Journal of Biochemistry and Molecular Biology, Vol. 35, No. 5, September 2002, pp. 513-517

JOURNAL OF Biochemistry and Molecular Biology

A Method for Identifying Splice Sites and Translation Start Sites in Human Genomic Sequences

Ki-Bong Kim[†,‡,*], Kiejung Park[†] and Eun Bae Kong[‡]

[†]Information Technology Institute, SmallSoft Co., Ltd., Daejeon 305-811, Korea

[‡]Department of Computer Engineering, Chungnam National University, Daejeon 305-764, Korea

Received 22 July 2002, Accepted 28 August 2002

We describe a new method for identifying the sequences that signal the start of translation, and the boundaries between exons and introns (donor and acceptor sites) in human mRNA. According to the mandatory keyword, ORGANISM, and feature key, CDS, a large set of standard data for each signal site was extracted from the ASCII flat file, gbpri.seq, in the GenBank release 108.0. This was used to generate the scoring matrices, which summarize the sequence information for each signal site. The scoring matrices take into account the independent nucleotide frequencies between adjacent bases in each position within the signal site regions, and the relative weight on each nucleotide in proportion to their probabilities in the known signal sites. Using a scoring scheme that is based on the nucleotide scoring matrices, the method has great sensitivity and specificity when used to locate signals in uncharacterized human genomic DNA. These matrices are especially effective at distinguishing true and false sites.

Keywords: GenBank, Independent nucleotide frequency, Scoring matrix, Signal site

Introduction

Begun in 1990, the Human Genome Project originally was planned to last 15 years, but effective resource and technological advances have accelerated the expected completion date to 2003. Several types of genome maps have already been completed. A working draft of the entire human genome sequence was announced in June 2000. This made it inevitable that large amounts of raw DNA sequences (without the knowledge of their mRNA and protein products for which they code) would emerge. Predicting the potential coding genes in these sequences will be imperative. It will be a major task that will involve several steps, such as identifying potential promoter sites, transcription and translation start sites, 5' and 3' splice sites (i.e., potential exons and introns), and polyadenylation sites. Undoubtedly, computer methodologies will play an important role in this task. Also, the existing information on the sequence characteristics of these different architectural elements of genes will have to be used.

Many previous studies attempted to characterize the sequences around the start, donor, and acceptor sites in eukaryotic organisms. Kozak (Kozak, 1987; Kozak, 1992) produced several comprehensive studies of the consensus sequence around the start of translation, and introduced the scanning model of ribosome progression that is now widely accepted. Senapathy *et al.* (Senapathy *et al.*, 1990) and Mount *et al.* (Mount *et al.*, 1992), among others, characterized the sequence patterns around splice junctions. Recent results described the patterns and processes for non-consensus (AT-AC) splicing (Mount, 1996). The consensus sequences that were uncovered in these previous studies have been most frequently described as matrices that contain the probabilities of the four bases in the positions immediately surrounding the sites. Specific computational systems for identifying splice junctions have been developed by many previous researchers. Fickett (Fickett, 1996) provides a review of the main work on identifying signals (both splice junctions and translation start sites), and points out that the best previous results used a combination of different types of evidence, both from the bases immediately surrounding the site and from sequences extending some distance away from the site. Fickett (Fickett, 1996) also describes a common variation on the consensus matrix that is known as the position weight matrix, which uses $P(b,i)/P(b)$ instead of just $P(b,i)$. $P(b,i)$ means the probability of each base b at position i. This gives the relative frequency of each base in each position, and may offer advantages in regions of high or low G + C content.

*To whom correspondence should be addressed.
Tel: 82-42-864-2524; Fax: 82-42-866-9241
E-mail: kbkim@bioinfo.smallsoft.co.kr

As part of an automated system for finding coding regions in uncharacterized human DNA, we developed a new method for finding the signals that indicate the start of translation, the beginnings of introns (donor sites), and the ends of introns (acceptor sites). The basis of the method is that the independent frequencies for each of the four bases in a fixed set of positions around each site are computed and modified with the expected frequency and relative weight in proportion to the probability of each base at each position. The resulting scoring matrices indicate that for several positions in all three types of sites (start of translation, donor, and acceptor sites), a sequence of several nucleotides is highly conserved. These conserved sequences are an essential part of the process of transcription initiation and exon splicing, and provide a specific molecular signal for the transcription initialization (Lee *et al*., 2001; Roytrakul *et al*., 2001) and RNA splicing machinery to identify the precise sites. It is unclear if different classes of organisms exhibit differences in the signals, and in the mechanism of initiation and splicing. As to the splicing mechanism, several small nuclear RNA's that are associated with the splicesomes appear to aid in the precise excision of introns (Green, 1986; Sharp, 1987). In this study, we focused on humans as the target organism. The consensus sequence can simply be read by choosing the value in each column of the scoring matrices, which complies with our simple rules. To generate the scoring matrices and consensus sequences for the start site, donor site, and acceptor site in human DNA, we used a large data set of each signal site, which was extracted from the ASCII flat file gbpri.seq in the GenBank release 108.0. The scoring matrices tabulate the information that summarizes the sequence information for each signal site.

Materials and Methods

Sequence data The signal sequence data for this study, or training data, were originally collected from ASCII flat file "gbpri.seq" in the GenBank Release 108.0 through complicated parsing and filtering works. The complete understanding of the GenBank format, including the complicated feature table, preceded the implementation of such parsing and filtering works. The keyword ORGANISM and feature table were employed to obtain the necessary sequence data. Database entries were discarded if the ORGANISM field did not correspond to Homo sapiens. The GenBank releases use a feature table format that is designed jointly by the GenBank, EMBL Nucleotide Sequence Data Library, and the DNA Data Bank of Japan. The format is in use by all three databases. The feature table contains information about genes and gene products, as well as regions of biological significance that are reported in the sequence. It also contains information on regions of the sequence that code for proteins and RNA molecules. The first line of the feature table is a header that includes the keyword, 'FEATURES', and the column header, 'Location/Qualifier'. Each feature consists of a descriptor line that contains a feature key and location. If the location does not fit on this line, then a continuation line may follow. The feature key begins in column 6 and may be no more than 15 characters in length. The location begins in column 22. Feature qualifiers begin on subsequent lines at column 22. Location, qualifier, and continuation lines may extend from columns 22 to 80. Feature tables are required, due to the mandatory presence of the source feature. The feature key CDS, which means sequence coding for amino acids in protein, was used to collect the translational start sites and splice sites. In the case of the start sites, the qualifiers, CAAT_signal and TATA_signal that mean 'CAAT box' and 'TATA box' in eukaryotic promoters respectively, were additionally employed to verify the first region in CDS as the initial exon. Since there are many patterns in the CDS line, we carefully filtered them to obtain clearer sequence data. The sequence data were discarded as follows: (1) If they were likely to contain erroneous annotation. (2) If they represented non-standard splicing mechanisms (e.g., alternative splicing or AT-AC splicing). (3) If they did not start with ATG for the start site. (4) If they did not have dinucleotide GT and AT for donor and acceptor sites. The resulting data set contains 311 start sites, 2456 acceptor sites, and 2375 donor sites. Among the data set that we collected, 161 start sites, 1456 acceptor sites, and 1375 donor sites were used as the training data set. The remainder of each signal site data was reserved to be used as true sites of the test data set. As compared with splice sites, the number of start sites is small since for the purpose of characterizing start sites, the sequences must contain more than 5 bases prior to the start of translation.

Scoring-matrix computation Two basic structures that have been used to represent the common sequence information on a specific signal site are consensus sequences, regular expressions and weight matrices, or profiles. Very efficient algorithms exist to search for consensus sequences or regular expressions. Nevertheless, when available, weight matrices are almost always preferred, because they are able, in general, to describe any pattern that a consensus sequence can, but in addition can capture much more subtle relationships. In its simplest form, a weight matrix describes a pattern of fixed width W; it is a rectangular array of numbers with W columns and one row for each possible base. The matrix in this study uses positions from –5 through +4 (*i.e.* 5 positions upstream of the start codon through 2 positions downstream, where the start codon itself occupies position 0-2). Even though many different methods for calculating the actual numbers to be used in weight matrices have been advanced, we used our own unique method. We began by constructing a table of statistics, $C_{b,i}$, where $C_{b,i}$ is the actual count of base b in position i in the known signal sites. From the table, we constructed a scoring matrix with the same axes with each entry denoting the "score" that a sequence obtains for having the corresponding base in the corresponding position. The scoring matrix is constructed to assign high scores to patterns with specific biological significance. The score $S_{b,i}$ is computed by the following method:

If $(C_{b,i} > E_{b,i})$

$$S_{b,i} = C_{b,i} - E_{b,i} + (C_{b,i} \times P_{b,i}) \quad (1)$$

else

$$S_{b,i} = C_{b,i} - E_{b,i} - (C_{b,i} \times P_{b,i}) \quad (2)$$

where $E_{b,i}$ is the expected count of base b in position i, and $P_{b,i}$ is the observed probability that the base b will occur in position i, computed over the known signal sites. This scoring matrix assigns a

Table 1. Scoring matrix for translation start site. The numbers at the top of each column mean positions from –5 through +4 (5 positions upstream of the start codon through 2 positions downstream, where the start codon itself occupies position 0-2). The matrix rows correspond to the bases A, T, G, C from top to bottom; consensus nucleotides are listed at the end of each column. Each entry denotes the "score" that a sequence obtains for having the corresponding base in the corresponding position.

	-5	-4	-3	-2	-1	0	1	2	3	4
A	-15.22	-14.01	100.06	18.89	-17.78	281.75	-40.25	-40.25	15.78	28.55
T	20.47	-16.47	-34.47	-21.26	-28.30	-40.25	281.75	-40.25	-21.26	-27.47
G	-11.22	-21.99	31.86	-17.78	-15.22	-40.25	-40.25	281.75	43.86	-15.84
C	17.33	75.54	-30.00	52.81	100.06	-40.25	-40.25	-40.25	-13.43	50.99
	N	C	R	A/C	C	A	T	G	R	A/C

Table 2. Scoring matrix for donor site. The numbers at the top mean positions from –5 through +4 (5 positions upstream of the boundary between exon and intron through 5 positions downstream, where intron begins at position 0). The matrix rows correspond to the bases A, T, G, C from top to bottom; consensus nucleotides are listed at the end of each column. Each entry denotes the "score" that a sequence obtains for having the corresponding base in the corresponding position.

	-5	-4	-3	-2	-1	0	1	2	3	4
A	-91.91	131.98	-96.63	921.71	-106.96	-343.75	-343.75	355.41	939.06	222.01
T	102.91	-137.94	142.82	-202.37	-225.23	-343.75	2406.25	-222.81	-207.00	185.12
G	-114.34	-109.21	-166.52	-131.67	1104.01	2406.25	-343.75	542.54	-159.63	1145.34
C	167.88	251.87	273.48	-207.78	-288.20	-343.75	-343.75	-235.88	-177.12	-206.22
	C	A/C	Y	A	G	G	T	R	A	G

Table 3. Scoring matrix for acceptor site. The numbers at the top mean positions from –5 through +4 (*i.e.* 5 positions upstream of the boundary between intron and exon through 5 positions downstream, where exon begins at position 0). The matrix rows correspond to the bases A, T, G, C from top to bottom; consensus nucleotides are listed at the end of each column. Each entry denotes the "score" that a sequence obtains for having the corresponding base in the corresponding position.

	-5	-4	-3	-2	-1	0	1	2	3	4
A	-200.96	115.18	61.28	2548.00	-364.00	704.10	-141.62	-166.73	-32.92	12.00
T	350.30	-106.85	-169.21	-364.00	-364.00	-271.80	1030.24	-136.53	46.62	9.15
G	-121.13	199.37	-191.29	-364.00	2548.00	346.86	-117.25	973.56	156.77	-91.07
C	430.93	175.58	838.00	-364.00	-364.00	-222.39	-170.44	-92.39	200.96	485.08
	Y	N	C	A	G	R	T	G	G/C	C

score to every sequence segment of width $W = 10$, calculated simply as the sum of the relevant matrix elements. The more the base is conserved in a position, the higher the positive score is that is assigned to the base in the position. In other words, the conserved base with high frequency is heavily weighted, as compared with the base with low and random frequencies. Tables 1, 2, and 3 are the result of the scoring matrices for the start, donor, and acceptor sites.

Consensus sequences To find the consensus sequence from the matrix, the following simple rule was used in arriving at a consensus sequence at each location. If the highest score that is computed for each nucleotide site equals or exceeds the expected count, then the corresponding nucleotide was chosen. If the second highest score at each position equals or exceeds 30% of the expected count, and the difference between it and the third highest score equals or exceeds 30% of expected count, then the corresponding nucleotide was also chosen. The consensus nucleotides that were found are listed at the end of each column in Tables 1, 2, and 3. In the consensus sequences, "N" means any nucleotide, "R" means A or G (purine), and "Y" means C or T (pyrimidine).

Detecting signals with scoring matrices The scoring matrices, obtained as described previously, can be used to find potential signal sites in a given sequence. For any pattern of anonymous DNA, one must compute a score for every subsequence of length 10, based on its corresponding scoring matrix. For each nucleotide position in a subsequence, the elements from appropriate scoring matrices are added to arrive at the total. The formula for scoring (giving a score between 0 and 100) is $Score = 100 \times (O_w - Min)/(Max - Min)$. *Min* and *Max* are the minimum and maximum possible totals (i.e., the sum of the lowest and highest scores in each of the ten positions in the scoring matrices), and O_w is the total of the scores for the ten nucleotides that occur in the subsequence that is being scored.

Table 4. Sensitivity and false-positive rates of start, donor, and acceptor site detections for a range of different threshold values, using the scoring matrices.

Signal Detection					
		True Sites Missed		False Sites Labeled True	
	Threshold	Num	%	Num	%
Start Site	60	0	0.0	6	4.0
Detection	70	4	2.7	1	0.7
(150 true sites	80	33	22.0	0	0.0
+ 150 false sites)	90	118	78.7	0	0.0
Donor Site	60	116	11.6	80	8.0
Detection	70	247	24.7	18	1.8
(1000 true sites	80	491	49.1	1	0.1
+ 1000 false sites)	90	789	78.9	0	0.0
Acceptor	60	12	1.2	10	1.0
Detection	70	118	11.8	5	0.5
(1000 true sites	80	404	40.4	2	0.2
+ 1000 false sites)	90	843	84.3	0	0.0

Results

Although comparisons are difficult to make without using identical data sets, a very rough comparison might be informative. The splice site detection method of Solovyev *et al.* (Solovyev *et al.*, 1994) is reported to be the best-known method, and their tests also used human-only data. Because their method was only used for the donor and acceptor sites, and not for start sites, we will only compare those numbers here. A description of their algorithm is beyond the scope of this discussion, but, briefly, it is a straightforward linear discriminant function that is based on a set of complex features. These features include the following: triplet composition in an 80-base window around donor sites, a 10-base consensus matrix, the number of *G* bases, *GG* pairs, and *GGG* triplets in a 50-base region of the intron, and octanucleotide frequency measures for a 114-base window around the site. They used a similar set of features for acceptor sites. They reported an overall accuracy for donor site prediction of 97%. However, they do not give a breakdown into false-positives and false-negatives, and since the number of pseudo-sites is far greater than the number of true sites (97.8% of their test data was pseudo-sites), it is hard to compare their numbers to the ones that are reported here. One of their figures indicates that they obtained a 3% false-positive rate for 96% sensitivity, which for their data would indicate approximately 900 false-positives. For acceptor sites, they report a 4% false-positive rate at 96% sensitivity, which would yield approximately 3600 false-positives (they had substantially more pseudo-acceptor sites). The false-positive rates in the Solovyev study only counted the sites that already contained a *GT* or *AT* as potential donor or acceptor sites (Solovyev *et al.*, 1994).

In order to verify sensitivity and false-positive rates of each signal site, we constructed a special test data set of subsequences with a length of 10; half were true sites and half were false sites. As mentioned before, the true sites in test data set are the non-overlapping part of all of the data that were collected from the GenBank. All of the false sites in the test data set originated from the upstream region of the real signal sites. For the donor sites, Table 4 shows that the scoring matrix has an 8.0% false-positive rate for 88.4% sensitivity at the threshold value of 60. For the acceptor sites, Table 4 shows that the scoring matrix has a 1.0% false-positive rate for 98.8% sensitivity at the threshold value of 60. For the start site, Table 4 shows that the scoring matrix has a 4% false-positive rate for 100% sensitivity at threshold 60. Even though the accuracy for donor sites is a little low when compared to those for the start and acceptor sites, this result could be considered a good one. This is as good as the linear-discriminant method, and it is surprisingly close since less information is used. Clearly, some non-local information can be useful. For example, the branch site occurs some distance upstream of the 3 acceptor; local matrices do not capture this site. In addition, the coding region side of any site cannot contain in-frame stop codons, so the presence of stop codons can be used to rule out many false-positives. Therefore, if the only goal is to identify splice sites, then a method that is based on both local and non-local information should be used.

Discussion

The main results of the current study, which are consistent with Ficketts conclusion that the best methods combine many different coding measures, offer an alternative point-of-view (Fickett, 1996). In these experiments, we used only one coding measure, the actual count of each base and relative weight in proportion to its probability in the immediate vicinity of the known of the splice and start sites. We obtained surprisingly good accuracy even though there was only limited information that was used. It is true that higher

accuracy may be obtained by using information from a larger window around the site (Solovyev *et al*., 1994), however, the bulk of the signal recognition ability comes from local information. In addition, non-local coding measures are already used elsewhere in gene assembly programs (Solovyev *et al*., 1994; Synder and Stormo, 1995; Xu *et al*., 1996), and the benefits of that information are reflected in better performance overall on the gene-finding problem. These results show that the majority of the required information is contained locally in the sequence pattern itself when it comes to identifying signals in DNA sequences. This makes good sense biologically, since translation and splicing machinery seems to operate primarily on mRNA that is near the sites.

The identification of sequence patterns is essential to understanding the machinery behind the translation and splicing of mRNA. Identification of the most likely bases at each position around a signal site is the first step in characterizing these patterns. The current study uses growing amounts of sequence data to move one step further towards characterizing signal sites. The scoring matrices that were computed in this study show very conserved bases around the start, donor, and acceptor sites. Although the overall consensus pattern changes only slightly with use of these new matrices, the ability to accurately detect true sites improves substantially. As more data accumulates, it should be possible to further refine these matrices and develop even better methods for site recognition. Besides providing better characterizing of the sites, these matrices should also helpful in improving the performance of gene finding systems. Finally, one can recommend that future efforts to characterize splice junctions and start sites should emphasize the collection of large, high-quality data sets for each organism of interest.

References

Fickett, J. (1996) The gene identification problem: an overview for developers. *Comput. Chem.* **20**, 103-118.

Green, M. R. (1986) PRE-mRNA SPLICING. *Ann. Rev. Genetics* **20**, 671-708.

Kozak, M. (1987) An analysis of 5-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res*. **15**, 8125-8148.

Kozak, M. (1992) A consideration of alternative models for the initiation of translation in eukaryotes. *Crit. Rev. Biochem. Mol. Biol.* **27**, 385-402.

Lee, S. H., Park, K. H., Kim, D. H., Choung, D. H., Suk, J. E., Kim, D. H., Chang, J., Sung, Y. C., Choi, K. Y. and Han K. (2001) Structural origin for the transcriptional activity of human p53. *J. Biochem. Mol. Biol*. **34**, 73-79.

Mount, S. (1996) AT-AC introns: An Attack on dogma. *Science* **271**, 1690-1692.

Mount, S., Burks, C., Hertz, G., Stormo, G., White, O. and Fields, C. (1992) Splicing signals in drosophila: intron size, information content, and consensus sequences. *Nucleic Acids Res.* **20**, 4255-4262.

Roytrakul, S., Eurwilaichitr, L., Suprasongsin C. and Panyim, S. (2001) A rapid and simple method for construction and expression of a synthetic human growth hormone gene in *Escherichia coli. J. Biochem. Mol. Biol*. **34**, 502-508.

Senathy, P., Shapiro, M. B. and Harris, N. L. (1990) Splice junctions, branch points, and exons: sequence statistics, identification, and applications to genome project. *Methods Enzymol.* **183**, 252-278.

Sharp, P. A. (1987) Splicing of Messenger RNA Precursors. *Science* **235**, 766-771.

Snyder, E. E. and Stormo, G. D. (1995) Identification of coding regions in genomic DNA. *J. Mol. Biol.* **248**, 1-18.

Solovyev, V. V., Salamov, A. A. and Lawrence, C. B. (1994) Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Res*. **22**, 5156-5163.

Xu, Y., Mural, R., Einstein, J. R., Shah, M. and Eberbacher, E. (1996) Grail: A multi-agent neural network system for gene identification. *Proc. IEEE* **84**, 1544-1552.