Rare codons in *E. coli* and *S. typhimurium* signal sequences

Dennis M. Burns and Ifor R. Beacham*

School of Science, Griffith University, Nathan, Brisbane, Qld. 4111, Australia

Received 17 May 1985; revised version received 24 July 1985

Codon usage has been examined in the signal sequences of 27 genes encoding proteins which possess leader peptides, and are inner-membrane located or exported. The results have been compared with codon usage in the corresponding coding sequences of most of the mature proteins. A bias is observed in the usage of rare codons for two of the three hydrophobic amino acids for which there are rare codons. Since hydrophobic residues are predominant in leader peptides, we suggest that a resulting concentration of rare codons in the signal sequence may play a role (or have played a role in the evolutionary past) in the secretion process by delaying translation.

Signal sequence *Codon usage* *Rare codon* *Protein export*

1. INTRODUCTION

All secreted (periplasmic), most outer membrane, and some inner membrane proteins in *Escherichia coli* and *Salmonella typhimurium* are initially synthesised as a precursor with an NH_2-terminal extension of about 25 amino acids [1,2]. These precursors, or pre-proteins, are subsequently processed by a membrane-bound signal peptidase [3] to yield the mature proteins. Other 'signals' relevant to secretion or translocation to the outer membrane are contained within the mature protein sequence [1,2,4,5], but the NH_2-terminal signal sequence is of paramount importance in the initial steps of protein export. It has been amply demonstrated that exported proteins are synthesised on membrane-associated ribosomes (see [6]); one of the initial steps in export must therefore be the interaction of the nascent signal sequence, which contains a hydrophobic 'core' of amino acid residues, with the cytoplasmic membrane. In the case of the *lamB* gene, it has been proposed, on the basis of genetic evidence, that a 'stop-translation' sequence is involved in curtailing translation at an early stage, thus allowing membrane interaction of the ribosome-polypeptide complex before chain elongation has progressed too far [7]. A similar arrest in translation has been suggested to occur in eukaryotic secretion mediated by a 'signal recognition protein' [8]. Similar recognition proteins may also be involved in prokaryotic secretion [9].

A second mechanism for curtailing translation is the occurrence of rare codons [10–17]. We have recently determined the complete nucleotide sequence of the *E. coli ush* gene (manuscript in preparation) which specifies periplasmic UDP-glucose hydrolase (5′-nucleotidase) and found a concentration of rare codons in the signal sequence of the Ush protein relative to its mature sequence. We therefore undertook a comparison of codon preferences in a collection of *E. coli* and *S. typhimurium* signal sequences of periplasmic, outer membrane and inner membrane proteins with those in the mature polypeptides. The results of this analysis indicate a clear bias in the use of leucine and proline rare codons in signal sequences which, we suggest, may play a role in the early stages of protein export.

2. MATERIALS AND METHODS

Codon usage was analysed using the computer program described by Delaney [18]. Seven files

* To whom correspondence should be addressed

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/85/$3.30

were created corresponding to (1) periplasmic signal sequences (PP), (2) outer membrane signal sequences (OM), (3) inner membrane signal sequences (IM), (4) combined total signal sequences (TSS), (5) periplasmic mature sequences (TOTPP), (6) outer membrane mature sequences (TOTOM), and (7) combined total mature sequences (TOTMAT). The 'mature sequences' are the DNA sequences corresponding to the mature (i.e. processed) proteins. These files are described in detail in table 1.

3. RESULTS AND DISCUSSION

Based on observations of their frequency of occurrence in 25 non-regulatory genes [15] and the

Table 1

Proteins which are synthesised in precursor form in *E. coli* and *Salmonella typhimurium*

Protein (gene)	Reference
(1) 12 periplasmic signal sequences (855 nucleotides, 285 codons)	
Alkaline phosphatase (*phoA*)	[19,20]
β-Lactamase, chromosome (*ampC*)	[21]
β-Lactamase, plasmid (*bla*)	[22]
5′-Nucleotidase (*ush*)	(a)
T-one receptor (*tonB*)	[23]
Arabinose-binding protein (*araF*)	[24]
Galactose-binding protein (*mglB*)	[24]
Histidine-binding protein (*hisJ*)	[25,26]
Leucine-specific binding protein (*livK*)	[27]
Lysine-arginine-ornithine-binding protein (*argT*)	[25]
Maltose-binding protein (*malE*)	[28]
Phosphate-binding protein (*phoS*)	[29]
(2) 12 outer membrane signal sequences (771 nucleotides; 257 codons)	
Outer membrane protein (*ompA*)	[30,31]
Outer membrane protein (*ompC*)	[32]
Outer membrane protein (*ompF*)	[33]
Outer membrane protein (*phoE*)	[34]
Outer membrane protein (*tolC*)	[35]
Lipoprotein (*lpp*)	[36]
Phage λ receptor protein (*lamB*)	[37]
Pap pili subunit (*papA*)	[38]
Heat-labile enterotoxin subunit A (*toxA*)	[39,40]
Heat-labile enterotoxin subunit B (*toxB*)	[40–42]
Heat-stable enterotoxin I (ST I)	[43]
Heat-stable enterotoxin II (ST II)	[44]
(3) 3 inner membrane signal sequences (171 nucleotides, 57 codons)	
M13 major coat protein (gene VIII)	[45]
M13 minor coat protein (gene III)	[45]
Serine receptor (*tsr*)	[46]
(4) 27 signal sequences (1797 nucleotides; 599 codons)	
The combined signal sequences listed (1), (2), (3)	
(5) 5 periplasmic mature sequences (4689 nucleotides; 1563 codons)	
β-Lactamase, plasmid (*bla*)	[22]
5′-Nucleotidase (*ush*)	(a)
T-one receptor (*tonB*)	[23]
Histidine-binding protein (*hisJ*)	[25,26]
Phosphate-binding protein (*phoS*)	[29]
(6) 8 outer membrane mature sequences (4656 nucleotides; 1552 codons)	
Outer membrane protein (*ompA*)	[30,31]
Outer membrane protein (*phoE*)	[34]
Outer membrane protein (*tolC*)	[35]
Lipoprotein (*lpp*)	[36]
Pap pili subunit (*papA*)	[38]
Heat-labile enterotoxin subunit B (*toxB*)	[41,42]
Heat-stable enterotoxin I (ST I)	[43]
Heat-stable enterotoxin II (ST II)	[44]
(7) 13 periplasmic and outer membrane mature sequences (9345 nucleotides; 3115 codons)	
The combined mature sequences listed (5), (6)	

The 7 groups of proteins are the source of the 7 nucleotide files described in section 2. (a) Burns and Beacham (in preparation)

relative abundance of the corresponding tRNA species [47], we define rare codons as those whose corresponding tRNA species occur with an abundancy of 0.3 or less, on a scale of ~0–1.0 (see [47]), and where their percentage use (defined as the number of times a codon is used in the 25 genes referred to above, divided by the number of times all of the codons specifying the same amino acid were used) is approx. 10% or less. One exception to this definition is the arg CGA codon; this codon corresponds to an abundant tRNA species [47] but is inefficiently recognized by the ICG-containing isoaccepting tRNA [47]. These rare codons are listed in table 2.

The frequency of use for each of the 64 codons in the sequence files described in section 2 is shown in table 3.

By considering a number of signal sequences together, the total number of codons examined approximates that of an average-sized gene. The relevant data from table 3 are shown in table 4, together with the result of a chi-squared analysis. Clearly there is a preferred usage of leucine and proline rare codons in total signal sequences (TSS) compared to the total mature protein sequences (TOTMAT; $p = 0.007$ and <0.001, respectively). However, the occurrences of rare codons for glycine, isoleucine, serine, threonine and arginine

Table 2

Rare codons

	tRNA content[a]	% use[b]
Leu UUA	0.25	6.1
Leu UUG	0.25	8
Leu CUU	0.3	9
Leu CUC	0.3	7
Leu CUA	minor	2
Ile AUA	0.05	1
Ser UCA	0.25	8
Ser UCG	0.25	11
Ser AGU	0.25	6
Pro CCU	minor	9
Pro CCC	minor	6
Thr ACA	minor	6
Arg CGG	minor	3
Arg AGA	minor	1
Arg AGG	minor	0.25
Gly GGA	0.15	5
Gly GGG	0.1	7
Arg CGA[c]	0.9	2

[a] From Ikemura and Ozeki [47]
[b] Number of times a codon used divided by number of times all of the codons specifying same amino acid was used, expressed as a percentage; see [15]
[c] A special case, see text

Table 3

Codon usage in signal and mature coding sequences

	TSS	PP	OM	IM	TOTPP	TOTOM	TOTMAT
Phe UUU	20	12	8	0	29	29	58
Phe UUC	13	5	5	3	22	25	47
Leu UUA	29	15	8	6	12	16	28
Leu UUG	7	3	3	1	8	13	21
Leu CUU	10	5	5	0	14	8	22
Leu CUC	10	6	2	2	6	3	9
Leu CUA	3	0	3	0	8	4	12
Leu CUG	34	15	16	3	69	65	134
Ile AUU	23	9	12	2	40	29	69
Ile AUC	9	3	5	1	39	31	70
Ile AUA	3	2	1	0	6	10	16
Met AUG	46	23	20	3	30	28	58

Table 3 (continued)

Val GUU	16	6	5	5	28	51	79
Val GUC	8	4	3	1	16	17	33
Val GUA	13	0	12	1	18	14	32
Val GUG	17	8	7	2	48	13	61
Ser UCU	21	5	12	4	16	17	33
Ser UCC	9	6	2	1	14	20	34
Ser UCA	4	3	1	0	9	15	24
Ser UCG	6	3	3	0	9	12	21
Pro CCU	7	3	2	2	14	5	19
Pro CCC	4	2	2	0	7	4	11
Pro CCA	2	0	2	0	22	19	41
Pro CCG	3	2	0	1	55	23	78
Thr ACU	12	6	6	0	17	35	52
Thr ACC	14	11	1	2	40	40	80
Thr ACA	7	5	2	0	15	23	38
Thr ACG	7	6	1	0	20	16	36
Ala GCU	30	14	14	2	28	51	79
Ala GCC	23	15	5	3	33	28	61
Ala GCA	40	17	22	1	35	44	79
Ala GCG	20	13	7	0	50	31	81
Tyr UAU	5	0	4	1	20	40	60
Tyr UAC	1	1	0	0	28	29	57
* UAA	0	0	0	0	5	6	11
* UAG	0	0	0	0	0	1	1
His CAU	2	2	0	0	12	8	20
His CAC	5	3	1	1	7	4	11
Gln CAA	3	3	0	0	18	31	49
Gln CAG	6	1	5	0	55	59	114
Asn AAU	9	2	7	0	27	36	63
Asn AAC	2	0	2	0	40	79	119
Lys AAA	32	13	13	6	85	79	164
Lys AAG	12	4	7	1	38	16	54
Asp GAU	1	1	0	0	65	57	122
Asp GAC	0	0	0	0	32	46	78
Glu GAA	1	1	0	0	66	49	115
Glu GAG	0	0	0	0	38	15	53
Cys UGU	1	0	1	0	2	14	16
Cys UGC	4	4	0	0	4	8	12
* UGA	0	0	0	0	0	0	0
Trp UGG	1	1	0	0	21	12	33

Table 3 (continued)

Arg CGU	3	2	0	1	23	30	53
Arg CGC	5	3	2	0	30	20	50
Arg CGA	0	0	0	0	3	2	5
Arg CGG	1	1	0	0	3	1	4
Ser AGU	6	4	2	0	12	15	27
Ser AGC	7	3	3	1	17	23	40
Arg AGA	0	0	0	0	3	5	8
Arg AGG	0	0	0	0	0	3	3
Gly GGU	12	6	6	0	53	59	112
Gly GGC	7	2	5	0	48	46	94
Gly GGA	1	0	1	0	12	10	22
Gly GGG	2	1	1	0	19	10	29

The abbreviations used for the sequence files, in the column headings, are given in section 2 (see also table 1)

Table 4

Use of rare codons in signal and mature coding sequences

	TSS	PP	OM	IM	TOTPP	TOTOM	TOTMAT	p
Leu: rare	59	29	21	9	48	44	92	0.007
non-rare	34	15	16	3	69	65	134	
% use rare	63.4	65.9	56.8	75.0	41.0	40.4	40.7	
Ile: rare	3	2	1	0	6	10	16	0.6
non-rare	32	12	17	3	79	60	139	
% use rare	8.6	14.3	5.6	0	7.1	14.3	10.3	
Ser: rare	16	10	6	0	30	42	72	0.3
non-rare	37	14	17	6	47	60	107	
% use rare	30.2	41.7	26.1	0	39.0	41.2	40.2	
Pro: rare	11	5	4	2	21	9	30	<0.001
non-rare	5	2	2	1	77	42	119	
% use rare	68.8	71.4	66.7	66.7	21.4	17.6	20.1	
Thr: rare	7	5	2	0	15	23	38	0.6
non-rare	33	23	8	2	77	91	168	
% use rare	17.5	17.9	20.0	0	16.3	20.1	18.5	
Arg: rare	1	1	0	0	9	11	20	0.6
non-rare	8	5	2	1	53	50	103	
% use rare	11.1	16.7	0	0	14.5	18.0	16.3	
Gly: rare	3	1	2	0	31	20	51	0.5
non-rare	19	8	11	0	101	105	206	
% use rare	13.6	11.1	15.4	0	23.5	16.0	19.8	

The abbreviations used for the sequence files, in the column headings, are given in section 2 (see also table 1). The data are summarised from table 3. Rare codons are defined as in the text, and are listed in table 2. The final column is the probability (p) value from a chi-squared analysis of the TSS and TOTMAT files

are very similar in both types of sequence (p = 0.3–0.6).

There are 3 hydrophobic amino acids (proline, leucine and isoleucine) for which there are rare codons; since signal sequences are rich in hydrophobic residues, then a bias in such sequences towards use of rare codons for 2 of these will presumably result in a degree of clustering of rare codons within signal sequences. It should be noted that the biased usage of proline and leucine rare codons is true for all proteins with processed leader sequences, regardless of the final cellular location of the mature protein (table 4), though the sequence file for inner-membrane proteins is relatively small (see section 2).

We suggest that a clustering of rare codons in signal sequences may play a role in the early steps of protein export, by delaying translation sufficiently (see section 1) to facilitate interaction with the membrane. This bias in rare codon usage may also represent an evolutionary vestige of a mechanism no longer of functional significance, and which has been superceded by, for example, a stop translation sequence [7] and/or a signal recognition-like protein [8,9]. A concentration of rare codons in signal sequences, to serve a purpose such as that proposed above, is likely to have arisen by a process of 'evolutionary tinkering' [48]; i.e. the efficiency of export has been enhanced by a gradual modification of the codon usage in pre-existing signal sequences.

ACKNOWLEDGEMENTS

We acknowledge the support of the Australian Research Grants Scheme. D.M.B. was the recipient of a Commonwealth Postgraduate Research Award.

REFERENCES

[1] Hall, M.N. and Silhavy, T.J. (1981) Annu. Rev. Genet. 15, 91–142.
[2] Silhavy, T.J., Benson, S.A. and Emr, S.D. (1983) Microbiol. Rev. 47, 313–344.
[3] Zwizinski, C., Date, T. and Wickner, W. (1981) J. Biol. Chem. 256, 3593–3597.
[4] Tommassen, J., Van Tol, H. and Lugtenberg, B. (1983) EMBO J. 2, 1275–1279.
[5] Benson, S.A., Bremer, E. and Silhavy, T.J. (1984) Proc. Natl. Acad. Sci. USA 81, 3830–3834.
[6] Randall, L.L. and Hardy, S.J.S. (1984) Microbiol. Rev. 48, 290–298.
[7] Hall, M.N., Gabay, J. and Schwartz, M. (1983) EMBO J. 2, 15–19.
[8] Walter, P. and Blobel, G. (1981) J. Cell Biol. 91, 557–561.
[9] Shultz, J., Silhavy, T.J., Berman, M.L., Fiil, N. and Emr, S.D. (1982) Cell 31, 227–235.
[10] Ikemura, T. (1981) J. Mol. Biol. 146, 1–21.
[11] Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. and Mercier, R. (1981) Nucleic Acids Res. 9, r43–r74.
[12] Smiley, B.L., Lupski, J.R., Svec, P.S., McMacken, R. and Godson, G.W. (1982) Proc. Natl. Acad. Sci. USA 79, 4550–4554.
[13] Gouy, M. and Gautier, C. (1982) Nucleic Acids Res. 10, 7055–7074.
[14] Grosjean, H. and Fiers, W. (1982) Gene 18, 199–209.
[15] Konigsberg, W. and Godson, G.N. (1983) Proc. Natl. Acad. Sci. USA 80, 687–691.
[16] Robinson, M., Lilley, R., Little, S., Emtage, J.S., Yarranton, G., Stephens, P., Millican, A., Eaton, M. and Humphreys, G. (1984) Nucleic Acids Res. 12, 6663–6671.
[17] Misra, R. and Reeves, P. (1984) Proc. Aus. Biochem. Soc. 16, 1.
[18] Delaney, A.D. (1982) Nucleic Acids Res. 10, 61–67.
[19] Inouye, H., Barnes, W. and Beckwith, J. (1982) J. Bacteriol. 149, 434–439.
[20] Kikuchi, Y., Yoda, K., Yamasaki, M. and Tamura, G. (1981) Nucleic Acids Res. 9, 5671–5678.
[21] Jaurin, B., Grundström, T., Edlund, T. and Normark, S. (1981) Nature 290, 221–225.
[22] Sutcliffe, J.G. (1978) Proc. Natl. Acad. Sci. USA 75, 3737–3741.
[23] Postle, K. and Good, R.F. (1983) Proc. Natl. Acad. Sci. USA 80, 5235–5239.
[24] Benjamin Scripture, J. and Hogg, R.W. (1983) J. Biol. Chem. 258, 10853–10855.
[25] Higgins, C.F. and Ferro-Luzzi Ames, G. (1981) Proc. Natl. Acad. Sci. USA 78, 6038–6042.
[26] Higgins, C.F., Haag, P.D., Nikaido, K., Ardeshir, F., Garcia, G. and Ferro-Luzzi Ames, G. (1982) Nature 298, 723–727.
[27] Oxender, D.L., Anderson, J.J., Daniels, C.J., Landick, R., Gunsalus, R.P., Zurawski, G. and Yanofsky, C. (1980) Proc. Natl. Acad. Sci. USA 77, 2005–2009.
[28] Bedouelle, H., Bassford, P.J., Fowler, A.V., Zabin, I., Beckwith, J. and Hofnung, M. (1980) Nature 285, 78–81.

[29] Magota, K., Otsuji, Miki, T., Horiuchi, T., Tsunasawa, S., Kondo, J., Sakiyama, F., Amemura, M., Morita, T., Shinagawa, H. and Nakata, A. (1984) J. Bacteriol. 157, 909–917.
[30] Movva, N.R., Nakamura, K. and Inouye, M. (1980) J. Mol. Biol. 143, 317–328.
[31] Movva, N.R., Nakamura, K. and Inouye, M. (1980) J. Biol. Chem. 255, 27–29.
[32] Mizuno, T., Chou, M.-Y. and Inouye, M. (1983) FEBS Lett. 151, 159–163.
[33] Mutoh, N., Inokuchi, K. and Mizushima, S. (1982) FEBS Lett. 137, 171–174.
[34] Overbeeke, N., Bergmans, H., Van Mansfeld, F. and Lugtenberg, B. (1983) J. Mol. Biol. 163, 513–532.
[35] Hackett, J. and Reeves, P. (1983) Nucleic Acids Res. 11, 6487–6495.
[36] Nakamura, K. and Inouye, M. (1979) Cell 18, 1109–1117.
[37] Hedgpeth, J., Clement, J.M., Marchal, C., Perrin, D. and Hofnung, M. (1980) Proc. Natl. Acad. Sci. USA 77, 2621–2625.
[38] Baga, M., Normark, S., Hardy, J., O'Hanley, P., Lark, D., Olsson, O., Schoolnik, G. and Falkow, S. (1984) J. Bacteriol. 157, 330–333.
[39] Spicer, E.K., Kavanaugh, W.M., Dallas, W.S., Falkow, S., Konigsberg, W.H. and Schafer, D.E. (1981) Proc. Natl. Acad. Sci. USA 78, 50–54.
[40] Yamamoto, T. and Yokota, T. (1982) J. Bacteriol. 150, 1482–1484.
[41] Dallas, W.S. and Falkow, S. (1980) Nature 288, 499–501.
[42] Yamamoto, T., Tamura, T.-A., Ryoji, M., Kaji, A., Yokota, T. and Takano, T. (1982) J. Bacteriol. 152, 506–509.
[43] So, M. and McCarthy, B.J. (1980) Proc. Natl. Acad. Sci. USA 77, 4011–4015.
[44] Picken, R.N., Mazaitis, A.J., Maas, W.K., Rey, M. and Heyneker, H. (1983) Infect. Immun. 42, 269–275.
[45] Van Wezenbeek, P.M.G.F., Hulsebos, T.J.M. and Schoenmakers, J.G.G. (1980) Gene 11, 129–148.
[46] Boyd, A., Kendall, K. and Simon, M.I. (1983) Nature 301, 623–626.
[47] Ikemura, T. and Ozeki, H. (1982) Symposia on Quantitative Biology, XLVII, pp.1087–1097, Cold Spring Harbour.
[48] Jacob, F. (1982) in: The Possible and the Actual, pp.27–46, Pantheon, New York.