BioSystems, 15 (1982) 105—109
Elsevier/North-Holland Scientific Publishers Ltd.

OBLIGATORY AMINO ACIDS IN PRIMITIVE PROTEINS

A.S. KOLASKAR and V. RAMABRAHMAM

School of Life Sciences, University of Hyderabad, Hyderabad 500 134, India

(Received June 16th, 1981)
(Revision received January 7th, 1982)

Conformational similarity among amino acid residues, a property derived by analysing (ϕ,ψ)-probability distributions of 20 proteinous amino acids from 38 different globular proteins, is used to arrive at a set of six 'obligatory' amino acids of primitive proteins. The amino acids *Ser*, *Val*, *Leu*, *Asp*, *Gly* and *Pro* have been argued to be 'obligatory' and to represent, conformationally, the remaining amino acids. The reasons for consideration of these six residues as 'obligatory' are discussed. Methods to check the validity of our proposition are suggested.

Genes of present day living systems can code for a maximum of 20 types of amino acids. In other words, polypeptides and proteins synthesised using the ribosomal machinery can, during synthesis, have at most 20 types of amino acids, termed 'proteinous' amino acids, though more than 250 types of amino acids occur in natural polypeptides (Mooz, 1976). Thus a natural and fundamental question that arises is why 20 amino acids only were selected as coded amino acids and whether all these 20 amino acids were present in proteins of primitive biosystems. The first part of the question, namely, the selection of 20 amino acids as coded amino acids, has received some attention recently and reasons for their existence in proteins have been suggested. (Röhlfing and Saunders, 1978; Weber and Miller, 1981). However, the latter part of the question; the types of amino acids present in primitive proteins is yet to be raised. In this communication we suggest that out of 20 proteinous amino acids only Ser, Val, Leu, Asp, Pro and Gly were obligatory and were present in proteins of primitive biosystems. The study of the composition, structure and function of such 'primitive proteins' is the subject of this communication.

In order to get some idea about the composition of these primitive proteins, one should find out the traits that might have remained almost invariant during evolution. One such trait seems to be the topology of domains of proteins, as only a restricted number of topologies of domains have been observed when crystal structure data of a large number of different globular proteins were analysed (Ptitsyn and Finkelstein, 1980). Domains are those regions of proteins which form very compact globules by having many internal contacts but a few contacts with other parts of the chain (Schulz and Schirmer, 1979). In general a single polypeptide chain protein can consist of two to three domains, each of about 70—80 amino acids (Rashin, 1981). The restricted number of observed topologies of domains might have beeen due to the availability of a restricted number of amino acids for synthesis of primitive proteins, which we assume consist of single domains, in contrast to proteins of more evolved systems which have multiple domains. If primitive proteins consist of a few types of amino acids and single domains in nature, then the topologies of these primitive proteins will be limited in number due to physical forces as argued by Ptitsyn and Finkelstein (1980). The conformational property of amino acids which determines the topology of domains and three dimensional structure

of proteins, can be used to find out the 'obligatory' amino acids in primitive proteins. Hence, we have developed a method, which is briefly discussed below, to find out the conformational similarity among proteinous amino acids. Using this property of amino acids a set of a minimum number of amino acids is derived which can represent the remaining proteinous amino acids conformationally.

Method

The extent of conformational similarity among amino acid residues can be studied by comparing the probability distribution of conformational states of each of the 20 proteinous amino acids with each other. The conformational states of amino acid residues in polypeptide chains can be characterised by two main chain dihedral angles ϕ and ψ. Hence, normalised (ϕ, ψ)-probability distribution maps were obtained for 20 proteinous amino acid residues using crystal structure data of 38 different globular proteins (Kolaskar and Ramabrahmam, 1981). The grid-wise comparison of the probability map of each residue, thus obtained, with the remaining 19 residues will indicate the extent of conformational similarity of these residues with the residue under consideration. A list of amino acid residues which are conformationally similar is given in Table 1. It can be seen from Table 1 that Ser, Glu and Lys are conformationally similar to Ala. In other words, when the (ϕ,ψ)-distributions of all 19 residues were compared with that of Ala only these three residues have similar (ϕ,ψ)-distribution as that of Ala. It should be noted here that since the reference residue during comparison varies, one will get a situation such as a residue A being conformationally similar to B, but B not being conformationally similar to A. We can cite the example of Ile being conformationally similar to Leu, but Leu not similar to Ile, but similar to some other residues such as Glu, Gly, etc., this, one can observe by looking at the right column of Table 1.

TABLE 1

The residues whose normalised (ϕ,ψ)-probability distributions are similar to those in the (left) column

Residue	Conformationally similar residues
Ala	Ser Glu Lys
Arg	Val Ala Asp Gln His Ile Leu Lys Phe Ser Thr
Asn	Asp Ser
Asp	Lys
Cys	Ser Phe Thr Val
Glu	Ala Leu
Gln	Leu Thr
Gly	—
His	Arg Ala Asp Cys Glu Gln Leu
Ile	Val
Leu	Ile Glu Gln Lys Phe Ser Thr Val
Lys	Asp
Met	Val Ala Arg Cys Gln Ile Leu Lys Phe Thr Tyr
Phe	Leu Gln Ile Lys Ser Thr Val
Pro	—
Ser	Ala Cys Leu Lys Thr
Thr	Ser Gln Ile Leu Phe Tyr Val
Trp	Ile Cys Phe Thr Val
Tyr	Val Thr
Val	Ile

— indicates that the (ϕ,ψ)-distribution in (ϕ,ψ)-map of the residue in the left column is very much distinct from those of the remaining 19 residues.

Analysis of Table 1 reveals that at least six amino acids are necessary to represent all 20 proteinous amino acids conformationally. The possible sets of six amino acids which can represent all 20 amino acids are given in Table 2. The derivation of this Table is discussed briefly below by taking the example of set III.

TABLE 2

Minimum number of amino acids which can represent conformationally all 20 proteinous amino acids

Set	Amino acids					
I	Ala	Val	Gln	Asp	Pro	Gly
II	Ala	Ile	Thr	Asp	Pro	Gly
III	Ser	Val	Leu	Asp	Pro	Gly

Set III we suggest as obligatory amino acids in proteins.

Ser occurs in the right column of Table 1 for Ala, Arg, Asn, Cys, Phe and Thr, indicating that the (ϕ,ψ)-probability distribution of Ser is similar to the (ϕ,ψ)-distribution of above mentioned residues. Thus it can represent any one of these residues with minimal conformational discrepancy. Similarly, Val represents Arg, Cys, Ile, Met, Phe, Thr, Trp and Tyr while Leu occurs in the right column of Table 1 for Arg, Gln, His, Met, Phe, Ser and Thr. Asp represents conformationally Asn, Lys, His and Arg. The remaining two of this set of six amino acids are Pro and Gly which have unique (ϕ,ψ)-probability distributions.

It can be seen that some residues of this set III represent more than one of remaining proteinous amino acids. For example, Arg can be represented by any one of Ser, Val and Leu. Sets I and II of Table 2 are derived in a similar fashion. Three amino acid residues, Gly, Pro and Asp are common to these three sets and only the first three amino acid residues are different. Therefore, we have looked at the possibility of synthesis of these amino acids in the laboratory. Though the synthesis of Ala, Val or Ala, Ile is not difficult, the synthesis of Gln or Thr is definitely not easy. However, this is not the case with Ser, Val or Leu. This prompts us to suggest that *Ser*, *Val*, *Leu*, *Asp*, *Gly* and *Pro* were obligatory amino acids in proteins by the start of biotic evolution.

Discussion

The set which we have proposed consists of a single, acidic amino acid, Asp, in place of the two, Asp and Glu, which occur in proteins. The presence of Asp rather than Glu is not surprising because even the chemical nature of Asp is more simple as compared to Glu which contains two $-CH_2$ groups in β and γ positions. The presence of an acidic amino acid is essential so as to prevent the protein or polypeptide from interacting with hydrated electrons (Scott, 1981). Thus, the necessity for acidic amino acids was enormous at the dawn of biotic evolution but one can not visualise such an important role for basic amino acids. Therefore, the absence of basic amino acids from the proposed set does not seem to be surprising.

In the proposed set there are three amino acids which are either polar or neutral-polar while the remaining three are hydrophobic in nature. Thus, these hydrophobic amino acids must have formed the interior of the globule of the protein while polar and neutral-polar amino acids might have acted to shield the protein from the effect of hydrated electrons which were present because of the aqueous environment. It is interesting to note that 50% of these obligatory amino acids, Val, Leu and Pro are hydrophobic and roughly the same percentage of hydrophobic amino acids is present in the present set of proteinous amino acids. The proposed set does not contain any aromatic or sulphur containing amino acids. Although aromatic and sulphur containing amino acids have a very important and crucial role in proteins and enzymes of evolved biosystems, their function seems to be specific. Similarly, the biosynthesis of aromatic amino acids suggests that they have been incorporated at a later stage as suggested by Wong (1981).

The proposed set consists of amino acids which prefer α-helix, β-sheet and chain reversals. For example Leu is an α-helix preferer, Val prefers β-sheet and Ser, Asp, Gly and Pro are known to occur frequently in chain reversals (Levitt, 1978; Kolaskar et al., 1980). All proteins which have enzymatic activity are known to have α-helix as one of the secondary structures. Thus, the presence of Leu seems to be essential.

Ser is known to be not only a β-bend preferer but also a site of the attachment of polysaccharide chains. Pro seems to be present in primitive proteins mainly because of its characteristic side chain property which gives rigidity to main chain conforma-

tion of the polypeptide chain. Gly not only plays an exactly opposite role to that of Pro, namely, providing flexibility to main chain conformation but also takes conformations which other amino acids can not take because their side chains are in L-configuration. In other words it avoids the need for the presence of amino acids in D-configuration. This has been very clearly shown from the crystal structure studies on insulin (Hodgkin, D., private communication). Thus, each of the amino acids suggested in the above set might have played an important role in primitive proteins which are assumed to be multifunctional and single domain in nature.

The experiments in which primordial conditions were simulated to study synthesis of amino acids have shown that less than half of the 20 proteinous amino acids were produced in more than trace amounts. But it is worth noting that the amino acids Ser, Val, Leu, Asp, Gly and Pro (Set III) are synthesized in non-negligible quantities (Dose, 1976; Wong, 1981). We have avoided comparing our set of obligatory amino acids with amino acids present in extra-terrestrial matter, or lunar dust, since there are no evidences for existence of life there.

It is further interesting to note that though we have arrived at the proposed set by taking into consideration only conformational properties of the amino acid residues, all but one (Asp) of the amino acids of Set III have four or more codons in the present genetic code and thus the presence of third letter for these amino acids seems to be immaterial. In other words, the doublet code for Gly, Val, Leu, Ser and Pro will be GG, GU, CU, UC and CC, respectively, where only the first two letters from the triple letter code are considered. In the case of Asp there are two codons in the present genetic code and the doublet code for Asp can be considered as GA. Thus, it appears quite plausible that the contemporary triplet code has evolved from a doublet code (Jukes, 1973).

One may check the validity of our arguments in the following fashion:

(i) Recently an algorithm has appeared which can be used to find out domains in the proteins (Rashin, 1981). Consider one such domain, say that of T4 lysozyme. In this domain, which consists of 74 residues, represent all amino acids by the proposed six amino acids using the property of conformational similarity. The sequence of this domain is given in Fig. 1. The topology of the polypeptide chain, the sequence of which is given in Fig. 1(b) will not be very much different from that of the observed topology of the first domain of T4 lysozyme. (Even if it is different one can find out the topology of this polypeptide chain and compare it with the topologies of the domain known from crystal structures of globular protein). The knowledge of topology will give an idea about the possible functions of the computer-simulated polypeptide chain which consists of only six types of amino acids, and thus about the range of choice of various substrates. The measured K_a-values for various substrates will indicate the multifunctional nature of this polypeptide chain and also its role in the primitive biosystems. The measurements of other physical constants will indicate the thermodynamical stability of this globule.

(ii) The same experiment of measurement of various properties of the polypeptide chain containing the proposed six amino acid residues can be carried out as follows. One can repeat in the laboratory Fox's type of experiment (1965) taking only the proposed six amino acids instead of 18 amino acids and then carry out the characterization of the synthesized polymers.

Thus, in short, we have discussed in this communication, that during evolution the topology of the domains remained nearly invariant and the restriction on these topo-

(1a)

Met-Asn-Ile-Phe-Glu-Met-Leu-Arg-Ile-Asp-
Glu-Gly-Leu-Arg-Leu-Lys-Ile-Tyr-Lys-Asp-
Thr-Glu-Gly-Tyr-Tyr-Thr-Ile-Gly-Ile-Gly-
His-Leu-Leu-Thr-Lys-Ser-Pro-Ser-Leu-Asn-
Ala-Ala-Lys-Ser-Glu-Leu-Asp-Lys-Ala-Ile-
Gly-Arg-Asn-Cys-Asn-Gly-Val-Ile-Thr-Lys-
Asp-Glu-Ala-Glu-Lys-Leu-Phe-Asn-Gln-Asp-
Val-Asp-Ala-Ala

(1b)

[Val/Leu]-[Asp/Ser]-[Val]-[Leu/Ser/Val]-[Leu]-[Val/Leu]-Leu-[Val/Asp/Leu/Ser]-[Val]-Asp-
[Leu]-Gly-Leu-[Val/Asp/Leu/Ser]-Leu-[Asp]-[Val]-[Val]-Asp-Asp-
[Ser/Leu/Val]-[Leu]-Gly-[Val]-[Val]-[Ser/Leu/Val]-[Val]-Gly-[Val]-Gly-
[Asp/Leu]-Leu-Leu-[Ser/Leu/Val]-[Asp]-Ser-Pro-Ser-Leu-[Asp/Ser]-
[Ser]-[Ser]-[Asp]-Ser-[Leu]-Leu-Asp-[Asp]-[Ser]-[Val]-
Gly-[Val/Asp/Leu/Ser]-[Asp/Ser]-[Ser/Val]-[Asp/Ser]-Gly-Val-[Val]-[Ser/Leu/Val]-[Asp]-
Asp-[Leu]-[Ser]-[Leu]-[Asp]-Leu-[Leu/Ser/Val]-[Asp/Ser]-[Leu]-Asp-
Val-Asp-[Ser]-[Ser]

Fig. 1. (a) Sequence of first domain of T4 lysozyme (three letter amino acid code has been used). This domain as pointed out by Rashin (1981) contains 74 amino acid residues. (b) Sequence of the above mentioned domain of T4 lysozyme in terms of the proposed six obligatory amino acids. The residues which are replaced from T4 lysozyme sequence by conformationally similar residues are put in blocks. Note that at many places more than one conformationally similar amino acid can replace the amino acid from T4 lysozyme sequence. It is expected that these sequences will have a topology which is similar to that of the domain of T4 lysozyme although the sequence and amino acid composition is very much different.

logies is because of the presence of few types of amino acids in primitive proteins, which are single domains in nature. Using only main chain conformational similarities among amino acids, we have arrived at a set of six amino acids and have argued that these are obligatory amino acids at the dawn of evolution of proteins.

References

Dose, K., 1976, Protein structure and function, J.L. Fox, Z. Deyl and A. Blazej (eds.) (Marcel Dekker, Inc., New York and Basel) pp. 149—184.

Fox, S.W., 1965, A theory of macromolecular and cellular origins, Nature 205, 328—340.

Jukes, T.H., 1973, Possibilities for the evolution of genetic code from a preceding form. Nature 246, 22—26.

Kolaskar, A.S., V. Ramabrahmam and K.V. Soman, 1980, Reversals of polypeptide chain in globular proteins. Int. J. Pept. Protein Res. 16, 1—11.

Kolaskar, A.S. and V. Ramabrahmam, 1981, Conformational similarity among amino acid residues — I Analysis of protein crystal structure data. Int. J. Biolog. Macromol. 3, 171—178.

Levitt, M., 1978, Conformational preferences of amino acids in globular proteins. Biochemistry (Wash.) 17, 4277—4285.

Mooz, E.D., 1976, Data on the naturally occurring amino acids, in: Handbook of Biochemistry and Molecular Biology, Proteins. G.D. Fasman (ed.), Vol. 1 (Chemical Rubber Co. Press, Cleveland) pp. 111.

Ptitsyn, O.B. and A.V. Finkelstein, 1980, Similarities of protein topologies: evolutionary divergence, functional convergence or principles of folding? Q. Rev. Biophys. 13, 339—386.

Rashin, A.A., 1981, Location of domains in globular proteins. Nature 291, 85—87.

Röhlfing, D.L. and M.A. Saunders, 1978, Evolutionary processes possibly limiting the kinds of amino acids in proteins to twenty: a review. J. Theor. Biol. 71, 487—503.

Schulz, G.E. and R.H. Schirmer, 1979, Principles of Protein Structure (Springer Verlag, New York, Heidelberg, Berlin).

Scott, J., 1981, Natural Selection in the primodial soup. New Sci. 89, 153—155.

Weber, A.L. and S.L. Miller, 1981, Reasons for the occurrence of the twenty coded protein amino acids. J. Mol. Evol. 17, 273—284.

Wong, J.T., 1981, Coevolution of genetic code and amino acid biosynthesis. Trends Biochem. Sci. 6, 33—36.