COOH-terminal decamers in proteins are non-random

Igor N. Berezovsky, Gelena T. Kilosanidze, Vladimir G. Tumanyan, Lev Kisselev*

Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow 117984, Russian Federation

Received 16 January 1997

Abstract We have undertaken an exhaustive statistical analysis of the amino acid sequences at the carboxyl-terminal (C) ends of proteins. The composition of the C-terminal decapeptides differs from that expected for the given proteins from the overall amino acid composition. For *E. coli*, yeast, and *H. sapiens* it was shown that positively charged amino acid residues are over-represented while Gly residues are under-represented. The C-terminal bias, a novel feature of protein structure, should be taken into account when molecular evolution, spatial structure, translational termination and protein folding are concerned.

Key words: Protein carboxy-terminal end; Statistical analysis; *E. coli*; Yeast; *H. sapiens*

1. Introduction

Proteins exhibit enormous variability in size, composition, spatial structure, cellular location, and a wide variety of biological functions. In spite of this tremendous diversity, all of them as linear copolymers share a common feature: they start with amino (N) termini and end up with carboxy (C) termini.

Cotranslational folding seems to be the major pathway for the majority of globular proteins to form a spatial structure [1–5]. In these cases, the terminal steps of protein synthesis are accompanied by the completion of folding, leading to the formation of a stable native protein globule. Since the translation process ends up at the C-termini of polypeptide chains, one may expect a particular role of the C-ends in the stabilization of protein globules and in the termination of translation. In fact, the modulation of translation termination efficiency by the two C-terminal amino acids in the nascent polypeptide chain has recently been revealed [6,7]. The non-random occurrence of some amino acid residues at the C-ends and certain codons in the last sense codon position [8–15] raises an important issue regarding the extent of this bias in terms of the polypeptide length.

We have undertaken an exhaustive statistical analysis of protein C-terminal sequences deduced from the coding nucleic acid (cDNA) 3′-terminal sequences from databases. This analysis was applied to *E. coli*, yeast and *H. sapiens* sequences to represent, respectively, prokaryotes, lower and higher eukaryotes.

2. Methods

The programs used in the database editing and subsequent statistical analysis were written in Borland C and run on an IBM/PC Pentium-100.

E. coli, yeast and *H. sapiens* coding sequences with length more than 150 nucleotide pairs were taken from the EMBL database, viewing the feature table according to CDS specification. The existence of either full coding sequences or long 5′-contexts to the stop codon was the prerequisite for incorporation of CDS into the set. Moreover, the preservation of open reading frames was monitored. All sequences under investigation included the integer number of codons. For all coding sequences, the corresponding amino acid sequences were deduced. Only one of the duplicate sequences was considered. To ensure the reliability of the cleaning procedure we performed it in two steps. First, the sequences with more than 50% identity of the C-terminal 30 amino acid residues were rejected. Second, the sequences with more than 50% identity in the last 10 amino acid residues were rejected. After database editing, the sets of unique coding sequences for *E. coli*, yeast and *H. sapiens* comprised 1918 sequences (643 165 residues), 3303 sequences (1 683 116 residues) and 3243 sequences (1 613 697 residues), respectively.

In each polypeptide from the (−1) up to the (−31) position, the deviation of the residue representation from the average residue usage in the sets of sequences deduced from cDNA 3′-terminal sequences was calculated. The χ^2 criterion has been used following the formula: (Obs−Exp)^2/Exp, where Obs and Exp are the observed and expected frequencies for amino acid residues. The sums of all 20 χ^2 values for each residue at a particular position gave a value of the total deviation for the given position with 19 degrees of freedom. The significance level was $P < 0.001$ and for it the Obs value was considered to surpass the Exp value if the χ^2 values exceeded 10.8 and 43.8 for 1 and 19 degrees of freedom, respectively. The χ^2 was not calculated if the Exp value was equal to or less than 2.

3. Results

The results of statistical analysis summarised in Table 1 clearly indicate that protein C-ends are biased: the frequencies of certain types of amino acids exceeded the expected levels, while the frequencies of some other residues were below the expected levels (over-represented and under-represented amino acids, respectively). This property is common to the representatives of all three groups of organisms. A common feature for the C-terminal decamers of prokaryotes and eukaryotes was the prevalence of positively charged amino acid residues (Lys and Arg) and the under-representation of Gly residues. Apart from this general pattern, the C-terminal decamers varied between the analysed groups in some detail: over-representation of Glu was found only for unicellular organisms, not for humans; Phe was over-represented in eukaryotes, not in *E. coli*; Pro, Thr, and Leu were under-represented in unicellular organisms, not in man. The bias for the extreme C-termini (positions from (−1) to (−5)) was stronger than for the upstream positions from (−5) to (−10).

In Fig. 1 the integral $\Sigma\chi^2$ values are shown for each of the 31 terminal positions irrespective of the others. These patterns clearly demonstrate the non-random amino acid composition of the C-terminal ends.

4. Discussion

The conservation of the described features of the C-termini for eubacterial, yeast, and human sequences (Table 1) repre-

*Corresponding author. Fax: (7) (095) 1351405.
E-mail: kissel@imb.imb.ac.ru

0014-5793/97/$17.00 © 1997 Federation of European Biochemical Societies. All rights reserved.
PII S0014-5793(97)00112-9

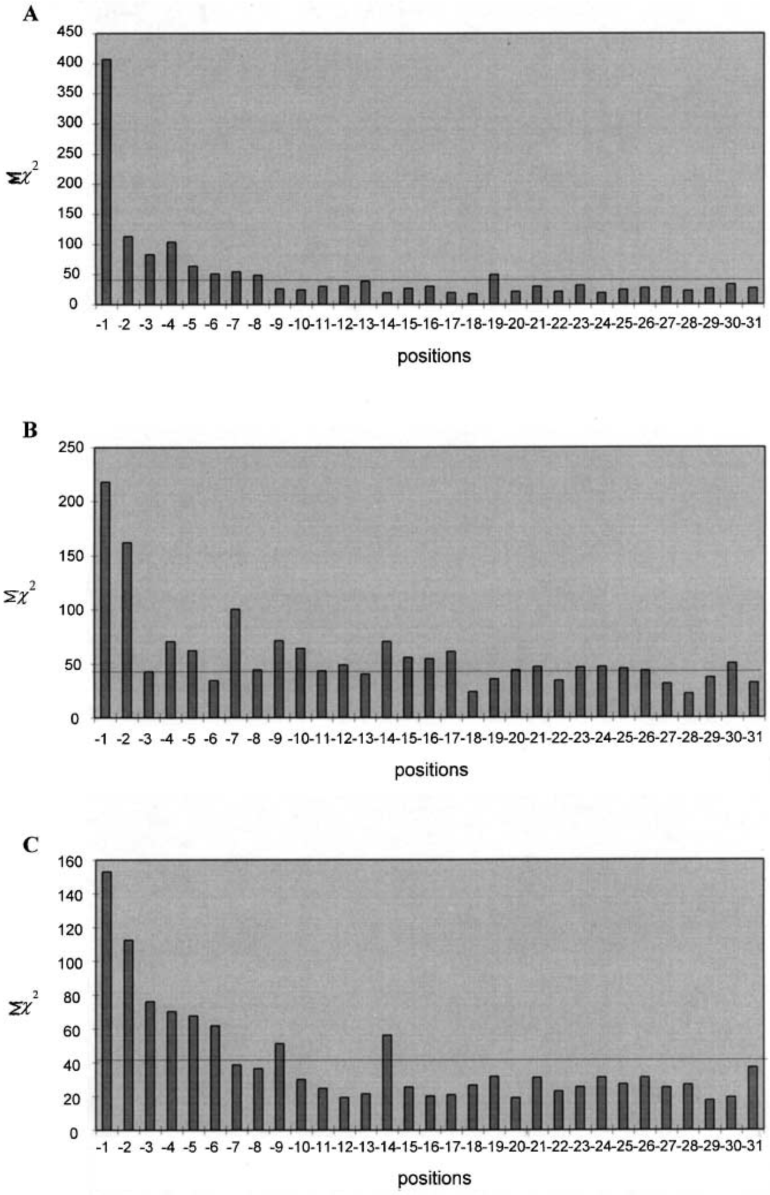

Fig. 1. $\Sigma\chi^2$ values for the last 31 positions at the C-terminal ends of proteins: (A) *E. coli*; (B) yeast; (C) *H. sapiens*. According to the chi-squared approach, we assume that the positions with $\chi^2 < 43.8$ for amino acids ($P < 0.001$) belong to one and the same general set. This level is shown in the panels by the horizontal thin line.

senting prokaryotes, lower and higher eukaryotes, respectively, implies that the bias in amino acid composition of the C-terminal decapeptides provided some evolutionary advantages and was preserved and enhanced during the course of evolution (Fig. 1, compare A–C).

The C-terminal bias seems to be hardly possible to interpret solely in terms of modulation of translation termination as has often been discussed [8–15] although the bias in the (−1) and (−2) positions should contribute to the efficiency of termination [6,7]. The main doubt arises from the fact that the C-terminal bias looks too extended for being directly involved in the mechanism of termination. Furthermore, a huge set of various combinations of over- and under-represented amino acid residues within the C-terminal decapeptides makes it very difficult to propose any chemically justified mechanism to control the efficiency of termination. The positions from (−5) to (−9) are too remote from the tRNA moiety of polypeptidyl-tRNA making unlikely the possibility of a direct interaction between the C-distal amino acids and the ribosomal peptidyltransferase centre.

The bias in amino acid composition of the C-terminal decamers could be associated with several factors: (i) the necessity to fix the C-terminal peptide fragment on the globular surface by non-covalent and covalent interactions with the protein core; electrostatic interactions involving the over-represented polar amino acids may stabilise the globule; (ii) the C-ends may be involved in the intramolecular subunit-subunit contacts to maintain quaternary protein structure by binding of the positively charged C-terminus of one subunit to the negatively charged amino acid cluster of the other subunit [16]; (iii) the positively charged cluster of the C-terminal decamers could serve to keep protein in the certain cellular compartment via protein-protein, or protein-nucleic acid, or other intermolecular interactions; and (iv) the bias at the last two

Table 1
COOH-terminal decamer bias in prokaryotic, lower and higher eukaryotic proteins

Position	*E. coli* (1918)		Yeast (3303)		*H. sapiens* (3243)	
	−	+	−	+	−	+
(−1)	Thr 64.8 Val 18.5 Leu 11.7 Met 11.3	Lys 194.5 Arg 36.2 Glu 19.9	Gly 41.3 Pro 30.1 Thr 20.5 Ser 12.1	Lys 68.5	Gly 51.8 Glu 15.1	Lys 17.1 Phe 14.6 Leu 11.5
(−2)	Pro 16.3	Lys 55.4	Pro 15.3	Lys 48.7 Phe 30.5 Arg 19.0		Arg 30.4 Lys 20.8 Ser 16.6
(−3)		Lys 38.2				Thr 16.8 Lys 14.7 Ser 13.7
(−4)		Lys 50.8 Arg 13.1	Ile 15.5	Lys 22.2		Cys 18.7
(−5)		Lys 13.5		Lys 39.3		Lys 22.9
(−6)		Lys 18.5				Arg 15.4
(−7)	Gly 12.1		Leu 14.3 Thr 12.3	Lys 28.7		
(−8)		Arg 12.4		Lys 11.2		Lys 14.7
(−9)				Glu 20.5		Arg 13.7
(−10)				Glu 20.5		

Over- (+) and under- (−) represented amino acids. Position (−1) corresponds to the last sense codon (the C-terminal amino acid). The χ^2 value for each residue was estimated with one degree of freedom and significance level $P < 0.001$ and is presented after the corresponding amino acid. Total number of the analysed sequences is placed in parentheses.

positions may play some role in the modulation of translation termination [7].

Acknowledgements: This work was partially supported by the Russian National Human Genome Program (grants to V.T. and to L.K.), by the Russian Foundation for Basic Research (grant to L.K.) and by EC INTAS-94 (grant to L.K.).

References

[1] Fedorov, A.N., Friguet, B., Djavadi-Ohaniance, L., Alakhov, Y.B. and Goldberg, M.E. (1992) J. Mol. Biol. 228, 351–358.

[2] Fedorov, A.N. and Baldwin, T.O. (1995) Proc. Natl. Acad. Sci. USA 92, 1227–1231.

[3] Hardesty, B., Kudlicki, W., Odom, O., Zhang, T., McCarthy, D. and Kramer, G. (1995) Biochem. Cell Biol. 73, 1199–1207.

[4] Kolb, V.A., Makeyev, V., Kommer, A. and Spirin, A. (1995) Biochem. Cell Biol. 73, 1217–1220.

[5] Brunak, S. and Engelbrecht, J. (1996) Proteins 25, 237–252.

[6] Mottagui-Tabar, S., Björnsson, A. and Isaksson, L.A. (1994) EMBO J. 13, 249–257.

[7] Björnsson, A., Mottagui-Tabar, S. and Isaksson, L.A. (1996) EMBO J. 15, 101–109.

[8] Brown, C.M., Stockwell, P.A., Trotman, C.N.A. and Tate, W.P. (1990) Nucl. Acids Res. 18, 2079–2086.

[9] Brown, C.M., Stockwell, P.A., Trotman, C.N.A. and Tate, W.P. (1990) Nucl. Acids Res. 18, 6339–6345.

[10] Buckingham, R.H. (1990) Experientia 46, 1126–1133.

[11] Kopelowitz, J., Hampe, C., Goldman, R., Reches, M. and Engelbergkulka, H. (1992) J. Mol. Biol. 225, 261–269.

[12] Arkov, A.L., Korolev, S.V. and Kisselev, L.L. (1993) Nucl. Acids Res. 21, 2891–2897.

[13] Brown, C.M., Dalphin, M.E., Stockwell, P.A. and Tate, W.P. (1993) Nucl. Acids Res. 21, 3119–3123.

[14] Alff-Steinberger, C. and Epstein, R. (1994) J. Theor. Biol. 168, 461–463.

[15] Arkov, A.L., Korolev, S.V. and Kisselev, L.L. (1995) Nucl. Acids Res. 23, 4712–4716.

[16] Hendsch, Z.S. and Tidor, B. (1994) Protein Sci. 3, 211–226.