



## Identifying the 3'-terminal exon in human DNA

Jack E. Tabaska<sup>1,2</sup>, Ramana V. Davuluri<sup>1</sup> and  
Michael Q. Zhang<sup>1,\*</sup>

<sup>1</sup>Cold Spring Harbor Laboratory, PO Box 100, Cold Spring Harbor, NY 11724, USA

Received on December 6, 2000; revised on January 30, 2001; accepted on February 2, 2001

### ABSTRACT

**Motivation:** We present JTEF, a new program for finding 3' terminal exons in human DNA sequences. This program is based on quadratic discriminant analysis, a standard non-linear statistical pattern recognition method. The quadratic discriminant functions used for building the algorithm were trained on a set of 3' terminal exons of type *3tuexon* (those containing the true STOP codon).

**Results:** We showed that the average predictive accuracy of JTEF is higher than the presently available best programs (GenScan and Genemark.hmm) based on a test set of 65 human DNA sequences with 121 genes. In particular JTEF performs well on larger genomic contigs containing multiple genes and significant amounts of intergenic DNA. It will become a valuable tool for genome annotation and gene functional studies.

**Availability:** JTEF is available free for academic users on request from <ftp://www.cshl.org/pub/science/mzhanglab/JTEF> and will be made available through the World Wide Web (<http://www.argon.cshl.org/>).

**Contact:** {mzhang, ramana}@cschl.org

### INTRODUCTION

Reliable detection of gene termini is critical to the success of computational gene discovery in genomic sequences. To resolve a string of exons into individual genes, one needs to know where the 5' and 3' terminal exons are. Although improvements in computer gene-finding programs have made it relatively easy to detect internal protein-coding exons in genomic sequences, terminal exon detection has remained a weak point for many gene prediction methods. These programs are frequently criticized for their propensity to skip terminal exons, thereby fusing several genes together (Claverie, 1997; Stormo, 2000). To help fill this gap in gene prediction technology, we have developed a standalone 3' terminal exon recognition module called JTEF (Jack's Terminal Exon Finder) as an addition to the MZEF internal exon

finding program (Zhang, 1997). Like MZEF, JTEF is based on quadratic discriminant analysis, a powerful statistical multivariate pattern-recognition method, which has been applied to many fields of science.

Our approach to gene finding is building optimized models for one entity at a time. As a first step towards this goal, MZEF was developed for finding the internal coding exons. As a next step in making MZEF into a complete gene-prediction system, we have developed JTEF for finding the 3' terminal exons of type *3tuexon* (3' translated & untranslated). A *3tuexon* is the 3' terminal exon having a 3'UTR following a CDS (see Zhang, 1998a,b, for further details). As a part of the development of JTEF, we earlier developed a program *polyadq* (Tabaska and Zhang, 1999) that finds polyA signals in human DNA sequences. This work extends the results of *polyadq* in building JTEF. The algorithm for JTEF simultaneously uses four Quadratic Discriminant Functions (QDFs) based on sequence context measures and polyA signal strength features. The QDFs for the acceptor, donor and terminal exons were trained on the standard data set of Kulp *et al.* (1996). The QDF for polyA signal was trained on *PolyA site database* (Tabaska and Zhang, 1999). We show that JTEF outperforms the existing best gene prediction programs in identifying the 3' terminal exons.

### METHODS

#### The algorithm

The major obstacle in detecting 3' terminal exons is the signal-to-noise ratio: only about 10% of coding exons are 3' termini. Furthermore, the consensus splice donor sequence contains a stop codon (G|GTRAGT), which is in the best coding frame 1/3 of the time, so roughly internal exons of this proportion look like 3' terminal exons. To help overcome this problem, JTEF uses a two-pass approach. First, the program scans input sequences for splice acceptor sites. Putative splice acceptors are evaluated using an acceptor site QDF. In addition, for each acceptor site the best corresponding stop codon is determined using a Bayesian scoring function (see below) that allows coding fragments of widely disparate lengths to be evaluated and compared. During this scan,

\*To whom correspondence should be addressed.

<sup>2</sup>Present address: Monsanto Company, 800 North Lindbergh, St Louis, MO 63167, USA.

no attempt is made to differentiate between internal and terminal exon acceptor sites; all acceptor–stop codon pairs are simply considered to be terminal exon candidates. An intermediate filtering step is performed after the acceptor site scan to reduce the candidate list by removing overlapping exons. A second pass through the sequence is then performed to find splice donors and polyA signals. This scan makes use of both donor site and polyA signal QDFs. When a donor site is found, it is matched up with upstream acceptor sites for evaluation as internal exons; when found, these are removed from further consideration. When a polyA signal is detected, it is matched up with upstream acceptor sites and evaluated using a terminal exon QDF. High-scoring terminal exons are then reported to the user.

### Quadratic discriminant analysis

Characterization of authentic splice sites, terminal exons, and polyA signals was performed by quadratic discriminant analysis. Interested readers may refer to Zhang (1997) and Venables and Ripley (1994) for a more complete description of quadratic discriminant analysis and Splus programs. The algorithm for JTEF uses four different QDFs (acceptor, donor, polyA and terminal exon) with different variables. The QDF variables were obtained by experimenting with many standard protein coding measures (see e.g. Fickett and Tung, 1992; Fickett, 1996, for more details), as well as some we developed specifically for terminal exon recognition.

### Bayesian scoring functions

Recognition of the coding portion of a terminal exon—the Terminal Coding Fragment (TCF)—presents some unique challenges. Many of these stem from the widely varying length of TCFs. Human TCFs range from only one or two bases up to several kilobases. Consider, then, the problem of identifying the correct stop codon given a terminal exon's acceptor site. In general, there will be three choices, namely the next downstream stop codon in each frame, and they may plausibly be anywhere from tens to thousands of bases away from the acceptor. Traditional scoring functions (such as codon or in-frame hexamer usage) would evaluate these using only the sequence characteristics of each frame. However, the length of a TCF is also informative: the fact that a frame remains open for thousands of bases is itself suggestive that it is the true coding frame. We therefore developed scoring functions that incorporate both length and sequence information.

Our approach can best be described in terms of a Bayesian coin-flipping problem. Suppose one is given a coin, and an estimated prior probability of 1% that the coin is not fair. The coin is flipped 5 times, resulting in 5 heads.

Application of Bayes' theorem gives:

$$p(\text{unfair}|5H) = \frac{p(5H|\text{unfair})p(\text{unfair})}{p(5H)} \\ = \frac{1 \cdot 0.01}{(0.5)^5 \cdot 0.99 + 1 \cdot 0.01} = 0.244$$

or a 24.4% posterior probability that the coin is unfair. Repeating the experiment with 50 flips, producing 50 heads, yields a nearly 100% probability that the coin is unfair. Clearly, the Bayesian posteriors incorporate both sequence (the results of the flips) and length (the number of flips) information, the characteristic we desire in a TCF scoring function.

The basic form of our Bayesian scoring functions is:

$$P(\text{TCF}|\text{seq}) = \frac{P(\text{seq}|\text{TCF}) \times P(\text{TCF})}{P(\text{seq})}$$

where  $P(\text{TCF})$  is the prior probability that any sequence is a TCF,  $P(\text{seq})$  is the marginal probability of observing the sequence of interest in anywhere the genome and  $P(\text{seq}|\text{TCF})$  is the probability that the observed sequence will be found in a terminal coding fragment. Priors were determined empirically. The various scoring functions differ in how  $P(\text{seq})$  and  $P(\text{seq}|\text{TCF})$  are calculated.

*Bayesian in-frame hexamer score.* Given a putative TCF,  $C_1C_2C_3 \cdots C_nS$ , where the  $C_i$ 's are internal codons and  $S$  is the stop codon,  $P(\text{seq})$  is calculated as:

$$P(\text{seq}) = P(C_1C_2C_3 \cdots C_nS) \\ = P(S) \times P(C_nS|S) \times P(C_{n-1}C_n|C_n) \\ \times \cdots \times P(C_2C_3|C_3) \times P(C_1C_2|C_2).$$

This is essentially a 5th order, 3-periodic Markov chain. Here, the individual hexamer and stop codon probabilities are determined from bulk genomic sequence.  $P(\text{seq}|\text{TCF})$  is calculated using probabilities determined from actual TCFs. Note that as a computational convenience, this score works backward from the stop codon; this allowed us to condition the stop codon frequency  $P(S)$  on TCF 3' ends rather than on bulk coding sequence, which improved predictive accuracy.

For a given acceptor site, in-frame hexamer scores are calculated for each frame. The highest scoring frame is chosen to be the acceptor site's best coding frame (which, in turn, determines the stop codon location).

*Bayesian bulk nmer score.* This score is calculated essentially as the in-frame hexamer score, except that for a given nmer size  $N$ , an  $N-1$ st order non-periodic Markov process is used.

**Bayesian autocorrelation score.** Here,  $P(\text{seq})$  and  $P(\text{seq}|\text{TCF})$  are the probabilities of observing a given number of 3-base identities of the form  $XNNX$  (where the  $X$ 's represent the same base) in bulk sequence and coding sequence, assuming a multinomial distribution.

**Bayesian branch point score.** Given a pentamer of the form  $\text{NTNAN}$ ,  $P(\text{seq}|\text{TCF})$  is the probability of observing that pentamer in the last 100 bases of an intron and  $P(\text{seq})$  the probability of observing the pentamer elsewhere in an intron. Note that this is a 'site' score, and not sensitive to length.

## DISCRIMINANT FUNCTIONS FOR SPLICE SITE RECOGNITION

We tried different scoring functions that incorporate various characteristics of splice sites, and the best scoring functions that discriminate real sites from pseudo sites were used for building the QDFs. The acceptor site QDF uses eight variables: (1) acceptor site weight matrix score; (2) in-frame hexamer transition score, with prior probability ( $P$ ) = 0.1; (3) bulk hexamer transition score,  $P = 0.5$ ; (4) bulk dimer transition score; with  $P = 0.5$ ; (5) downstream bulk trimers, with  $P = 0.5$ ; (6) best branch point score, with  $P = 0.1$ ; (7) upstream  $G$  run count (minimum length = 2); and (8) period 3 autocorrelation transition score, with  $P = 0.5$ . Descriptions of scoring functions not discussed herein may be found in Fickett and Tung (1992) and Fickett (1996). Upstream scores are calculated on the interval  $[-100, -1]$  with respect to the acceptor site, and downstream scores between  $+1$  and the end of the best coding frame. For this and all other discriminants, transition scores are calculated by scoring both the upstream and downstream region of the site, and subtracting the score of the non-coding side of the site from that of the coding side.

The Donor site QDF uses six variables: (1) donor site weight matrix score; (2) in-frame hexamer transition score, with  $P = 0.1$ ; (3) bulk hexamer transition score, with  $P = 0.5$ ; (4) log of exon length; (5) Fourier period three transition score; and (6) bulk trimer transition score, with  $P = 0.5$ . Upstream scores are calculated on the interval between the candidate exon's acceptor site and the last base of the exon, and downstream scores between  $+1$  and the acceptor site's best stop codon.

## DISCRIMINANT FUNCTION FOR POLYA SIGNAL (PAS) RECOGNITION

The polyA signal QDF uses three variables: (1) downstream element weight matrix average score (see Tabaska and Zhang, 1999); (2) downstream element weight matrix average position; and (3) signal-to-site region bulk dimers, with  $P = 0.5$ . The downstream element scoring function was applied to a sequence of region  $(+16, +100)$  and that

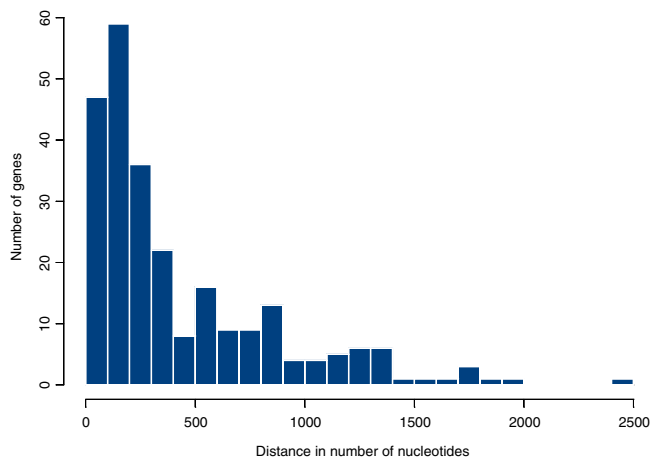


Fig. 1. Distance between STOP codon and polyA signal.

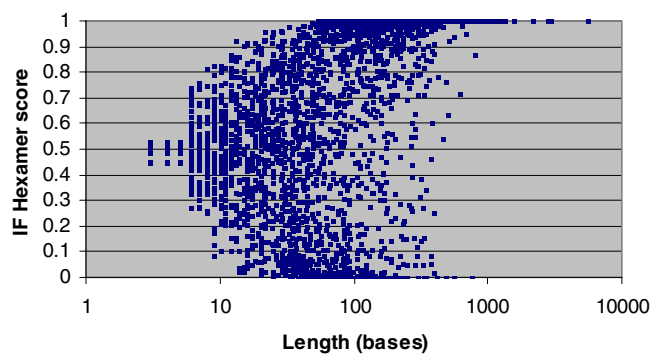


Fig. 2. Bayesian in-frame hexamer score versus coding fragment length.

of signal-to-site function was applied to a sequence of region  $(+6, +30)$ , where  $+1$  is the 5'-most base of PAS. We have also calculated the distance between STOP codon and polyA signal for a set of 253 well-annotated genes (Figure 1). However, this was not considered as one of the feature variables in the PAS QDF as it did not improve the discriminatory power significantly.

## DISCRIMINANT FUNCTION FOR TERMINAL EXON RECOGNITION

The terminal exon QDF uses three variables: (1) acceptor site QDF score; (2) polyA signal QDF score; and (3) terminal coding fragment bulk dimer transition score (across the stop codon) with  $P = 0.1$ .

### The data

**The training sets.** A representative benchmark data set of human DNA sequences generated by Kulp *et al.* (1996) and available at <http://www.fruitfly.org/sequence/humandatasets.html> was used for training acceptor site

**Table 1.** JTEF prediction test results on a test set of 65 sequences with 121 genes

Program	Completely correct exons	Partially correct exons			Total exons correct	Partially predicted as other type of exons	False positive*
		Acceptor match	Stop match	Partial			
JTEF	82	4	13	2	101	–	15
GenScan	68	3	5	1	77	20	7
Genemark.hmm	61	3	7	1	72	24	12

Program	$S_n$	$S_p$	Average
JTEF	83.5	87.1	85.3
GenScan	63.6	91.2	77.7
Genemark.hmm	59.5	85.7	72.6

Total number of terminal exons: 121.

$S_n$  = Sensitivity = the proportion of true exons that are correctly predicted by the program,

$S_p$  = Specificity = the proportion of predicted exons that are actually true exons, average =  $(S_n + S_p)/2$ .

\*False positives are counted within 2 kb upstream and 2 kb downstream regions of the terminal exon.

and donor site QDFs. This set consists of 353 multi-exon genes with 2107 exons. Only those sequences that follow the following criteria are included in the training set:

- every gene should have a *realistic* annotated polyA signal, as judged by manual inspection; and
- every 3' terminal exon should be of type *3tuxon* with at least 100 bases of clear sequence upstream and downstream of the exon boundaries.

The polyA signal QDF was trained on *polyA site database* prepared for our earlier program *polyadq*, a polyA signal finder (Tabaska and Zhang, 1999). The background sequence statistics were worked out on a set of 764 high throughput genomic sequences of human genome. We considered only completed sequences with size larger than 100 kb. Total size of these genomic sequences is ~100 Mb.

*The test set.* For testing the program, a test sequence set of human DNA sequences dated May, 1999 and later (to avoid the training data) were retrieved from GenBank. Only those sequences with proper annotations that are supported by either experimental evidence or exact cDNA matches were considered. Further, the 3' terminal exon of every gene in the test set was of type *3tuxon*. These sequences were augmented with the genomic clone sequence data of Chromosome 1 and 13 (BRCA 2 region) from Annotated Genome Sequence Repository (AGSR) maintained at the Sanger Centre, UK (<http://www.sanger.ac.uk/HGP/Genes/prediction/AGSR.shtml>). Only true genes with *3tuxon* as 3' terminal exon were considered. The test set consists of 65 genomic sequences with 121 3' terminal exons.

## DISCUSSION

The Bayesian scoring functions were very useful in building the quadratic discriminant functions for JTEF. In particular, the Bayesian in-frame hexamer score was proved to be very helpful in recognizing the terminal coding fragment and its frame. Figure 1 shows that this score varies with coding fragment length. We calculated this score for a set of coding and non-coding sequences that began with a true or pseudo acceptor and ended with a TGA, TAG or TAA. It can be observed that the score for the shorter sequences is very close to the prior probability of 0.5. As the sequence gets longer, the scores migrate away from the prior probability toward +1 or 0. This measure becomes virtually a binary function for very long sequences: it is practically certain whether a sequence is TCF (+1) or not (0).

We evaluated the performance of JTEF on a test set of 65 sequences with 121 genes (*3tuxon* as terminal exons) and compared with GenScan (Burge and Karlin, 1997) and Genemark.hmm (Lukashin and Borodovsky, 1998). Even though it may not be completely fair to compare a single module like JTEF with an integrated gene-finder like Genscan, we compared the performance with respect to the 3' terminal exons only. The false positives were counted within 2 kb upstream and 2 kb downstream regions of the terminal exon. This was done in view of the fact that most of the false positives normally occur within the vicinity of the real ones (Thanaraj, 2000). Also JTEF was developed as a module for MZEF. We are in the process of developing 5' terminal exon prediction program. We will eventually integrate all the three modules and we expect the integrated program to eliminate the false positives in the other regions of the genome. False positives for Genscan and Genemark.hmm

**Table 2.** JTEF prediction test results on a test set of 25 multi-gene contigs with 81 genes

Program	Completely correct exons	Partially correct exons			Total exons correct	Partially predicted as other type of exons	False positive
		Acceptor match	Stop match	Partial			
JTEF	53	0	7	2	62	–	13
GenScan	46	1	2	1	50	17	5
Genemark.hmm	42	1	3	1	45	17	8

Program	$S_n$	$S_p$	Average
JTEF	76.5	82.7	79.6
GenScan	61.7	90.9	76.3
Genemark.hmm	55.6	84.9	70.2

**Table 3.** JTEF prediction test results on a test set of 17 multi-gene contigs with 64 genes

Program	Completely correct exons	Partially correct exons			Total exons correct	Number of genes fused with other	False positive
		Acceptor match	Stop match	Partial			
JTEF	41	0	6	2	49	–	9
GenScan	37	1	2	1	41	21	5
Genemark.hmm	35	0	2	1	38	7	7

Program	$S_n$	$S_p$	Average
JTEF	76.6	84.5	80.5
GenScan	64.1	89.1	76.6
Genemark.hmm	59.4	84.4	71.9

include only the predicted terminal exons. If the real terminal exons were predicted as single or internal exons then those were considered as wrong predictions and reported separately (Tables 1–3). The performance results of JTEF in comparison with Genscan and Genemark.hmm are shown in Tables 1–3. Table 1 presents the performance of JTEF in comparison with GenScan and Genemark.hmm on all the 65 sequences. The sensitivity ( $S_n$ ) of JTEF (83.5%) is much higher than both GenScan (63.6%) and Genemark.hmm (59.5%) with comparable specificity ( $S_p$ ). Also, the average of  $S_n$  and  $S_p$  for JTEF (85.3%) is higher than both GenScan and Genemark.hmm. These results show that JTEF has higher average accuracy than both GenScan and Genemark.hmm.

Gene modelers are often criticized for their poor performance on larger genomic contigs containing multiple genes, particularly those cases with two or more genes on the same strand. When genomic sequences with two or more genes on the same strand are submitted to gene modelers they often fail to detect the terminal exons, thereby fusing the genes. We analyzed these three programs on

a subset of sequences that have more than one gene per sequence. This subset contains 25 sequences with 81 genes. Table 2 shows the performance results of JTEF in comparison with GenScan and Genemark.hmm. JTEF has outperformed both GenScan and Genemark.hmm with higher average accuracy. We also analyzed the performance of the three programs on another subset of 17 sequences containing 64 genes (with two or more genes on the same strand). Table 3 shows the results on this subset. Here too, JTEF has outperformed the other two programs with higher average accuracy. GenScan and Genemark.hmm fused 21 (34%) and 7 (11%) of the 64 genes with the adjacent genes respectively. Though Genemark.hmm did not fuse as many as GenScan, it either predicted many of the terminal exons as internal exons or missed them completely thereby giving a wrong gene model. On the other hand JTEF predicted most of these (77%) exons correctly.

The performance of JTEF based on average of  $S_n$  and  $S_p$  is ~80% for all the three cases. This shows that the accuracy of JTEF was not affected by large genomic

sequences containing multiple genes. The only drawback of JTEF is that it is specifically designed to recognize 3' exons (as defined in Zhang, 1998a,b), i.e. majority of 3' terminal exons that contain the stop codon. The 3' exons (<10%, Zhang, 1997) would have to be detected by other methods. JTEF is developed for using it in combination with other internal exon finders such as MZEF. When JTEF is used in combination with these other modules the false positives would be drastically reduced thereby increasing the specificity of overall gene model. Our group is currently working on developing a module for predicting 5' terminal exon and we will integrate all the modules once it is ready. It was demonstrated recently, in re-analyzing human chromosome 21, that even one of the most difficult gene (APP) structure could be accurately predicted in silico (Zhang, 2000), by combining CorePromoter (Zhang, 1998a,b), CpG-promoter (Ioshikhes and Zhang, 2000), MZEF (Zhang, 1997) and JTEF. The performance results of JTEF are encouraging and it will become a valuable addition to the gene finding repertoire.

## ACKNOWLEDGEMENTS

This work was supported by a grant from Merck Genome Research Institute and by a grant from NHGRI/NIH.

## REFERENCES

- Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.
- Claverie,J.M. (1997) Computational methods for the identification of genes in vertebrate genomic sequences. *Hum. Mol. Genet.*, **6**, 1735–1744.
- Clogan,D.F. and Manley,J.L. (1997) Mechanism and regulation of mRNA polyadenylation. *Genes Devel.*, **11**, 2755–2766.
- Fickett,J.W. (1996) The gene identification problem: an overview for developers. *Comput. Chem.*, **20**, 103–118.
- Fickett,J.W. and Tung,C.S. (1992) Assessment of protein coding measures. *Nucleic Acids Res.*, **20**, 6441–6450.
- Ioshikhes,I.P. and Zhang,M.Q. (2000) Large-scale human promoter mapping using CpG islands. *Nature Genet.*, **26**, 61–63.
- Kulp,D., Haussler,D., Reese,M.G. and Eeckman,F.H. (1996) A generalized hidden Markov model for the recognition of human genes in DNA. *ISMB*, **4**, 134–142.
- Lukashin,A. and Borodovsky,M. (1998) GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.*, **26**, 1107–1115.
- Pesole,G., Fiormarino,G. and Saccone,C. (1994) Sequence-analysis and compositional properties of untranslated regions of human messenger-RNAs. *Gene*, **140**, 219–225.
- Pesole,G., Liuni,S., Grillo,G. and Saccone,C. (1997) Structural and compositional features of untranslated regions of eukaryotic mRNAs. *Gene*, **205**, 95–102.
- Stormo,G.D. (2000) Gene-finding approaches for eukaryotes. *Genome Res.*, **10**, 394–397.
- Tabaska,J.E. and Zhang,M.Q. (1999) Detection of polyadenylation signals in human DNA sequences. *Gene*, **231**, 77–86.
- Thanaraj,A.T. (2000) Positional characterization of false positives from computational prediction of human splice sites. *Nucleic Acids Res.*, **28**, 744–754.
- Venables,W.N. and Ripley,B.D. (1994) *Modern Applied Statistics with S-Plus*. Springer, New York.
- Zhang,M.Q. (1997) Identification of protein coding regions in the human genome by quadratic discriminant analysis. *PNAS*, **94**, 565–568.
- Zhang,M.Q. (1998a) Statistical features of human exons and their flanking regions. *Hum. Mol. Genet.*, **7**, 919–932.
- Zhang,M.Q. (1998b) Identification of human gene core promoters in silico. *Genome Res.*, **8**, 319–326.
- Zhang,M.Z. (2000) Discriminant analysis and its application in DNA sequence motif recognition. *Briefings in Bioinformatics*, **1**, 331–342.