Prediction of Exonic and Intronic Regions with Variants of Coding Measures Based on Kasiski Method

Laura Cruz, Héctor Fraire, Yolanda González, Javier González, Manuel Aguilar*, and Alejandro Macías

Departamento de Posgrado e Investigación
Instituto Tecnológico de Ciudad Madero
Juventino Rosas y Jesús Urueta, Colonia los Mangos
Ciudad Madero, Tamaulipas, México
Tel-Fax:(833) 215-85-44
lauracruzreyes@yahoo.com

Abstract. Prediction of exonic and intronic regions is an important problem of bioinformatics, which has been solved with a set of medium accuracy coding measures. This paper presents a novel methodology to increase the prediction accuracy. We propose to adapt the cryptanalytic Kasiski Method to the analysis of a DNA sequence Cr. The result of this process is incorporated into coding measures, to predict regions of a DNA sequence C, such as $C \in Cr$. To validate our approach we obtained variants of three classical coding measures: codon usage, amino acid usage, and codon preference. The new measures were applied to the β-globin gene. We obtained an increase of 30% in the prediction.

1 Introduction

Eukaryote gene structure mainly contains protein-coding regions called exons and non-protein-coding regions called introns [1]. The use of coding measures is one of the most useful approaches to characterize a DNA sequence. There are applications which predict protein-coding regions, those use coding measures, dynamic programming, neural networks[2], Fourier analysis [3] or combinations of them. Some examples of those applications are GeneMark [4], Genie [5], and GenScan [6].

In the context of information security, cryptanalysis methods are used to decode a ciphertext without the necessary information, one of them is the Kasiski Method (KM). In this paper we propose a methodology that combines coding measures with KM to improve the prediction accuracy of exons and introns of a sequence C. That can be possible by replacing in each coding measure the random model of coding DNA with the codons distribution of a sequence Cr, such as $C \in Cr$, based on the assumption that non-coding DNA is not completly random. We improve the prediction of exons and introns with a variant of codon

* This project was supported by Conacyt

© A. Gelbukh, C. Yáñez Márquez, O. Camacho Nieto (Eds.)
Advances in Artificial Intelligence and Computer Science
Research on Computing Science 14, 2005, pp. 323-334

preference measure. In the following sections we describe three coding measures, KM, the analysis used to adapt the KM to coding measures, the methodology used in the proposed approach, experimentation, and conclusions.

2 Coding Measures

A coding statistic measure can be defined as follows: Given a codon sequence C, to calculate a real number which implies the likelihood of the given sequence be coding for a protein. There are many coding statistic measures such as codon usage bias, base compositional bias between codon positions, and periodicity in base occurrence. In this section we describe three coding measures: codon usage, amino acid usage, and codon preference; this description is based on [7].

2.1 Codon Usage(CU)

According to the global codon frequency of a species genome, it is possible to calculate the coding potential of a DNA sequence zone by the comparison of its codon frequency read in a given frame [8]. Let $F(c)$ be the probability of codon c in the considered species (F is the probability in a usage codon table), be in a sequence that code for a protein. Given an adjacent codons sequence $C = C_1C_2C_3...C_m$ of size m, the probability $P^i(C)$ to find the codons sequence C read in a frame i, knowing that C codes for a protein, is determined with Equation 1.

$$P^i(C) = F(C_1^i)F(C_2^i)F(C_3^i)\dots F(C_m^i)\ 1 \leq i \leq 3 \tag{1}$$

On the other hand, let $F_0(c)$ be the probability of codon c in a non-coding sequence. Then the probability $P_0(C)$ to find C, knowing that C does not code for a protein, is calculated with Equation 2. Like in a random model of coding DNA, for each codon is given an equal probability $F_0(c) = \frac{1}{64}$ (there are 64 codons of the genetic code).

$$P_0(C) = F_0(C_1)F_0(C_2)F_0(C_3)\dots F_0(C_m) \tag{2}$$

In Equation 3, the log-likelihood ratio of codons sequence C representing its coding potential in frame i is shown.

$$LP^i(C) = \log \frac{P^i(C)}{P_0(C)} \tag{3}$$

If $LP^i(C) > 0$, then the likelihood of C to be coding in the frame i is higher than assuming that C to be non-coding in frame i. If $LP^i(C) < 0$, then the likelihood of C to be non-coding in frame i is higher than assuming that C to be coding in frame i.

2.2 Amino Acid Usage(AAU)

This coding measure is computed by amino acid bias based on the observed frequencies of a single amino acid in an existing protein [9]. The probability $F_A(c)$ of the a codon c to be traslated to an oligopepetide, is obtained by summing up the probabilities of all the synonymous codons to c, as is shown in Equation 4.

$$F_A(c) = \sum_{\forall c' \equiv c} F(c') \tag{4}$$

Where $\forall c' \equiv c$ means for all codon c' synonymous to c. In Equation 5, it is shown the probability P_A^i to find the amino acids sequences resulting of the translation C in the frame i, knowing that C codes for a protein.

$$P_A^i(C) = F_A(C_1^i)F_A(C_2^i)F_A(C_3^i)\ldots F_A(C_m^i)\ 1 \leq i \leq 3 \tag{5}$$

Like in a non-coding DNA model, we assume the probability of each amino acid to be proportional to the number synonymous codons is $F_{A0}(c) = \frac{n_c}{64}$, where n_c is the number of synonymous codons to c. With this, it is possible to calculate the probability $P_{A0}(C)$ to find C, knowing that C does not code for a protein in a similar way of Equation (2). The coding potential $LP_A^i(C)$ of C in a given frame i can be calculated in a similar way of Equation (3).

2.3 Codon Preference(CP)

In [10] a coding statistic to measure the uneven usage of synonymous codons is introduced. This value can be taken directly from the codon usage table. The relative probability of each synonymous codon to code for a given amino acid can also be computed. Let $F_R(c)$ be the relative probability in coding regions of codon c among all the synonymous codons to c, see Equation 6.

$$F_R(c) = \frac{F(c)}{F_A(c)} \tag{6}$$

The probability $P_R^i(C)$ to find C given the particular amino acids sequence coded by C in the frame i is computed with the Equation 7. Like in a non-coding DNA model, there is no preference between synonymous codons to code a given amino acid. Hence, the probability of codon c in non-coding DNA is $F_{R0}(c) = \frac{1}{n_c}$. The probability $P_{R0}(C)$ to find C, knowing that C does not code for a protein is computed in a similar way of Equation (2). The coding potential $LP_R^i(C)$ of C in a given frame i can be calculated in a similar way of Equation (3).

$$P_R^i(C) = F_R(C_1^i)F_R(C_2^i)F_R(C_3^i)\ldots F_R(C_m^i)\ 1 \leq i \leq 3 \tag{7}$$

3 Adaptation of Kasiski Method

A brief description: The KM can be used to identify the length of the keyword used to create a chiphertext message encrypted by Vigenere cipher. The main idea of this method is the observation that any two identical substrings of the plaintext will encrypt to identical ciphertext substrings when they are both aligned equally relative to the keyword boundaries. That is to say that the distance between the starting position of each substring is a multiple of the keyword length. In converse, if we identify multiple substrings of the ciphertext that are identical and of a sufficient length (say three or more letters long) it is highly probable that they are encryptions of identical plaintext. Thus, the length of the keyword must divide the greatest common divisor of the differences in starting position of each of the ciphertext substrings [11].

As mentioned in Section 2, the selected coding measures work with codons of C to compute their coding potentials respectively. On the other hand, DNA constitutes information to be decoded but it is not approached by cipher-decipher algorithm or keywords. We are going to use the exhaustive search of trigrams of KM to get the distribution of codons of Cr to combine with the coding measures. Therefore, we have to get the probabilities $F'(c)$, $F'_A(c)$ and $F'_R(c)$ from the resulting information of the KM applied to Cr (Table 2). But in this case, $F'(c)$ represents the probability of the codon c appears into Cr, $F'_A(c)$ and $F'_R(c)$ are homologous to $F_A(c)$ and $F_R(c)$ respectively, in addition, they are realtive to $F'(c)$.

To get the distribution of codon of Cr we are going to analize the following three main steps of KM:

- Step 1. To analyze the repeated triplets in the ciphertext, keeping their starting position.
- Step 2. To calculate the difference between starting position of the repeated triplets of the ciphertext.
- Step 3. To calculate the greatest common divisor of the calculated differences, will be the length of the ciphertext's key.

In Table 1, it is shown the output of the step 1 of the KM applied to a ciphertext. After that, it is necessary to apply steps 2 and 3 to compute the length of the keyword. To adapt the searching of KM to coding measures, we need to realize that the KM will analyze a DNA sequence instead of ciphertext and it will not compute the length of any keyword, because there is no keyword in DNA, then steps 2 and 3 are deleted. We are going to use the distribution of codons of Cr to replace the random model of coding DNA from the coding measures with the purpose to improve the estimation of coding potential of DNA sequence. To get the Table 2 we have to modify the Table 1, the Triplet and Occurrence columns are required (the label Triplet is changed by Codon), but the Positions column is removed because the position of a codon in the coding measures is indifferent. Moreover, a column called Amino Acid is added manually according to the Codon column, resulting in the new Table 2, making

it possible to calculate $F'(c)$, $F'_A(c)$ and $F'_R(c)$, which are used in the coding measures variants described in the next section.

Table 1. Output of step 1 of KM applied to a ciphertext.

Triplet	Ocurrence	Positions
T_1	O_1	$P_1, ...P_i; i \geq 1$
..	..	..
T_N	O_N	$P_1, ...P_i; i \geq 1$

Table 2. Output of KM applied to DNA sequence.

Codon	Ocurrence	Amino acid
C_1	O_{C_1}	A_{C_1}
..	..	..
C_{64}	$O_{C_{64}}$	$A_{C_{64}}$
	$TotalOcurrence$	

4 The Variants of Coding Measures based on KM

Our proposal: Given two codons sequences C and Cr, such as, $C \in Cr$, given one of the coding measures. For example, in case of codon usage, to replace the random model of coding DNA $P_0(C)$ by the probability $P^i_{Cr}(C)$ to find C and being just a sequence into Cr. In the the other two cases $P_{A0}(C)$ and $P_{R0}(C)$ are replaced in an homologous way. These variants are designed to discriminate C between being a sequence that codes for a protein or being just a sequence into Cr.

4.1 Variant of Codon Usage based on KM(VCUKM)

VCUKM similary to codon usage [8], according to the global codon frequency of a species genome, it is possible to calculate the coding potential of a DNA sequence zone by the comparison of its codon frequency read in a given frame into Cr. The probability $F'(c)$ to appear the codon c into Cr is described in Equation 8. The probability $P^i_{Cr}(C)$ to find C read in the frame i into Cr is computed as in Equation 9 (where $TotalOcurrence$ and O_c are taken from Table 2).

$$F'(c) = TotalOcurrence/O_c \tag{8}$$

$$P^i_{Cr}(C) = F'(C^i_1)F'(C^i_2)F'(C^i_3)\ldots F'(C^i_m)\ 1 \leq \mathrm{i} \leq 3 \tag{9}$$

In Equation 10, the log-likelihood ratio of C representing its coding potential in frame i is shown.

$$LP^i_{Cr}(C) = \log \frac{P^i(C)}{P^i_{Cr}(C)} \tag{10}$$

If $LP^i_{Cr}(C) > 0$, then the likelihood of C to be coding in the frame i is higher than assuming that C to be just a sequence into Cr in frame i. If $LP^i_{Cr}(C) < 0$, then the likelihood of C to be just a sequence into Cr in frame i is higher than assuming that C to be coding in frame i.

4.2 Variant of Amino Acid Usage based on KM(VAAUKM)

Similarly to amino acid usage [9], VAAUKM is a coding measure that is computed by amino acid bias based on the observed frequencies of a single amino acid in Cr. The probability $F'_A(c)$ of the codon c, to be traslated to an oligopeptide into Cr, is the summing up of the probabilities of all the synonymous codons to c to appear into Cr, as described in Equation 11.

$$F'_A(c) = \sum_{\forall c' \rightarrow c} F'(c') \tag{11}$$

The probability $P^i_{A-Cr}(C)$ to find C read in the frame i into Cr is computed as in Equation 12. $LP^i_{A-Cr}(C)$ is computed in a similar way of Equation (10).

$$P^i_{A-Cr}(C) = F'_A(C^i_1)F'_A(C^i_2)F'_A(C^i_3)\ldots F'_A(C^i_m)\ 1 \leq i \leq 3 \tag{12}$$

4.3 Variant of Codon Preference based on KM(VCPKM)

Similarly to codon preference [10], VCPKM is a coding statistic to measure the uneven usage of synonymous codons into Cr. Let $F'_R(c)$ be the relative probability in Cr of the codon c among all the synonymous codons to c, see Equation 13. The probability $P^i_{R-Cr}(C)$ to find C read in the frame i into the Cr is computed as in Equation 14. $LP^i_{R-Cr}(C)$ is computed in a similar way of Equation (10).

$$F'_R(c) = \frac{F'(c)}{F'_A(c)} \tag{13}$$

$$P^i_{R-Cr}(C) = F'_R(C^i_1)F'_R(C^i_2)F'_R(C^i_3)\ldots F'_R(C^i_m)\ 1 \leq i \leq 3 \tag{14}$$

5 Design of a Methodology to Predict Exons and Introns

Our proposed methodology has the main objective to predict exons and introns in a DNA sequence by the introduced variants in Section 4. This methodology has the following 4 steps:

1. To Give two codons sequence C and Cr, such as $C \in Cr$.
2. To Apply the first step of the KM to Cr (See Fig. 2).
3. To compute the coding potentials of the codons sequence C with the variants of coding measures proposed in Section 4, by adding the additional information of KM applied to Cr (See Fig 4).

Details of Pseudo-codes

In this section the pseudo-code of the Fig. 1, 2, 3 are described. In Fig. 1 the Kasiski procedure with the parameters Cr and N is shown. It is an exhaustive search of overlapped N-grams with a complexity of $O(n^2)$, in this case, $N = 3$ because we are going to search for triplets in Cr.

```
1 Kasiski(Cr, N)
2 begin
3 For index1 = 1 to length − N
4     unique ← true
5     If index1 = length − N
6         Call Search_and_Save_N-gram(index1, ,Cr,N)
7     End If
8     For index2 = index1 + 1 to length − N
9         counter ← 0
10        While Cr[index1 + counter] = Cr[index2 + counter] and counter! = −1
11            Increment counter
12            If counter = N
13                Call Search_and_Save_N-gram(index1,index2,Cr,N)
14                unique ← false
15                counter ← −1
16            End If
17        End While
18    End For
19    If unique = true
20        Call Search_and_Save_N-gram(index1, ,Cr,N)
21    End If
22 End For
23 end
```

Fig. 1. Pseudo-code of exhaustive search of Kasiski Method applied to Cr

Where *length* is the number of nucleotides of Cr, in lines 3 and 8, $index_1$ and $index_2$ are used as indexes to compare the nucleotides along Cr. That means, for

each $index_1$ from 1 to $length - N$ and for each $index_2$ from $i+1$ to $length - N$ are compared $Cr[index_1]$ with $Cr[index_2]$, $Cr[index_1 + 1]$ with $Cr[index_2 + 1]$ and $Cr[index_1 + 2]$ with $Cr[index_2 + 2]$, to know if $Cr[index_1 \ldots index_1 + 2]$ and $Cr[index_2 \ldots index_2 + 2]$ are the same triplet. There is a while loop in line 10, to find overlapped N-grams with the composed condition $Cr[index_1 + counter] = Cr[index_2 + counter]$ and $counter! = 0$, if it is satisfied to *true*, the *counter* is incremented, when $counter = N$ there are two N-grams starting in the positions $index_1$ and $index_2$. Then in line 13, the procedure **Search_and_Save_N-gram**($index_1, index_2, Cr, N$) is invoked to keep those N-grams and their positions $index_1$ and $index_2$. In lines 5-7, the procedure **Search_and_Save_N-gram**($index_1, , Cr, N$) without the second index is invoked to keep the last N-gram of Cr in the position $length - N$, in case of this N-gram has not been saved yet. Also, in lines 19-20, the same procedure without the second index is invoked to keep a N-gram that is not repeated along Cr because *unique* had never changed to false in line 14. It means that this N-gram is unique in Cr.

```
 1 Search_and_Save_N-gram(index1,index2,Cr,N)
 2 begin
 3 If Cr[index1 + N] ∈ list.N-gram
 4     If index2 != NULL
 5         If index2 ∈ actual-node.positions-list
 6             For actual-node
 7                 Add position index2 to positions-list
 8                 Increment ocurrence
 9             End For
10         End If
11     End If
12 Else
13     list ← new-node
14     For new-node
15         N-gram ← Cr[index1 + N]
16         Add position index1 to positions-list
17         Increment ocurrence
18         If index2 != NULL
19             Add position index2 to positions-list
20             Increment ocurrence
21         End If
22     End For
23 End If
24 end
```

Fig. 2. Pseudo-code of the Search and Save N-gram procedure

In Fig. 2 the pseudo-code of **Search_and_Save_N-gram** procedure with the parameters $index_1$, $index_2$, Cr and N is shown. This procedure allows to keep

the N-grams and their starting positions avoiding to keep a N-gram previously saved. If $index_2$ is not NULL there are two N-grams to keep. Otherwise, there is only one N-gram to keep. *list* is a linked list with the fields *N-gram*, *ocurrence* and another linked list *positions-list*. In line 3, if $Cr[index_1 \ldots index_1 + N]$ is found in any node in *list* with $N\text{-}gram = Cr[index_1 \ldots index_1 + N]$, and if $index_2$ is not NULL and it is in *positions-list*, then for the *actual-node* the position $index_2$ is added to the *positions-list* and *ocurrence* is increased. If $Cr[index_1..index_1+N]$ is not found in any node in *list* with $Cr[index_1 \ldots index_1+N]$, then in line 13 a *new-node* for *list* is created. For this *new-node*, in lines 15-17 respectively, $Cr[index_1 \ldots index_1 + N]$ is assigned to the field *N-gram*, position $index_1$ is added to *positions-list* and *ocurrence* is increased. Additionally for *new-node* in lines 18-21, if $index_2$ is not NULL, then $index_2$ is added to *positions-list* and *ocurrence* is increased again. At the end, the first and second column of Table 2 will be filled with the information contained in each node from *list* with its fields *N-gram* and *ocurrence*. The third column of Table 2 is filled manually.

```
 1 Coding_Potencials(C,size-window,n)
 2 begin
 3 For each frame i of C
 4   For each codon j of C
 5     P, P_A, P_R, P_Cr, P_A-Cr, P_R-Cr ← 1
 6     While j % size-window != 0
 7       For each n codons
 8         P ← ×F(C_j^i)
 9         P_A ← ×F_A(C_j^i)
10         P_R ← ×F_R(C_j^i)
11         P_Cr ← ×F'(C_j^i)
12         P_A-Cr ← ×F'_A(C_j^i)
13         P_R-Cr ← ×F'_R(C_j^i)
14       End For each
15       LP[i][j] ← log(P / P_Cr)
16       LP_A[i][j] ← log(P_A / P_A-Cr)
17       LP_R[i][j] ← log(P_R / P_R-Cr)
18       Positions[i][j] ← (j × 3) + (i − 1)
19     End While
20   End For
21 End For each
22 end
```

Fig. 3. Procedure to compute the coding potencials with additional information of Kasiski method.

In Fig. 3 is shown the pseudo-code of Coding Potentials procedure with the parameters C, *size-windows*, and n. This procedure computes the coding

potentials of the variants of the coding measures. For each frame $1 \leq i \leq 3$ (line 3) of the codons sequence C and for each codon of C (line 4), but measured in windows with size of *size-window* codons (line 6) and steps of n bp (line 7). In lines 8, 9, and 10 the probabilities P^i, P^i_A and P^i_R are computed respectively, as in Section 2. In lines 11, 12, and 13 the probabilities P^i_{Cr}, P^i_{A-Cr} and P^i_{R-Cr} are computed respectively, as in Section 4. In the lines 15, 16, and 17 the coding potentials of variants of coding measures are computed. To know if a region is coding for the given LP, you must compute the maximum value among its three frames, if this is higher than 0 the probability to be a coding region is higher than assuming it like a non-coding region, otherwise the probability to be a non-coding region is higher than assuming it like a coding region. Finally in line 18 keep the final position of each window and frame.

6 Experiments

Table 3 shows the accuracy in percentages of prediction of exons and introns of a sequence of Human β-globin region(which is procesed by KM) on chromosome 11 from posistions 62001-64000 from the Genbank [12].

Table 3. Acurracy Prediction of exons and introns

Cr	Human β-globin region on chromosome 11=73308 bp		
C	sequence from positions 62001-64000 containing the β-globin gen		
	Accuracy Prediction of		
	Exon 1: 62137-62278	Exon 2: 62409-62631	Exon 3: 63482-63742
CU	**91%**	**100%**	86%
AAU	66%	63%	68%
CP	**100%**	**100%**	**100%**
VCUKM	50%	**95%**	72%
VAAUKM	41%	70%	59%
VCPKM	**100%**	**95%**	**90%**
	Intron 1: 62279-62408	Intron 2: 62632-63481	Intron 3: 63482-64000
CU	33%	82%	80%
AAU	22%	61%	20%
CP	10%	32%	65%
VCUKM	20%	**97%**	**95%**
VAAUKM	10%	86%	85%
VCPKM	60%	86%	**95%**

The accuracy of prediction of three exons and three introns of original coding measures and their variants are shown. We considered a threshold of false positive to prediction of exons and introns of 10%. CP and VCPKM predicted the three exons, CU predicted two exons. VCPKM predicts the third intron and CP and VCUPK predicts two introns. Guigó [7] plots the coding potencial of several

coding measures, of which the best one was codon usage, although he considered the first exon on the position 62187 instead of the real position 62137 obtaining a fairly spectrum of the three exons. In Fig 4. we plot the coding potentials of all coding measures. We plotted each coding measure with *window-size* = 120pb and n = 12pb.

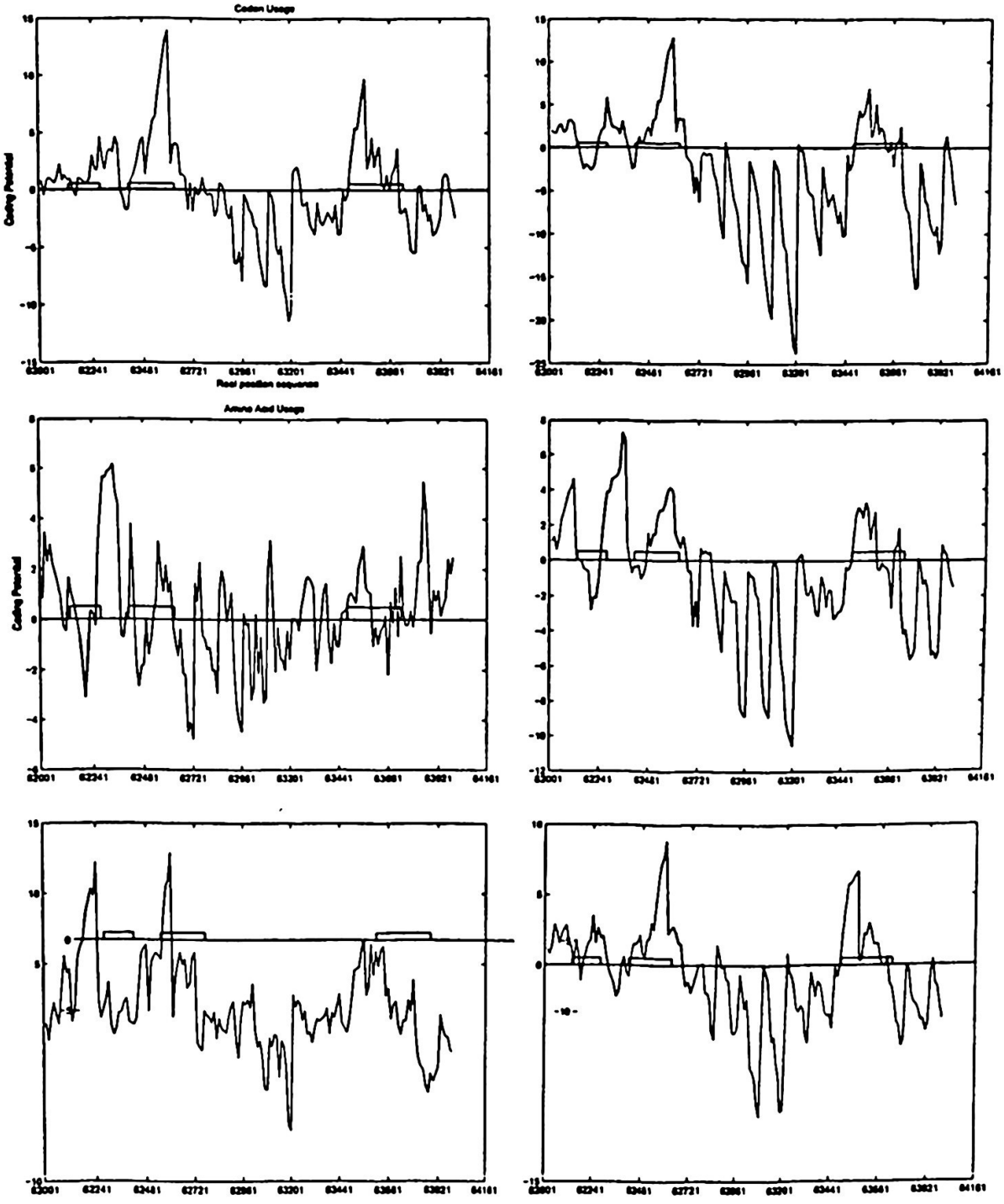

Fig. 4. Plot of original coding potentials and variants based on KM

7 Conclusions

In this article, we propose a new approach to solve the problem of predicting exonic and intronic regions from a DNA sequence. The main contribution is a methodology to obtain variants of existent coding measures. The new variants incorporate the result of applying the kasiski method to the exhaustive search of triplets on a sequence data. With this approach it is possible to code a DNA sequence with a Gaussian model of Cr, and get a better characterization of the introns than the original measures, which use a random model. We consider that a good characterization of introns is directly related to a high accuracy in prediction. For test purposes the β-globin gene was analyzed. The prediction was made using three new measures obtained with our methodology. The experiments of this paper showed an improved accuracy of 30% using the VCPKM coding measure, which is a codon preference variant. VCPKM predicted three exons and one intron. To process a whole chromosome, the exhaustive kasiski method can be scalable by using a high-performance computer. We consider that the principles followed in this research can be applied for obtaining other variants of coding measures.

References

[1] Griffiths, A., Gelbart W., Lewontin R., Miller J.: *Modern Genetic Analysis.* Second Edition. (2002).

[2] Snyder, E., Stormo, G.: Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks. *Nucleic Acids Research* 21. (1993) 607-613.

[3] Hall R., Stern L.: A rapid method of whole genome visualization illustrating features in both coding and non-coding regions. *The Second Asia-Pacific Bioinformatics Conference.* (2004).

[4] Lukashin , A., Borodovsky, M.: GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Research* 26. (1998). 1107-1115.

[5] Reese, M., Kulp, D., Tammana, H., Haussier D.: Gene Finding in Drosophila melanogaster. *Genome Research* 10. (2000). 529-538.

[6] Burge, C., Karlin, S.: Prediction of Complete Gene Structure in Human Genomic DNA. *Journal of Molecular Biology* 268. (1997). 78-94.

[7] Guigó, R.: DNA Composition, Codon Usage and Exon Prediction. *In Genetic Databases, M.J. Bishop Edition, Academic Press.* (1999).

[8] McCaldon, P., Argos, P.: *Proteins: Structure, Function and Genetics* 4. (1988) 99-122.

[9] Fickett, J.W., Tung,C.S.: Assessment of protein coding measures. *Nucleic Acids Research* 20. (1992) 6441-6450.

[10] Gribskov, M., Devereux, J., Burgess, R. B: The codon preference plot: graphic analysis of protein coding sequences and prediction of gene expression. *Nucleic Acids Research* 12. (1984) 539-549.

[11] Gebbie, S.: A survey of the Mathematics of Cryptology. Masther's Thesis. University of the Witwatersrand, Johanesburg. (2003).

[12] Human beta globin region on chromosome 11. http://www.ncbi.nlm.nih.gov /entrez/viewer.fcgi?db=nucleotide&val=455025. Entrez Molecular Database.