UTR Reconstruction and Analysis Using Genomically Aligned EST Sequences

Zhengyan Kan[1], Warren Gish[2], Eric Rouchka[1], Jarret Glasscock[2], and David States[1]

[1]Institute for Biomedical Computing, Washington University
700 S. Euclid Ave., St. Louis, MO 63110
zkan@ibc.wustl.edu, ecr@ibc.wustl.edu, states@ibc.wustl.edu
[2]Genome Sequencing Center, Washington University School of Medicine
4444 Forest Park Blvd., St. Louis, MO 63108
jglassco@sapiens.wustl.edu, gish@watson.wustl.edu

Abstract

Untranslated regions (UTR) play important roles in the post-transcriptional regulation of mRNA processing. There is a wealth of UTR-related information to be mined from the rapidly accumulating EST collections. A computational tool, UTR-extender, has been developed to infer UTR sequences from genomically aligned ESTs. It can completely and accurately reconstruct 72% of the 3' UTRs and 15% of the 5' UTRs when tested using 908 functionally cloned transcripts. In addition, it predicts extensions for 11% of the 5' UTRs and 28% of the 3' UTRs. These extension regions are validated by examining splicing frequencies and conservation levels. We also developed a method called polyadenylation site scan (PASS) to precisely map polyadenylation sites in human genomic sequences. A PASS analysis of 908 genic regions estimates that 40-50% of human genes undergo alternative polyadenylation. Using EST redundancy to assess expression levels, we also find that genes with short 3' UTRs tend to be highly expressed.
Availability: [http://www.ibc.wustl.edu/~zkan/UTR/]

Keywords: UTR reconstruction, gene prediction, mRNA, EST, human genome, alternative polyadenylation.

Introduction

The regulatory role of untranslated regions (UTR) in the eukaryotic transcript is becoming better appreciated as experimental studies discover more and more UTR signals. They function in various post-transcriptional events, such as mRNA turnover, polyadenylation, localization and translational initiation (Jackson 1993). During the early development of *Drosophila*, cis-acting elements in the 3' UTR were found to mediate the subcellular localization of maternal mRNAs and the translational repression of morphogenic genes. (Decker and Parker 1995, Grunert and Johnston 1996). In the *C. elegans* heterochronic gene pathway, regulatory RNA genes such as *lin-4* and *let-7* were found to control the expression of downstream genes by interacting with complementary sites in the 3' UTRs (Lee, Feinbaum and Ambros 1993; Reinhart et al. 2000). In mammalian cells, iron response element (IRE) is known to fine tune intracellular iron homeostasis by controlling the translational initiation of ferritin and mRNA stability of transferrin receptor (Jacobson 1996). In addition, AU-rich element (ARE), a common determinant of mRNA stability, is invariably found in the 3' UTRs (Chen and Shyu 1995). Computational sequence analyses have further suggested the functional importance of UTRs on a genomic scale. By comparing orthologous sequences in different classes of vertebrates, three separate studies found that UTRs are highly conserved in numerous genes (Duret et al. 1993; Makalowski and Boguski 1998; Jareborg et al. 1998). However, apart from the anecdotal evidences, very little is understood about the regulatory events occurring in these regions.

There are growing interests in using a computational approach to model UTR motifs and identify novel regulatory elements (Dandekar and Hentze 1995; Dandekar et al. 1997). The current state of the transcript databases is ill prepared for such undertakings. In Genbank, the UTRs are often poorly annotated, incomplete or missing. For example, most records do not provide annotation concerning polyadenylation sites. Gautheret et al. (1998) noted that 359 out of 720 published sequences examined appear to be incomplete in the 3' UTR. Our own survey of GenBank found that about 10% of human transcript sequences labeled as "complete cds" consist of only the coding sequences. The ongoing genome annotation efforts also neglect the UTRs as few gene predictions contain information about the non-coding regions. Furthermore, statistical gene-finders generally perform poorly in defining the UTRs (Claverie 1997). These problems could pose serious obstacles to UTR-oriented studies. The missing UTR sequences may contain key regulatory elements, such as those in control of polyadenylation activities. Moreover, the vast majority of the human transcriptome would be unavailable for large-scale analyses that search for novel regulatory elements or new genes carrying known UTR motifs. Hence, there is clearly a need for methods that can identify and analyze the untranslated regions for both experimentally cloned and predicted genes.

Expressed sequence tags (EST) provide an abundant resource for analysis of transcripts. There are 1.7 million human EST sequences in the current release (020400) of

dbEST. Moreover, since most human ESTs are primed from the 3' termini of genes, the EST resource is particularly enriched in 3' UTR sequences. Jiang and Jacob (1998) have developed an EST-based gene prediction tool, EbEST, and estimated it can predict 60-70% of 3' UTRs in human genes. However, their method has an error rate of ~10% and ignores important information related to expression level and alternative polyadenylation. Redundancy of ESTs reflects the relative abundance of transcripts and has been used to measure the expression patterns of genes in *C. elegans*, *D. melanogaster* and *A. thaliana* (Duret and Mouchiroud 1999). Thus, an EST-based method can utilize the redundancy information to sort out the complex patterns resulted from alternative forms of transcript. To molecular biologists, alternative polyadenylation is a frequently encountered, yet poorly understood phenomenon. It has been suggested that alternative polyadenylation modulates differential gene expression by producing condition-specific transcript forms with different stability or translatability from the normal transcript (Edwards-Gilbert, Veraldi, and Milcarek 1997). Recently, one genome-wide EST clustering study found evidence of alternative polyadenylation in 189 of the 1000 EST clusters examined (Gautheret et al. 1998). The prevalence of alternative polyadenylation and the growing interest in understanding their functional roles demand a gene prediction method capable of precisely defining polyadenylation sites.

We have developed a tool, UTR-extender, for predicting UTR sequences by clustering ESTs onto a genomic template. UTR-extender consists of a method called **P**oly**A**denylation **S**ite **S**can (PASS) that can pinpoint poly-A sites using EST alignment. This algorithm also uses EST redundancy information to delineate the predominant gene structure. UTR-extender is evaluated by reconstructing the UTRs of 908 functionally cloned sequences. We also use the tool to predict novel extensions to known genes and to analyze the extent of alternative polyadenylation in the human transcribed genome. A web-based interface for UTR-extender and our UTR reconstruction results are available at http://www.ibc.wustl.edu/~zkan/UTR/.

Methods

Sequence Data

6,072 *H. sapiens* mRNA sequences were retrieved from the NCBI RefSeq database (Maglott et al. 2000) dated September, 1999. Coding sequences are extracted according to the "CDS" annotations. *H. sapiens* and *M. musculus* EST sequences were screened from the dbEST database released on March 23, 1999. The *H. sapiens* and *M. musculus* EST datasets consist of 1,285,084 and 431,503 entries respectively. *H. sapiens* genomic contig sequences were retrieved from the Genome Contigs Database dated September, 1999 (Rouchka and States, 1999). The dataset consists of 2,189 contigs and 418,118,168 bases of finished sequence.

Database Searching and Genomic Alignment

Database searches are performed by WU-BLASTN2 (Gish, unpublished). Sim4 (Florea et al. 1998) is used to align cDNA sequence to the genomic sequence. Near-identity thresholds are empirically determined to filter out paralogous or chimeric hits while allowing for a background error rate in cDNA sequencing. *H. sapiens* coding sequences are first searched against the contig database. The high scoring sequences are aligned to the matching genomic contigs. The genomic locus is considered native if the alignment is > 95% identity over the entire gene length. The resulting 908 mRNA sequences make up the REFSEQ dataset. The genomic sequence including each genic region and up to 20 kb extensions at two ends is extracted. The full-length sequence is also aligned to this template. The genomic template is searched against the EST database. High scoring ESTs are aligned to the genomic sequence using sim4. A true EST hit must be > 95% identical over > 90% of the length. Poorly aligned terminal regions are trimmed. EST sequences possibly derived from repetitive elements are filtered by searching against a repeat element library (Smit and Green, unpublished) using WU-BLASTN2. 5' ESTs aligned in the minus orientation and 3' ESTs aligned in the plus orientation are removed as they may be derived from genes on the opposite strand.

Polyadenylation Site Scan (PASS)

The locations of 3' EST alignments are used to precisely define the 3' termini of genes. Only EST sequences annotated as 3' and aligned in the minus direction with respect to the seed mRNA are considered "true" 3' ESTs. The redundancy requirement is used to weed out mislabeled ESTs. The 3' end of a 3' EST alignment is considered to indicate a possible poly-A site. EST alignments that terminate within 10 nts of each other are clustered and represent one possible site. On the genomic template, a 30-nt region upstream of a possible site is searched for canonical poly-A patterns, AATAAA and ATTAAA. 50 nts of downstream sequence is searched for A-rich region, defined as windows of 10 nts containing 8 or more A's. All possible sites are scored. The score is based on 3 factors - size of the cluster, poly-A signal, and downstream A-rich region. By default, the score (S) is the natural log of the cluster size (R). The presence of poly-A signal (P) confers a 2-point addition and the presence of A-rich region (A) results in a 2-point penalty.

$$S = \log R + 2 \times P - 2 \times A$$

$$A = \{0, 1\} \& P = \{0, 1\}$$

The threshold value used is 1.1, requiring a minimum cluster size of 4 for a positive identification. However, the presence of poly-A signal reduces the threshold to 1, whereas possible internal priming raises the threshold to 23. For a positive cluster, all 3' end positions are sorted and the center one is taken to represent the poly-A site. The 3' termini of 129 REFSEQ sequences were annotated as "poly-A site". The genomic locations aligned to these termini make up a set of "known" poly-A sites. Poly-A trimming is often imprecise for both EST and mRNA sequences. Thus, when comparing a predicted site with a known site, two sites are considered to agree when they are within 20 nts of one another.

Gene Boundary Determination

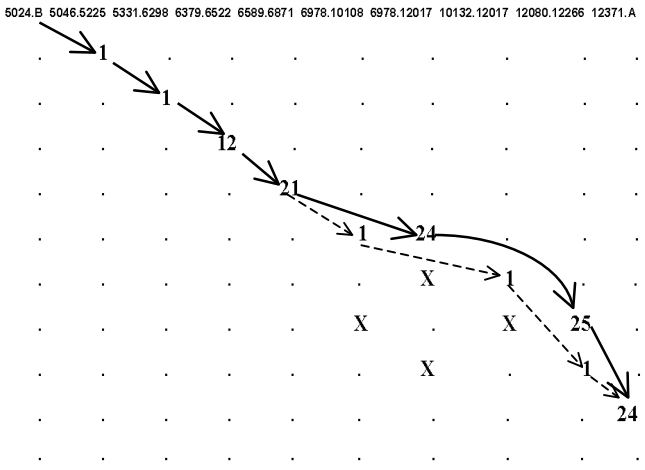
The objective of this experiment is to extend an incomplete transcript, coding sequence in this case, by using EST alignments to highlight the exonic regions on the genomic sequence. To determine the gene boundaries, the coding region is expanded transitively to incorporate adjacent EST alignments. First, overlapping EST alignments are grouped into clusters. Second, the initial boundaries expand toward 5' and 3' directions to include clusters. The boundaries are changed whenever an additional cluster is added. During the expansion, gaps smaller than 500 nts in size are tolerated. However, when expanding downstream, gaps as large as 1.5 kb are allowed before a poly-A site is reached. The 3' terminal has to be polyadenylated unless no site is found within the clustered region. In the case of multiple poly-A sites, the most distal site is chosen. On the other hand, the upstream expansion is terminated prior to incorporating a poly-A site. The 5' termini can be anywhere the EST alignments end.

Gene Structure Prediction

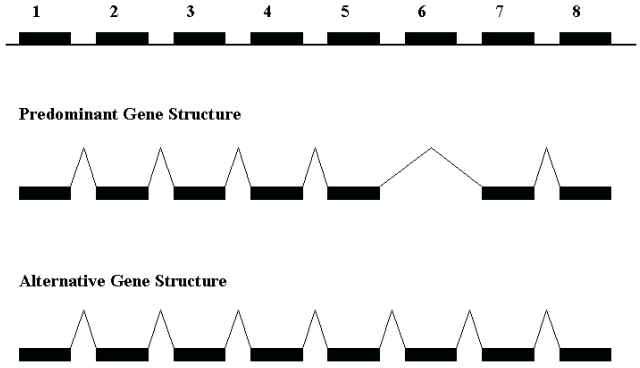
After the boundaries are mapped, the entire gene structure is inferred using an intron-centered strategy different from most gene prediction algorithms. The candidate introns are first inferred from segmentations of EST alignments. Each intron is represented by its start and end coordinates on the genomic sequence. A "legitimate" intron must contain the consensus "GT..AG" signal or have more than 5 EST "carriers" that share the same splicing pattern.

The connections between two introns are classified into 4 types: conflicting, contiguous, transitive and gapped. Conflicting connections arise when two introns overlap with one another but have different coordinates. Such a pattern is assumed to result from alternative splicing. A conflicting connection between two introns, i and j, is given a conflict score A(i,j) = -∞. If i and j are not in conflict, A(i,j) is set to 0. The connection is contiguous when at least one EST carries two introns i and j in neighboring positions. This type of connection is scored by B(i,j), set to the number of ESTs that carry introns i and j

A)



B)



	Exon Boundaries (Predominant)	Exon Boundaries (Alternative)
1	5024-5046	5024-5046
2	5225-5331	5225-5331
3	6298-6379	6298-6379
4	6522-6589	6522-6589
5	6871-6978	6871-6978
6		10108-10132
7	12017-12080	12017-12080
8	12266-12371	12266-12371

Figure 1. Connectivity Matrix

Shown in (A) is a sample connectivity matrix used for reconstructing human UXT mRNA (NM_004182). It is a 10x10 matrix. The 10 nodes shown in the first row include the 5' beginning (5024.B), the 3' poly-A site (12371.A) and 8 predicted introns. Each intron is labeled by its start and end coordinates on the genomic template. In the cell M(i,j), a numerical value indicates the number of ESTs that connect the i^{th} and j^{th} nodes. The "X" sign indicates a conflict between the i^{th} and j^{th} nodes. A null value is indicated by the "." sign. The arrows with a solid line represent a trace of the predominant gene structure. The arrows with a dashed line delineate the alternative gene structure. Both gene structures are shown in (B).

"contiguously". Two neighboring introns may not be seen on the same EST. In that case, a transitive clustering mechanism is used to examine if there are overlapping EST coverage between two introns. If there is overlapping coverage, the connection is "transitive", and the transitive score C(i,j) is set to 1. Otherwise, the connection is "gapped", and C(i,j) is set to 0. A matrix M is constructed to record the connectivity between introns (Fig. 1). The score M(i,j) for the connection between two introns i and j is the sum of A(i,j), B(i,j) and C(i,j).

Once the M matrix is filled, it is traced to assemble a gene structure, starting from the 5' terminal and ending at the 3' terminal. At each step, the trace considers all nodes j downstream of the current node i. The node with the maximum connectivity score is selected. Following this rule, the trace never connects two introns in conflict. When multiple downstream introns are contiguously connected, the trace chooses the one with the most redundant EST coverage. In the absence of contiguous connections, the trace proceeds to the intron immediately downstream that is transitively connected. In the absence of transitive connections, the trace chooses the next intron with a gapped connection.

In this algorithm, redundancy of ESTs is used to identify the predominant gene structure from a myriad of alternative forms. A recent study has found frequent alternative splicing in human genes, particularly in the 5' UTR (Mironov et al. 1999). We believe it is necessary to develop an algorithm that takes into account alternative forms of a transcript.

Evaluation of UTR Reconstruction

The reconstructed UTRs are predicted genic regions that extend beyond the initial boundaries obtained from aligning CDS to the genomic template. To evaluate the accuracy of prediction, the predicted UTR regions are compared with the genomic alignment of known UTRs. To be consistent with established evaluation standards (Burset and Guigo 1996), we calculate Sn (sensitivity), Sp (specificity) and AC (approximate correlation) at the nucleotide level and by sequence. However, UTR-extender does not make a binary decision in "gapped" regions due to a lack of EST information. These sequences are therefore classified as "unknown" and excluded when calculating the accuracy measures. In addition, predicted exonic regions beyond the known boundaries are ignored. These extensions are characterized using different approaches.

Following the above rule, all nucleotides evaluated are either "exonic" or "intronic". For each sequence, TP (true positive) is calculated as the number of exonic nucleotides (nts) predicted as exonic. Likewise, TN (true negative) is the number of intronic nts correctly predicted. FP (false positive) is the number of intronic nts predicted as exonic, and FN (false negative) is the number of exonic nts predicted as intronic. Sn measures the proportion of exonic nts correctly predicted. Sp measures the proportion of predicted exonic nts that are validated. We introduce ISn and ISp to measure the sensitivity and specificity in predicting introns. AC (approximate correlation) is calculated from Sn, Sp, ISn and ISp.

$$Sn = \frac{TP}{TP + FN} \qquad Sp = \frac{TP}{TP + FP}$$

$$ISn = \frac{TN}{TN + FP} \quad ISp = \frac{TN}{TN + FN}$$

$$AC = 2 \times (\frac{1}{4} \times [Sn + Sp + ISn + ISp] - 0.5)$$

In UTR regions without introns, the ISn and ISp are set to 0, and AC is modified to the following:

$$AC = 2 \times (\frac{1}{2} \times [Sn + Sp] - 0.5)$$

Two novel measures, GAP and GET, are introduced. GAP measures the proportion of known UTR sequences considered "unknown". GET measures the proportion of known UTR sequences correctly predicted. Len represents the length of known UTR. The GAP and GET scores are calculated as follows:

$$GAP = \frac{Len - (TP + FN)}{Len}$$

$$GET = \frac{TP}{Len}$$

Sn, Sp, AC values are averaged over sequences for which a reconstruction is made. GAP and GET are calculated from and averaged over all sequences. For sequences without reconstruction, GAP is set to 100% and GET is set to 0%.

Characterization of Extensions

The poly-A site density is defined as the number of predicted poly-A sites per kb of exonic sequence. The density values for 5' and 3' UTRs as well as CDS were estimated from 129 genic regions analyzed by PASS. Only 1 site was found in 258.5 kb of coding sequences. 150 sites were found in 85.2 kb of 3' UTR sequences.

The splicing frequency is measured by the number of splice junctions per kb of exonic sequence. The reference splicing frequencies are calculated from the genomic alignment of 908 REFSEQ sequences. The splice junctions in the extension regions are inferred from EST alignment.

To assess conservation levels, the extension sequences are extracted from the genomic templates and searched against mouse ESTs using WU-BLASTN2. A significant hit must have a minimum length of 50 nts, and its percent

identity must exceed 70%. For each extension, the EST hit with the highest percent identity is taken to represent the degree of conservation. For comparison, the UTR and CDS of the same genes are also searched. For the 3' extensions, an equal-length segment is extracted from the 3' UTR and CDS. The entire region is used if it is shorter than the extension length. For the 5' extensions, sequences are taken from the 5' UTR.

Results

Our method is designed to infer the transcript sequence from genomically aligned ESTs assuming the coding region has been characterized by functional cloning or genomic annotation. It consists of three major steps. The first step is the construction of EST alignments to the genomic sequence. In the UTR reconstruction experiment, we use a "REFSEQ" set consisting of 908 genomically mapped human transcript sequences from the RefSeq database. For each transcript, the coding sequence (CDS) is extracted and searched against the genomic contigs using WU-BLASTN2 (Gish, unpublished). That allows the genomic template – sequence containing the CDS and flanking regions – to be extracted. The genomic template is searched against dbEST. High scoring EST hits are aligned to the genomic template using sim4 (Florea et al. 1998). These EST alignments are then screened for cognate hits. The second step is to define the exact gene boundaries. The 3' polyadenylation sites are predicted to map the 3' termini. Starting from boundaries of the coding region, ESTs are transitively clustered to infer the entire genic region. Finally, the gene structure is inferred. On the genomic template, the 5' and 3' exonic extensions to the coding region make up the reconstructed UTRs. They are evaluated by comparing with the known UTRs (Fig. 2).

Defining the Polyadenylation Sites

Genes on opposite strands frequently overlap in the human genome. Since most human ESTs are derived from the 3' termini of genes, indiscriminate clustering of ESTs tend to falsely associate two genes overlapping in the 3' UTR (Tsai et al. 1994; Hillier et al. 1996; Burke et al. 1998). Furthermore, it is difficult to demarcate two genes closely positioned on the same strand, as statistical gene finders are known to frequently miscall gene boundaries (Claverie 1997). Hence, mapping the 3' polyadenylation sites helps to accurately define the 3' boundaries, a critical step in gene prediction.

We developed a procedure called **P**oly**A**denylation **S**ite **S**can (PASS) to define the poly-A sites using information encoded in the EST alignments. The 3' end of a 3' EST should align to a poly-A site if it is not internally primed. All ESTs labeled as 3' and aligned to the mRNA at the minus direction are considered a true 3' read. All genomic locations aligned to the 3' ends of true 3' ESTs are therefore putative poly-A sites.

After the putative sites are defined, three factors are considered. **(1) EST redundancy.** We observed that true 3' end frequently give rise to a large number of 3' ESTs terminating at approximately the same location. This type of "pileup" effect is considered to be a strong signal of true 3' termini. **(2) Canonical polyadenylation signal.** A region of 30 nucleotides (nts) upstream of the putative 3' end is searched for the hexamer AATAAA or ATTAAA. **(3) Internal priming.** A poly-A stretch internal to an mRNA can anneal to the primer and generate 3' ESTs. It was estimated that internal priming events occur at a frequency of 1.5% (Hillier, 1996). A 50-nt sequence immediately downstream of the putative site is searched for an A-rich region, the presence of which signals internal priming.

To test the accuracy of PASS, 129 mRNA sequences with known 3' polyadenylated ends were selected from the REFSEQ dataset. After scanning the genomic templates, PASS detected 104 of the 129 known sites. That indicates the genomic detection of poly-A sites by PASS has about 80% sensitivity. Since the program needs at least 4 EST hits to score a positive, it was also evaluated by requiring a minimum EST coverage. When 108 sequences with more than 5 EST hits were surveyed, the sensitivity rose to 91.7%.

Since alternative polyadenylation is poorly annotated, we have little evidence to determine whether an unconfirmed site is a false positive or an alternative poly-A site. Two measures are used to evaluate the specificity of the method. The first assumes that polyadenylation does

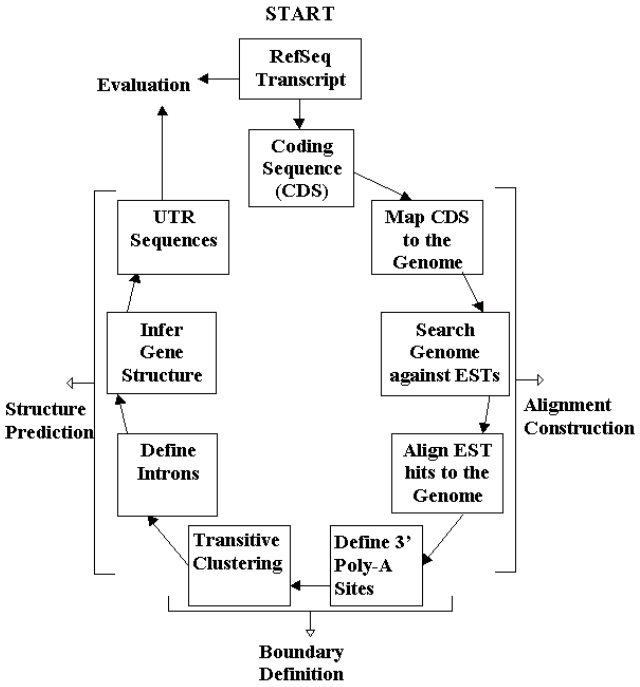


Figure 2. Overview of UTR Reconstruction

not occur in the coding regions. Since PASS scans the entire genic region, any positives identified in the coding region must be false. Only one positive was accepted out of 61 potential sites within the coding regions. 305 potential sites were found in the 3' UTR, and 150 were accepted. A nearly 100% rejection rate in the coding region indicates that PASS prediction is specific.

The downstream element (DE), a G/T rich region with poorly characterized patterns, is frequently observed downstream of the polyadenylation sites (Tabaska and Zhang, 1999). To further test the specificity of PASS predictions, the DE serves as an unbiased evaluation measure since PASS does not use it in the scoring system. A 10-nt region that is rich in G/T and contains the pattern "(C|T) T (G|T|C) T (T|G) (T|G)" is considered a DE. Random simulation estimates that it occurs in a 50-nt window roughly at a 16% frequency. DE elements are found in the downstream regions for 55% of 104 true sites. 48% of unconfirmed sites that are located in the genic regions are found to contain DE elements.

	5' UTR	3' UTR
Sequences	724	780
Reconstruction	288	674
Accurate	238	650
Successful	111	558
Sn (%)	93.21	98.32
Sp (%)	95.89	98.48
AC	0.87	0.96
GAP (%)	68.50	18.86
GET (%)	28.91	79.71

Table 1. Evaluation Results of UTR Reconstruction

Sn and Sp measure sensitivity and specificity in predicting the exonic sequences. AC (approximate correlation) summarizes the sensitivity and specificity measures in both exonic and intronic regions. GAP measures the proportion of known UTR sequence classified as "unknown". GET measures the proportion of known UTR sequence correctly predicted. "Accurate" indicates the number of predictions with perfect Sn, Sp and AC scores. "Successful" indicates the number of "accurate" predictions that fully reconstruct the known UTR, with 100% GET.

UTR Reconstruction

The objective of this experiment is to test the performance of UTR-extender by reconstructing the UTRs of functionally cloned genes. The CDS for 908 genomically mapped REFSEQ genes serve as initial seeds. For each gene, the predicted UTR is compared with the known UTR at the nucleotide level. Conventional criteria for evaluating gene prediction programs – Sn (sensitivity), Sp (specificity) and AC (approximate correlation) - are calculated as well as two additional measures - GAP and GET. GAP measures the proportion of the known UTR that is considered unknown due to a lack of EST coverage. GET measures the proportion of the known UTR correctly predicted.

In REFSEQ, there are 724 5' UTRs and 780 3' UTRs with length longer than 10 nts (Table 1). Overall, reconstruction in the 3' UTR is far more successful than in the 5' UTR. 288 5' UTRs and 674 3' UTRs were reconstructed. 238 5' UTR reconstruction and 650 3' UTR reconstruction are 100% specific, but not necessarily complete. 111 5' UTRs and 558 3' UTRs are completely and accurately reconstructed.

When averaged over all 3' UTR sequences, GET is 80% and AC is 0.96. For 5' UTR, we "GET" approximately 29% of the sequences with 0.87 AC. GAP in the 5' UTR is 69%, compared with 19% in the 3' UTR. The low sensitivity in the 5' UTR can be attributed to poor EST coverage. As measured by Sn, Sp and AC, the prediction accuracy is high in both UTRs. Alternative splicing was found to be the main source of error. The drop in accuracy in the 5' UTR is attributable to the fact that splicing occurs more often in the 5' UTR than 3' UTR.

Characterization of UTR Extensions

28% of 5' UTR reconstruction and 33% of 3' UTR reconstruction consist of extensions. In 73% of the predicted 3' UTRs, the known 3' end is located in proximity to a predicted poly-A site, but 20% of them have downstream extensions (Table 2). These are likely due to alternative polyadenylation. In addition, a substantial

	5' UTR	3' UTR	3' end defined	3' end undefined
Total	288	674	495	179
Same site	123	410	396	14
Incomplete	84	43	0	43
Extension	81	221	99	122
% Extension	**28.1**	**32.8**	**20.0**	**68.2**

Table 2. Summary of UTR Extensions

A predicted end that extends the known end by > 20 nts is classified as an "extension". A predicted end that falls short of the known end by > 20 nts is called "incomplete". Others are considered to terminate at the "same site". The 3' end of the reference sequence is "defined" when a poly-A site is predicted within 20 nts of the known end. However, additional poly-A sites may exist downstream and result in extension.

number of functionally cloned transcripts may be incomplete in the UTRs. However, neighboring genic regions could be incorporated in the transitive clustering process. To investigate this possibility, we analyzed several characteristics of the extension regions - continuity of EST coverage, density of polyadenylation sites, splicing frequency, and conservation level.

The average length of the extended genomic regions is 614 nts for 5' UTR and 832 nts for 3' UTR. The maximal extension is 6,902 nts for 5' UTR and 9,712 nts for 3' UTR. 72% of the 3' extensions and 100% of the 5' extensions have overlapping EST coverage in the extension region. In 62 gapped 3' extensions, the largest gap is ~1000 nts and the median is 278 nts. The overlapping coverage and small gap sizes indicate these extensions are either UTRs or nearby genes in very close proximity.

There are 350 predicted poly-A sites in the extension regions (Table 3). 93% of the 3' extensions end at a predicted poly-A site. Since few positive predictions are made in the coding regions, the density of polyadenylation sites suggests that these extensions are 3' UTR. It has been noted that 3' UTR contains distinctively fewer splicing junctions than 5' UTR and coding region (Pesole et al. 1997; Makalowski and Boguski 1998). We define splicing frequency as the number of splice sites per kb of exonic sequence. A survey of the entire REFSEQ set estimates the splicing frequency at 0.18 for 3' UTR, 2.33 for 5' UTR and 4.94 for coding regions. For the 3' extension regions, 20 splice sites were found in 137,135 nts of exonic sequences. That yields a low splicing frequency of 0.15, characteristic of 3' UTR. The 5' extension regions have a frequency of 3.18, comparable to that of 5' UTR.

Makalowski and Boguski (1998) estimated that the degrees of conservation between orthologous transcripts of human and rodents are ~70% for UTRs and ~90% for CDS. Since rodent orthologs are not available for all REFSEQ sequences, we assessed conservation levels by searching them against the mouse ESTs. The maximal percent identity found in EST hits is used to represent the degree of conservation of each sequence. For comparison, we also searched UTR and coding sequences extracted from the same genes. It was found that the extension regions share similar conservation properties with their respective UTRs. For example, 75% of 3' extensions and 68% of 3' UTRs have blast hits ≥ 70% identity (Table 3). Only 12% of 3' extensions and 13% of 3' UTRs have hits ≥ 90% identity. The median conservation is 77% for both 3' extensions and 3' UTRs, 72% for the 5' extensions, and 73% for the 5' UTRs. At 87%, the coding sequences have a notably higher conservation level. Moreover, 38% of coding sequences have hits ≥ 90% identity. The conservation levels of 5' UTR and 3' UTR are not comparable, as EST coverage in human and mouse are biased toward the 3' termini, but against the 5' termini of genes. However, the conservation properties clearly suggest a link between extensions and their corresponding UTRs. Furthermore, the significant conservation levels in these extension regions suggest functional importance.

Analysis of Alternative Polyadenylation

PASS analysis of 129 genic regions reveals 31 genes with additional poly-A site(s) upstream of the annotated poly-A site. Furthermore, UTR-extender predicts downstream extensions for 33% of genes. These observations suggest that alternative polyadenylation is common in the human transcribed genome. To probe this hypothesis, we carried out a large-scale analysis of alternative polyadenylation.

We defined "the frequency" as the proportion of sequences with multiple poly-A sites, and "the ratio" as the average number of poly-A sites per sequence. Predicted poly-A sites were mapped to the exonic regions of 780 genes with known 3' UTRs. The resulting frequency is 27%, and the ratio is 1.23. However, it is possible to under-estimate the frequency of alternative polyadenylation at low EST coverage. As we increment the cutoff of minimum 3' EST coverage for the dataset, both frequency and ratio increase initially but soon

	5' Extension	5' UTR	CDS	3' UTR	3' Extension
Density of poly-A site	0	0	~0	1.76	2.55
Splicing frequency	3.18	2.33	4.94	0.18	0.15
Median conservation	72%	73%	87%	77%	77%
Percent identity ≥ 70%	24 (48)	16 (28)	159 (183)	124 (183)	137 (183)
Percent identity ≥ 90%	2 (48)	1 (28)	70 (183)	23 (183)	21 (183)

Table 3. Characterization of extension regions

The poly-A site density is the number of poly-A sites predicted per kb of exonic sequence. The splicing frequency is measured by the number of splice junctions per kb of exonic sequence. The highest similarity among EST hits represents the conservation level for each sequence. The median conservation is the median in all sequences. Also shown are the numbers of sequences with conservation ≥ 70% and 90% identity, the estimated conservation levels for UTR and coding region between human and rodents. The numbers in parenthesis are the 7total numbers of sequences searched.

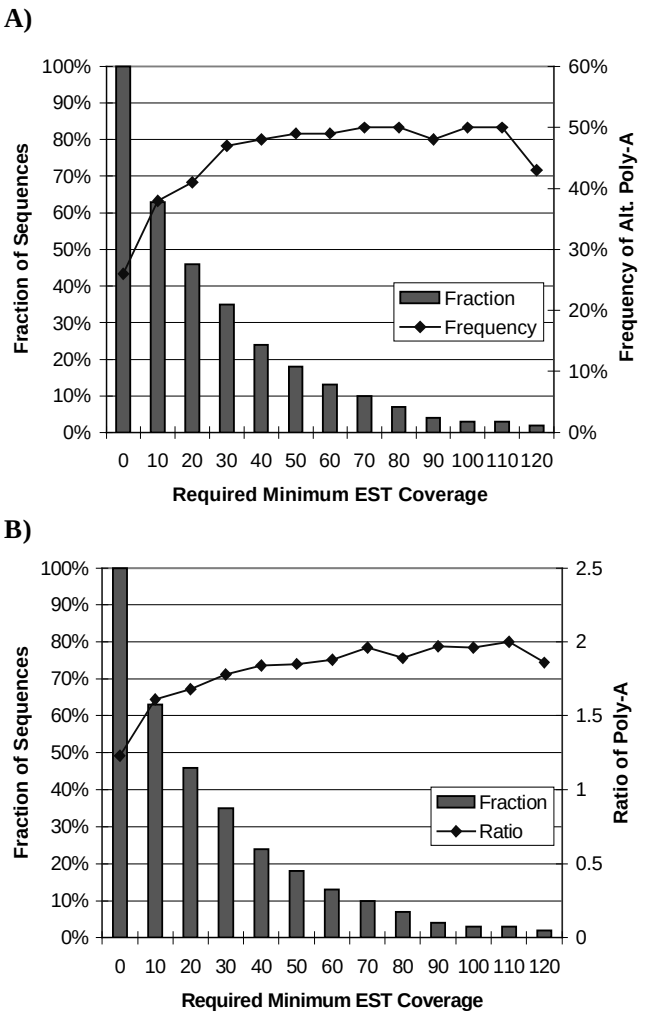


Figure 3. Correlation between Frequency of Alternative Polyadenylation and EST Coverage

A threshold on minimum EST coverage, the number of 3' ESTs per sequence, is imposed on each subset of sequences analyzed. (A) As the threshold is raised from 0 to 120, the fraction of sequences (bar) that satisfy the threshold is decreasing. The frequency of alternative polyadenylation (line) initially increases but stabilizes around 47%. (B) Likewise, the ratio of polyadenylation (line) also increases to a plateau around 1.8.

reach a plateau (Fig. 3). For sequences with more than 30 3' EST hits, the frequency increases to 47% and the ratio increases to 1.78. These values stay about the same at higher EST coverage. Hence, the initial increase in frequency and ratio is likely resulted from an influx of polyadenylation-related information, not a side effect of having more EST hits. Moreover, the data suggests that up to 50% of the human transcripts may undergo alternative polyadenylation.

We also examined the expression patterns of poly-A sites by measuring the number of 3' ESTs terminating at each site. A transcript generally contains a "predominant" site receiving notably higher EST coverage than "alternative" sites. In 210 alternatively polyadenylated sequences, the predominant site on average has 33 EST carriers, whereas the alternative site only has 6. 162 of these have a known 3' end matching a predicted poly-A site, which is predominant 74% of the time. Hence, it is possible to infer the predominantly expressed transcript form, which tends to be functionally cloned, from the redundancy of EST hits.

When we measure the gene expression level by the number of 3' EST hits at the poly-A site, a correlation is discovered between short 3' UTRs and high expression levels. All 3' UTRs are required to have at least one defined poly-A site, so they are likely to be complete. It is found that 7 out of the 10 most highly expressed genes contain 3' UTRs shorter than 100 nts. The correlation is not entirely due to a smaller size of highly expressed proteins. When only genes with short CDS (≤800 nts) are considered, the group with the shortest UTRs (≤80 nts) enjoys the highest average expression level (Fig. 4).

Discussion

We have developed a computational tool called UTR-extender to infer UTR sequences from genomically aligned ESTs. In a test set of 908 functionally cloned sequences, UTR-extender can successfully reconstruct about 72% of the 3' UTRs and 15% of the 5' UTRs. On average, 80% of a 3' UTR sequence and 29% of a 5' UTR sequence is accurately predicted. UTR-extender also found extensions for 11% of the 5' UTRs and 28% of the 3' UTRs. Splicing frequency and conservation levels in these extension regions were characterized. UTR-extender consists of a method, PASS, which can pinpoint the polyadenylation sites in the 3' UTR. Using PASS in a large-scale analysis of polyadenylation, we estimate there are 40-50% of human genes containing alternative poly-A sites. Furthermore, using EST redundancy to estimate gene expression level, we find that genes with short 3' UTR tend to be highly expressed. Our results suggest that EST resources have great potential for sequence mining and functional analysis in the untranslated regions.

The decision to cluster ESTs onto the genome instead of clustering ESTs de novo is based on the common observation that EST is an error-prone resource (Hillier et al. 1996; Wolfsberg and Landsman 1997, Bouck et al. 1999). Without rigorous filtering, the hypothetical transcripts assembled from ESTs can contain a high error rate and mislead subsequent functional analyses. In our study, the high-quality genomic sequence serves as a reference that flags any EST not identically or entirely aligned. A significant number of ESTs filtered are found to be chimeric, low quality or contaminated at termini (data not shown). All of these would be hard to recognize in EST self-clustering. Moreover, the transcript is inferred from the genomic sequence, not from the multiple alignment of ESTs, which at low redundancy is prone to

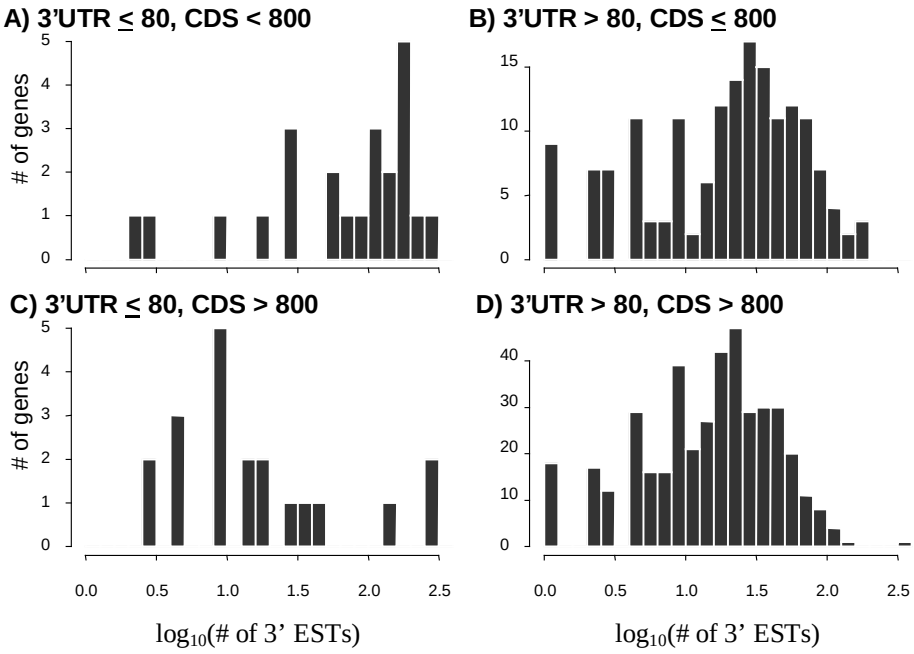


Figure 4. Correlation between Expression Level and 3' UTR Length

Shown in the figure are distributions of estimated gene expression level ($\log_{10}$ # of 3' ESTs) based on the length of the coding region (A&B vs. C&D) and length of the 3'UTR (A&C vs. B&D). (A) The subset of transcripts having both short CDS (≤800 nts) and short 3'UTRs (≤80 nts) tend to be highly expressed. For transcripts with a longer 3' UTR (>80 nts), the expression level of genes with short CDS (B) tends to be higher than that for genes with longer CDS (D). Relatively few transcripts were found with long coding sequences and short 3' UTRs (C).

incorporate sequencing errors or paralogous sequences. A contiguous template also provides a context for EST clustering and allows it to bridge gaps in EST coverage. We observe that ESTs derived from two far-apart poly-A sites cannot overlap since the length of EST is limited. As a result, ESTs from the same 3' UTR but different poly-A sites tend to be dissociated into separate clusters. Hence, self-clustering of ESTs is likely to miss distal alternative poly-A sites, under-estimate the true extent of alternative polyadenylation, and over-estimate the size of the human transcriptome.

Our analyses revealed extensive quality problems in the transcript databases with regards to the UTRs. Of 908 sequences examined, 20% have little 5' UTR and 14% have little 3' UTRs. Of the remaining sequences, UTR-extender can extend 11% of the 5' UTRs and 28% of the 3' UTRs. Some extensions can be attributed to biological phenomena such as alternative polyadenylation. However, there is likely a large number of incomplete transcripts. Polyadenylation is also poorly annotated. In 73% of the reconstruction, the known 3' termini agree with poly-A site predictions, but only about 20% of these are annotated as having a polyadenylated end. Our findings suggest an urgent need for curation efforts, such as the ongoing RefSeq project, to establish a gold standard for the transcript sequences. We believe UTR-extender can be helpful in such undertakings. For example, it would be impractical to go back and experimentally investigate the completeness for all of the transcripts in the database. However, an EST-based software tool can easily process large amounts of sequence data and earmark problem cases for expert review.

It remains a mystery whether the use of alternative poly-A sites serves a regulatory role or is due to non-specific selection by the polyadenylation machinery at the 3' end of a pre-mRNA. Our comparative analysis using mouse ESTs reveal significant conservation for 3' UTR extensions, most of which result from alternative polyadenylation events at a site downstream of the canonical site. We believe the conservation level is a strong hint that alternative polyadenylation activity is functional.

Acknowledgments

We sincerely thank Dr. I. Korf for stimulating discussions about genome annotation and Dr. M. Lovett for information related to cDNA cloning. This work is supported in part by grants from the Dept. of Energy (DE-FG02-94ER61910) and the National Institute of Health (R01-HG01391).

References

Bouck, J., Yu, W., Gibbs, R., and Worley, K. 1999. Comparison of Gene Indexing Databases. *Trends in Genetics* 15:159-161.

Burge, C., and Karlin, S. 1997. Prediction of Complete Gene Structures in Human Genomic DNA. *J. Mol. Biol*. 268:78-94.

Burke, J., Wang, H., Hide, W., and Davison, D. B. 1998. Alternative Gene Form Discovery and Candidate Gene Selection from Gene Indexing Projects. *Genome Res* 8:276-90.

Burset, M., and Guigo, R. 1996. Evaluation of Gene Structure Prediction Programs. *Genomics* 34:353-367.

Chen, C-Y. A., and Shyu A-B. 1995. AU-rich Elements: Characterization and Importance in mRNA Degradation. *Trends Biochem. Sci.* 20:465-70.

Claverie, J. M. 1997. Computational Methods for the Identification of Genes in Vertebrate Genomic Sequences. *Hum. Mol. Genet.* 6:1735-1744.

Dandekar, T., Beyer, K., Bork, P., Kenealy, M. R., Pantopoulos, K., Hentze, M., Sonntag-Buck, V., Flouriot, G., Gannon, F., Keller, W., and Schreiber, S. 1997. Systematic Genomic Screening and Analysis of mRNA in Untranslated Regions and mRNA Precursors: Combining Experimental and Computational Approaches. *Bioinformatics* 14:271-278.

Dandekar, T., and Hentze, M. W. 1995. Finding the Hairpin in the Haystack: Searching for RNA Motifs. *Trends Genet.* 11:45-50.

Decker, C. J., and Parker, R. 1995. Diversity of Cytoplasmic Functions for the 3' Untranslated Region of Eukaryotic Transcripts. *Curr. Opin. in Cell Biol.* 7:386-392.

Duret, L., Dorkeld, F., and Gautier, C. 1993. Strong Conservation of Non-coding Sequences During Vertebrate Evolution: Potential Involvement in Post-transcriptional Regulation of Gene Expression. *Nucleic Acids Res.* 21:2315-2322.

Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M., and Miller, W. 1998. A Computer Program for Aligning a cDNA Sequence with a Genomic DNA Sequence. *Genome Res.*8:967-974

Gautheret, D., Poirot, O, Lopez, F., Audic, S., and Claverie, J. M. 1998. Alternative Polyadenylation in Human mRNAs: A Large-scale Analysis by EST Clustering. *Genome Res.* 8:524-530.

Gish, W. 1998. WU-BLAST2. http://blast.wustl.edu/

Grunert, S., and Johnston, D. S. 1996. RNA Localization and the Development of Asymmetry During *Drosophila* Oogenesis. *Curr. Opin. Genet. Dev.* 6:395-402.

Hillier, L., Lennon, G., Becker, M., Bonaldo, M. F., Chiapelli, B., Chisoe, S., Dietrich, N, Dubuque, T., Favello, A., Gish, W. et al. 1996. Generation and Analysis of 280,000 Human Expressed Sequences Tags. *Genome Res.* 6:807-828.

Jackson, R. J. 1993. Cytoplasmic Regulation of mRNA Function: The Importance of the 3' Untranslated Region. *Cell* 74:9-14.

Jacobson, A. 1996. Interrelationships of the pathways of mRNA Decay and Translation in Eukaryotic Cells. *Annu. Rev. Biochem.* 65:693-739.

Jareborg, N., Birney, E., and Durbin, R. 1999 Comparative Analysis of Non-coding Regions of 77 Orthologous Mouse and Human Gene Pairs. *Genome Res.* 9:815-824.

Jiang, J., and Jacob, H. J. 1998. An Automated Tool for using Expressed Sequence Tags to Delineate Gene Structure. *Genome Res.* 8:268-275.

Lee, C., Feinbaum, R., and Ambros, V. 1993. The Product of the Heterochronic Gene *lin-4* is a Small RNA with Antisense Complementarity to *lin-14*. *Cell* 75:843-854.

Maglott, D. R., Katz, K. S., Sicotte, H., and Pruitt, K. D. NCBI's LocusLink and RefSeq. *Nucleic Acids Res.* 28:126-128.

Makalowski, W., and Boguski, M. S. 1998. Evolutionary Parameters of the Transcribed Mammalian Genome: an Analysis of 2,820 Orthologous Rodent and Human Sequences. *Proc. Natl. Acad. Sci. U. S. A.* 95:9407-9412.

Mironov, A. A., Fickett, J. W., and Gelfand, M. S. 1999. Frequent Alternative Splicing of Human Genes. *Genome Res.* 9:1288-1293.

Pesole, G., Liuni, S., Grillo, G., and Saccone, C. 1997. Structural and Compositional Features of Untranslated Regions of Eukaryotic mRNAs. *Gene* 205:95-102.

Rouchka, E. C., and States, D. J. 1999. Assembly and Analysis of Extended Genomic Contig Regions, Technical Report, WUCS-99-10, Dept. of Computer Science, Washington Univ. http://stl.wustl.edu/contigs/HUMAN/

Smit, A. F. A., and Green, P. 1996. RepeatMasker. http://ftp.genome.washington.edu/RM/RepeatMasker.html

Tabaska, J. E., and Zhang, M. Q. 1999. Detection of Polyadenylation Signals in Human DNA Sequences. *Gene*. 231:77-96.

Tsi, J. Y., Namin-Gonzales, M. L., and Silver, L. M. 1994. False Association of Human ESTs. *Nature Genet.*2:321-322.

Wolfsberg, T. G., and Landsman, D. 1997. A Comparison of Expressed Sequence Tags (ESTs) to Human Genomic Sequences. *Nucleic Acids Res.*25:1626-1632.