Protein Engineering vol.12 no.1 pp.23–30, 1999

Amino acid composition of protein termini are biased in different manners

Igor N.Berezovsky[1], Gelena T.Kilosanidze, Vladimir G.Tumanyan and Lev L.Kisselev

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

[1]To whom correspondence should be addressed. E-mail: ber@imb.imb.ac.ru

An exhaustive statistical analysis of the amino acid sequences at the carboxyl (C) and amino (N) termini of proteins and of coding nucleic acid sequences at the 5′ side of the stop codons was undertaken. At the N ends, Met and Ala residues are over-represented at the first (+1) position whereas at positions 2 and 5 Thr is preferred. These peculiarities at N-termini are most probably related to the mechanism of initiation of translation (for Met) and to the mechanisms governing the life-span of proteins via regulation of their degradation (for Ala and Thr). We assume that the C-terminal bias facilitates fixation of the C ends on the protein globule by a preference for charged and Cys residues. The terminal biases, a novel feature of protein structure, have to be taken into account when molecular evolution, three-dimensional structure, initiation and termination of translation, protein folding and life-span are concerned. In addition, the bias of protein termini composition is an important feature which should be considered in protein engineering experiments.

Keywords: amino acid composition/nucleotide composition/protein structure/protein termini/statistical analysis

Introduction

In the cytoplasm of eukaryotic cells, protein synthesis is initiated by methionyl-tRNA placing the first amino acid residue, Met, versus the first coding position of mRNA. In prokaryotes, mitochondria and chloroplasts, the first amino acid incorporated into a growing polypeptide chain is fMet (for a review, see Sherman *et al.*, 1985). However, owing to the presence in both eukaryotic and prokaryotic cells of aminopeptidase activities, the N-terminal fMet and Met are often split, leaving the other amino acid residues as N-termini in processed (mature) proteins. Careful examination of these *N*-exopeptidases led to discovery of the rules governing this process (Yoshida and Lin, 1972; Sherman *et al.*, 1985): the N-terminal (f)Met is split if the adjacent amino acid (position 2) is either Ala, Cys, Gly, Pro, Ser, Thr or Val. If the second triplet encodes for Arg, Asp, Asn, Glu, Gln, Ile, Leu, Lys or Met, the initial (f)Met remains as the first amino acid of the mature protein. From the comparison of the properties of these two sets of amino acids, it was suggested (Bachmain *et al.*, 1986; Varshavsky, 1992) that the physical background of this observation is related to radii of hydration of the amino acid side chains. The chemical structure of the N-terminal amino acid considerably influences the half-life of a protein (Stewart *et al.*, 1995; Grigoryev *et al.*, 1996): from 20 h up to 3 min. This ‘N-end rule’ was confirmed by changing the life-span of proteins via site-directed mutagenesis of their N-termini. In addition to truncation from the N-termini, proteins are also modified by post-translational addition of certain amino acid residues catalyzed by a tRNA-dependent N-terminal aminotransferase (Varshavsky, 1996). Taken together, all these data suggest that the N-terminal amino acid composition should be non-random. This assumption was examined in this work by a statistical analysis of the mature protein structures available in databanks.

The carboxyl (C) termini of proteins possess some distinct properties: (i) at the ribosomal P-site, they neighbor the tRNA moiety in polypeptidyl-tRNA at the terminal step of protein synthesis and by interacting with it they may modulate the efficiency of translation termination; further, the C-terminal amino acid residues may also interact with class I polypeptide chain release factors (RF1, RF2 or eRF1), which occupy the ribosomal A-site at translation termination (Mottagui-Tabar *et al.*, 1994; Björnsson *et al.*, 1996; Nakamura *et al.*, 1996; Tate and Mannering, 1996; Buckingham *et al.*, 1997; Drugeon *et al.*, 1997); (ii) if it is known for the given protein that it folds co-translationally, the nascent C-terminus may interact with the already formed part of the protein globule (Fedorov *et al.*, 1992; Fedorov and Baldwin, 1995; Hardesty *et al.*, 1995; Kolb *et al.*, 1995; Brunak and Engelbrecht, 1996); (iii) in general, one may assume that the C-end may serve as a lock stabilizing the spatial protein structure, as for example has already been shown for collagen (Prockop and Kivirikko, 1995).

The non-random occurrence of certain codons in the last (–1) sense codon position and certain amino acid residues at the C-terminus is well known (Buckingham *et al.*, 1990; Brown *et al.*, 1990a,b, 1993; Kopelowitz *et al.*, 1992; Arkov *et al.*, 1993, 1995; Alff-Steinberger and Epstein, 1994). The modulation of translation termination efficiency by the two C-terminal amino acids of the nascent polypeptide chain [positions (–1) and (–2)] has recently been revealed (Mottagui-Tabar *et al.*, 1994; Björnsson *et al.*, 1996). These data raise an important issue regarding this C-terminal bias in terms of the polypeptide length. If the last two amino acids affect the termination process, only the (–1) and (–2) positions might be biased, leaving the amino acid composition of other C-terminal positions random. However, if the bias extends beyond the last two positions, it may indicate that the amino acid composition of the C-terminal fragments is governed by other factors rather than by requirements of the translation termination machinery.

The above considerations prompted us to examine the extended N- and C-terminal regions of proteins and the 3′ ends of the coding DNA sequences preceding the stop codons by means of an exhaustive statistical analysis. The compulsory prerequisite for application of such an approach was a preliminary cleansing of the protein and nucleic acid databases and a sufficient total number of available sequences to ensure a high significance level.

Materials and methods

Context databases for proteins and nucleic acids

Continuous protein sequences and the C-terminal peptides longer than 50 residues were extracted from the SWISS-PROT database, according to Sequence specification in the feature table. In total, 123 *Escherichia coli* C-terminal protein sequences and peptides (30 966 amino acid residues) were analyzed. The set of 516 mammalian proteins (129 745 amino acid residues) was also considered. *E.coli* and *Homo sapiens* coding sequences longer than 150 base pairs were taken from the EMBL database, viewing the feature table according to Coding Sequence (CDS) specification. The existence of either complete coding sequences or long 5′ contexts to the stop codons was the prerequisite for introducing the CDS to the set. In addition, preservation of open reading frames was also monitored. All sequences under investigation included an integral number of codons. All duplicates of the sequences were excluded from further consideration. The main goal of our work was the analysis of local peculiarities of terminal regions of the sequences and the reproducibility of the obtained peculiarities in different sequences. We fulfilled the procedure for excluding sequences with high and average similarity. Therefore, we performed the cleaning procedure in three steps. First, the sequences with more than 60% identity of the 3′-terminal 50 codons were rejected. Second, the sequences with more than 50% identity in the last 30 codons were rejected. Finally, for the last 10 codons the sequences with more than 60% identity were also rejected. This procedure decreases similarity at the terminal regions of sequences. After database editing, the sets of unique coding sequences for *E.coli* and *H.sapiens* comprised 2003 sequences (668 374 codons) and 3640 sequences (1 772 364 codons), respectively. These two sets were further subdivided into three subsets for UAA, UGA and UAG stop codons; for *E.coli*, 1247 sequences (420 381 codons), 601 sequences (197 237 codons) and 155 sequences (50 753 codons), respectively; for humans, 1132 sequences (545 276 codons), 1729 sequences (827 012 codons) and 779 sequences (400 076 codons), respectively.

Protein spatial structure database

Proteins with known 3-D structure were taken from the protein data base (PDB). In total, 486 unique amino acid sequences were considered, composed of 116 298 amino acid residues. In some cases (41) when the N-terminal amino acids were not resolved by X-ray analysis and were not present in the PDB, they were taken from ENTREZ to complete the protein sequence from the N-termini.

Statistical analysis

The expected (Exp) frequencies of amino acid residues were calculated from the average residue usage in the above-mentioned sets of amino acid sequences. The expected frequency for an amino acid residue of type A at position i will be Exp = $(NA/N)M$, where NA = total number of amino acid residues of type A in the analyzed set of sequences, excluding position i, N = total number of all amino acid residues in the analyzed set of sequences, excluding position i and M = total number of sequences, i.e. the sum of ith positions in the analyzed set of sequences. The expected frequencies for codons were calculated similarly.

For each amino acid residue (codon) at a given position, the deviation of the observed (Obs) values from the Exp values was estimated by the χ^2 criterion according to the formula (Obs – Exp)^2/Exp. For each residue or codon, the χ^2 value was estimated separately with one degree of freedom. The sums of all 20 (61) χ^2 values for each residue (codon) at the given position gave the total deviation for the given position with 19 (60) degrees of freedom. To evaluate the range of differences between the C-terminal regions and the neighboring fragments, a pairwise comparison between them was performed. For this purpose, each position in the sequence was treated as a set containing 20 groups of data and the difference between them was calculated by the χ^2 criterion using the following formula (Borovkov, 1984):

$$\sum_{i=1}^{K} [(m_i/M - n_i/N)^2 MN/(m_i + n_i)]$$

where m_i and n_i are frequencies of amino acid residues in the two positions of the sequence under comparison, M and N are total numbers of amino acid residues in the compared positions and K is equal to 20 because each position may be occupied by any of 20 different amino acids. At the significance level <0.001, Obs was considered to be different from Exp if the χ^2 exceeded 10.8, 43.8 and 99.6 for one, 19 and 60 degrees of freedom, respectively. χ^2 was not calculated for Exp $\leqslant 2$.

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

Results

Amino acid composition of the polypeptide termini for proteins with known 3-D structure

For many proteins there is no complete coincidence between the coding sequence at genomic or mRNA levels and the amino acid sequence of the mature proteins. This holds true for the polypeptide termini where post-translational processing may significantly alter the N ends, as mentioned in the Introduction. The cellular carboxypeptidases may affect the composition of the C end in mature proteins. For these reasons, we first analyzed the N and C termini of mature proteins with known 3-D structure taken from the PDB (Table I).

Strong over-representation of Met is observed at the first sense position (+1) as anticipated from the known mechanism of initiation of protein synthesis. Besides this anticipated bias, two other amino acid residues are over-represented: Ala (position +1) and Thr (positions +2 and +5). Several amino acids are under-represented at position (+1) while at position (+2) only Gly is under-represented (Table I). The over-representation of Ala and Thr seems to agree with the N-end rule (Grigoryev *et al.*, 1996; Varshavsky 1992, 1996).

It should be noted that the protein set deposited in the PDB is certainly non-random. It is composed of the most abundant and stable proteins whereas proteins with short half-lives (e.g. heat shock proteins) or rare proteins (e.g. transcription factors) are not represented in the PDB. For this reason, the bias documented in Table I is typical for a certain set of proteins, namely for the most abundant and stable molecules.

The C-ends in the protein set with known 3-D structures are also biased (Table I): Lys and Cys are over-represented whereas Thr is under-represented.

From the data summarized in Table I, we conclude that both termini of the polypeptide chains of the most abundant and stable proteins in various groups of organisms are non-random. The main limitation of this conclusion derives from the analyzed set of proteins, which is not sufficient for exhaustive

Table I. Over- (+) and under- (–) representation of the N- and C-terminal amino acid residues in the set 3-D protein structures

Position		Residue	Obs	Exp	χ^2
N-end	(1)	Met	142	9.2	1924.5
		Ala	98	40.7	80.6
		Leu	11	38.3	19.5
		Phe	3	18.7	13.2
		Ile	8	25.4	11.9
		Asn	6	22.0	11.6
		Tyr	3	16.9	11.4
	(2)	Thr	48	28.8	12.8
		Gly	14	39.9	16.9
	(5)	Thr	52	29.1	18.1
C-end	(–1)	Lys	50	29.5	14.3
		Ala	64	42.3	11.1
		Thr	11	29.2	11.3
	(–2)	Lys	52	29.5	17.1
	(–3)	Lys	51	29.5	15.6
		Cys	18	7.9	12.8
	(–8)	Cys	23	7.9	28.9

Over- (+) and under- (–) represented amino acids. Position (1) corresponds to the first position (the N-terminal amino acid) and position (–1) corresponds to the last position (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 analyzed sequences = 486.

statistical analysis and at the same time the set is not representative because only crystallizable proteins were analyzed. Further, the PDB set is composed of proteins of both prokaryotes and eukaryotes and if these groups possess different biases at both termini they are not recognized if only PDB sequences are considered.

For all these reasons, we continued to analyze the C-terminal bias by taking into account many more protein and nucleic acid sequences from databanks.

Statistical bias for the E.coli and mammalian C-terminal sequences

We found a bias in the C-terminal amino acid frequencies for both *E.coli* and mammalian polypeptide chains. Even for a relatively limited set of protein sequences from *E.coli*, it was possible to consider over-representation of Lys in the (–1) position ($\chi^2 = 28.1$; Obs and Exp values for this position are 22 and 7.5, respectively), confirming at the protein level the over-representation of Lys codons in *E.coli* demonstrated earlier (Brown *et al*., 1990a,b, 1993; Kopelowitz *et al*., 1992; Arkov *et al*., 1993; Alff-Steinberger and Epstein, 1994). Position (–2), Obs = 15, Exp = 7.5, and position (–3), Obs = 16, Exp = 7.5, also prefer Lys residues over the expected values calculated from the average frequencies of amino acid residues in the set of 123 protein sequences containing 30 966 residues. For mammalian proteins (Table II) it was found that there were prominent peculiarities in the C-terminal sequences: Lys and Cys residues were over-represented, in agreement with the Lys codon over-representation in *H.sapiens* (Arkov *et al*., 1995). However, more refined analysis was hindered by the insufficient total number of well characterized protein sequences: for statistical analysis, it is critically important to deal with large sets of sequences to obtain reliable results. Therefore, we had to extend our analysis to the coding regions of genes, because in this case the sets available after appropriate cleansing (see Materials and methods) were much more numerous and allowed high-fidelity analysis.

Table II. Over- and under-representation of the C-terminal amino acid residues in the set of mammalian protein sequences

Position from end of sequence	Residue	Obs	Exp	χ^2	$\Sigma\chi^2$
(–1)	Lys	56	33.1	15.8	79.9
	Gly	14	39.0	16.0	
(–2)	Cys	28	14.0	14.0	47.2
	Lys	**51**	**33.1**	**9.7**	
(–3)	Cys	31	14.0	20.6	57.9
(–5)	Cys	28	14.0	13.9	31.7
(–6)	Lys	59	33.1	20.2	39.1
(–7)	Ile	**35**	**23.5**	**5.7**	**38.2**
(–9)	Lys	53	33.2	11.8	45.0
(–14)	Cys	34	14.0	28.4	45.7

The χ^2 value for each amino acid residue is estimated with one degree of freedom and significance level $P < 0.001$. The $\Sigma\chi^2$ values are estimated with 19 degrees of freedom and significance level $P < 0.001$. The expected (Exp) and observed (Obs) values and the corresponding χ^2 values for amino acid residues and the $\Sigma\chi^2$ values for those positions that do not reach 10.8 and 43.8 (for one and 19 degrees of freedom, respectively) are given in bold. In positions 4, 8 and 10–13 and other non-mentioned positions within the last 31 positions over- or under- representation of amino acid residues is insignificant or not found.

Statistical bias for the 3′-terminal coding nucleotide sequences and deduced protein C ends

The frequencies for the over- and under-represented 3′-terminal codons for the *E.coli* and human sequences are presented in Table III. The χ^2 values demonstrated the bias from the average frequencies estimated by the analysis of the codon context database (see Materials and methods). The codon distribution (2003 sequences from *E.coli*) differed significantly ($P < 0.001$) in the 3′-terminal coding region, mainly from the (–1) to (–8) positions, counting from the stop codon. The over- or under-representation of one or more codons for charged amino acid residues was observed from the (–1) to (–8) positions. The preferred codons for the C-terminal octamer were those for Lys and Arg. Strongly under-represented were codons for Thr (two out of four), Met (AUG) and Gly (GGC and GGU). For *H.sapiens* (3640 sequences), the codon bias covered the C-terminal nonamer versus the octamer for *E.coli* codons. The majority of the different codon types as biased at the (–1) position, where seven amino acids were over-represented, as opposed to five residues noted earlier (Arkov *et al*., 1995). In the (–4) position the Cys residue was over-represented.

Our data regarding the prevalence and deficiency of certain types of amino acids at the protein C ends partially support earlier data (Trifonov, 1987; Buckingham *et al*., 1990; Brown *et al*., 1990a,b, 1993; Kopelowitz *et al*., 1992; Arkov *et al*., 1993, 1995; Alff-Steinberger and Epstein, 1994). We found 14 and 16 over-represented codons while seven and eight codons were under-represented in the *E.coli* and *H.sapiens* coding sequences, respectively. We divided all *E.coli* and human sequences into three subsets for UAA, UGA and UAG stop codons. In *E.coli*, the over- and under-represented codons were distributed similarly in these subsets, except for the most biased (–1) position in the UAA subset. For human sequences, the largest bias of codon representation was shown for the UGA subset (complete data for all subsets are available through e-mail: ber@imb.imb.ac.ru). For the amino acid frequencies in the sequences ‘translated’ from the coding sequences, we noticed that the over- and under-representation of some codon types was eliminated. Consequently, the total number of

Table III. Codon representation at the 5′ side of stop codons

Position from the stop codon	*E.coli*					*H.sapiens*				
	Codon	Amino acid residue	Obs	Exp	χ^2	Codon	Amino acid residue	Obs	Exp	χ^2
(–1)	AAG	Lys	87	24.8	155.9	UCC	Ser	102	63.0	24.2
	AAA	Lys	153	71.3	93.6	CUC	Leu	104	68.0	19.0
	AGG	Arg	18	3.6	58.8	UUC	Phe	109	74.6	15.9
	GAG	Glu	79	38.6	42.4	AAC	Asn	109	76.3	14.0
	CAA	Gln	57	28.1	29.7	UUG	Leu	67	42.6	14.0
	CGA	Arg	20	7.0	24.0	UAC	Tyr	88	59.3	13.9
	ACC	Thr	6	46.6	35.4	CAC	His	77	52.4	11.5
	CUG	Leu	55	103.4	22.5	AAA	Lys	122	89.9	11.5
	ACG	Thr	7	26.5	14.3	GGA	Gly	20	63.9	30.1
	GGC	Gly	29	57.8	14.3	GGG	Gly	23	60.0	22.8
	GUG	Val	24	49.3	13.0	ACC	Thr	39	72.9	15.7
	AUG	Met	29	53.0	10.9	GGU	Gly	19	42.7	13.2
						GAG	Glu	107	147.6	11.2
(–2)	AAG	Lys	73	24.8	93.3	AGA	Arg	78	39.7	37.0
	AGG	Arg	21	3.5	86.1	AGG	Arg	71	38.6	27.2
	AGA	Arg	16	5.8	17.8	AAA	Lys	129	89.9	17.1
	GGA	Gly	33	16.3	17.2	CUG	Leu	95	141.6	15.3
	CGG	Arg	24	10.7	16.4	GUG	Val	71	105.4	11.2
	UCA	Ser	30	15.2	14.5					
	ACA	Thr	30	15.6	13.3					
	ACC	Thr	19	46.6	16.3					
(–3)	AAG	Lys	55	24.9	36.4	AGC	Ser	104	69.6	17.1
	GAG	Glu	75	38.6	34.4	AUC	Ile	49	83.7	14.4
	AGA	Arg	19	5.8	29.9					
	ACA	Thr	30	15.6	13.3					
	AAA	Lys	102	71.5	13.0					
	AGG	Arg	10	3.6	11.6					
	AUG	Met	28	53.0	11.8					
(–4)	AAG	Lys	62	24.9	55.4	UGU	Cys	73	35.8	38.7
	CGA	Arg	24	7.0	41.2	UGC	Cys	74	45.2	18.4
	AAA	Lys	108	71.5	18.7					
	GGA	Gly	33	16.3	17.2					
	CUG	Leu	62	103.4	16.6					
	GGU	Gly	24	52.9	15.8					
(–5)	AAG	Lys	57	24.9	41.4	AAG	Lys	171	125.4	16.6
	AGG	Arg	14	3.6	30.6					
	ACU	Thr	40	20.1	19.6					
(–6)	UUG	Leu	57	24.1	44.8	AGA	Arg	70	39.7	23.1
	AUA	Ile	28	10.2	31.0	AGG	Arg	62	38.6	14.1
	AGG	Arg	12	3.6	19.9	UCA	Ser	63	40.5	12.5
	AAA	Lys	10	71.5	16.7	CAA	Gln	67	44.1	11.9
(–7)	AGG	Arg	10	3.6	11.6					
(–8)	AAG	Lys	44	24.9	14.6	AAA	Lys	124	89.8	13.0
						AAG	Lys	164	125.4	11.9
(–9)						AGA	Arg	67	39.7	18.8
						GGU	Gly	21	42.7	11.0

The χ^2 value for each codon was estimated with one degree of freedom and significance level $P < 0.001$.

deviations from the average residue representation was reduced. The resulting ‘+’(over-represented) and ‘–’(under-represented) consensuses are shown in Table IV. In terms of amino acid composition, the C-termini are enriched mainly in positively charged and deficient in bulky aliphatic residues at the last position. These features extend for the C-terminal octamers (*E.coli*) and nonamers (*H.sapiens*). For *E.coli*, the preference for charged amino acids is revealed at the following positions: Lys, (–1; Obs = 260, Exp = 96.1) and from (–3) to (–6) [157, 170, 139, 137 – Obs values for positions (–3)- (–6), respectively; 96.3 – Exp value for position (–4); 96.4 – Exp value for positions (–3), (–5) and (–6)]; Arg, (–1), (–4) and (–8) [177, 155, 152 – Obs values for positions (–1), (–4) and (–8), respectively; 114.6 – Exp value for position (–1); 114.7 – Exp value for positions (–4) and (–8)]; and Glu, (–1; Obs = 118, Exp = 86.6). For humans, similarly, there is a tendency for over-representation of charged residues: Lys, from (–1) to (–3), (–5) and (–8) [278, 281, 277, 282, 288 – Obs values for

Table IV. The C-terminal bias of amino acid residues

Position	*E.coli*				*H.sapiens*			
	Residue	Obs	Exp	χ^2	Residue	Obs	Exp	χ^2
(–1)	Lys	240	96.1	215.4	Phe	189	135.3	21.3
	Arg	177	114.6	33.9	Lys	278	215.3	18.3
	Glu	118	86.6	11.4	Leu	410	344.2	12.6
	Thr	22	109.0	69.4	Gly	123	252.5	66.5
	Val	85	140.6	22.0	Glu	192	255.1	15.6
	Leu	151	201.8	12.8	Thr	146	198.8	14.0
	Ile	82	118.2	11.1				
	Met	29	53.0	10.9				
(–2)					Arg	277	197.3	32.2
					Lys	281	215.3	20.1
					Ser	353	283.2	17.2
(–3)	Lys	157	96.4	38.1	Thr	264	198.6	21.5
	Met	28	53.0	11.8	Ser	358	283.1	19.8
					Lys	277	215.3	17.7
(–4)	Lys	170	96.3	56.3	Cys	147	80.9	53.9
	Arg	155	114.7	14.2				
(–5)	Lys	139	96.4	18.8	Lys	282	215.2	20.7
(–6)	Lys	137	96.4	17.1	Arg	271	197.3	27.5
					Glu	211	166.7	11.8
(–7)	Gly	102	148.0	14.3				
(–8)	Arg	152	114.7	12.1	Lys	288	215.2	24.6
(–9)					Arg	246	197.3	12.0

The χ^2 value for each residue was estimated with one degree of freedom and significance level $P < 0.001$.

positions (–1)- (–3), (–5) and (–8), respectively; 215.3 – Exp values for positions (–1)- (–3); 215.4 – Exp values for positions (–5) and (–8)]; Arg, (–2), (–6) and (–9) (Obs = 277, 241, 246, respectively; Exp = 197.3).

C-terminal nona/octapeptides differ from internal peptides in amino acid composition. The integral $\Sigma\chi^2$ values for each of the 31 terminal positions for the codons and amino acid residues (Figure 1) clearly demonstrate the non-random amino acid composition of the C ends and the non-random codon representation in the 3′-terminal coding regions. The integral χ^2 values correspond to the integral frequencies of all codons and amino acids at the given position and illustrate the non-randomness in the codon and amino acid frequencies for *E.coli* and human sequences. In order to demonstrate that the amino acid contexts were unusual in the downstream terminal regions of coding sequences we performed a pairwise comparison of the residue representations between all pairs of positions for the last 31 positions (Figure 2). Obviously, the downstream terminal regions had the largest bias. A difference between positions within the terminal region and between positions from the terminal region and other positions within the last 31 amino acid residues was also shown. The non-terminal positions had residue frequencies corresponding to the average values.

In spite of the huge evolutionary distance between *E.coli* and *H.sapiens*, the general tendency for over-representation of certain types of amino acids at the C-terminal ends of polypeptides was evident (Table IV). At the same time, the terminal codon and amino acid patterns are peculiar for each evolutionary group: (a) the non-random C-terminal peptides in humans were slightly longer than in *E.coli*; (b) eight different codons corresponding to seven different amino acid residues were over-represented at the (–1) position in humans, and six and five, respectively, in *E.coli*; (c) in humans, strong under-representation of codons for non-polar amino acids (Leu, Ile, Gly and Val) from the (–1) to (–3) positions was observed, not clearly defined in *E.coli*; (d) in humans, the statistical bias was larger upstream from the UGA stop codon whereas in *E.coli* there was no preference for any of the stop codons. Further, in humans UGA was the most abundant stop-codon, whereas in *E.coli* it was UAA, as noted earlier by others.

Discussion

The analysis of amino acid sequences from databases of spatial protein structures revealed peculiarities of N-terminal amino acid compositions. Over-representation of Met at the first position is an evident consequence of beginning of translation from Met or fMet in eukaryotic and prokaryotic cells. The excess of Ala is favorable for proteins which follow the ‘N-end rule’ (Varshavsky, 1992, 1996). Indeed, there is over-representation of amino acid residues which provide protein long life-spans and under-representation of residues which serve as signals for rapid degradation. Interestingly, the analysis of codons of over-represented (excluding Met) and under-represented amino acid residues results in the following consensus: for under-represented residues the triplet is (nonG, U or A, N); for over-represented Ala codons it is (G, C, N). It is known (Trifonov, 1987) that (G, nonG, N) codons a characteristic pattern for coding nucleotide sequences. Alternatively, it is tempting to speculate that the uncovered pheno-

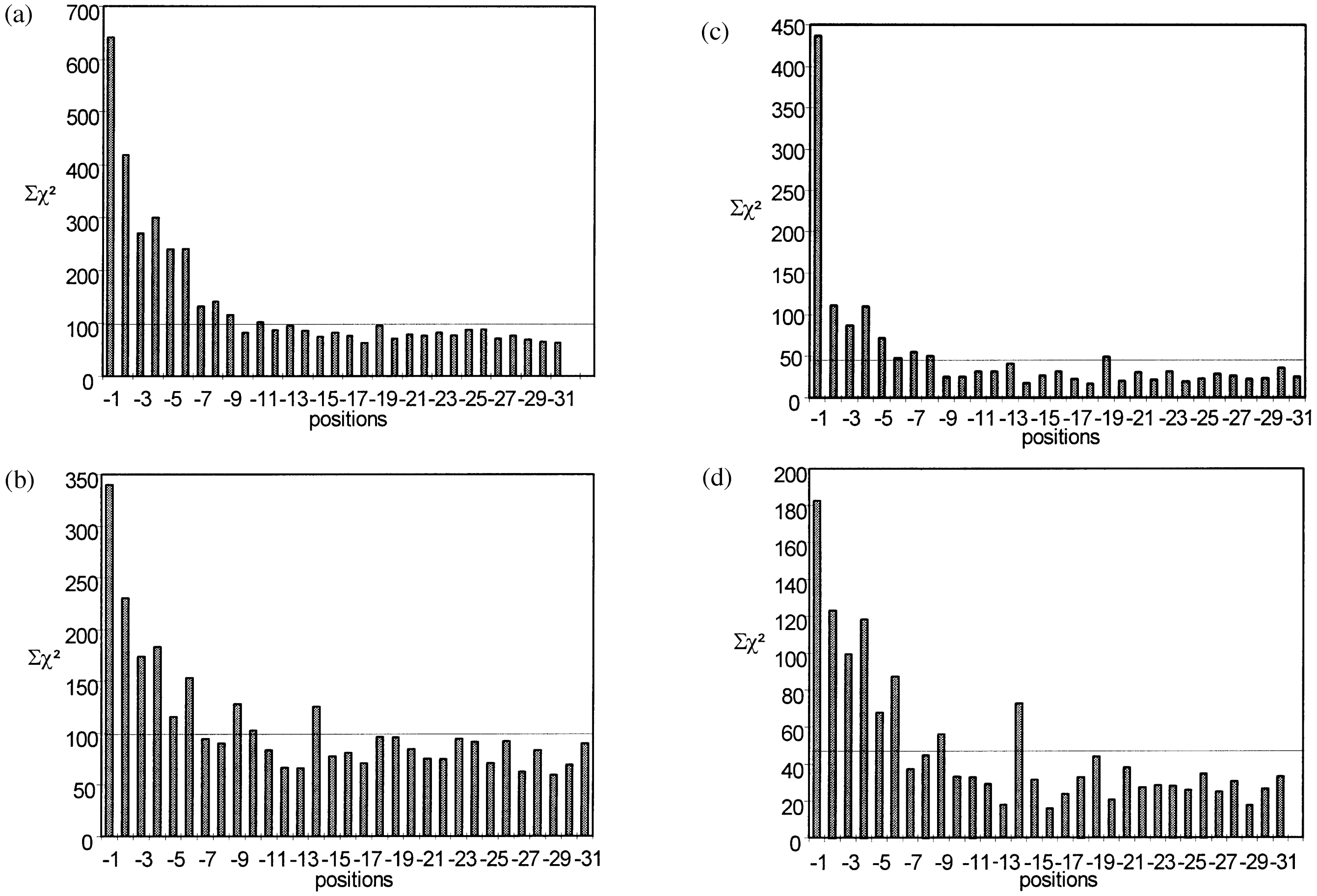

Fig. 1. $\Sigma\chi^2$ values for the last 31 positions of nucleotide sequences at the 5′ side of a stop codon (a and b) and at the C termini of proteins (c and d). According to the chi-squared approach, we assume that the positions with $\chi^2 < 99.6$ for codons and $\chi^2 < 43.8$ for amino acids ($P < 0.001$) belong to the same general set. This level is shown with the horizontal line. *E.coli* (a and c) and humans (b and d).

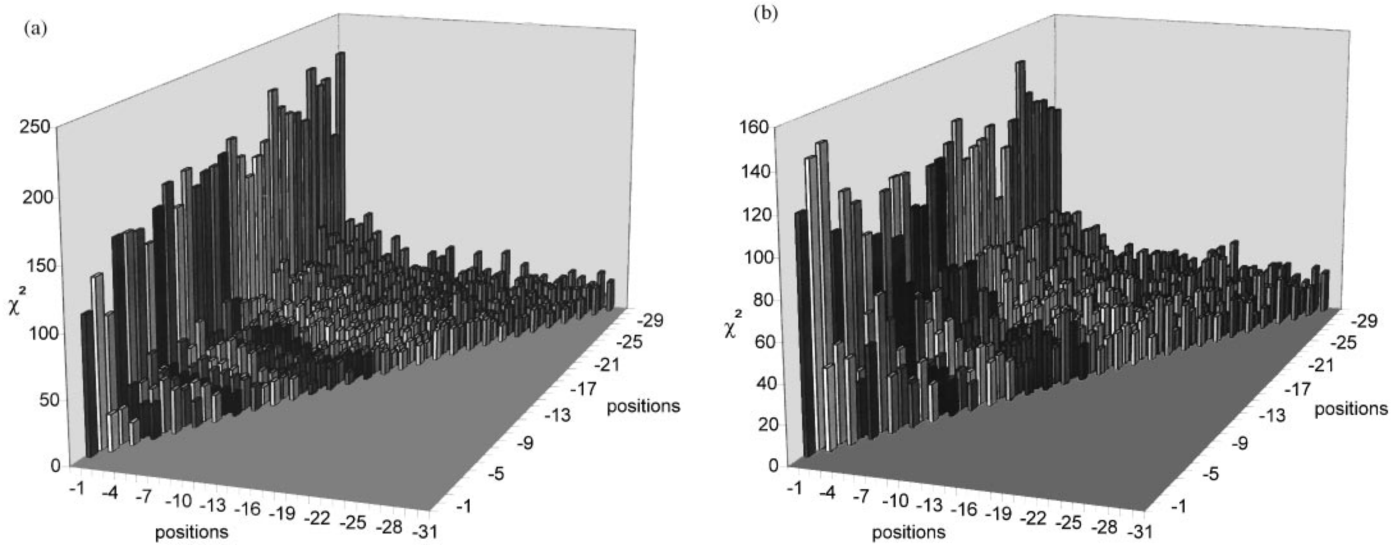

Fig 2. χ^2 values for comparison of two positions in the C-terminal regions of proteins from *E.coli* (a) and humans (b). See text for details.

menon is related to a possible role of the triplet after AUG as additional initiation signal. It is noteworthy that there are objective limitations for creating a representative set of sequences for the N-ends because there is no specific initiation signal as, for example, stop-codon to map the C-end.

The general similarity between the amino acid composition of the protein C ends calculated from the PDB database of spatial structures and the amino acid composition of proteins deduced from their coding nucleotide sequences provides indirect evidence that carboxypeptidases are not involved in processing of a considerable part of C-termini. In other words, we assume that the majority of protein C ends in the cell are truncated in the course of maturation and/or proteolysis.

Most proteins have a quaternary organization ensured by intersubunit interactions. The positively charged C end of one subunit may interact with the negatively charged internal amino acid cluster belonging to the other subunit of the same protein molecule. Since most of the C ends have a surface location, one may anticipate that these fragments could serve as substrates for post-translational modification(s) (acetylation, methylation, etc.). Moreover, positively charged C-terminal clusters on the globular surface may govern the binding to other macromolecules charged negatively, e.g. nucleic acids.

The observed bias at the N and C ends may be considered in the frame of the general concepts of protein globule formation and stabilization (Cantor and Schimmel, 1980; Creighton, 1993). Indeed, protein folding is often initiated at the N end of the polypeptide chain (Fedorov *et al*., 1992; Fedorov and Baldwin, 1995; Hardesty *et al*., 1995; Kolb *et al*., 1995). The C-terminal octa/nonamers with the bias in their amino acid composition shown in this work should better fix their positions at the globular surface via non-covalent and covalent interactions. Ion pairs formed at the surface contribute to the stability of protein 3-D structure while salt bridges inside the globule tend to destabilize it (Barlow and Thornton, 1983; Dill, 1990; Horovitz *et al*., 1990; Serrano *et al*., Šali *et al*., 1991; Hendsch and Tidor, 1994; Starich *et al*., 1996). In addition to salt bridges, H-bonds may be involved in positioning of the C termini on the globular surface. Moreover, the covalent cross-links between Cys residues (S–S bridges) could also be involved in stabilization of the protein surface structure.

The C-terminal octa/nonamer bias is considered to have little effect on translation termination. It is known that the (–3) amino acid position has a very weak influence on translation termination in *E.coli* at the UGA stop codon wherease the C-terminal dipeptide affects the termination efficiency (Mottagui-Tabar *et al*., 1994; Björnsson *et al*., 1996), but this effect is not related to the strong bias at the last two positions. For example, as shown in Table IV, the (–2) position in *E.coli* proteins has no bias at all towards three termination codons whereas basic amino acids ensure efficient termination at UGA versus acidic residues that are inefficient (Björnsson *et al*., 1996). For the (–1) position, many amino acid residues are favorable for efficient termination in *E.coli* at UGA (Björnsson *et al*., 1996), although in this position only polar (charged) amino acids are over-represented (Table IV).

In the present work, neither protein nor nucleic acid sets were subdivided depending on the protein abundance, size, subunit or domain composition, isoelectric point, globular or filamentous shape, loop regions, etc. At the moment, the total number of available sequences seems to be insufficient for such a kind of refined statistical analysis of protein families. However, in the near future, when many more sequences become available such analysis may reveal more sharp biases for certain groups of proteins. On the other hand, it may appear that some of the protein group(s) is (are) random at their C-termini (for instance, filamentous or membrane proteins, where the role of C ends is probably less critical for the maintenance of the overall protein structure).

In conclusion, we explain the bias at the protein termini shown in this work by: (i) putative role of N- and C-termini in protein spatial structure (Thornton and Sibanda, 1983; Christopher and Baldwin, 1996); (ii) involvement in maintenance of the quaternary protein structure; (iii) specific targeting to certain cell compartments and/or a substrate for certain types of post-translational modification(s) and/or degradation; (iv) modulation of the translation initiation and termination by the marginal amino acids (Varshavsky,1992, 1996; Mottagui-Tabar *et al*., 1994; Björnsson *et al*., 1996; Nakamura *et al*., 1996; Tate and Mannering, 1996). It is evident that the bias of protein ends is one of the constraints which one should take into account in the design of new protein molecules or in engineering of the existing proteins.

Acknowledgements

The authors are grateful to M.Gelfand, O.Ptitsyn, E.Gupalo and A.Galkin for critical reading of the manuscript and valuable comments. This work was partially supported by the Human Frontier Science Program (L.L.K.), the Russian National Human Genome Program (grant to V.G.T.), the Russian Foundation for Basic Research (grant to L.L.K.), the EC (INTAS-96-1706) and the State Program of Support of Russian Scientific Schools (L.L.K.). Part of this work was presented at the Workshop 'Mathematical Analysis of Biological Sequences', August 4–6, 1996, Trondheim, Norway and partly and briefly described elsewhere (Berezovsky *et al*., 1996, 1997).

References

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

Arkov,A.L., Korolev,S.V. and Kisselev,L.L. (1993) *Nucleic Acids Res*., **21**, 2891–2897.

Arkov,A.L., Korolev,S.V. and Kisselev,L.L. (1995) *Nucleic Acids Res*., **23**, 4712–4716.

Bachmain,A., Finley,D. and Varshavsky,A. (1986) *Science*, **234**, 179–186.

Barlow,D.J. and Thornton,J.M. (1983) *J. Mol. Biol.*, **168**, 867–885.

Berezovsky,I.N., Kilosanidze,G.T., Tumanyan,V.G. and Kisselev,L.L. (1996) *Folding Des.*, **1,** Supplement, 9–10.

Berezovsky,I.N., Kilosanidze,G.T., Tumanyan,V.G. and Kisselev,L. (1997) *FEBS Lett*., **404,** 140–142.

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

Borovkov, A.A. (1984) *Mathematical Statistics. Additional Chapters.* Nauka, Moscow, pp. 13–15.

Brown,C.M., Stockwell,P.A., Trotman,C.N.A. and Tate,W.P. (1990a) *Nucleic Acids Res*., **18**, 2079–2086.

Brown,C.M., Stockwell,P.A., Trotman,C.N.A. and Tate,W.P. (1990b) *Nucleic Acids Res*., **18**, 6339–6345.

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

Brunak,S. and Engelbrecht,J. (1996) *Proteins*, **25**, 237–252.

Buckingham,R.H., Sörensen,P., Pagel,F.T., Hijazi,K.A., Mims,B.H., Brechemier-Baey,D. and Murgola,E.J. (1990) *Biochim. Biophys. Acta*, **1050**, 259–260.

Buckingham,R.H., Grentzmann,G. and Kisselev,L. (1997) *Mol. Microbiol*., **24,** 449–456.

Cantor,C.R. and Schimmel,P.R. (1980) *Biophysical Chemistry.* Freeman, San Francisco.

Christopher, J.A. and Baldwin, T.O. (1996) *J. Mol. Biol.*, **257**, 175–187.

Creighton,T.E. (1993) *Protein Structures.* Freeman, San Francisco.

Dill,K.A. (1990) *Biochemistry*, **29**, 7133–7155.

Drugeon,G., Jean-Jean,O., Frolova,L., Le Goff,X., Philippe,M., Kisselev,L. and Haenni,A.-L. (1997) *Nucleic Acids Res*., **25,** 2254–2258.

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

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

Grigoryev,S., Stewart,A.E., Kwon,Y.T., Arfin,S.M., Bradshaw,R.A., Jenkins, N.A., Copeland,N.G. and Varshavsky,A. (1996) *J. Biol. Chem.*, **271**, 28521–28532.

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

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

Horovitz,A., Serrano,L., Avron,B., Bycroft,M. and Fersht,A.R. (1990) *J. Mol. Biol.*, **216**, 1031–1044.

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

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

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

Nakamura,Y., Ito,K. and Isaksson,L.A. (1996) *Cell*, **87**, 147–150.

Prockop, D.J. and Kivirikko, K.I. (1995) *Annu. Rev. Biochem.*, **64**, 403–434.

Šali,D., Bycroft,M. and Fersht,A.R. (1991) *J. Mol. Biol.* **220**, 779–788.

Serrano,L., Horovitz,A., Avron,B., Bycroft,M. and Fersht,A.R. (1990) *Biochemistry*, **29**, 9343–9352.

Sherman,F., Stewart,J.W. and Tsunasawa,S. (1985) *BioEssays*, **3**, 27–31.

Starich,M.R., Sandman,K., Reeve,J.N. and Summers,M.F. (1996) *J. Mol. Biol.*, **255**, 187–203.

Stewart,A.E., Arfin,S.M. and Bradshaw,R.A. (1995) *J. Biol. Chem.*, **270**, 25–28.

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

Thornton, J.M. and Sibanda, B.L. (1983) *J. Mol. Biol.*, **167**, 443–460.

Trifonov,E.N. (1987) *J. Mol. Biol.*, **194**, 643–652.

Varshavsky,A. (1992) *Cell*, **69**, 725–735.

Varshavsky,A. (1996) *Proc. Natl Acad. Sci. USA*, **93**, 12142–12149.

Yoshida,A. and Lin,M. (1972) *J. Biol. Chem.*, **247**, 952–957.

Received February 20, 1998; revised September 10, 1998; accepted September 14, 1998