

UTR Reconstruction and Analysis Using Genomically Aligned EST Sequences

Zhengyan Kan¹, Warren Gish², Eric Rouchka¹, Jarret Glasscock², and David States¹

¹Institute 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

²Genome Sequencing Center, Washington University School of Medicine
4444 Forest Park Blvd., St. Louis, MO 63108
jglassco@sapiens.wustl.edu, gish@watson.wustl.edu

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, Decker and Parker 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). There are growing interests in developing computational methods to identify and analyze the untranslated regions. Moreover, the highly abundant and rapidly accumulating EST collections store a wealth of UTR-related information to be mined.

We have developed a method called PolyAdenylation Site Scan (PASS) to precisely map polyadenylation sites in human genomic sequences. PASS looks for candidate poly-A sites highlighted by 3' EST clusters. The true poly-A sites are determined after considering three factors – EST redundancy, poly-A signal and internal priming. In 129 genes with known poly-A sites, our program scores a sensitivity of 80% regardless of EST coverage. An algorithm has been developed to infer the predominant gene structure from the complex and often conflicting splicing patterns in genomic EST alignment. Combined with the predicted 3' termini information, our program UTR-extender is able to accurately infer UTR sequences giving the coding sequence as a seed.

When tested using 908 functionally cloned transcripts, UTR-extender can completely and accurately reconstruct 72% of the 3' UTRs and 15% of the 5' UTRs. 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. Furthermore, a PASS analysis of 908 genic regions estimates that 40-50% of human genes undergo alternative polyadenylation. This result reveals extensive variability in the 3' termini of human genes and invokes an intriguing question about the functional role of alternative polyadenylation.

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.

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

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.

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.