 Nucleic Acids Research, 2000, Vol. 28, No. 1 **293–295**

Transterm: a database of messenger RNA components and signals

Grant H. Jacobs, Peter A. Stockwell, Mark J. Schrieber, Warren P. Tate and Chris M. Brown*

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

Received October 7, 1999; Accepted October 11, 1999

ABSTRACT

Transterm facilitates studies of messenger RNAs and translational control signals. Each messenger RNA (mRNA) from GenBank is extracted and broken into its functional components, its coding sequence, initiation context, termination context, flanking sequence representing its 5′ UTR (untranslated region), 3′ UTR and translational signals. In addition, numerical parameters characterising each coding region in Transterm, including codon and GC bias, are available. For each species in Transterm, the initiation and termination regions are aligned by their start or stop codons and presented as base frequency matrices and tables of the information content of the bases in the alignments. Users can obtain summaries of characteristics of the mRNAs for species of their choice and search for translational signals both in the Transterm database and in their own sequence. The current release contains data from over 10 000 species, including the complete genomes of 20 prokaryotes and three eukaryotes. Both flat-file and relational database forms of Transterm are accessible via the WWW at http://biochem.otago.ac.nz/Transterm/

SEQUENCE DATA IN Transterm

Originally designed for studies of control of translation termination, recent versions of Transterm are aimed at studying translational control signals wherever they might occur in a messenger RNA (1,2). Messenger RNA (mRNA) sequences in Transterm are considered to be composed of several functional components (Fig. 1), which can be analysed by species and are searchable (see below). Experimentally determined untranslated regions (UTRs) are a small subset of the known sequence data (3). We have created a large sequence database of 5′ and 3′ flanks in Transterm which can be searched for translational signals in lieu of experimentally-determined UTRs. These 5′ and 3′ flanks are considered to span from the base immediately outside the coding region for 1000 or 3000 bases, respectively, or until another coding sequence (CDS) is encountered. Likewise, initiation and termination regions consist of up to 100 bases on either side of the start and stop codons, respectively, which are truncated if an adjacent CDS is encountered <100 bases from the start or stop codon.

Three sequence datasets are available (Table 1), being (i) sequences with no duplicates removed (‘redundant’), (ii) with duplicates removed by the Cleanup algorithm (4) and (iii) with duplicates removed by previous Transterm criteria, using the immediate start and stop contexts (1). The numerical parameters of each sequence (Table 2) of the latter dataset are stored as tables which users can access by species via the WWW interface.

SUMMARY PARAMETERS FOR THE CODING REGIONS OF EACH ENTRY AND SPECIES IN Transterm

In additional to numerical parameters for each sequence entry, the alignment of ‘Transterm non-redundant’ initiation and termination regions (Fig. 1) for each species is summarised as a consensus matrix (Table 3). New to this release is the inclusion of the tables describing the information content (5) of the initiation and termination regions of each species in Transterm (Table 3). These data include indications of sequence bias, for example as in start and stop contexts and sequence bias due to motifs associated with translation initiation [e.g., Shine–Dalgarno (5), Kozak (6) and termination (7) motifs]. Codon usage tables are also available for each species.

TOOLS FOR SEARCHING FOR SEQUENCE MOTIFS OR FEATURES IN Transterm

A collection of patterns describing known motifs or features found in mRNAs are available for users. Using a graphical WWW interface which interacts with the Transterm sequence data files and relational database, users can search all or portions of the Transterm database for each of these patterns or for user-defined patterns. Each of the mRNA regions defined in Figure 1 and Table 1 can be searched, allowing the user to choose both the region of the mRNAs and the sequence dataset (which differ in the treatment of redundant sequences) to be searched. Users may also search their own sequences with one or all of the patterns defined in Transterm, or a pattern they have defined.

THE CURRENT RELEASE

Transterm comprises a large number of files summarising sequence parameters, sequence datafiles and a relational database accessed via a WWW interface. The current release of

*To whom correspondence should be addressed. Tel: +64 3 479 5201; Fax: +64 3 479 7866; Email: chris.brown@otago.ac.nz

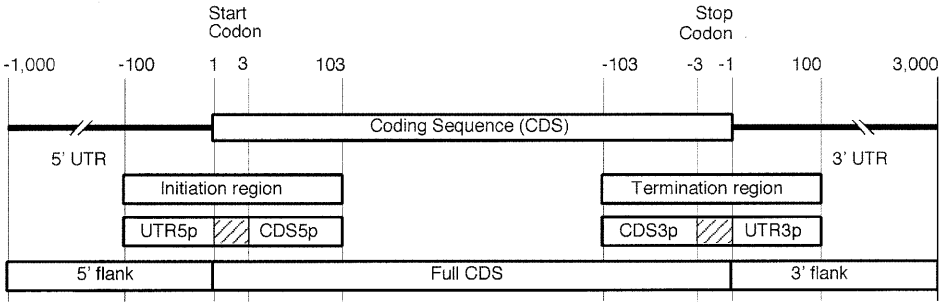

Figure 1. Functional regions of mRNAs which may be searched for motifs or patterns. Within Transterm a coding sequence from GenBank is broken into 5′ flank, Initiation, Full CDS, Termination and 3′ flank. The Initiation and Termination regions are further divided into their UTR, start or stop codons (hatched box) and CDS portions. The scale at the top gives the size of these regions. The regions flanking a given CDS will be shorter if 'limited' by the presence of an adjacent CDS within the number of bases shown (see text).

Table 1. Number of entries in Transterm (excluding separate complete genome datasets)

Dataset	Initiation region	Termination	5′ flank	3′ flank	CDS
All sequences	237 898	264 377	237 898	264 377	237 898
Non-redundant by Cleanup (4)	n.d.	n.d.	188 713	194 469	178 431
Non-redundant on termini	175 884	192 005	175 884	192 005	175 884

n.d., not determined.

Table 2. Numerical parameters for each coding sequence

Parameter	Meaning
Coding sequence length	Length of coding sequence from first base of start codon to last base of stop codon, inclusive
Nc (10)	Effective number of codons
GC3 (10)	Fraction of third codon position G+C
CAI (11,12)	Codon adaptation index

Transterm was built using GenBank (8) release 112.0. In previous releases, Transterm included only those species with more than 40 complete CDS entries. The current release contains all species with any CDS entries, providing better coverage of less well-studied species. Transterm now includes data from 10 320 species. The number of entries of each sequence region for each dataset are summarised in Table 3.

Table 3. Summary tables available for each species

Parameter	Sequence region (see Fig. 1)
Base frequencies (13)	Initiation region and termination region
Base information content (5)	Initiation region and termination region
Codon usage	CDS

Separately, 20 prokaryotic and three eukaryotic genomes have been processed to yield complete datasets for these genomes. For genomes the numerical parameters are summarised for all entries, rather than just the non-redundant sequences, so that genomes are treated as a whole rather than a collection of individual entries. The genomes are also present as individual entries in their respective species, where duplicate entries are discarded.

GenBank LOCUS and Accession indices are associated with each entry in Transterm. With these identifiers, users can access other data associated with entries of interest via Entrez (9) or other databases available via the WWW. Links to other transcript-related databases are available at the Transterm site.

Species in Transterm are now defined by the species designations of the NCBI Taxonomy database (3) (http://www.ncbi.nlm.nih.gov/Taxonomy/tax.html), represented in GenBank by 'taxids'. This, with better filtering of all steps in creating Transterm entries, has improved the quality of the database as a whole. The SSN (short species name) abbreviations for a species have been updated to a form which should be more stable from one release to another.

Further improvements in this release are mostly data quality improvements which are described in online documentation on the Transterm WWW site. The source code for Transterm is

available online, including Perl scripts for various tasks, for example processing the NCBI Taxonomy database into species indices.

ACKNOWLEDGEMENTS

This work is supported by a NZ Health Research Council grant to W.P.T. and Marsden grants to C.M.B. and W.P.T. M.J.S. is the recipient of an Otago University targeted research scholarship.

REFERENCES

1. Dalphin,M.E., Stockwell,P,A., Tate,W.P. and Brown,C.M. (1999) *Nucleic Acids Res.*, **27**, 293–294.
2. Brown,C.M., Dalphin,M.E., Stockwell,P.A. and Tate,W.P. (1993) *Nucleic Acids Res.*, **21**, 3119–2123.
3. Pesole,G., Liuni,S., Grillo,G., Ippedico,M., Larizza,A., Makalowski,W. and Saccone,C. (1999) *Nucleic Acids Res.*, **27**, 188–191. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 193–196.
4. Grillo,G., Attimonelli,M., Liuni,S. and Pesole,G. (1996) *Comput. Appl. Biosci.*, **12**, 1–8.
5. Gorodkin,J., Heyer,L.J., Brunak,S. and Stormo,G.D. (1997) *Comput. Appl. Biosci.*, **13**, 583–586.
6. Shine,J. and Dalgarno,L. (1974) *Proc. Natl Acad. Sci. USA*, **71**, 1342–1346.
7. Kozak,M. (1986) *Cell*, **44**, 283–292.
8. Benson,D.A., Boguski,M.S., Lipman,D.J., Ostell,J., Ouellette,B.F., Rapp,B.A. and Wheeler,D.L. (1999) *Nucleic Acids Res.*, **27**, 12–17. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 15–18.
9. McEntyre,J. (1998) *Trends Genet.*, **14**, 39–40.
10. Wright,F. (1990) *Gene*, **87**, 23–29.
11. Sharp,P.M. and Li,W.H. (1987) *Nucleic Acids Res.*, **15**, 1281–1295.
12. Lloyd,A.T. and Sharp,P.M. (1991) *Mol. Gen. Genet.*, **230**, 288–294.
13. Devereux,J., Harberli,P. and Smithies,O. (1984) *Nucleic Acids Res.*, **12**, 387–395.