Transcript Sequence Registry

Junping Jing junping_2_jing@sbphrd.com
Jian Jiang jiangj00@hotmail.com
Jeffrey S. Aaronson jeffrey_s_aaronson@sbphrd.com
SmithKline Beecham, 709 Swedeland Road, King of Prussia, PA 19406, USA

Keywords: transcript, mRNA

1 Introduction

In many computational genomic studies, including large-scale expression profiling, it is useful to have a comprehensive collection of transcript sequences. The Reference Sequence [NCBI, 1999] project at NCBI is an effort to identify a non-redundant set of full-length mRNA sequences contained within GenBank [Benson *et al.*, 1999] entries. Currently, RefSeq consists of approximately 6,300 human mRNA sequences and approximately 2,000 mouse mRNA sequences. Since each RefSeq entry requires human review in order to meet desired quality standards, the rate of introduction of new RefSeq sequences is limited. At the same time, high-throughput genomic sequencing efforts, along with high-throughput annotation, are increasing the rate of introduction of (putative) new genes, full-length and partial, into GenBank. As such, it is unlikely that RefSeq will provide a comprehensive collection of transcript sequence from GenBank for some time. In order to provide a comprehensive repository of transcript sequences to support our internal projects, we have developed a system, the Transcript Sequence Registry (TSR), which automatically excises transcript sequences from both RefSeq and GenBank, and then stores the sequences, along with associated information, in a relational database. The Transcript Sequence Registry is intended to be the foundation of a gene-oriented resource; though unlike UniGene [Wheeler *et al.*, 2000], expressed sequence tags (EST) are excluded.

2 Methods and Results

Human, mouse and rat entries in GenBank and RefSeq are collected, and transcript sequence within these entries are identified. Currently, we consider transcript sequence to be indicated by the following features within the feature table: mRNA, CDS, 5'UTR, 3'UTR, and exon. For RNA entries, as indicated on the LOCUS line, the source feature is used as well. Each transcript sequence is stored in Sybase relational tables, together with all its feature qualifiers. The transcript sequence is spliced out according to instructions in the Location field. The TSR is updated daily from GenBank non-cumulative update files and at each full GenBank release. A checksum for each GenBank entry is calculated and stored, so modified GenBank entries can be recognized and transcripts derived from these modified entries can be updated. A checksum for each transcript sequence is computed, and compared with previous sequence checksums in order to recognize, and record, multiple instances of the same sequence. As of Jan. 11, 2000, the TSR contains 120,480, 47,101 and 18,406 unique transcript sequences, respectively for human, mouse and rat. We employ an internally-developed program *Redundant3* [Mott, R., unpublished results] to identify pairwise sequence similarity relationships within each transcript set. Thresholds for matches are stringent, in order to ensure that each pair of matching sequences represent a single gene. From these relationships, we create BLAST-searchable [Altschul *et al.*, 1990] non-redundant transcript databases for human, mouse and rat transcript sequences.

The TSR is a data infrastructural resource that provides a comprehensive collection and collation of publicly available transcript sequences. We are currently exploiting the TSR to aid in the biological characterization of cDNA clones on microarray grids in transcript profiling studies.

References

[1] Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipmann, D.J., Basic local alignment search tool, *J. Mol. Biol.*, 215(3):403–410, 1990.

[2] Benson, D.A., Boguski, M.S., Lipmann, D.J., Ostell, J., Ouellette, B.F., Rapp, B.A., and Wheeler, D.L. GenBank. *Nucleic Acid Res.*, 27(1):12–17, 1999.

[3] NCBI RefSeq web page: http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html.

[4] Wheeler, D.L., Chappey C., Lash A.E., Leipe, D.D., Madden, T.L., Schuler, G.D., Tatusova, T.A. and Rapp, B.A. Database resources of the National Center for Biotechnology Information, *Nucleic Acid Res.*, 28(1):10–14, 2000.