.2	87.6	85.7	80.5	75.8	71.5

<sup>a</sup> See the note in Table V.

able, i.e., those of lengthen-shuffling FFT<sup>17</sup> and the Markov chain model,<sup>8</sup> are compared. We rewrite the computer programs for both algorithms. The training set consists of 2000 positive and negative samples, and the test set consists of another 2000 positive and negative samples, respectively. There are various sequence lengths in both sets; the details are listed in Table VII. Because the limited size of the training sets does not allow to train higher order model, only the fourth-order Markov chain model is used. Refer to Table VII. Compared with the lengthen-shuffling FFT algorithm, the accuracy of the present algorithm increases greatly, 11% higher than that of the former for the fragment length of 192 bp. The coding sequences in the database are selected from the sense strands of genes at the above website. The nonhomogeneous periodic Markov chain model is utilized for training and testing coding sequences in the sense strands and the homogeneous Markov chain model for noncoding sequences. The accuracy of the present algorithm is slightly higher than that of the Markov chain model for sequence lengths of 192 and 162 bp. For the shorter sequences, the accuracy of the present algorithm is relatively lower. This is the result that the way of extracting information from sequences for the two algorithms is different. The fourth-order Markov chain model makes full use of the local statistical characteristics of base distribution in three frames of coding sequences. The model uses more than five thousands parameters to describe local coding characteristics. On the other hand, the present algorithm uses only four parameters, mainly the stop-codon variable derived from the statistics of three stop

codons in the three phases along the sequence. Obviously, the present algorithm is much more simple and reliable. The biological meanings of the four variables are clear. The major contribution of this paper is to emphasize the importance of the statistics of stop codons in gene-finding issue. In collaboration with the stop-codon variables into other algorithms, it would be possible to raise their gene-finding accuracy.

### Apply the Algorithm to Recognize Human Genes

As an example, we would like to show how the present algorithm is used to recognize coding regions of human DNA sequences. In order to have a more reliable result, the training set consists of all 4000 coding and noncoding sequences, respectively, in the database. The linear discriminative coefficients  $C_i$  ( $i = 1, 2, 3, 4$ ) are calculated, and the threshold  $U_0$  can be determined. For convenience, instead of the linear discriminative function  $F(\mathbf{X})$ , a  $F$  value is introduced. For a given sequence, the criterion to make the decision of coding/noncoding is decided by the value of  $F$ , defined as

$$F = \frac{[F(\mathbf{X}) - F_{\min}]}{(F_{\max} - F_{\min})} \quad (1)$$

where  $F(\mathbf{X}) = \mathbf{C} \cdot \mathbf{X}$ , and  $F_{\min}$  and  $F_{\max}$  are the minimum of the function  $F(\mathbf{X})$  for positive samples and the maximum of the function  $F(\mathbf{X})$  for negative samples respectively. To make sure that  $F \in [0, 1]$ , let

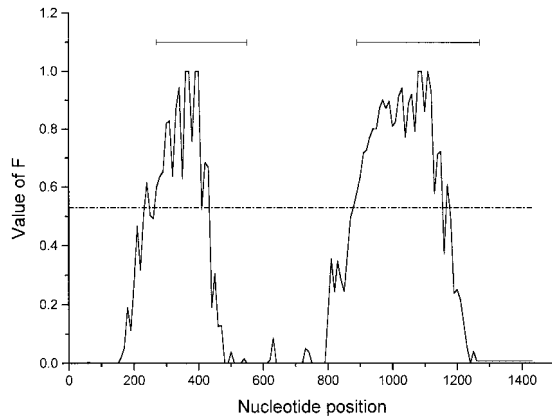
**Table VII The Average Accuracy of Three Algorithms with Various Lengths**

Fragment length (bp)	192	162	129	108	87	63	42
Accuracy <sup>a</sup> (%)	92.4	90.8	87.2	85.5	82.7	75.1	72.1
Accuracy <sup>b</sup> (%)	81.6	80.7	78.1	77.00	76.00	70.6	67.3
Accuracy <sup>c</sup> (%)	91.4	90.1	89.5	87.5	85.0	81.4	74.6

<sup>a</sup> The present algorithm.

<sup>b</sup> The lengthen-shuffling FFT algorithm.<sup>17</sup>

<sup>c</sup> The fourth order Markov chain model.



**FIGURE 4** Example for recognizing coding regions of human gene HSODF2. The  $x$  axis indicates the base positions referred to the start points of 162 bp windows, the  $y$  axis the  $F$  value. The dash-dot line shows the threshold  $F_0$  value (0.53). The short lines represent the coding regions predicted.

$F(\mathbf{X}) = F_{\max}$ , if  $F(\mathbf{X}) > F_{\max}$ ; otherwise, let  $F(\mathbf{X}) = F_{\min}$ , if  $F(\mathbf{X}) < F_{\min}$ . Now use a sliding window technique to recognize coding regions in human DNA sequences. The window size is 162 bp, and moves across the sequence at 10 bp each step. For a given sequence, the 162 bp windows are used successively starting from the first base of the sequences. The four variables of the algorithm are calculated for each window, then the  $F$  value is computed. Consequently, a series of  $F$  values are obtained along the sequence. A threshold  $F_0$  can be used to make the decision of coding/non-coding, depending on  $F > F_0$  or  $F < F_0$ , where  $F_0$  is given by

$$F_0 = \frac{(U_0 - F_{\min})}{(F_{\max} - F_{\min})} \quad (2)$$

Using the values of  $U_0$ ,  $F_{\max}$ , and  $F_{\min}$ , we find  $F_0 = 0.53$ . Furthermore, the successive points with  $F > F_0$  form some regions. Then, search the three stop codons in these regions in all three frames, and the longest fragments without stop codons in one frame are predicted as coding regions. The graphic output of recognizing coding regions in the human gene HSODF2 sequence is shown in Figure 4. The sequence contains two exons (280–599; 843–1275). Consequently, two coding fragments are predicted, which lie in (270–552; 890–1272).

## CONCLUSION

Despite of the great efforts devoted to developing algorithms of gene recognition for more than

twenty years, the problem remains unsolved completely. Recently, the test of recognizing human genes from newly sequencing genome sequences indicates that more than 50% exons cannot be correctly identified, even using the best algorithms currently available. Because the average length of exons of vertebrate genes is only 137 bp,<sup>23</sup> shorter exons are “un-visible” by many of the current algorithms. There is an anxious need to look for new and powerful algorithms for recognizing shorter coding regions in the human genes. This paper represents an attempt to solve the issue in some sense. It can be seen that the recognition accuracy by the present algorithm is greatly raised compared with those designed specially for recognizing shorter coding regions available now. The success seems to come from the utilization of the statistics of stop codons in all three phases along the sequence. To our knowledge, this would be for the first time that such an idea is proposed in the gene-finding field. It is hoped that the proposed algorithm would be useful to improve the accuracy of the algorithms widely used nowadays, if the statistical distributions of stop codons could be incorporated appropriately into them.

We thank Ren Zhang for helping in organizing the database used here. The present study was supported in part by the 973 Project of China (grant G1999075606).

## REFERENCES

1. Fickett, J. W. *Nucleic Acids Res* 1982, 10, 5303–5318.
2. Staden, R. *Nucleic Acids Res* 1984, 12, 505–519.
3. Staden, R.; McLachlan, A. D. *Nucleic Acids Res* 1982, 10, 141–156.
4. Farber, R.; Lapedes, A.; Sirotkin, K. *J Mol Biol* 1992, 226, 471–479.
5. Tsonis, A. A.; Elsner, J. B.; Tsonis, P. A. *J Theor Biol* 1991, 151, 323.
6. Tiwari, S.; Ramachandran, S.; Bhattacharya, A.; Bhattacharya, S.; Ramaswamy, R. *Comp Appl Biosci* 1997, 13, 263–270.
7. Fickett, J. W. *Comput Chem* 1996, 10, 103–118.
8. Borodovsky, M.; Mcininch, J. *Comput Chem* 1993, 17, 123–133.
9. Bursset, M.; Guigo, R. *Genomics* 1996, 34, 353–367.
10. Zhang, M. Q. *Proc Natl Acad Sci USA* 1997, 94, 565–568.
11. Salzberg, S. L.; Delcher, A.; Kasif, S.; White, O. *Nucleic Acids Res* 1998, 26, 544–548.
12. Salzberg, S. L.; Delcher, A.; Fasman, K.; Henderson, J. *J Comput Biol* 1998, 5, 667–680.

13. Lukashin, A. V.; Borodovsky, M. *Nucleic Acids Res* 1998, 26, 1107–1115.
14. Li, Wentian *Comput Chem* 1999, 23, 283–301.
15. Zhang, C.-T.; Wang, J. *Nucleic Acids Res* 2000, 28, 2804–2814.
16. Thanaraj, T. A. *Nucleic Acids Res* 2000, 28, 744–754.
17. Yan, M.; Lin, Z.-S.; Zhang, C.-T. *Bioinfo* 1998, 14, 1–5.
18. Zhang, C.-T.; Chou, K.-C. *J Protein Chem* 1993, 12, 329–335.
19. Chou, K.-C.; Zhang, C.-T. *AIDS Res Human Retro* 1992, 8, 1967–1976.
20. Silverman, B. D.; Linsker, R. *J Theor Biol* 1986, 118, 295–300.
21. Trifonov, E. N. *J Mol Biol* 1987, 194, 643–652.
22. Lio, P.; Ruffo, S.; Buiatti, M. *J Theor Biol* 1994, 171, 215–223.
23. Hawkins, J. D. *Nucleic Acids Res* 1988, 16, 9893–9908.
24. Zhang, M. Q. *Human Mol Genetics* 1998, 7, 919–932.
25. Mardia, K. V.; Kent, J. T.; Bibby, J. M. *Multivariate Analysis*; Academic Press: London, 1979.
26. Guigo, R.; Fickett, J. W. *J Mol Biol* 1995, 253, 51–60.
27. Suckow, J. M.; Amano, N.; Ohfuku, Y.; Kakinuma, J.; Koike, H.; Suzuki, M. *FEBS Letters* 1998, 86, 86–92.
28. Harrison, P. M.; Echols, N.; Gerstein, M. B. *Nucleic Acids Res* 2001, 29, 818–830.
29. Goncalves, I.; Duret, L.; Mouchiroud, D. *Genome Res* 2000, 10, 672–678.