13.20.5. Translation termination sequences.

The Translational Termination Signal Database May 1994

(This is an edited version of the TransTerm documentation.)

Reference: Brown *et al.* (1994) *Nucleic Acids Res.* **22:**3620-3624.

The TransTerm database of termination codon contexts has been extended to include measures of sense codon usage, and initiation codon contexts. The database contains:

a) the sequence around the termination codon (-10, +10);
b) the sequence around the initiation codon (-20, +10);
c) the length ,'G+C%' of the third position of codons (GC3), the codon adaptation index (CAI) and the 'effective number of codons' statistic (Nc);
d) summary tables including total codon usage, stop codon and tetranucleotide stop-signal usage, and matrices tallying base frequencies at each position around the initiation and termination codons.

The data are arranged to facilitate investigation of the relationships between the three phases of protein synthesis. This edition of the database is also described in the September 1994 Database issue of Nucleic Acids Research, and investigators using the database should cite this article.

It is well established that the identities of bases around the initiation (1-4) and termination codons (5, 6) in many organisms are not random. Sense codon usage is frequently biased also (7-9). The signals and contexts actually found depend on both functional constraints and also genomic G+C biases (10-14).

Initiation and termination contexts.

Initiation and termination codon contexts were extracted using the information in the feature tables of GenBank entries for organisms which had over 40 valid sequences available (Flat file format Release 82, April 1994) (15-17). The locus names were selected by searching the "ORGANISM" line of the entry; the exact strings searched are listed. Only the appropriate divisions of Genbank were searched. The data are listed under three letter keys e.g. dro for D. melanogaster. Each "CDS" or "mat_peptide" described in the feature table was interpreted using feature locations, qualifiers and join specifications by the program FISH_TERM. For valid coding regions the sequences twenty bases before (-20) and ten after (+10) the initiation codon, and ten before (-10) and ten after (+10) the stop codon were extracted. Sequences and identifiers are found in the ***.dat files. Identifiers are in the form LOCUS n, where n refers to the nth "CDS" or mat_peptide feature table entry for that Locus. Entries were rejected if:

a) they were duplicates in the termination region (duplicate is defined as less than two mismatches over the window of 21 nucleotides). If the sequences were duplicates, the one with the longer termination region was retained. If the termination region lengths were identical, the one with the longer initiation region was retained,
b) they had no stop codon,
c) the stop codon was not preceded by a valid open reading frame,
d) the open reading frames were shorter than 100 bases.

Partial sequences with valid stop codons were retained, leaving between 3% and 13% of entries without initiation regions. Sequences were truncated to include only noncoding sequences if the feature table described a 5' or 3' coding sequence.

Measures of synonymous sense codon bias.

Three measures of synonymous sense codon usage are found in the ***.dat files.

(i) The Codon Adaptation Index (CAI) measures the match between sense codon usage of a coding sequence and that of a set of highly expressed genes from that organism (18, 19). A value of 1 indicates that the codon usage is identical to that of the highly expressed genes. These values are also listed in the ***.dat files. For each organism, the group of genes with the highest CAI scores are included as a separate file (***_h.dat files) comprising the highest 10% or 40 scoring coding sequences. Groups of genes with high CAI scores tend to be highly expressed and have biased termination signals.

(ii) The G+C% of the third positions of sense codons (GC3) (20).

(iii) The 'effective number of codons' (Nc) (20). Nc can vary from 20, where one codon is used for each amino acid, to 61 where all synonymous codons are used equally. This measure of the codon bias was calculated for coding sequences over 300 nucleotides long. The values for Nc listed sometimes differ slightly from those in reference 20, due to a difference in interpretation of the adjustment for absent amino-acids.

Summaries for each organism.

Codon usage tables in the GCG format are included for all 93 organisms (***.cod).

The total frequency of each triplet stop codon expressed as a count and a percentage are on the 's' line in the file SPECIES_TRI.DAT. For comparison the frequency and percentage of the same trinucleotides in any frame in the non-coding region immediately following the stop codon is also shown- on the 'n' line, as is the GC3 for this region. A similar file, SPECIES_TETRA.DAT tallies the frequencies of 4-base stop signals and the corresponding noncoding regions.

As an example the *D. melanogaster* high CAI set (dro_h) has a strong preference for G (52%) in the fourth position of the stop signal. This stands in striking contrast to the generally low G+C content in the noncoding regions of this set of genes (G+C = 42%). Many of the organisms analysed show such biases. However in some organisms, particularly vertebrates and plants, the biases in the use of termination signals are less prominent (13).

For the regions immediately around the initiation and termination codons the incidence of the bases in each position were derived from the ***.dat files using the GCG program Consensus (21). This required slight modification of the ***.dat files to GCG format. The consensus matrices are found in the ***.initmatrix and ***.termmatrix files. For example in dro_h.initmatrix and dro_h.termmatrix files there are extreme biases in the contexts of both initiation and stop codons the most significant biases are in the four positions prior to the initiation codon (CAAMATG) and within the stop signal (TAAG).

Please send comments or requests for additional information to the authors e-mail address: biocwpt@otago.ac.nz, or FAX +64 3 4797866.

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

Acknowledgments.

W.P.T. is an International Research Scholar Award of the Howard Hughes Medical Research Institute. This work was supported in part by a grant from the New Zealand Health Research Council

References.

1. Kozak, M. (1992) *J. Cell Biol.*, **115:** 887-903.
2. Tzareva, N.V., Makhno, V.I. & Boni, I.V. (1994) *FEBS Lett.*, **337:** 189-194.
3. De-Smit, M.H. & Vanduin, J. (1994) *J. Mol. Biol.*, **235:** 173-184.
4. Cavener, D.R. & Ray, S.C. (1991) *Nucleic Acids Res.*, **19:** 3185-3192.
5. Tate, W.P. & Brown, C.M. (1992) *Biochemistry*, **31:** 2443-50.
6. Brown, C.M., Dalphin, M.E., Stockwell, P.A. & Tate, W.P. (1993) *Nucleic Acids Res.*, **21:**,3119-3123.
7. Wada, K.N., Wada, Y., Ishibashi, F., Gojobori, T. & Ikemura, T. (1992) *Nucleic Acids Res.*, **20:** 2111-2118.
8. Sueoka, N. (1992) *J. Mol. Evol.*, **34:** 95-114.
9. Kurland, C.G. (1993) *Biochem. Soc. Trans.*, **21:** 841-846.
10. Collins, D.W. & Jukes, T.H. (1993) *J. Mol. Evol.*, **36:** 201-213.

11. Sharp, P.M., Stenico, M., Peden, J.F. & Lloyd, A.T. (1993) *Biochem. Soc. Trans.*, **21:** 835-841.
12. Eyre-Walker, A. (1994) *Mol. Biol. Evol.*, **11:** 88-98.
13. Martin, R. (1994) *Nucleic Acids Res.*, **22:** 15-19.
14. Pedersen, W.T. & Curran, J.F. (1991) *J. Mol. Biol.*, **219:** 231-241.
15. Benson, D., Lipman, D.J. & Ostell, J. (1993) *Nucleic Acids Res.*, **21:** 2963-2965.
16. Brown, C.M., Stockwell, P.A., Trotman, C.N.A. & Tate, W.P. (1990) *Nucleic Acids Res.*, **18:** 6339-6345.
17. Brown, C.M., Stockwell, P.A., Trotman, C.N.A. & Tate, W.P. (1990) *Nucleic Acids Res.*, **18:** 2079-2086.
18. Sharp, P.M. & Li, W. (1987) *Nucleic Acids Res.*, **15:** 1281-1295.
19. Lloyd, A.T. & Sharp, P.M. (1991) *Mol. Gen. Genet.*, **230:** 288-294.
20. Wright, F. (1990) *Gene,* **87:** *23-29.*
21. *Devereux, J. Haberli, P. & Smithies, O. (1984) Nucleic Acids Res.*, **12**,387-395.
22. Rice, C.M., R., F., Higgins, D.G., Stoehr, P.J. & Cameron, G.N. (1993) *Nucleic Acids Res.*, **21** 2967-2971.

end