ON THE FREQUENCY OF ARGININE IN PROTEINS AND ITS IMPLICATIONS FOR MOLECULAR EVOLUTION

Michael Wallis

School of Biological Sciences, University of Sussex, Falmer,
Brighton, BN1 9QG, Sussex, U.K.

Received December 3,1973

SUMMARY

Evidence is presented against the concept that arginine appeared later in the evolution of life that the other common amino acids, as an 'evolutionary intruder'. Alternative explanations for the relatively low frequency of arginine in proteins are considered, based on the proposition that there has been selection for such low frequency because of special properties of the amino acid.

Arginine occurs less frequently in proteins than might be expected on the basis of the number of codons available to it in the genetic code (1, 2). Jukes (2) has proposed the interesting hypothesis that this discrepancy is because arginine is a relatively new amino acid, which replaced another, probably ornithine, during the course of evolution. He suggests that the replacement occurred because arginine happened to have greater affinity for the ornithine t-RNA-aminoacyl-ligase system than did ornithine itself. However, several arguments can be opposed to such a theory. Here I summarize the evidence against this 'intruder hypothesis' and propose alternative explanations for the relatively infrequent occurrence of arginine.

Evidence Against the Intruder Hypothesis

The 'evolutionary intruder' hypothesis (2) proposes that arginine was absent from the earliest organisms, and that the codons now serving it originally specified ornithine. Subsequently, arginine appeared as the product (or intermediate) of a metabolic pathway, chanced to have a higher affinity for t-RNA^{orn} (or the corresponding activating enzymes) than ornithine itself, and therefore replaced ornithine in proteins. The introduction of arginine had some disadvantages, which were offset by selection against its occurrence in proteins, and its partial replacement by lysine. The result was a relative scarcity of arginine and a relative abundance of lysine - as seen in contemporary proteins.

An underlying assumption of this hypothesis appears to be that the initial function of arginine was not its role in protein synthesis, and that the advantages gained from its availability for this original function outweighed disadvantages which were encountered as it replaced ornithine in proteins. Of the known roles of arginine, outside its role in protein synthesis, the urea cycle is most obviously important, but this is largely confined to (some) eukaryotes; arginine is of course found in all known prokaryotes, suggesting that the urea cycle did not originate before the occurrence of arginine as a constituent of proteins. In prokaryotes there appears to be no generally important role for arginine outside its role in protein synthesis. The absence of such a role is an obstacle to the acceptance of the 'intruder hypothesis'.

Even if such a role for arginine could be found, it seems unlikely that it would have led to the replacement of ornithine by arginine in proteins. A more likely result would perhaps have been the coevolution of increased specificity of ornithine selection in protein synthesis, via increased specificity and discrimination of activating enzymes (or their equivalent in the earliest organisms). The six arginine codons are served by more than one type of t-RNA, and it would be remarkable if arginine had succeeded in capturing all six from ornithine.

The 'intruder hypothesis' is supported by the fact that the arginine:lysine ratio in the invariable region of immunoglobulins is lower than that of the variable region (2). Maybe the need for

Table 1. Rates of molecular evolution and arginine:lysine ratios for some proteins.

Protein	Arg/Lys* ratio	Rate of Evolution**
Pancreatic ribonuclease	0.52	33
Immunoglobulins		
kappa chain, constant region	0.33	39
kappa chain, variable region	1.32	33
Lactalbumin	0.11	25
Haemoglobin (α & β chains)	0.28	14
Myoglobin	0.15	13
Pancreatic trypsin inhibitor	0.71	11
Lysozyme (animal)	1.87	10
Myelin membrane encephalitogenic		
protein	1.46	7
Trypsinogen	0.13	5
Cytochrome c	0.15	3
Glyceraldehyde 3-PO_4 dehydrogenase	0.36	2
Histone IV	1.38	0.06

* derived from the data of Jukes (2).

** derived from the data of Dayhoff (ref 3 - Table 6-1). The rate of evolution is expressed as accepted point mutations/10^8 yr/100 residues.

Peptides containing less than 50 residues have not been included.

There is no apparent correlation between Arg/Lys ratio and rate of evolution (correlation coefficient, $r = - 0.019$; $p > 0.7$).

stability in the invariable portion tended to lead to exclusion of arginine, with its supposed structural disadvantages, while the requirement for innovation in the variable portion has been accompanied by less rigorous evolutionary conservatism. Such an argument should be extendable to other proteins. Those proteins with a low rate of evolution (because selection is conserving the structure) should, by this argument, have a lower arginine:lysine ratio than those with a high rate of evolution. The arginine:lysine ratios of those proteins for which rates of evolution have been calculated (3) are shown in Table 1. There appears to be no correlation between rate of evolution and lysine:arginine ratio.

Alternative Explanations for the Relative Scarcity of Arginine

The observation that the frequency of occurrence of an amino acid in proteins is generally correlated with the number of codons for that amino acid has been used to support the hypothesis that most species variation of proteins is due to incorporation of neutral mutations (1). This hypothesis has been contested (e.g. refs. 4,5). The argument will not be continued here, but a relevant aspect is that most residues in most proteins do appear to be strongly conserved - either they show no changes during evolution or they show only conservative changes. The amino acid compositions of proteins must also be largely the result of natural selection, and not necessarily the result of a 'random drift' towards a composition reflecting the nature of the genetic code.

Nevertheless, whatever the basic forces controlling observable protein evolution, the composition of proteins does correspond to frequencies of codons in the genetic code, and the divergence of arginine from this general rule needs explanation. If we assume that during the earliest stages of the evolution of life there was a correspondence between frequency of amino acids and frequency of codons, the divergence of arginine must be due to the subsequent operation of selection. At least 3 possible bases for such selection can be proposed, as alternatives to selection resulting from the replacement of ornithine by arginine.

1) Arginine as a factor in controlling the degradation of proteins.

Little is known about the mechanisms by which protein turnover is regulated, but proteolytic degradation is certainly involved. Arginine is frequently a point of attack for proteases with specific physiological functions (e.g. activation of fibrinogen by thrombin, conversion of proinsulin to insulin) and it is possible that physiological turnover of proteins is regulated to some extent by their arginine contents. This could have led to a selection pressure against the occurrence of arginine as degradation mechanisms evolved.

2) Possible role of arginine codons in control of protein synthesis.

The fact that an amino acid is served by several different codons provides interesting possibilities for the control of protein synthesis at the translational level (6-8). It is possible that one or more of the codons available for arginine is involved in control of translation, functioning as a 'modulating triplet' (6). It is interesting that prevailing evidence (from RNA bacteriophage mRNA sequences and from haemoglobin mutants - ref. 3) provides many instances of the use of arginine codons of the group CGX (where X is A, G, C or U), but no definitive evidence for the use of AGG or AGA. Further, in *Escherischia coli*, tRNA recognizing AGA and AGG represents only a tiny fraction (about 2%) of the total unfractionated $tRNA^{Arg}$ (8).

3) Arginine and protein structure.

A unique aspect of the arginine side-chain is its very high pKa; of all the common amino acids it is probably the only one whose charged nature can never be suppressed under physiological conditions. It may be the most difficult amino acid to fit into the interior of a protein. If the evolution of proteins involved an increase in size and in subunit interactions, and therefore an increase in the ratio of 'internal' to 'external' residues, this could have resulted in a selection pressure against arginine.

These three possibilities for selection against arginine are not mutually exclusive - all three might have been operating, and other possibilities undoubtedly exist. Selection against arginine may have resulted in selection for lysine (leading to the observed high frequencies of this amino acid in proteins). In addition, there is at least one feature which may have led to direct selection for lysine. Lysine is frequently found in modified form (e.g. ϵ-N-acetylated, ϵ-N-methylated) in proteins. Mechanisms for enzymatic modification of proteins may have evolved after the evolution of the genetic code, and have resulted in a demand for a higher frequency of lysine than was originally provided for by its two codons.

REFERENCES

1. King, J.L. and Jukes, T.H. (1969) Science, 164, 788-798.
2. Jukes, T.H. (1973) Biochem. Biophys. Res. Commun. 53, 709-714.
3. Dayhoff, M.O. (1972) Atlas of Protein Sequence and Structure, 1972, (Vol. 5) National Biomedical Research Foundation, Washington, D.C.
4. Clarke, B. (1970) Science, 168, 1009-1011.
5. Richmond, R.C. (1970) Nature, 225, 1025-1028.
6. Ames, B.N. and Hartman, P.E. (1963) Cold Spring Harbour Symp. Quant. Biol. 28, 349-356.
7. Boyer, S.H. (1970) In 'Modern Trends in Human Genetics', Vol. 1 pp. 1-48 (Ed. Emery, A.E.H.) Butterworths, London.
8. Anderson, W.F. (1969) Proc. Natn. Acad. Sci. (U.S.), 62, 566-573.