Recognizing translation initiation sites of eukaryotic genes based on the cooperatively scanning model

Yonghong Wang, Jinwei Liu, Tongju Zhao and Qing Ji

1Department of Physics, Hebei University of Technology, Tianjin 300130, China and 2Department of Physics, Tianjin University, Tianjin 300072, China

Received on December 25, 2002; revised March 24, 2003; accepted April 17, 2003

ABSTRACT

Motivation: Translation initiation sites (TISs) of genes are the key points of protein synthesis. Exact recognition of TISs in eukaryotic genes is one of the most important tasks in gene-finding algorithms. However, the task has not been satisfactorily fulfilled up to the present. Here, we propose a cooperatively scanning model for recognizing TISs and the first exons of eukaryotic genes on the basis of the structural characteristics of multi-exon genes.

Results: The model was employed to cooperatively scan the TISs and 3′ splicing sites in eukaryotic genes, and the TISs and the first exons of 132 mammalian gene sequences are identified to evaluate the model. Accuracy of exactly recognizing the TISs and the first exons has been found to amount respectively to 64.4 and 51.5%. We believe that the model will be a useful tool for genome annotation and that it can be easily incorporated into other algorithms to achieve higher accuracy in recognizing TISs and the first exons.

Availability: The program is available upon request.

Contact: yhwangwz@263.net

INTRODUCTION

As the Human Genome Project enters its large-scale sequencing phase, a large amount of uncharacterized DNA sequences needs to be annotated. Recognizing genes by computational methods becomes more and more important. In the past decade, a number of gene-finding algorithms have been developed, including recognition of coding regions (Borodovsky and McIninch, 1993), internal exons (Zhang, 1997) and complete genes (Guigo et al., 1992; Solovyev et al., 1995; Burge and Karlin, 1997; Krogh, 1997; Lukashin and Borodovsky, 1998), but accurate recognition of genes remains a goal yet to be achieved. This is especially the case in the translation initiation sites (TISs) and the first exons of eukaryotic genes. In contrast with prokaryotic genes, eukaryotic genes lack such conservative sequences as the S-D sequence of prokaryotic genes in upstream regions. Another disadvantage is that coding fragments in the first exons are short. These characteristics of eukaryotic genes make their TISs hard to recognize. The several algorithms (Salzberg, 1997; Pedersen and Nielsen, 1997; Zien et al., 2000) that have been developed for distinguishing the TISs from false TISs in eukaryotic DNA sequences, have shed insights on the issue. Salzberg used conditional probabilities to describe the signal of the context at TISs of eukaryotic genes. Pedersen and Nielsen applied the neural network method to capture characteristics of the cDNA sequences around TISs. Zien et al. employed SVM technique to recognize TISs in cDNA sequences. Other researchers have attempted to employ various scanning models to locate TISs. Kozak, e.g. proposed a scanning model for recognizing the TISs of eukaryotic genes in mRNA molecules, in which the first ATG with appropriate context in a mRNA molecule is regarded as the TIS (Kozak, 1989, 1996). Similarly, Agarwal and Bafna used the context of 5′ end capped sequences and TISs to predict genes and detect full-length cDNA (Agarwal and Bafna, 1998). Salamov et al. (1998) developed the model of ATGpr to recognize start codon in human cDNA sequences by some statistical parameters. In addition, some algorithms of finding complete gene structure can be applied to recognize the TISs of eukaryotic genes (Rogic et al., 2001). In spite of all these solid work and insightful findings, recognition of TISs of eukaryotic genes remains elusive, including the notable gene-finding algorithms (Stormo, 2000).

In this paper, we attempt to extract the characteristics of shorter flanking sequences around TISs of eukaryotic genes and propose a cooperatively scanning model for identifying TISs and the first exons of multi-exon genes. The approach applies the information of TISs, the 3′ splicing sites of the first and internal exons of eukaryotic genes. TIS prediction is performed in two steps. For a given gene sequence, the 3′ splicing sites of exons are first scanned to locate some internal position in the gene as end point for scanning TIS. Then, the TIS and 3′ splicing site of the first exon are cooperatively scanned and determined. The approach makes full use of information of
Recognizing translation initiation sites

complete genes and has been found to be able to effectively recognize TISs of multi-exon genes.

DATABASE

In order to recognize the TISs and the first exons of eukaryotic genes, a dataset of three versions is constructed from multi-exon genes. Version-1 consists of 2050 positive and 5000 negative samples for recognizing TISs from false TISs. The positive samples are partly extracted from the human multi-exon entries at http://www.fruitfly.org/seq_tools/datasets/Human and partly Guigo’s 570 vertebrate genes (Burset and Guigo, 1996). The entries are used as the positive samples, which have the first exons longer than 50 bp and upstream sequences from TISs longer than 80 bp. The negative samples are extracted from the intron files at the above website. Each sample comprises a start codon or a triplet ATG, 80 bases upstream and 60 bases downstream sequences from the base A of start codon. Version-2 consists of 1000 positive and 1000 negative samples for recognizing the 3’ splicing sites of exons. Each positive sample contains a 3’ splicing site, 90 bases upstream and downstream from the site. Similarly, each negative sample contains a dinucleotide GT at the middle of the sequence, 90 bases upstream and downstream. The positive samples are extracted from the first exons of the multi-exon genes and the negative samples from the intron sequences at the above website. Version-3 consists of 132 multi-exon genes selected from the database (Rogic et al., 2001), in which the entries are applied to evaluate the cooperatively scanning model. There is no overlap between the three versions. In addition, the file 4813_hum_CDS.fa at the above website is also used to count the codon usage of the first ten codons succeeding TISs, internal codon usage and base distribution in three frames of coding regions.

ALGORITHM

In this section, the linear discriminant function (LDF) and the cooperatively scanning model are described, respectively. The LDF is used to distinguish real TISs, and 3’ splicing sites from the negative samples, respectively. The cooperatively scanning model is used to recognize TISs and the first exons of eukaryotic genes.

The linear discriminant function

The $N$ feature variables, indicated by a vector $X$, are used to describe a sample for a given recognition task. Each sample can be represented as a point in a $N$-dimensional space. The positive and negative samples are associated with two kinds of points, respectively. The LDF is adopted to distinguish the positive samples from the negative samples. The LDF is actually a $N$-dimensional super plane, described by a vector $C$. The procedure to determine the vector $C$ can be found in any text-book of multivariable statistics (Mardia et al., 1979). For a recognition task, the vector $C$ can be calculated on the basis of the samples in the training set. The LDF is defined as $f(X) = C \cdot X$. The decision of the positive/negative samples is made by the criterion of $f(X) > U_0$, $f(X) < U_0$, where $X = (x_1, x_2, \ldots, x_N)^T$, and ‘$\cdot$’ is a transpose operator for a matrix. Using the samples in the training set, the threshold $U_0$ is uniquely determined by making the false positive and false negative rates identical.

Cooperatively scanning model

To recognize TISs of the complete genes in version-3 of the dataset, two scanning procedures are performed. First, 3’ splicing sites in a gene sequence are detected to determine the internal position of the gene, which is used as the end point for scanning TIS. Then the TIS and 3’ splicing site of the first exon are cooperatively scanned along the sequence up to the end point. When a triplet ATG is found, the LDF is calculated, and the dinucleotide GTs are searched within the open reading frame succeeding the ATG. With the position with maximum of LDF found, the fragment from ATG to GT is regarded as a candidate of the first exon. The scanning procedure continues until the end point is reached. If a new ATG is found and the values of the LDFs for the triplet ATG and the potential 3’ splice site are larger than those of the former candidate, the new fragment is regarded as the new candidate, which, surviving up to the end point, is predicted as the first exon and the TIS and 3’ splice site of the first exon is predicted.

RECOGNIZING TISs FROM NEGATIVE SAMPLES

In this section, the real/false TIS decision is made for recognizing TISs from negative samples. First, the characteristics of flanking sequences around TISs are described, and then recognizing variables are defined. Finally, the LDF is applied to recognize TISs in the samples of version-1.

The characteristics of flanking sequences around TISs

The file 4813_Hum_cds.fa at the website mentioned above contains complete coding sequences of 4183 human genes. The large collection of the complete genes is analyzed as that: (1) the codon usage of the first 10 codons succeeding TISs is calculated, the occurrence frequencies of codons are distinctly biased; (2) the distributions of A, C, G and T at three frames of coding sequences are analyzed, and the statistical result shows that the base distributions are asymmetric (see Wang et al., 2002). The contents of base T, G and A are poor at the first, second and third codon positions respectively; (3) the G+C content at the third codon positions within 60 bp window succeeding TISs is relatively high. The 5’-UTRs and the consensus sequences around TISs of eukaryotic genes play an important role in gene expression (Davuluri et al., 2000). Analyzing 5’-UTR sequences shows that the G+C content of the fragments from −80 to −15 base is relatively rich, where the
base A of start codon is defined as +1. Salzberg (1997) suggested that the score of dinucleotide conditional probabilities are more powerful than that of weight matrix for recognizing TISs of vertebrate genes. The conditional probability matrix is used to recognize TISs of eukaryotic genes here.

**Recognition variables**

Based on the above characteristics, five variables for recognizing TISs of vertebrate genes are defined as follows: (1) the first variable \( x_1 \) is related with the dinucleotide conditional probabilities. For a given sequence \( S = (x_1, x_2, \ldots, x_L) \), the score \( P(S) \) is calculated as

\[
P(S) = P(x_1) \prod_{i=2}^{L} P(x_i | x_{i-1})
\]

where \( P(x_i | x_{i-1}) \) is the conditional probability matrix and \( P(x_1) \) is the prior conditional probabilities. The variable is defined as log-odds score in the window \((-12, 7)\)

\[
x_1 = \log \frac{P(S)}{0.25^L} = \log P(S) - L \log 0.25
\]

(2) variable \( x_2 \) is defined as 64 times the product of the content of the bases T, G and A at the first, second and third codon positions, respectively; (3) variable \( x_3 \) is defined as the slope of the \( z \) component of the specific \( Z \) curve (Zhang and Wang, 2000) at the third codon positions within windows \((1, 60)\); (4) variable \( x_4 \) is defined as the product of the occurrence frequencies of the first ten codons in logarithm; (5) variable \( x_5 \) is defined as the slope of \( z \) component of the \( Z \) curve (Zhang and Wang, 1994) in windows \((-80, -15)\).

In the above five variables, variable \( x_1 \) cites Salzberg’s work, the variables \( x_3 \) and \( x_5 \) use the \( Z \) curve to represent G+C content. The variables \( x_2 \) and \( x_4 \) are designed to represent the asymmetry of base distribution and bias of codon usage in coding regions succeeding TISs of genes.

**The accuracy over 6-fold cross-validation test**

To evaluate the algorithm of recognizing TISs, the 6-fold cross-validation test is adopted. In version-1 of the dataset, 1000 positive and 1000 negative samples are randomly selected as the training set, and the other as the test set. The discriminant coefficients \( C \) and the threshold \( U_0 \) can be determined on the basis of the samples in the training set. Accuracy is then calculated with reference to the samples in the test set. The procedure is repeated six times. Six-fold cross-validation test is performed. The average sensitivity, specificity and accuracy over the 6-fold cross-validation test are calculated and listed in Table 1, which shows that the accuracy is up to 89%, higher than that of the presently available algorithms.

All the samples are also trained to obtain the linear discriminant coefficients, which are used in the cooperatively scanning model (Table 2).

| Table 1. The sensitivity, specificity and accuracy over 6-fold cross-validation test |
|----------------------------------|-----------------|-----------------|
|                                  | Training set (%)| Test set (%)     |
| Accuracy                        | 89.05           | 88.96           |
| Specificity                     | 89.05           | 89.09           |
| Average accuracy                | 89.05           | 89.02           |

<table>
<thead>
<tr>
<th>Table 2. The linear discriminant coefficients and threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear discriminantive coefficients</td>
</tr>
<tr>
<td>C1</td>
</tr>
<tr>
<td>3’SS(^a)</td>
</tr>
<tr>
<td>TIS(^b)</td>
</tr>
</tbody>
</table>

\(^a\)The 3’SS represents recognition of 3’ splicing sites.  
\(^b\)The TISs represents recognition of TISs.

**RECOGNIZING 3’ SPLICING SITES**

**Recognition of 3’ splicing sites**

In order to determine the internal positions of genes, 3’ splicing sites need to be recognized. Five recognizing variables are applied to distinguish 3’ splicing sites from false sites in the samples of version-2. The five variables are defined as follow: (1) Variable \( x_1 \) is derived from the specific \( Z \) curve (Zhang and Wang, 2000). The nine components in two sides of a splicing site are denoted as \( y_i \) and \( z_i (i = 1–9) \), respectively. The variable is defined as

\[
x_1 = \sum_{i=1}^{9} (y_i - z_i)^2
\]

(2) Variable \( x_2 \) is related to the contents of base T, G and A at the first, second and third codon positions in coding regions, respectively. Assuming that the first, second and third bases of the 62 bases upstream from 3’ splicing sites are at the first codon position, respectively, we proceed to calculate the products of contents of bases T, G and A at the first, second and third codon positions in 60 bp upstream, respectively. Variable \( x_2 \) is defined as 64 times the minimum of the products. (3) Variable \( x_3 \) is associated with codon and synonymous codon usage. The 4813 human genes at the above website are used to calculate human codon and synonymous codon usage (data not listed here, but available on request). For the upstream 20 triplets from a 3’ splicing site, the product of codon occurrence frequency and synonymous codon frequency is calculated for each triplet. Parameter \( L \) is defined as the sum of the products for the 20 triplets. Similarly, parameter \( R \), describing downstream 20 triplets from the 3’ splicing site, is represented as the sum of the products of codon occurrence frequency and synonymous codon frequency.
frequencies and synonymous codon frequencies. Suppose a parameter \( P \) to be \((L − R)/(L + R)\). Assuming the first, second and third bases of the 62 bases upstream fragment from the 3′ splicing site at the first codon position, respectively, we define variable \( x_2 \) as the maximum of the Ps in three cases. (4) Variable \( x_4 \) is concerned with the consensus sequences around 3′ splicing sites. The weight matrix of the consensus sequence of GeneID is used. Variable \( x_4 \) is defined as the sum of base occurrence frequencies in the window \((-3, 6)\), where each potential 3′ splicing site is defined as \((-1, 1)\). (5) Variable \( x_5 \) is defined as the dinucleotide GG content in the 61 base window succeeding 3′ splicing sites. The 1000 positive and 1000 negative samples in version-2 of the database set are used as the training set, and distinguished by the LDF. The accuracy amounts to 93.6%, the discriminant coefficients are listed in Table 2.

**RECOGNIZING TISs AND THE FIRST EXONS BASED ON THE COOPERATIVELY SCANNING MODEL**

In this section, the cooperatively scanning model is applied to recognize TISs and the first exons of complete mammalian genes in version-3. The procedure consists of two steps. In the first step, 3′ splicing sites are predicted to determine the positions in genes. In the second step, the TISs and 3′ splicing sites of the first exons are cooperatively scanned and predicted.

**Detecting the internal positions of genes**

In DNA sequences, a large number of the triplet A TGs exist in intergenic regions, coding and intronic regions. To reduce the number of the false TISs in a given sequence, some position in a gene is determined by recognizing 3′ splicing sites, which is used as the end point for scanning TISs in the second step.

The mean length of internal exons is 137 bp (Hawkins, 1988) and the mean length from 5′ splicing sites to the first stop codon in upstream introns is 61 bp (see discussion). In order to locate the internal position of a gene, the 3′ splicing sites are searched along the sequence. If the length from the sites previous to dinucleotide GTs to the first stop codon in the upstream region is longer than 200 bp, its LDF values are calculated based on the discriminant coefficients in Table 2. The site with the maximum value of the LDF is regarded as the end point for scanning TISs.

**Recognizing TISs and the first exons based on the cooperatively scanning model**

The cooperatively scanning procedure means that the TIS and 3′ splicing site of the first exon in a gene are cooperatively searched before the end point (see above). The procedure is described as follows: (1) to search ATG beginning from the 12th base along a given sequence. (2) While a ATG is found, the value \( F_1 \) of its LDF is calculated, the ATG is taken as a candidate of TIS. Once position P1 and F1 are determined, dinucleotide GTs are searched, with the ATG within the open reading frame as the beginning point. Among the dinucleotide GTs identified, the position with LDF maximum \( f_1 \) is taken as a potential 3′ splicing site. The fragment from ATG to the 3′ splicing site is regarded as a candidate of the first exon, denoted as \((F_1, f_1)\). (3) To search better candidates, ATG is further scanned along the sequence. If a new candidate, denoted as \((F_2, f_2)\), is found, it will be compared with the former one. If \( F_2 > F_1, f_2 > f_1 \), the new one is accepted as the candidate of the first exon. The scanning procedure continues to the end point.

Rogic et al. (2001) constructed a dataset containing 195 mammalian genes to evaluate seven gene-finding algorithms, of which 152 entries are multi-exon genes. With the entries with 5′-UTR shorter than 12 bp and the first exons shorter than 30 bp ignored, the other 132 multi-exon genes are applied to evaluate the cooperatively scanning model, in which 64.4% of TISs, 59.1% of 3′ splicing sites and 51.5% of the first exons are exactly recognized, respectively. In addition, 15% of the first exons predicted overlap with the real ones.

**DISCUSSION**

**The structure of vertebrate genes and recognition of TISs**

In contrast to prokaryotic genes, most eukaryotic genes are multi-exon genes. The length of exons is relatively shorter, especially the first exons. Its length distribution has a peak around 60 bp. Some algorithms for recognizing TISs (Pedersen and Nielsen, 1997; Zien et al., 2000) used the information of 150 bp long coding sequences to recognize TISs of vertebrate genes, but the fraction of the first exons longer than 150 bp is very small. It is more significant to identify TISs of vertebrate genes using shorter flanking sequences around TISs. With an attempt to predict the TISs using shorter flanking sequences, we did a cross-validation test, which shows that the present algorithm can effectively extract information and recognize TISs.

**The properties of flanking sequences and importance of the variables**

The TISs of genes are the start point of protein synthesis. Sometimes, gene expression is regulated in translation stage. The flanking sequences play an important role in translation initiation. The length of 5′-UTRs is closely related with translation procedure, long 5′-UTRs hamper the initiation process, whereas the appropriate ATGs is skipped over in short 5′-UTR, such as those shorter than 17 bp. The rate of translation is increased in direct proportion with the length of 5′-UTRs, ranging from 17 to 80 bp. The base C content in 5′-UTRs of human genes is high (Louis and Ganoza, 1988). Further analyzing 5′-UTRs, the G+C content is relatively rich, the feature is described by Z curve.
The efficient translation is related with the specific position’s bases, the purine at the −3 position and base G at +4 position are very important for efficient translation initiation. It is well known that there are the consensus sequences around TISs. Salzberg’s work (Salzberg, 1997) shows that the method of conditional probabilities is superior to that of weight matrix for recognizing TISs, the score of conditional probabilities is used as a recognizing variable here.

In general, the coding regions succeeding TISs do not affect the translation initiation, but the asymmetric base distribution in the regions is very important for recognizing TISs, even though the coding fragments of the first exons are fairly short. In the present algorithm, the two variables extracted from coding fragments make full use of the low contents of bases in three frames and high G+C content at the third codon positions.

Most proteins comprise the signal peptides, the N-terminal part of the polypeptides. Generally, a signal peptide consists of three regions: n-region, h-region and c-region (Nielsen and Krogh, 1998), all of which are conservative. We attempt to partially extract the features of signal peptides by the first 10 codon usage in human genes. The variable is very powerful for recognizing TISs.

The five variables in the algorithm describe the characteristics of different sequences around TISs. To evaluate the importance of them, all samples of version-1 are used as the training set to test recognizing capacity of each variable. The result shows that the variable related to the first 10 codon usage is the most powerful. This is followed by the variable derived from base contents of T, G and A at the first, second and the third codon positions, the variable using conditional probabilities, the variable related to the specific Z curve and the variable of the z component of Z curve.

Detecting the internal positions of genes

There are a large number of ATGs in the upstream and downstream from TISs in genes. Some of them have stronger signals than those of TISs. In some algorithms, the triplet ATG with the maximum score is regarded as a TIS through scanning each ATG in full-length sequences, the structural features of multi-exon genes are not fully used. Instead of taking the strategy, we attempt to reduce the number of the negative samples by determining the internal positions of genes, such as 3’ splice sites. The negative samples, which lie in the downstream sequences at the positions, are not considered. For the goal, the lengths from 5’ splicing sites to the first stop codons in upstream intron regions along coding frame are analyzed using Guigo’s 570 sequences (Burset and Guigo, 1996). The distribution is presented in Figure 1. The mean length is 61 bp.

With consideration of the mean length, together with the mean length of internal exons, the dinucleotide GT, which have longer than 200 bp upstream opening reading frame and the maximum score of its LDF, is predicted as a 3’ splicing site of a gene. The site is used as the end point for scanning TISs.

The cooperatingly scanning model

The internal points of genes make the negative samples drastically reduced, but a great number of negative samples still remain in the upstream regions. Previous study shows that the appropriate ATG closest to 5’ end of a mRNA molecule is regarded as TIS of a eukaryotic gene (Kozak, 1989, 1996), but Pedersen’s analysis shows that about 40% of mRNAs have one or more upstream AUGs (Pedersen and Nielsen, 1997). Therefore, it might be concluded that the first ATGs in mRNA molecules are not necessarily the start codon of genes. Under such circumstances, it is still difficult to definitely recognize TISs and 3’ splicing sites of the first exons. It is our belief that the accuracy of recognizing TISs and the first exons can be improved with consideration of the features of gene structure. The first exons are assumed to be those that possess strong signals at both ends, or those that possess strong signals at one end and moderate signals at the other. Therefore, the fragments from ATG to GT, which have very weak signals at one end and strong signals at the other, are dismissed. When the first exon is searched along a given sequence, the fragments coinciding with the assumption for the first exons are taken as candidates. Whenever a more ideal candidate appears, it replaces the former one as the new candidate. In general, the first exon can replace the candidates in the upstream and it is not substituted by the candidates in the downstream, i.e. the first exon survives up to the end point for scanning TISs.

Compared with other scanning models for recognizing TISs, the present model is more sophisticated, with the inclusion of more information of gene structure. It is of interest to compare the present model with others, such as Kozak’s

Fig. 1. The distance distribution of 5′ splicing sites to the first stop codon in the upstream intron regions along coding frame. The x-axis represents the distances, whereas y-axis indicates the fraction of sequences.
scanning model. With the same variables and linear discriminant function, the TISs of the 132 entries in version-3 are recognized by Kozak’s scanning model, only 45% of TISs are identified. Compared with the result, the accuracy of the present model is much higher than that of Kozak’s scanning model.

The present model covers more features of gene structure, about 64.4% of TISs and 51.5% of the first exons are exactly identified and about 15% of the first exons predicted overlap with the real ones, but about 33% of the first exons are not recognized fully. In the entries where the first exons are erroneously identified, the signals at both ends are weak. Chances are that either the first exons cannot replace the candidates in the upstream, or they are substituted by the candidates in the downstream, most of which are the part of internal exons. This indicates that it is needed to further extract characteristics of local and global signals from DNA sequences for recognizing TISs and the first exons.

CONCLUSION

Despite the great efforts in TIS recognition of vertebrate genes, the task is still not satisfactorily accomplished. Recently, a detailed comparison and evaluation on seven notable algorithms have been performed, using 195 mammalian sequences. However, the accuracy for recognizing TISs is far from ideal (Rögic et al., 2001). It is therefore imperative to extract new features, propose new models and develop new and powerful algorithms. This paper attempts to make some efforts in the respect. The greater accuracy of the algorithm shows that the cooperatively scanning model is a powerful method to recognize TISs and the first exons in the multi-exon genes. The model effectively takes good advantage of both information of flanking sequences around TISs and structural features of multi-exon genes. We believe the model can also be applied to recognize other tasks in bioinformatics. Moreover, accuracy of the approach can be further improved, if the CG islands, 5′-UTRs and 3′ splicing sites can be well recognized.

ACKNOWLEDGEMENT

The present study was supported in part by the Education Foundation of Hebei Province under Grant No. 2002121.

REFERENCES


