POSSIBLE BASE SEQUENCES IN THE AMINO ACID CODE

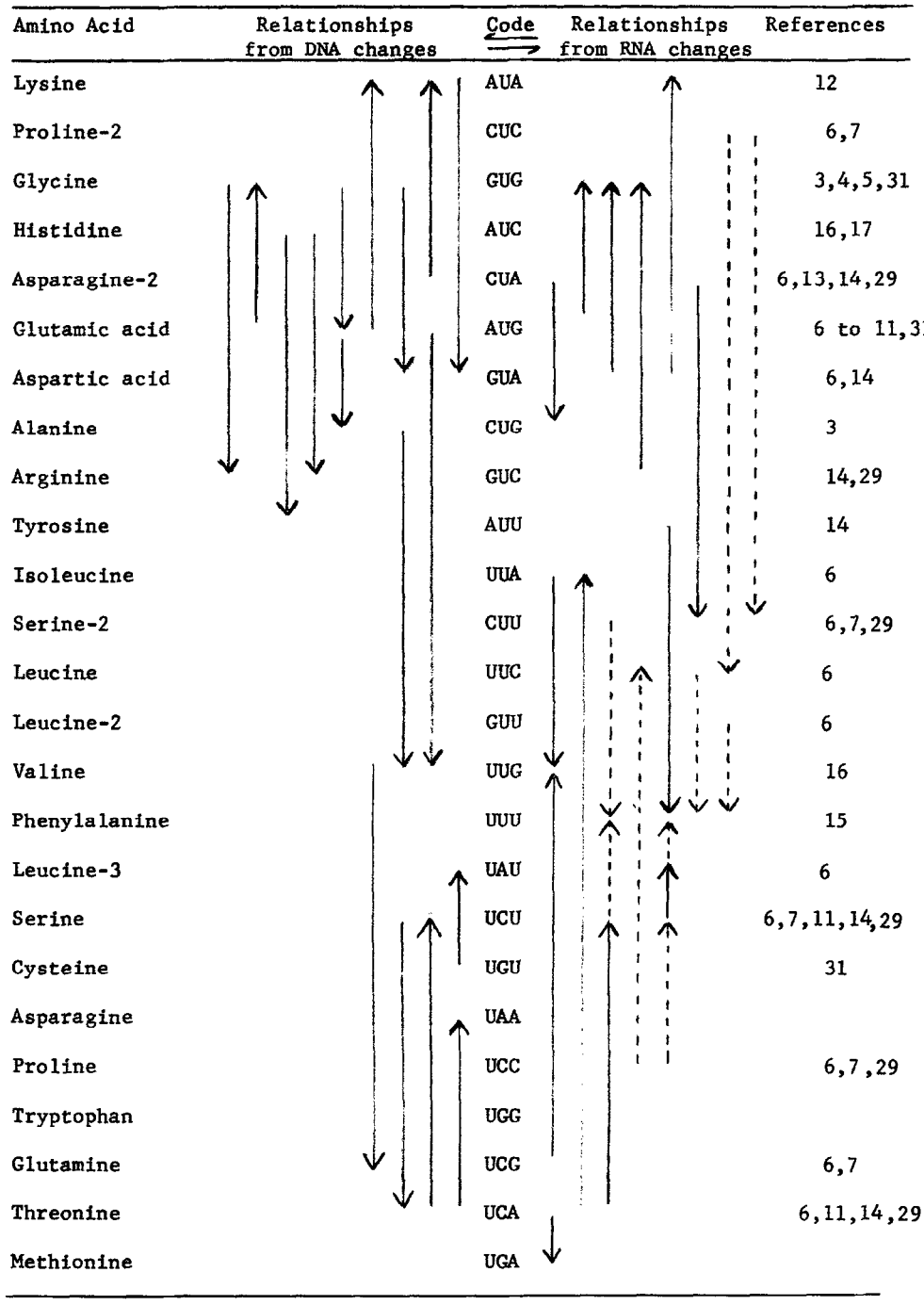
Thomas H. Jukes

American Cyanamid Company, Agricultural Division, Princeton, New Jersey

Received April 23, 1962

In a previous communication (1) the base sequences in messenger RNA codes (2) for 15 amino acids were postulated as a result of comparing various mutations which lead to amino acid replacements. Further consideration of the available information on such mutations enables the base sequences to be listed for 22 of the 23 proposed codes (2) with alternate possibilities for the other one, and sequences are proposed also for two additional alternate (degenerate) codes.

The postulation depends on the observation (2) that certain mutations involving a change in a single amino acid at an established locus in a polypeptide chain can be translated to the change of a single base in the triplet code in messenger RNA for the amino acid which changes. This deduction, extended to other mutations, leads to the assembly of an interlocking series of base sequences derived from thirty-four single-amino-acid mutations that have been described as occurring in various proteins. These are shown in the table. The series, when arranged vertically as shown in the table, can be placed in any one of six possible horizontal sequences, reading the triplets from left to right; for example, if the alanine code is written as CUG, GUC, CGU, UGC, GCU, UCG, the corresponding and related sequences for glutamic acid become respectively AUG, GUA, AGU, UGA, GAU, and UAG. The table lists the base sequences in messenger-RNA codes for the amino acids; resulting from the base triplets proposed by Speyer and co-workers (2), from information on single-amino-acid mutations, and from the assumption that CUG is the code for alanine. The arrows refer to mutations described in the references cited in the table; thus, glycine ⟶



aspartic acid is reported in reference 4. Broken arrows refer to two possibilities for the same change. Changes in RNA all refer to studies with tobacco mosaic virus (TMV) protein. The terms asparagine-2, leucine-2 and

leucine-3 refer to the alternate codes described by Speyer et al. (2) for these two amino acids. Serine-2 and proline-2 are explained below.

The postulation that CUG is the code for alanine was based on the concept (1), unsupported by experimental examination of the peptides involved, of a simultaneous change in two sequential bases on the DNA strand at a locus involved in messenger production for synthesis of beta-lactoglobulin in cow's milk. A recent publication (18) indicates that this supposition must be withdrawn because the val/ala difference is located on a different peptide from the gly/asp difference. In this case, the vertical relationship between the letters representing the bases in each of the three columns formed by the triplet sequences in the table remains undisturbed; however, the three vertical columns formed by the letters are now subject to placement in any of the series of six possible juxtapositions.

In the series of triplet codes described by Speyer and co-workers (2), there are nine which each contain three different bases against a possibility of eighteen such combinations. It is interesting that these nine appear in the table in the following repetitive pattern;

CUA	GUA	CUG
AUC	AUG	GUC
UCA	UGA	UCG

This pattern, in which all three-different-letter triplets containing a middle U are included, appears only if the two possibilities depending on CUG and GUC for alanine are selected from the six juxtapositions for the series. An examination of the table reveals its content as being (A) All possible triplets with U in the middle, (B) All other possible triplets containing two U's, (C) All other triplets which contain U followed by two identical bases, (D) All triplets consisting of UC or UG followed by a purine letter. This pattern leads to the assignment of UGG for tryptophan and UCC and CUC for proline by selection from the alternate choices for each of these codes. The proposed alternate code of CUU for serine-2 would occur undetectably as a mixture with UCU in the experiments of Speyer et al.(2), and it explains the asp(aspNH_2) to ser mutation (6,29). The pair UCC/CUC

both may be proline in the absence (2) of assignments other than proline for 1U - 2C. It is not possible at present to choose between GUU and UGU for cysteine; the arrow from serine to leucine-2 indicates a possibility that leucine-2 may be UGU. However, this mutation may alternatively be UCU to UAU. Assuming that all triplets containing a middle U should be included, the correct horizontal sequence for the code should therefore be that listed in the table or its mirror image. If this assumption is disregarded, the other four juxtapositions are additional possibilities.

The observations summarized in the table lead to the conclusion, which is implicit in the review by Crick, (19), that biological evolution through mutation may proceed from a series of single-base substitutions in the chain of the genetic DNA molecule. These will lead to changes in proteins when the substitutions are in regions where messenger RNA molecules are formed that direct protein synthesis, so that the substitutions will produce changes in the triplet base codes for the individual amino acids. Some of the substitutions are transitional changes which produce A to G, G to A, C to U and U to C changes in messenger RNA. These changes may reflect adenine to hypoxanthine to guanine, and cytosine to uracil to thymine changes in DNA, if one can analogize from similar mutational changes that have been produced by nitrous acid (20). The other base changes are transversions between purine and pyrimidine bases (21). Changes that produce a decrease in messenger U are possibly less common than other changes. The spontaneous aspNH_2 to arg change reported by Wittmann (6) and the hemoglobin mutation glu to gluNH_2 (30) do not fit the table.

It is not necessary to expect a stereochemical relationship between the coding triplet and the amino acid. Transfer-RNA molecules may originally have been a random series containing all possible coding triplets in combination with all possible amino-acid-recognition sites. Then evolution, in a trend from bases to DNA to RNA to protein, could lead to the discarding of those transfer-RNA molecules which did not provide

the correct combination of coding triplet and specific amino acid for uniting with messenger RNA to produce a biologically useful protein. On this basis the alternate codes may be primitive rather than "degenerate".

Interspecific changes in proteins have also been described that are identical with certain mutational changes in the table. These include ser to leu-2 or leu-3 in corticotropin (22), glu to lys and thr to ser in ribonuclease (23), ileu to val in insulin and hypertensin (24,25), and thr to ileu in insulin (26). In addition, ala to gly in silk fibroin has been described (27). Two other changes in insulin, thr to ala and ser to gly (24), are in conflict with the table.

The vertical sequences in the table differ from any of those listed by Smith (28) solely as a result of his adopting the glu to $gluNH_2$ sequence. This adoption brings about conflict with the $aspNH_2$ to ala and glu to ala sequences that were followed in constructing the table (1).

References

(1) Jukes, T.H., Biochem. & Biophys. Res. Comm., 7...(1962)
(2) Speyer, J.F., Lengyel, P., Basilio, C. and Ochoa, S., Proc. Nat. Acad. Sci., 48,441 (1962)
(3) Gordon, W.G., Basch, J.J. and Kalan, E.B., J. Biol. Chem. 236, 2908 (1961)
(4) Baglioni, C., J. Biol. Chem., 237,69 (1962)
(5) Helinski, D.R. and Yanofsky, C., Proc. Nat. Acad. Sci., 48,173(1962)
(6) Wittmann, H.G., Naturwissenschaft., 48,729 (1961)
(7) Tsugita, A., Protein, Nucleic Acid, Enzyme (Tokyo) 6,385 (1961), cited by Speyer et al. (2)
(8) Hunt, J.A. and Ingram, V.M., Nature, 184,640 (1959)
(9) Hill, R.C. and Schwartz, H.C., Nature, 184,641 (1959)
(10) Hunt, J.A. and Ingram, V.M., Nature, 184,870 (1959)
(11) Ingram, V.M. and Stretton, A.O.W., Nature, 190,1079 (1961)
(12) Murayama, M., Fed. Proc., 19,78 (1960)
(13) Baglioni, C., and Ingram, V.M., Biochim. and Biophys. Acta, 48,253 (1961)
(14) Tsugita, A., cited in (2)
(15) Nirenberg, M.W. and Matthaei, J.H., Proc. Nat. Acad. Sci., 47,1588 (1961)
(16) Gerald, P.A. and Efron, M.L., Proc. Nat. Acad. Sci., 47,1758 (1961)
(17) Muller, C.J. and Kingma, S., Biochim. and Biophys. Acta, 50,595 (1961)
(18) Kalan, E.B., Gordon, W.G., Basch, J.J. and Townend, R., Arch. Biochem. & Biophys., 96,376 (1962)
(19) Crick, F.H.C., Growth in Living Systems, Basic Books, Inc., N.Y., (1961) p. 3
(20) Vielmutter, W. and Schuster, H., Biochem. & Biophys. Res. Comm., 2,324 (1960)

(21) Freese, E. B., Proc. Nat. Acad. Sci., 47,540 (1961)
(22) Li, C. H. and Dixon, J.S., Science, 124,934 (1956)
(23) Anfinsen, C.B., Aqvist, S.E., Cooke, J.P. and Jonsson, B., J. Biol. Chem., 234,1118 (1959)
(24) Harris, J. I., Sanger, F. and Naughton, M. A., Arch. Biochem. and Biophys., 62,427 (1956)
(25) Peart, W. S., Biochem. J., 62,520 (1956); Skeggs, L. T., Jr., Kahn, J. R. and Shumway, N. P., J. Exp. Med., 103,301 (1956)
(26) Ishihara, Y., Saito, T., Ho, Y. and Fujiro, M. Nature, 181,1468 (1958)
(27) Kay, L. M., Schroeder, W. A., Munger, N. and Burt, N. J. Am. Chem. Soc., 78,2430 (1956)
(28) Smith, E. L., Proc. Nat. Acad. Sci., 48...(1962) (in the press)
(29) Tsugita, A. and Fraenkel-Conrat, H., J. Mol. Biol. 4,73 (1962)
(30) Swenson, R. T., Hill, R. L., Lehmann, H. and Jim, R. T. S. J. Bio. Chem. 237..(1962) (in the press)
(31) Yanofsky, C., Presented at Federation Meetings, Atlantic City, N. J. April 14, 1962.