D72–D76 Nucleic Acids Research, 2009, Vol. 37, Database issue Published online 4 November 2008
doi:10.1093/nar/gkn763

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

Grant H. Jacobs[1,2], Augustine Chen[1], Stewart G. Stevens[1], Peter A. Stockwell[1], Michael A. Black[1], Warren P. Tate[1] and Chris M. Brown[1,*]

[1]Biochemistry Department and Webster Centre, University of Otago, PO Box 56 and [2]Bioinfotools, 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).

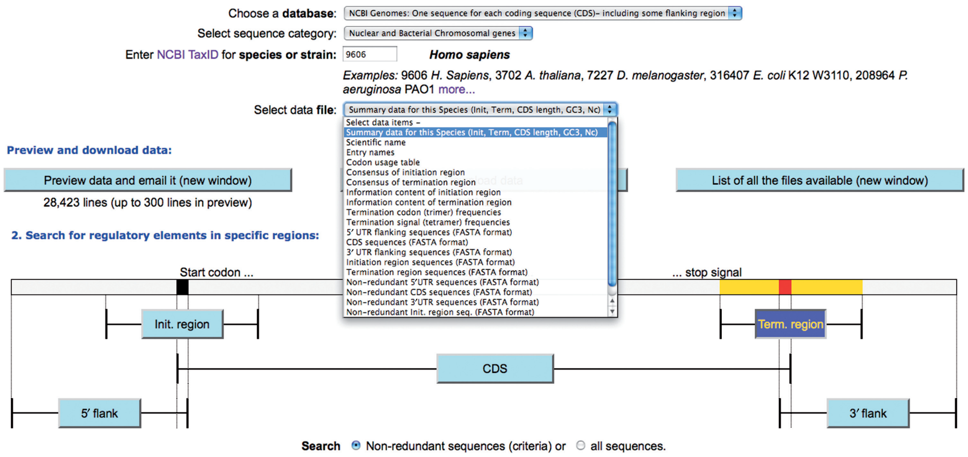
DATABASE GENERATION AND CONTENT

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

*To whom correspondence should be addressed. Tel: +643 4795201; Fax: +643 4797866; Email: chris.brown@otago.ac.nz



Choose a **database**: NCBI Genomes: One sequence for each coding sequence (CDS)– including some flanking region

Select sequence category: Nuclear and Bacterial Chromosomal genes

Enter NCBI TaxID for **species or strain**: 9606 ***Homo sapiens***

Examples: 9606 *H. Sapiens*, 3702 *A. thaliana*, 7227 *D. melanogaster*, 316407 *E. coli* K12 W3110, 208964 *P. aeruginosa* PAO1 more...

Select data **file**: Summary data for this Species (Init, Term, CDS length, GC3, Nc)

28,423 lines (up to 300 lines in preview)

Search ● Non-redundant sequences (criteria) or ○ all sequences.

Select data items –
Summary data for this Species (Init, Term, CDS length, GC3, Nc)
Scientific name
Entry names
Codon usage table
Consensus of initiation region
Consensus of termination region
Information content of initiation region
Information content of termination region
Termination codon (trimer) frequencies
Termination signal (tetramer) frequencies
5′ UTR flanking sequences (FASTA format)
CDS sequences (FASTA format)
3′ UTR flanking sequences (FASTA format)
Initiation region sequences (FASTA format)
Termination region sequences (FASTA format)
Non-redundant 5′UTR sequences (FASTA format)
Non-redundant CDS sequences (FASTA format)
Non-redundant 3′UTR sequences (FASTA format)
Non-redundant Init. region seq. (FASTA format)
Preview and download data:
Preview data and email it (new window)
List of all the files available (new window)
2. Search for regulatory elements in specific regions:
Start codon ...
... stop signal
Init. region
Term. region
CDS
5′ flank
3′ flank

Figure 1. Part of the new Transterm user interface. Users select data to analyse from four datasets, e.g. 'NCBI Genbank—One sequence for each coding sequence entry'. A taxomic group is selected by NCBI 'TaxId' number (e.g. 9606), then a particular type of output (listed in Table 1) can be selected by using the pull down menu (e.g. Consensus of initiation region, Figure 2). Data selected can be for all the sequences or a non-redundant set (for *H. sapiens* 96 417 versus 32 763 sequences). This data can also be searched using Blast or Scan for matches.

```
Pos   -20 -19 -18 -17 -16 -15 -14 -13 -12 -11 -10  -9  -8  -7  -6  -5  -4  -3  -2  -1  +1  +2  +3  +4  +5  +6  +7  +8  +9 +10 +11 +12 +13
 A:    22  22  21  22  24  24  27  30  29  30  28  29  33  30  25  22  22  22  16  11  89   0   0  44  29  15  29  32  17  25  25  15  26
 C:    36  36  36  35  33  33  31  27  21  15  11  10  15  21  30  35  38  33  50  60   0   0   0  25  33  47  29  25  39  34  29  43  34
 G:    23  21  23  24  24  26  27  29  40  48  54  54  41  33  28  24  23  31  20  21 9.8   0 100  18  24  19  31  19  27  24  16  26  25
 T:    19  21  19  19  20  17  15  14 10.0 7.6 7.2 6.7 11  16  17  20  17  15  14 8.3 1.4 100   0  12  14  19  11  24  17  16  30  16  15
Cons    N   N   N   N   N   N   N   N   R   R   R   R   R   N   N   N   N   N   N   N   A   T   G   M   N   N   N   N   S   N   N   S   N

Pos   -20 -19 -18 -17 -16 -15 -14 -13 -12 -11 -10  -9  -8  -7  -6  -5  -4  -3  -2  -1  +1  +2  +3  +4  +5  +6  +7  +8  +9 +10 +11 +12 +13
 A:    30  30  29  30  29  27  29  29  27  29  31  28  30  35  33  31  33  34  15  23  83   0   0  37  27  28  41  33  27  32  26  24  29
 C:    21  21  22  22  21  22  20  20  19  20  20  19  21  19  16  21  22  21  48  31   0   0   0  16  34  25  17  25  25  24  31  27  26
 G:    17  15  15  17  16  18  21  21  23  27  24  25  21  17  18  16  17  20  11  12  17   0 100  33  14  14  23  14  15  21  13  15  23
 T:    31  34  35  31  34  33  30  29  30  24  24  28  28  30  33  32  28  25  26  34 .39 100   0  15  25  34  19  29  32  23  30  34  22
Cons    N   N   N   N   N   N   N   N   N   N   N   N   N   N   W   N   N   N   Y   N   A   T   G   R   N   N   N   N   N   N   N   N   N
```

Figure 2. The 'Consensus of initiation region' files for Synechocystis PCC6803 (NBSynePCC_2-1148.initmatrix) and Pseudomonas aeruginosa PAO1 (NBPseuaeru-208964.initmatrix). 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).

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) initiation codons (*.initmatrix). The upper panel shows a typical Shine-Dalgarno (SD) like pattern for a high GC% genome (for example purines at −13 to −7, whereas the lower panel PC6803 has an atypical pattern for a bacterium (less purine bias at −13 to −7, pyrimidine bias at −2, −1). Further investigation of this observation using Transterm data could utilise alternative representations of the same data, see Table 1 (Panel C) (*.initnrttbit, *.initnrttcvs), the aligned sequences themselves (*.init, *.dat) or summaries of the data (*.sum). As suggested by this data cyanobacteria have been shown to use a combination of SD-dependent and SD-independent initiation (13,14).

A list of the key classes of output files are shown in Table 1. More detail of the content of each of these files in an online help document on the website. Many of these analyses are newly available in this release.

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 form 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

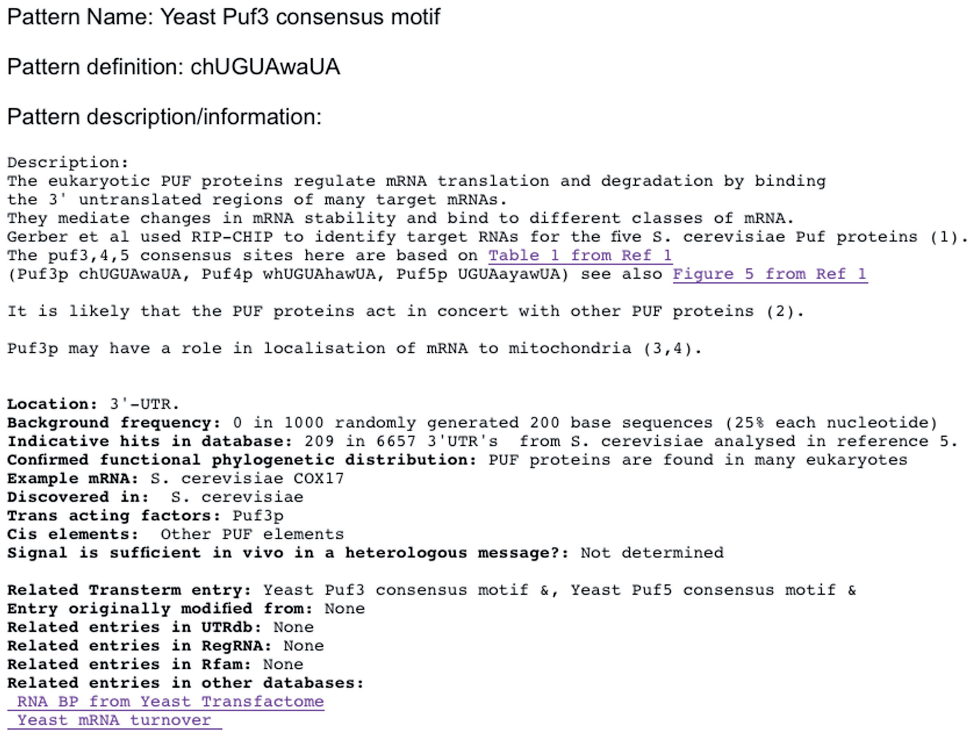
Table 1. The key output files and a brief description of the contents of each. Further descriptions are available through the online help 'Main Transterm Datafiles'

ClassSSN-TaxID.complete	Entries with complete CDS (have both inits + terms)
A: Lists of entries and identifiers in the redundant and non-redundant sets	
*.dat	Data: LOCUS, AccNo, Init [-20, +20], Term [−10, +10], Len, GC3, Nc
*.entry	Genbank names without descriptions
*.names	List of GenBank names (original input file)
*.text	Feature table outputs of TEXT information
*.TTSelected	Entries selected by reject_dups criteria
B: 5′-UTRs	
*.5UTR	5′-UTRs/flanks, transterm format
*.5UTRnrtt	5′-UTRs/flanks, non-redundant
*.5UTRnrtt.fa	5′-UTRs/flanks, FASTA sequences, non-redundant
*.5UTR.fa	5′-UTRs/flanks, FASTA sequences
C: Initiation codon context	
*.InitEntries	Entries in.init
*.init.fa	Initiation region, FASTA sequences
*.init	Initiation region
*.initmatrix	GCG consensus output for initiation region (NR)
*.initnrttbit	Bit scores for NR initiation region
*.initnrttchi	Chi scores for NR initiation region
*.initnrttcvs	CVS scores for NR initiation region
*.initnrtt.fa	Initiation region, FASTA sequences, non-redundant
*.initnrttver	Schneider info. scores, init. region, non-redundant
*.initver	Schneider information scores, init. region
D: CDS (coding sequences)	
*.CDS.fa	Full CDS entries, FASTA sequences
*.CDS	Full CDS entries
*.CDSnrtt.fa	Full CDS entries, FASTA sequences, non-redundant
*.CDSnrtt	Full CDS entries, non-redundant
E: Codon usage and bias	
*.cod	GCG format of codon usage
*.rscu	Output rscu table
*.sum	Summary of all the key values
F: Termination codon context	
*.TermEntries	Entries in.term
*.term.fa	Termination region, FASTA sequences
*.term	Termination region
*.termmatrix	GCG consensus output for termination region (NR)
* _termnr.summary	Count_signal of tetramer freq (readable output)
* _termnr.tet_tab	Termination tetramer (codon + 3′ base) frequencies
* _termnr.tri_tab	Termination trimer (codon) frequencies
*.termnrttbit	Bit scores for NR termination region
*.termnrttchi	Chi scores for NR termination region
*.termnrttcvs	CVS scores for NR termination region
*.termnrtt.fa	Termination region, FASTA sequences, non-redundant
*.termnrtt	NR version of.term, by old reject_dups criteria
*.termnrttver	Info. scores, term. region, non-redundant
*.termver	Information scores, term. region
G: 3′-UTRs	
*.3UTR.fa	3′-UTRs/flanks, FASTA sequences
*.3UTR	3′-UTRs/flanks
*.3UTRnrtt.fa	3′-UTRs/flanks, FASTA sequences, non-redundant
*.3UTRnrtt	3′-UTRs/flanks, non-redundant

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



Pattern Name: Yeast Puf3 consensus motif

Pattern definition: chUGUAwaUA

Pattern description/information:

Description:
The eukaryotic PUF proteins regulate mRNA translation and degradation by binding
the 3' untranslated regions of many target mRNAs.
They mediate changes in mRNA stability and bind to different classes of mRNA.
Gerber et al used RIP-CHIP to identify target RNAs for the five S. cerevisiae Puf proteins (1).
The puf3,4,5 consensus sites here are based on Table 1 from Ref 1
(Puf3p chUGUAwaUA, Puf4p whUGUAhawUA, Puf5p UGUAayawUA) see also Figure 5 from Ref 1

It is likely that the PUF proteins act in concert with other PUF proteins (2).

Puf3p may have a role in localisation of mRNA to mitochondria (3,4).

Location: 3'-UTR.
Background frequency: 0 in 1000 randomly generated 200 base sequences (25% each nucleotide)
Indicative hits in database: 209 in 6657 3'UTR's from S. cerevisiae analysed in reference 5.
Confirmed functional phylogenetic distribution: PUF proteins are found in many eukaryotes
Example mRNA: S. cerevisiae COX17
Discovered in: S. cerevisiae
Trans acting factors: Puf3p
Cis elements: Other PUF elements
Signal is sufficient in vivo in a heterologous message?: Not determined

Related Transterm entry: Yeast Puf3 consensus motif &, Yeast Puf5 consensus motif &
Entry originally modified from: None
Related entries in UTRdb: None
Related entries in RegRNA: None
Related entries in Rfam: None
Related entries in other databases:
RNA BP from Yeast Transfactome
Yeast mRNA turnover

Figure 3. An example of Transterm element description (Puf3p-binding site). Elements may be described by strings, regular expressions, matrices or RNA secondary structure rules. In this case the element is simply described as a string. Users may construct more complex descriptions of the element based on the referenced literature, for example allowing mismatches, insertions or deletions.

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) fRNAdb: 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.