ARGININE AS AN EVOLUTIONARY INTRUDER INTO PROTEIN SYNTHESIS

by

Thomas H. Jukes

Space Sciences Laboratory
University of California, Berkeley, Calif. 94720

Received June 4, 1973

SUMMARY

The arginine content of proteins is, on the average, far less than would be anticipated from the fact that 6 of the 61 codons for amino acids in the genetic code are for arginine. In contrast, lysine is more abundant than would be expected from its 2 codons. It is suggested that arginine replaced ornithine in protein synthesis during the evolution of the genetic code.

The frequency of occurrence of the different amino acids in proteins is not uniform, and it has been noted that there is a fair correlation between the abundance of amino acids and the number of codons for each in the genetic code. It is of interest to explore divergences from this correlation, and to examine the possible reasons for such divergences.

The distribution of amino acids in 53 vertebrate proteins was compared with the number of codons for each amino acid by King and Jukes (1969). A good correlation was found except for the conspicuous case of arginine, which was present to the extent of 4.2% as compared with an anticipated 10.7%. Arginine and lysine are the amino acids that are primarily responsible for basicity in proteins. It appears that lysine is used preferentially to arginine in most proteins (Table 1), despite the fact that lysine has 2 codons and arginine has 6. The basicity of arginine (pK'_3 = 12.5) is much greater than that of lysine (pK'_3 = 10.5).

Subak-Sharpe et al (1966) noted the scarcity of CpG in the DNA of certain organisms as indicating a selective process against this doublet, which is present in four arginine codons. Apparently, arginine is to some extent rejected in evolution. It therefore appears that the concept

that "all is for the best in this best possible of all codes" may not be valid.

Perhaps arginine entered the mechanism of protein synthesis by charging the tRNAs for another amino acid, which was thus displaced from the genetic code. This might happen if arginine had a high affinity for the tRNA-aminoacyl-ligase system in question. An example of such a phenomenon in enzyme chemistry is the far greater affinity of folic acid antagonists containing a 4-amino group than the normal substrate, dihydrofolic acid, has for dihydrofolic reductase (Delmonte and Jukes, 1962).

Ornithine gives rise to arginine in a regular metabolic pathway. The formation of this pathway during evolution would have resulted in arginine entering the biochemical arena. This could have led to arginine replacing ornithine in protein synthesis if arginine had a stronger affinity than ornithine for ornithine tRNA. The entry of arginine into proteins could have conveyed certain special advantages, but, at the same time, lysine (homo-ornithine) might have been more acceptable than arginine for many purposes as a substitute for the missing and similar amino acid, ornithine. The living system would then tend to select lysine codons for such purposes in preference to arginine codons, with the results shown in Table 1.

The proteins in Table 1 are of varying lengths and have different functions. The important parameter is the proportion of arginine and lysine codons in the total number of codons, and this is shown without further statistical treatment.

The immunoglobulins are proteins in which two genes are responsible for a single polypeptide chain. The chains, light and heavy, contain variable or specificity (S) and constant (C) regions. These regions are assumed to have descended from a single common ancestral protein by gene duplication (Hill et al, 1966). Following this, the light chain evolved as a molecule containing S and C regions of approximately equal length. The heavy chain has an S region of about the same length as that of the

TABLE 1. ARGININE AND LYSINE CONTENT OF 83 CLASSES OF PROTEINS CONTAINING 50 OR MORE AMINO ACID RESIDUES.

Proteins, Eukaryotic	No. of Proteins In Class	Average No. of Residues	Arg	Lys	Residues Percent Arg	Lys
Trypsinogen, Bovine	1	229	2	15	0.9	6.6
Lactalbumin	3	123	1.3	11.7	1.1	9.5
Ferredoxin - plants	5	96.6	1.2	4.6	1.2	4.8
Leghemoglobin - plants	1	140	2	14	1.4	10.0
Antitrypsin - human	1	422	7	32	1.7	7.6
Chymotrypsinogen	2	245	4.5	12.5	1.8	5.1
β-Lactoglobulin, Bovine AB	1	162	3	15	1.9	9.3
Myoglobin	8	152.9	2.9	19.6	1.9	12.8
Hemoglobin Alpha Chain	10	141.1	3.0	11.6	2.1	8.2
Globin, chironomus	1	136	3	10	2.2	7.4
Cytochrome c	40	107	2.4	15.6	2.2	14.6
Hemoglobin Beta Chain	10	145.2	3.3	11.3	2.3	7.8
Immunoglobulins, light chains:						
Constant, Kappa, human & mouse	2	106	2.5	7.5	2.4	7.1
Constant, Lambda, human & mouse	2	105	1	7	1.0	6.7
Variable Kappa, human & mouse	18	107.8	4.9	3.7	4.6	3.4
Variable, Lambda, human & mouse	10	109	4.0	3.2	3.6	2.9
Immunoglobulins, heavy chains:						
Constant, human	1	336	7	27	2.1	8.0
Constant, rabbit	1	326	13	19	3.9	5.7
Variable, human	6	109.7	5.8	4.2	5.3	3.8
Protease Inhibitor - Lima Bean	1	84	2	4	2.4	4.8
Caseins	2	204	5	12.5	2.5	6.1
Haptoglobin Alpha Chain	1	84	2	8	2.6	10.1
Carbonic anhydrase, human	1	260	7	18	2.7	6.9
Hemerythrin, worm	1	113	3	11	2.7	9.7
Trypsin inhibitor, soybean	1	71	2	5	2.8	7.0
Globin, lamprey	2	146	4.5	13	3.1	8.9
Glyceraldehyde 3-PO_4 Dehydrogenase	2	332.5	9.5	27	2.9	8.1
Thyrotropin α and luteinizing hormone α	3	96	3	10	3.1	10.4
Alcohol Dehydrogenase horse	1	374	12	30	3.2	8.0
Muscle triose PO_4 isomerase	1	248	8	20	3.2	8.1
Chorionic gonadotropin, human α	1	92	3	6	3.3	6.5
Adrenodoxin, bovine	1	118	4	5	3.4	4.2
Carboxypeptidase, bovine	1	307	11	15	3.6	4.9
Prophospholipase A_2, pig	1	130	5	9	3.8	6.9
Neurotoxin, scorpion	2	63.5	2.5	5.5	3.9	8.7

Proteins, Eukaryotic	No. of Proteins In Class	Average No. of Residues	Average Arg	Average Lys	Residues Percent Arg	Residues Percent Lys
Ribonuclease, pancreatic	3	125	5	9.6	4.0	7.7
Cytochrome B_5	6	87.2	3.6	6.6	4.1	7.8
Keratin, High-sulfur protein	2	153.5	6.5	0	4.2	0.0
Trypsin inhibitor, pancreatic	2	56	2.5	3.5	4.5	6.2
Trypsin inhibitor, ascaris	1	66	3	7	4.5	10.6
Bee venom phospholipase A	1	128	6	10	4.7	7.8
Proinsulin	3	83.6	4	2	4.8	2.4
Elastase, porcine	1	240	11	2	5.0	1.2
Lipotropin β and γ	2	90.5	4.5	10	5.0	11.0
Prolactin	2	194	10.5	9	5.6	4.5
Papain	1	212	12	10	5.7	4.7
Mouse nerve growth factor	1	120	7	8	5.8	6.7
Glutamate Dehydrogenase	1	500	30	32	6.0	6.4
Parathyroid Hormone	1	84	5	9	6.0	10.7
Neurotoxins, snake venom	11	61.2	4.2	5.3	6.2	8.4
Avidin, chicken	1	128	8	9	6.3	7.0
Histone II B2	1	125	8	20	6.4	16.0
Growth Hormone	2	189.5	12	10.5	6.4	5.6
Monkey amyloid protein A	1	76	5	4	6.6	5.3
β chorionic gonadotropin, human	1	139	11	4	7.9	2.9
Thyrotropin β and luteinizing hormone β	2	116.5	10	6.5	8.6	5.6
Lysozyme - vertebrates	5	129	11.6	6.2	9.0	4.8
Trypsin inhibitor, basic, bovine	1	58	6	4	10.3	6.9
Myelin Membrane Encephalitogenic protein	1	170	18.5	12.5	10.9	7.4
Trypsin Inhibitor - Maize	1	65	8	1	12.3	1.5
Histone III, calf thymus	1	135	18	13	13.3	9.6
Histone IV	2	102	14.5	10.5	13.7	10.8
TOTALS		9,656	402.7	668.7	4.17	6.92
Proteins, Prokaryotic						
Rubredoxin	2	52.5	0	3	0.0	5.7
Cytochrome c_2	1	112	0	17	0.0	15.2
Cytochrome c_{551}	3	82	0.3	8.3	0.3	10.2
Cytochrome c_3	1	109	0.5	18.5	0.6	18.5
50S Ribosomal protein A_2, *E. coli*	1	120	1	12	0.8	10.0
Neocarzinostatin *Streptomyces*	1	109	1	0	0.9	0.0
Thioredoxin	1	108	1	10	0.9	9.3
Ribonuclease T_1	1	104	1	1	1.0	1.0

light chain, and a C region about three times as long, representing a tripling of the ancestral protein. I suggested in 1969 that most point mutations in the specificity (S)(variable) regions of immunoglobulins are advantageous, and are rapidly incorporated as evolutionary changes, because "it is immunologically advantageous to have a large available assortment of different antibodies to cope with various antigenic determinants" while "those in the constant (C) regions are usually deleterious, thus accounting for the variability of S and the constancy of C sequences," (Jukes, 1969). If functional constraints are removed from a protein, the greater abundance of arginine codons should lead it to evolve in the direction of higher arginine and lower lysine than of one whose composition is constrained by functional reasons. From the above considerations, the S regions should be higher in arginine and lower in lysine than the C regions as a result of randomly-occurring point mutations, accepted by the S regions and rejected by the C regions. Table 1 shows that there is such a difference between the S and C regions.

Proteins, Prokaryotic	No. of Proteins In Class	Average No. of Residues	Arg	Lys	Residues Percent Arg	Lys
Ferredoxin - clostridial type	5	54.8	0.2	0.8	1.0	1.3
Subtilisin	2	274.5	3	10	1.1	3.6
Azurin	4	125.8	1.2	12.8	1.2	9.3
Acyl carrier protein	1	77	1	4	1.3	5.2
Coat protein - turnip yellow mosaic virus	1	188	3	7	1.6	3.7
Penicillinase Staph. aureus	1	257	4	43	1.6	16.7
Ferredoxin, Chromatium	1	81	2	2	2.5	2.5
Thermolysin, Bacillus thermoproteolyticus	1	316	10	11	3.2	3.5
Nuclease, staphylococcal	1	149	5	23	3.4	15.4
Cytochrome B_{562}	1	110	4	16	3.6	14.5
Tryptophan Synthetase A	1	267	11	13	4.1	4.9
Aspartate transcarbamylase R Chain	1	152	8	10	5.3	6.6
Penicillinase, B. licheniformis	1	265	15	24	5.7	9.0
Coat Protein - Tobacco Mosaic Virus	5	157.4	9.6	1.6	6.1	1.3
Alpha lytic protease, myxobact.	1	198	12	2	6.1	1.0
Lysozyme, bacteriophage	2	160.5	12.5	12.5	7.9	7.9
TOTALS		3,630	106.3	262.5	2.93	7.23

The prokaryotic proteins in Table 1 average even lower in arginine than those of the eukaryotes. Perhaps this is because many eukaryotes, in contrast to prokaryotes, can dispose of arginine by converting it to excretory products such as urea and creatinine, and therefore eukaryotes are somewhat "less intolerant" of arginine than are the prokaryotes.

ACKNOWLEDGMENT

Support by NASA grant NGR 05-003-460 to the University of California, Berkeley, is acknowledged.

REFERENCES

Delmonte, L. and Jukes, T. H. (1962) Pharmacol. Reviews, 14:91.
Hill, R. L., Delaney, R., Fellows, R. E., Jr. and Lebovitz, H. E. (1966) Proc. Natl. Acad. Sci. USA, 56:1762.
Jukes, T. H. (1969) Biochemical Genetics, 3:109.
King, J. L. and Jukes, T. H. (1969) Science, 164:788.
Subak-Sharpe, H., Burk, R. R., Crawford, L. V., Morrison, J. M., Hay, J. and Keir, H. M. (1966) Cold Spr.Harb. Symp.Quant Biol., 31:737.