Transterm: a database to aid the analysis of regulatory sequences in mRNAs

Grant H. Jacobs¹,², Augustine Chen¹, Stewart G. Stevens¹, Peter A. Stockwell¹, Michael A. Black¹, Warren P. Tate¹ and Chris M. Brown¹,*

¹Biochemistry Department and Webster Centre, University of Otago, PO Box 56 and ²Bioinfotools, PO Box 6129, Dunedin, New Zealand

Received September 16, 2008; Revised and Accepted October 6, 2008

ABSTRACT

Messenger RNAs, in addition to coding for proteins, may contain regulatory elements that affect how the protein is translated. These include protein and microRNA-binding sites. Transterm (http://mRNA.otago.ac.nz/Transterm.html) is a database of regions and elements that affect translation with two major unique components. The first is integrated results of analysis of general features that affect translation (initiation, elongation, termination) for species or strains in Genbank, processed through a standard pipeline. The second is curated descriptions of experimentally determined regulatory elements that function as translational control elements in mRNAs. Transterm focuses on protein binding sites, particularly those in 3'-untranslated regions (3'-UTR). For this release the interface has been extensively updated based on user feedback. The data is now accessible by strain rather than species, for example there are 10 Escherichia coli strains (genomes) analysed separately. In addition to providing a repository of data, the database also provides tools for users to query their own mRNA sequences. Users can search sequences for Transterm or user defined regulatory elements, including protein or miRNA targets. Transterm also provides a central core of links to related resources for complementary analyses.

INTRODUCTION

Messenger RNAs are translated into proteins, directed by specific signals in the mRNA. The genetic code and codon usage may differ between species. Translation in specific organisms may also require that they make efficient use of elements around the initiation and termination codons and also use a codon bias for that organism’s set of tRNAs. The preferred, often most efficient set of signals, in a particular organism can often be inferred from that most commonly used in that organism. For example, Homo sapiens has a strong bias prior to initiation codons (Kozak’s consensus) (1), whereas Escherichia coli has a G/U bias following termination codons. These have been associated with efficiency of initiation and termination respectively (2,3).

In addition to this general bias reflecting overall translation, individual mRNAs may contain regulatory elements within the mRNA that affect mRNA localization, stability or translation of the associated coding region (4–6). These function most frequently in the 3'-UTR but also in 5'-UTRs or coding regions (7,8). Key known elements are protein and miRNA-binding sites (9,10). Mutations and variations in these regulatory elements have been shown experimentally to affect their function and to be underlying contributors to genetic disease (11).

DATABASE GENERATION AND CONTENT

Transterm sequences and summaries

The detail of how Transterm 2008 was generated, and software used is available on the web site. A summary including major changes in this release is presented below. Data is parsed from NCBI Genbank or NCBI Genomes entries using CDS (coding sequence) fields, and mRNA fields when available. Key regions (CDS, 5'-UTRs and 3'-UTR, Init, Term) or flanks are extracted using this CDS or mRNA information. Eight sets of data are provided for each taxonomic strain with over 40 CDS or mRNAs. The strains are identified from the TaxID (NCBI taxonomy database identifier) in the Genbank entry. Data collected can differ in experimental support and redundancy.

For ‘Genomes’ sets reducing redundancy is not done, as genomes are considered to be complete datasets, but for Genbank data redundancy is removed according to our published procedure (12). This results in redundant and
non-redundant sets of regions: users choose which is appropriate to their needs. These sets of data are processed to generate summary data for each TaxID.

In previous releases of Transterm, data was 'mapped up' to the species level. With the increasing number of specific strains of a particular species now present in Genbank, we now use the strain as the taxonomic unit to collate and organize the data. For example, the 10 complete E. coli strains are processed separately, rather than combined. The sets of data are then processed as described previously to give a comprehensive set of analyses for each dataset. A view of part of the new interface is shown in Figure 1.

Two files summarizing initiation codon context for two complete bacterial genomes are shown in Figure 2. This is a comparison between a section of data from the context of two eubacteria, *Synechocystis* PCC6803 (TaxID: 1148) and *Pseudomonas aeruginosa* PAO1 (TaxID: 208964). A count of the percentage of each base in each position is shown (see text for analysis). The position (Pos) in the matrix is shown above -20 to +13, the ATG is at +1 to +3. The consensus (Cons) (>65%) is shown below. For these datasets the upper sequences were 41.7% GC3 and lower 65.8% GC3. More comprehensive descriptions of the data are also available (Table 1).

Transterm elements

Published literature was surveyed for descriptions of new elements. New elements would be included as they become available through published literature or feedback from users. Criteria for inclusion in Transterm are that it must be experimentally verified and published in a peer reviewed journal, and that it must be sufficiently well defined to be converted into a computer readable form (regular expression, matrix, secondary structure, or discrete sequence). Some elements, e.g. the Puf3-binding site from *Saccharomyces cerevisiae* are currently in this format in Transterm only. The format of an example (Puf3 protein-binding site) is shown in Figure 3.

Where appropriate, elements reported in other databases, have been included after an independent literature search.
review. In a similar fashion, several databases include reformatted Transterm elements (15,16). Some elements e.g. the well-studied Iron Responsive Element (IRE) are available as computer readable descriptor in several online databases, in these cases hyperlinks are provided from Transterm to allow the user to choose the most appropriate tool for analysis. Large highly structured RNA elements (e.g. riboswitches, IRESs) are not included, but are described in Rfam, ncRNA and IRESsite (17,18). The focus of Transterm is on protein-binding sites.

COMPARISON WITH OTHER TRANSLATIONAL CONTROL DATABASES

Several other databases provide some specific data, tools or services that complement those of Transterm. There is a list of Table 1. The key output files and a brief description of the contents of each. Further descriptions are available through the online help 'Main Transterm Datalfiles'.

ClassSSN-TaxID.complete

| Data: LOCUS, AccNo, Init [-20, +20], Term [-10, +10], Len, GC3, Nc
| Genbank names without descriptions
| List of GenBank names (original input file)
| Feature table outputs of TEXT information
| Entries selected by reject_dups criteria
| 5'-UTRs/flanks, transterm format
| 5'-UTRs/flanks, non-redundant
| 5'-UTRs/flanks, FASTA sequences, non-redundant
| 5'-UTRs/flanks, FASTA sequences
| 5'-UTRs/flanks, non-redundant
| Initiation region, FASTA sequences
| Initiation region
| GCG consensus output for initiation region (NR)
| Bit scores for NR initiation region
| Chi scores for NR initiation region
| CVS scores for NR initiation region
| Initiation region, FASTA sequences, non-redundant
| Schneider info. scores, init. region, non-redundant
| Schneider information scores, init. region
| Full CDS entries, FASTA sequences
| Full CDS entries
| Full CDS entries, FASTA sequences, non-redundant
| Full CDS entries, non-redundant
| GCG format of codon usage
| Output rscu table
| Summary of all the key values
| Entries in.init
| Initiation region, FASTA sequences
| initiation region
| GCG consensus output for termination region (NR)
| Count_signal of tetramer freq (readable output)
| Termination tetramer (codon + 3' base) frequencies
| Termination trimer (codon) frequencies
| Bit scores for NR termination region
| Chi scores for NR termination region
| CVS scores for NR termination region
| Termination region, FASTA sequences, non-redundant
| Schneider information scores, term. region, non-redundant
| Information scores, term. region
| 3'-UTRs/flanks, FASTA sequences
| 3'-UTRs/flanks
| 3'-UTRs/flanks, FASTA sequences, non-redundant
| 3'-UTRs/flanks, non-redundant
|
of resources referenced in the Transterm help online but the most relevant are summarized here. Rfam—the database of RNA families contains some cis-regulatory elements common to Transterm—these are cross-referenced. The elements are described in a different way (covariation models) and therefore are suitable for different types of analyses. RegRNA (15), UTRdb (19), Recode (20) all have related functionality but have not been updated since 2006.

Update frequency
Translational control elements are updated regularly and the sequence datasets annually.

ACKNOWLEDGEMENTS
Thanks to users who made suggestions for improvement or gave feedback.

FUNDING
Health Research Council (HRC05/195 to W.P.T., C.M.B., L.P. and R.T.P.); REANNZ and TelstraClear Capability build fund grant (CB611 to C.M.B., M.A.B.); and utilizes the NZ Biomirror and Bestgrid resources.

Conflict of interest statement. None declared.

References
1. Kozak,M. (1999) Initiation of translation in prokaryotes and eukaryotes. Gene, 234, 187–208.
2. Poole,E.S., Brown,C.M. and Tate,W.P. (1995) The identity of the base following the stop codon determines the efficiency of in vivo translational termination in Escherichia coli. EMBO J., 14, 151–158.
3. Cridge,A.G., Major,L.L., Mahagaonkar,A.A., Poole,E.S., Isaksson,L.A. and Tate,W.P. (2006) Comparison of characteristics and function of translation termination signals between and within prokaryotic and eukaryotic organisms. Nucleic Acids Res., 34, 1959–1973.
4. Sonenberg,N. and Hinnebusch,A.G. (2007) New modes of translational control in development, behavior, and disease. Mol. Cell, 28, 721–729.
5. Dahm,R., Kiebler,M. and Macchi,P. (2007) RNA localisation in the nervous system. Semin. Cell Dev. Biol., 18, 216–223.
6. Balvay,L., Lopez Lastra,M., Sargueil,B., Darlix,J.L. and Ohlmann,T. (2007) Translational control of retroviruses. Nat. Rev. Microbiol., 5, 128–140.
7. Chen,A., Kao,Y.F. and Brown,C.M. (2005) Translation of the first upstream ORF in the hepatitis B virus pregenomic RNA modulates translation at the core and polymerase initiation codons. Nucleic Acids Res., 33, 1169–1181.
8. Paquin,N. and Chartrand,P. (2008) Local regulation of mRNA translation: new insights from the bud. Trends Cell Biol., 18, 105–111.
9. Shyu,A.B., Wilkinson,M.F. and van Hoof,A. (2008) Messenger RNA regulation: to translate or to degrade. EMBO J., 27, 471–481.
10. Griffiths-Jones,S., Grocock,R.J., van Dongen,S., Bateman,A. and Enright,A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res., 34, D140–D144.
11. Chen,J.M., Ferec,C. and Cooper,D.N. (2006) A systematic analysis of disease-associated variants in the 3' regulatory regions of human protein-coding genes II: the importance
of mRNA secondary structure in assessing the functionality of 3′ UTR variants. *Hum. Genet.*, **120**, 301–333.

12. Jacobs, G.H., Stockwell, P.A., Tate, W.P. and Brown, C.M. (2006) Transterm—extended search facilities and improved integration with other databases. *Nucleic Acids Res.*, **34**, D37–D40.

13. Juntarajumnong, W., Incharoensakdi, A. and Eaton-Rye, J.J. (2007) Identification of the start codon for sphS encoding the phosphate-sensing histidine kinase in *Synechocystis* sp. *PCC 6803*. *Curr. Microbiol.*, **55**, 142–146.

14. Mutsuda, M. and Sugiura, M. (2006) Translation initiation of cyanobacterial rbcS mRNAs requires the 38-kDa ribosomal protein S1 but not the Shine-Dalgarno sequence: development of a cyanobacterial in vitro translation system. *J. Biol. Chem.*, **281**, 38314–38321.

15. Huang, H.Y., Chien, C.H., Jen, K.H. and Huang, H.D. (2006) RegRNA: an integrated web server for identifying regulatory RNA motifs and elements. *Nucleic Acids Res.*, **34**, W429–W434.

16. Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S.R. and Bateman, A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, **33**, D121–D124.

17. Kin, T., Yamada, K., Terai, G., Okida, H., Yoshinari, Y., Ono, Y., Kojima, A., Kimura, Y., Komori, T. and Asai, K. (2007) RNAdb: a platform for mining/annotating functional RNA candidates from non-coding RNA sequences. *Nucleic Acids Res.*, **35**, D145–D148.

18. Mokrejs, M., Vopalensky, V., Kolenaty, O., Masek, T., Feketova, Z., Sekyrova, P., Skaloudova, B., Kriz, V. and Pospisek, M. (2006) IRESite: the database of experimentally verified IRES structures (www.iresite.org). *Nucleic Acids Res.*, **34**, D125–D130.

19. Mignone, F., Grillo, G., Licciulli, F., Iacono, M., Liuni, S., Kersey, P.J., Duarte, J., Saccone, C. and Pesole, G. (2005) UTRdb and UTRsite: a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. *Nucleic Acids Res.*, **33**, D141–D146.

20. Baranov, P.V., Gurvich, O.L., Hammer, A.W., Gesteland, R.F. and Atkins, J.F. (2003) Recode 2003. *Nucleic Acids Res.*, **31**, 87–89.