Nucleic Acids Research, 1999, Vol. 27, No. 1 293–294

TransTerm, the translational signal database, extended to include full coding sequences and untranslated regions

Mark E. Dalphin*, Peter A. Stockwell, Warren P. Tate and Chris M. Brown

Department of Biochemistry, University of Otago, PO Box 56, Dunedin, New Zealand

Received October 9, 1998; Revised October 14, 1998; Accepted October 23, 1998

ABSTRACT

TransTerm is a database of mRNA sequences and parameters useful for detecting translational control signals in general. TransTerm-98 has been expanded beyond previous years to include full coding sequences and UTRs, while retaining the original small contexts about the coding sequence start- and stop-codons. The database contains more than 130 000 non-redundant coding sequences with associated untranslated regions (UTRs) from over 450 species. This includes the complete genomes of 12 prokaryotic and one eukaryotic organism. Several coding sequence parameters are available: coding sequence length, Nc, GC3 and, when it is computable, Codon Adaptation Index (CAI). Codon usage tables and summaries of start- and stop-codon contexts are also included. TransTerm-98 has both a relational database form with a WWW interface and a flatfile format, also available by Internet browser. TransTerm is available at: http://biochem.otago.ac.nz:800/Transterm/homepage.html

TransTerm, initially a database of mRNA sequence contexts flanking start- and stop-codons, has been expanded to include both the full coding sequence (CDS) and surrounding upstream and downstream untranslated regions (UTR). Furthermore, at users' requests, we have retained both our original flatfile format (1) with the newer relational database format reported last year (2). Figure 1 shows the sections of an idealised mRNA sequence which are contained in TransTerm. TransTerm still includes the coding sequence parameters: effective number of codons (Nc), fraction of codon third position G+C (GC3) and Codon Adaptation Index (CAI) where it can be calculated (3,4). Summaries of base frequencies in the start- and stop-codon regions as well as codon usage tables are also available.

TransTerm-98 has been prepared from GenBank, release 106. It includes data on 454 species, including 12 complete prokaryotic genomes available from GenBank and one complete eukaryotic genome. Species are selected on the basis that they contain at least 40 coding sequences which do not duplicate each other. Organelle and bacteriophage genomes which often have fewer than 40 coding sequences, have less stringent requirements to be included. Coding sequences are extracted from the GenBank entry using the 'CDS' and 'mat_peptide' Feature Table entries. Untranslated regions are included in the database when documented in the GenBank Feature Table; otherwise 500 nucleotides in the 5′-direction and 2000 nucleotides in the 3′-direction are included as UTRs. If a UTR of this length overlaps with another coding region, the UTR is truncated at the interface with that coding region. There are currently over 136 000 non-redundant coding sequences in TransTerm-98 with substantially more

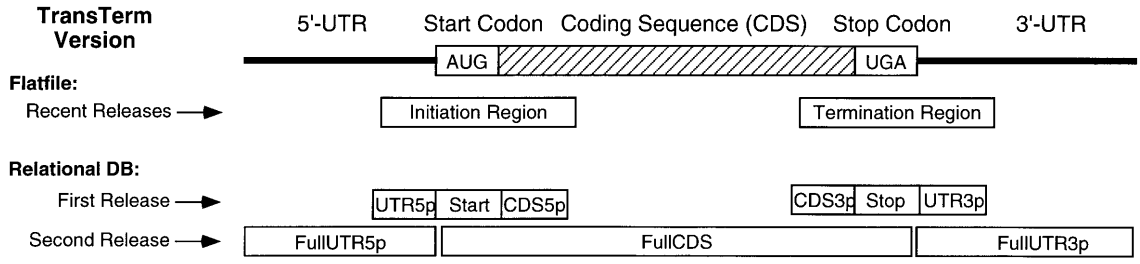

Figure 1. An idealised mRNA, showing the coding sequence bracketed by a start codon (AUG) and a stop codon (UGA). The untranslated regions (UTR), 5′ and 3′ to the coding sequence, are shown as well. The flatfile releases of TransTerm include sequence from the initiation region and the termination region in the *.dat files (1). The relational database form of TransTerm includes these same regions broken into the six columns: UTR5p, Start, CDS5p, CDS3p, Stop and UTR3p. The latest release of Transterm also includes the full coding sequence (FullCDS) and full UTRs (FullUTR5p, FullUTR3p).

*To whom correspondence should be addressed. Tel: +64 3 479 7841; Fax: +64 3 479 7866; Email: mdalphin@sanger.otago.ac.nz

redundant coding sequences included in the relational database format.

The relational database format of TransTerm now has a World Wide Web (WWW) interface to allow retrieval of selected portions of the database by species. We are investigating methods of extending the versatility of the interface to allow selection over a wider set of parameters. Additionally, an experimental WWW interface has been added to allow users to scan patterns across the CDS and UTR sequence data, looking for signals. The interface uses the program, scan_for_matches (5), to search the sequence data for patterns, returning matched sequences by Email. The sequence data searched are the redundant sequences from GenBank. Searching the redundant data aids users to locate documentation which may be scattered among several GenBank entries. The patterns used by scan_for_matches include simple sequence matches as well as more complex patterns which may express variable spacing between elements or elements like stem–loops.

TransTerm is available on the WWW at: http://biochem.otago.ac.nz:800/Transterm/homepage.html

REFERENCES

1 Brown,C.M., Dalphin,M.E., Stockwell,P.A. and Tate,W.P. (1993) *Nucleic Acids Res.*, **21**, 3119–3123.
2 Dalphin,M.E., Brown,C.M., Stockwell,P.A. and Tate,W.P. (1998) *Nucleic Acids Res.*, **26**, 335–337.
3 Sharp,P.M. and Li,W.H. (1987) *Nucleic Acids Res.*, **15**, 1281–1295.
4 Wright,F. (1990) *Gene*, **87**, 23–29.
5 Dsouza,M., Larsen,N. and Overbeek,R. (1997) *Trends Genet.*, **13**, 497–498.