The translational signal database, TransTerm: more organisms, complete genomes

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

Department of Biochemistry and Centre for Gene Research, University of Otago, Dunedin, New Zealand

Received October 11, 1996; Accepted October 15, 1996

ABSTRACT

TransTerm is a database of initiation and termination sequence contexts from more than 250 organisms listed in GenBank, including the four complete genomes: *Haemophilus influenzae, Methanococcus jannaschii, Mycoplasma genitalium,* and *Saccharomyces cerevisiae*. For the current release, more than 60 000 coding sequences were analysed. The tabulated data include initiation and termination contexts organised by species along with quantitative parameters about individual coding sequences (length, %GC, GC3, Nc and CAI). There are also tables of initiation- and termination-region nucleotide-frequencies, codon usage tables and summaries of stop signal usage. TransTerm is available on the World Wide Web at: http://biochem.otago.ac.nz:800/Transterm/homepage.html

INTRODUCTION

Questions about the sequence context of stop codons first prompted the construction of the TransTerm Database four years ago (1). Since then there have been many studies verifying the importance to translation of the signals surrounding the stop codon (2–9). Sequence bias is also observed in the initiation regions (10–13) and in codon usage (14). TransTerm contains information on these three areas, recording individual start and stop coding sequence contexts for over 60 000 genes, codon usage for over 250 organisms and sequence parameters such as length, GC3, Nc and CAI (15,16) for all the coding sequences. Summaries of the %GC for each organism's coding regions are also available.

DATABASE ORGANISATION

TransTerm contains a detailed description of the database preparation, so only a brief discussion will be presented here. Individual species from GenBank (17) release 96.0, dated August 15, 1996, were screened to locate those with more than 40 independent coding regions. In addition, some mitochondria and bacteriophage were selected, although the basis of their inclusion was somewhat arbitrary, for example, ØX174 has only 11 genes, but it was included because this is a complete genome. The complete genomes of *Haemophilus influenzae*, *Methanococcus jannaschii*, *Mycoplasma genitalium*, and *Saccharomyces cerevisiae* (18–20,23) were also included in a separate section of the database. We have excluded the GenBank viral division as we do not yet have a method to screen against viral variants. The complete list of organisms included in this release of TransTerm is too large to specify here, however, it is available on-line at our World Wide Web site.

Coding regions that are well defined in the GenBank Feature Table entry were included; those that are not correctly defined (e.g. stop codons in the 'open reading frame' or no stop codon) or were shorter than 100 nucleotides or duplicated sequences already found for that species were excluded. We extracted the stop codon context (–10 to +10) and start codon contexts (–20 to +10) for each coding sequence. In addition, for each sequence, we calculated the parameters: sequence length, GC3, Nc and CAI (where possible).

These sequence contexts and parameters were placed in files, named by species: *****.dat, where '*****' is a five letter code made from the first two letters of the organism's genus and the first three letters of the organism's species, as defined in the GenBank ORGANISM field. Mitochondria names are prefixed with an 'M' and whole genome names are prefixed with a 'G'. For example, the whole genome of *Saccharomyces cerevisiae* is named 'GSacer'. An example from the species *Homo sapiens*, called Hosap.dat is shown in Table 1. Codon usage tables in GCG format (21), *****.cod and summaries of the initiation and termination contexts created by the GCG program 'Consensus', *****.initmatrix and *****.termmatrix are also available.

Summaries of all the termination data for all species are available in the files species_tri.dat and species_tet.dat.

DATABASE AVAILABILITY

The TransTerm database is easily accessible from our World Wide Web site: http://biochem.otago.ac.nz:800/Transterm/homepage.html . It is also available by anonymous ftp from the University of Otago: ftp.otago.ac.nz in the directory pub/biochemistry/Transterm and by anonymous ftp from the EMBL database server: ftp.ebi.ac.uk (22). Please send comments, corrections and requests for additional information to us by email at: mdalphin@sanger.otago.ac.nz or Fax +64 3 479 7866.

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

Table 1. A section of the *Homo sapiens* contexts and parameter file Hosap.dat

Locus#CDS	Acc.No	Initiation Start	Term. Stop	Len	GC3	Nc
D00017#1	D00017	ACGGCCCAGCTTCCTTCAAAATGTCTACTGTTC	TGGAGATGACTGAAGCCCGACAC	1020	0.587	48.4
D00723#1	D00723	ACCCCCGCACCCCTGCGAACATGGCGCTGCGAG	TATTGAGGAGTGAAAATGGAACT	522	0.413	57.7
D11428#1	D11428	GAGCAGAACTTGCCGCCAGAATGCTCCTCCTGT	GAAACGCGAATGAGGCGCCCAGA	483	0.737	43.7
D13752#1	D13752	AGGGTGGAGGGAGCATTGGAATGGCACTCAGGG	AGCGATTAACTAGTCTTGCATCT	1512	0.775	42.2
D29634#1	D29634	CTGGAGAGCCCAGACCTGGGATGGCGGATTCGT	CTCCCTCTGCTGACATTTCAAGC	1161	0.887	33.4
D31784#1	D31784	GACTCGACGGTGCCATCAGCATGAGAACTTACC	CAAAGACTCCTAATCTGTTGCCT	2373	0.488	57.0
D32131#1	D32131	ATTCTCCCCAGACGCCGAGGATGGCCGTCATGG	TTGTAAAGTGTGAGACAGCTGCC	1098	0.799	38.9
D42106#1	D42106	GAGCCCGGAAGATTTCAGCCATGCCTCACAGCT	CTTCCAGCCCTGAGCTTCCGATG	378	0.766	31.6
D49817#1	D49817	..TCGGGCGCAGCCGCGAAGATGCCGTTGGAAC	CAGGAAACACTGAGGCAGACGTG	1563	0.794	40.4

Locus#CDS refers to the GenBank LOCUS field, followed by a number representing the ‘n-th’ ‘CDS’ or ‘mat_peptide’ for that LOCUS. AccNo contains the GenBank ACCESSION number. Initiation shows the context of the start codon from 20 nucleotides before the start codon to 10 nucleotides past the start codon; Start is placed to help locate the start codon. Term marks the beginning of the termination context which extends from 10 nucleotides before the stop codon to 10 nucleotides beyond the stop codon. Len is the length in nucleotides of the coding sequence; GC3 and Nc are parameters that characterise the coding region.

REFERENCES

1 Brown, C.M., Dalphin, M.E., Stockwell, P.A. and Tate, W.P. (1993) *Nucleic Acids Res.*, **21**, 3119–3123.

2 Tate, W.P., Poole, E.S. and Mannering, S.A. (1996) In Cohn, W.E. and Moldave, K. (ed.) *Progress in Nucleic Acid Research and Molecular Biology.* Academic Press Inc, San Diego, CA. Vol 52, pp. 293–335.

3 Tate, W.P. and Mannering, S.A. (1996) *Mol. Microbiol*, **21**, 213–219.

4 Low, S.C. and Berry, M.J. (1996) *Trends Biochem. Sci.*, **21**, 203–208.

5 Gesteland, R.F. and Atkins, J.F. (1996) *Annu. Rev. Biochem.*, **65**, 741–768.

6 Phillips-Jones, M.K., Hill, L.S.J., Atkinson, J. and Martin, R. (1995) *Mol. Cell Biol.*, **15**, 6593–6600.

7 Bjornsson, A., Mottaguitabar, S. and Isaksson, L.A. (1996) *EMBO J.*, **15**, 1696–1704.

8 Zhang, S.P., Rydenaulin, M. and Isaksson, L.A. (1996) *J. Mol. Biol.*, **261**, 98–107.

9 Bonetti, B., Fu, L.W., Moon, J. and Bedwell, D.M. (1995) *J. Mol. Biol.*, **251**, 334–345.

10 Pain, V.M. (1996) *Eur. J. Biochem.*, **236**, 747–771.

11 Kozak, M. (1996) *Mamm. Genome*, **7**, 563–574.

12 Kozak, M. (1992) *Annu. Rev. Cell Biol.*, **8**, 197–225.

13 Mccarthy, J.E.G. and Brimacombe, R. (1994) *Trends Genet.*, **10**, 402–407.

14 Sharp, P.M., Stenico, M., Peden, J.F. and Lloyd, A.T. (1993) *Biochem. Soc. Trans.*, **21**, 835–841.

15 Sharp, P.M. and Li, W.-H. (1987) *Nucleic Acids Res.*, **15**, 1281–1295.

16 Wright, F. (1990) *Gene*, **87**, 23–29.

17 Benson, D., Lipman, D.J. and Ostell, J. (1993) *Nucleic Acids Res.*, **21**, 2963–2965.

18 Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.-F., Dougherty, B.A., Merrick, J.M., *et al.* (1995) *Science*, **269**, 496–512.

19 Fraser, C.M., Gocayna, J.D., White, O., Adams, M.D., Clayton, R.A., Fleischmann, R.D., Bult, C.J., Kerlavage, A.R., Sutton, G., Kelley, J.M., *et al.* (1995) *Science*, **270**, 397–403.

20 Bult, C.J., White, O., Olsen, G.J., Zhou, L., Fleischmann, R.D., Sutton, G.G., Blake, J.A., FitzGerald, L.M., Clayton, R.A., Gocayne, J.D., *et al.* (1996) *Science*, **273**, 1058–1073.

21 Genetics Computer Group. (1994) *Program manual for the Wisconsin Package, Version 8*. Genetics Computer Group, Madison, Wisconsin, USA 53711.

22 Rice, C.M., R., F., Higgins, D.G., Stoehr, P.J. and Cameron, G.N. (1993) *Nucleic Acids Res.*, **21**, 2967–2971.

23 Cherry, J.M., Adler, C., Ball, C., Dwight, S., Chervitz, S., Juvik, G., Weng, S. and Botstein, D. (Sept. 1996). ‘Saccharomyces Genome Database’, SacchDB4.6.1, http://genome-www.stanford.edu/Saccharomyces/ .