The Occurrence in Amino Acid Sequences of Extensive Informational Symmetries Based on Possible Codon–Codon Complementarity in the Encoding Polynucleotides

By GIDEON M. POLYA and D. R. PHILLIPS

Department of Biochemistry, La Trobe University, Bundoora, Vic. 3083, Australia

(Received 26 August 1975)

1. A procedure is described for the detection and assessment of informational complementarity in an amino acid sequence; it is based on possible autocomplementarity in the mRNA, and involves codon-to-codon matching. 2. This procedure was applied to myelin basic protein, a variety of protamines, histone IV, silk fibroin, rat skin collagen $\alpha 1$ chain and a sheep keratin. A multiplicity of extensive low-probability informational symmetries, based on codon-to-codon matching, were detected. 3. These low-probability orderings, which are independent of the actual mRNA codons, are rationalized in terms of the evolutionary ordering of the amino acid sequences concerned, in such a way that constraints on the secondary structure of the coding polynucleotides were satisfied. This possible interpretation is supported by a number of significant common properties of the protein sequences analysed.

The non-randomness of amino acid sequences can be assessed from the point of view of protein chemical function (Nishikawa & Ooi, 1974; Wu *et al.*, 1974; Levitt & Warshel, 1975) or by comparative analyses, which can be used to reconstruct possible evolutionary events (Urbain, 1969; Dayhoff, 1972; Wu *et al.*, 1974). A further way of assessing the significance of an amino acid sequence is to examine the extent to which the sequence may have been determined by requirements for extensive secondary structure in the coding polynucleotides.

The sequencing of mRNA molecules has revealed extensive base-paired secondary structures, which may have significant biological functions (Adams *et al.*, 1969; Sanger, 1971; Min Jou *et al.*, 1972; Proudfoot & Brownlee, 1974; Fiers *et al.*, 1975). A functional requirement for such RNA loops could impose a major restriction on the possible amino acids in a functional protein (Ball, 1972). Analyses of the MS2-virus coat-protein RNA sequences have suggested that such constraints have been involved in this case (Ball, 1973*a,b*). However, this type of assessment is complicated by the fact that the genetic code is such that a high degree of secondary structure can be obtained in random nucleotide sequences (Gralla & De Lisi, 1974; Fitch, 1974).

The triplet nucleotide codons in the opposite arms of an RNA loop can be either exactly matched (codon-to-codon matching) or can be mismatched in two ways. Analyses involving consideration of all possible codon phasings have been applied to detect amino acid sequences that can be theoretically coded for by looped RNA sequences (White *et al.*, 1972;

Riley, 1973; Laux *et al.*, 1973; Fitch, 1974). However, no assessments of the improbability of the 'successful' amino acid orderings were made in these studies, and further, this approach is not very useful for the prediction of actual mRNA sequences (Mark & Petruska, 1972; Fitch, 1972, 1974; Wu *et al.*, 1974). Nevertheless, these types of analyses can provide an alternative assessment of the non-randomness of an amino acid sequence, as in the case of the low-probability informational symmetry, based on codon-to-codon matching, found for somatostatin (Polya, 1975). Although no evidence in favour of the occurrence of codon–codon matching symmetry ('anticomplementary repetition') has been found in immunoglobulins and ferredoxins (Urbain, 1969), this type of symmetry can be found in other proteins. The present paper describes the occurrence and significance of extensive informational symmetries, based on codon-to-codon matching, in some major structural and chromosomal proteins.

Experimental

Detection of codon-to-codon-matching symmetries

Informational complementarities in amino acid sequences, assuming codon-to-codon matching and ignoring G·U base-pairing, were determined by using the matrix-display scheme of Polya (1975). The amino acid complements considered in this analysis are listed in Table 1. A Fortran IV computer program was designed to handle large amino acid sequences, and computations were carried out by using a PDP-10

Table 1. *Amino acid complements considered in the codon-to-codon-matching analysis*

All pairs of 'complementary amino acids' are listed, for which a possible codon for one amino acid could be completely complementary (in an antiparallel and exact codon-to-codon-matching sense) to a possible codon for the complementary amino acid. The single-letter code used for the amino acids (Dayhoff, 1972) is as follows: A, alanine; C, cysteine; D, aspartic acid; E, glutamic acid; F, phenylalanine; G, glycine; H, histidine; I, isoleucine; K, lysine; L, leucine; M, methionine; N, asparagine; P, proline; Q, glutamine; R, arginine; S, serine; T, threonine; V, valine; W, tryptophan; Y, tyrosine.

F-K	S-R
F-E	S-G
L-Q	P-R
L-K	P-G
L-E	P-W
	T-S
I-N	T-G
I-D	T-C
I-Y	T-R
M-H	A-S
V-N	A-G
V-D	A-C
V-Y	A-R
V-H	

computer. The matrices were examined for extensive and contiguous complementary sequences, which could then be arranged as codon-codon-matched loops (informational-complementarity loops), as previously described (Polya, 1975).

Application of this restrictive analysis can be used to determine those amino acid sequences that are ordered in such a way that it is possible to construct theoretical coding mRNA sequences, which are looped in a fashion involving extensive codon-to-codon matching. The restrictions greatly simplify the analysis to a simple matrix-display procedure of the kind used by Tinoco *et al.* (1971) to determine the minimum-energy secondary structure of RNA sequences. These restrictions also minimize the number of statistically probable orderings, which result from more general matching analyses, which consider all possible codon phasings. However, as described below, a codon-to-codon matching arrangement has, in addition, a peculiar significance, in comparison with other codon-phasing possibilities, in the context of the evolution of looped mRNA sequences involving maximal retention of the general chemical character of the amino acid sequences thus encoded.

The triplet nucleotide codons in the base-paired opposite arms of an RNA loop can be either exactly matched (phase 2) or mismatched in two ways (phases 1 and 3). (We are here defining phases 1, 2 and 3 as matching arrangements, in which the first, second

and third nucleotides respectively of opposite codons can be base-paired.) One can assess the significance of these three codon-matching arrangements in the context of gene evolution directed towards optimizing RNA autocomplementarity, with the retention of the functional amino acid sequence. Since the maximum variability in codons for amino acids with a multiplicity of codons is located in the 3'-terminal nucleotide position of codons, matching of these positions (phase 3) is least advantageous, and use of phase 1 or phase 2 matching arrangements offers the best strategy for retention of an amino acid sequence with optimized RNA looping. However, the middle base of a triplet codon is the base most responsible for the general character (e.g. hydrophobicity) of the corresponding amino acid. Accordingly, the use of phase 2 matching would yield the best conservation of the general character of the amino acids of the sequence in this evolutionary scheme (North, 1972). This conservation of the general character of amino acids gains added significance if one considers the importance of internal (more non-polar) rather than surface (more polar) amino acids in the determination of homologous tertiary structures by non-homologous amino acid sequences in the two halves of a protein such as elastase (Shotton & Watson, 1970) or chymotrypsin (Nishikawa & Ooi, 1974).

Assessment of the significance of symmetries

The codon-codon-matched loops (informational-complementarity loops) were assessed statistically, by determining the probability (P) that a random amino acid sequence, of the same length as the sequence concerned, would have an equal or greater degree of symmetry (c.f. Brezinski, 1975). For an odd-numbered amino acid sequence of length $(2n+1)$, matched head to tail, this probability can be determined from the relationship:

$$P = \sum_{i=m}^{n-1} \binom{n-1}{i} \left(\frac{3}{20}\right)^{i+1} \left(\frac{17}{20}\right)^{n-1-i}$$

where P is the probability of obtaining a $(1, 2n+1)$ match and at least m out of the remaining $(n-1)$ pairs matched, assuming an average of three complementary amino acids per amino acid (Polya, 1975). For even-numbered sequences of length $(2n+2)$ around the diagonal of the matrix, one can for simplicity ignore the two central codons in the loop (since we have a requirement for unpaired nucleotides in the head of RNA loops), and apply the same probability estimation indicated above. Values of P for a large number of values of n and $(m+1)$ were computed by using a Fortran IV program.

The probabilities discussed above relate to the 'concentration' of informational complementarity within a discrete amino acid sequence, and give the

probability of finding a sequence of a certain size, with at least a specified degree of matching at a particular point in a much larger overall amino acid sequence. If P is the probability of finding a $(2n+1)$ or $(2n+2)$ loop with at least $(m+1)$ out of n possible matches at a particular point in an overall sequence of N amino acids in a discrete polypeptide, then the probability (P') of finding an arrangement of this size and with at least this degree of matching anywhere in the overall sequence is given by:

$$P' = 1 - (1 - P)^{N-(2n+1)} \text{ (odd case)}$$

or

$$P' = 1 - (1 - P)^{N-(2n+2)} \text{ (even case)}$$

If P is low, then P' approximates to $P(N-2n-1)$ (odd case) or $P(N-2n-2)$ (even case).

Note that in our analysis we have restricted ourselves to consideration of contiguous amino acid sequences that are informationally complementary in a codon-to-codon-matching sense. Such contiguous sequences could be theoretically coded for by an RNA sequence determining a hairpin-loop second-

ary structure. Non-contiguous sequences that are informationally related as described above were also found. Such arrangements could be theoretically coded for by RNA sequences determining, for example, extensive complementarity between sequences at the 5'- and 3'-ends of the mRNA. For simplicity we have not considered such non-contiguous 'anti-complementary repetitions' (Urbain, 1969) in our analysis.

Results

The most significant informational-complementarity loop derived from human myelin basic protein is shown in Fig. 1(a). The N -terminal half of this sequence is of considerable biological significance, since it is evolutionarily conservative and contains the main determinant of the encephalitogenic potency of the protein, as well as residues that can be glycosylated, methylated or phosphorylated *in vivo* (Dunkley & Carnegie, 1974). Conversely, the C -terminal half of this sequence has diminished functional significance, since the homologous rat S (small) myelin basic

Fig. 1. Informational-complementarity loops for sequences from myelin basic protein and histone IV

The data are taken from Dayhoff (1972). The letters represent amino acids (see Table 1), and the complementarities are represented by vertical lines. (a) Loop derived from the sequence threonine-98 to serine-132 of human myelin basic protein. For this sequence $P = 2.4 \times 10^{-5}$ and $P' = 0.003$. (b) Loop derived from histone IV. This sequence in pea histone IV differs from the bovine and pig sequences in having isoleucine instead of valine in position 60, as indicated (Dayhoff, 1972). The probability of obtaining at least this degree of complementarity in a sequence of this size anywhere in a random 102-amino acid sequence is approx. 0.0003.

Fig. 3. Informational-complementarity loops for clupeines YI, YII and Z

(a)–(d), Clupeine YI; (e) and (f), clupeine YII; (g) and (h), clupeine Z. The values for P and P' (P' values in parentheses) are: (a) 2.0×10^{-5} (0.0001); (b) 2.1×10^{-4} (0.002); (c) 2.0×10^{-5} (0.0001); (d) 4.0×10^{-4} (0.002); (e) 2.1×10^{-4} (0.001); (f) 1.8×10^{-4} (0.001); (g) 1.1×10^{-5} (0.0003); (h) 4.0×10^{-4} (0.002).

two such sequences of arginine-complementary amino acids is such that a multiplicity of extensive informational-complementarity loops can be constructed, as shown in Fig. 2. A multiplicity of very-low-probability

orderings can also be derived from the amino acid sequences of clupeines YI, YII and Z (Fig. 3). The values of P and P' indicate very-low-probability orderings of the amino acids in these protamines, and

Fig. 4. Informational-complementarity loop for the major repeating sequence of silk fibroin

the significance of these estimates is enhanced if one considers that arginine residues (which of course cannot be matched with each other) represent two-thirds of the overall amino acid composition. The amino acids are so ordered that near-maximal matching of complementary amino acids is possible. Thus in the arrangement in Fig. 2(d), for example, the only matching possibility that is unsatisfied is between tyrosine-17 and the valines 11, 29 and 30. If one considers all the matchings shown in Fig. 2(d), the probability of obtaining at least this degree of complementarity in a random amino acid sequence of the same length is 0.0005.

The most significant informational-complementarity loop found in this study is that for the major repeating sequence in silk fibroin. A major peptide sequence in silk fibroin, which comprises 60% of the protein, has the sequence:

where n is usually 2 and has a mean value of 2 (Lucas & Rudall, 1968). Riley (1973) showed that a codon-to-codon-matching arrangement could provide maximal theoretical autocomplementarity for the silk-fibroin mRNA. Riley (1973) attempted to predict the silk-fibroin mRNA sequence, with limited success in relation to the mRNA sequence data of Suzuki & Brown (1972). As discussed above, this type of analysis is not useful for the prediction of mRNA sequences, but it can be used to quantify the non-randomness of an amino acid sequence in an informational sense. The repeating sequence Ser-Gly-Ala-Gly-Ala-Gly could be completely complementary to itself in a codon-to-codon-matching sense. Since the sequence is repeated, one can then arrange the whole of the sequence under study in the form of a huge

informational-complementarity loop (Fig. 4). The probability of obtaining this degree of complementarity in a random amino sequence of the same length is approx. 2×10^{-23} , this very low value being derived from the extensive and faithful repetition of the sequence Ser-Gly-Ala-Gly-Ala-Gly.

Since the codon-to-codon matching analysis applies so dramatically to silk fibroin, it was decided to apply our analysis to other major structural proteins that involve a homogeneous secondary structure and contain repeated amino acid sequences. Collagen is a major structural protein with an extensive triple-helix structure (Rich & Crick, 1961). Inspection of the *N*-terminal sequence of the rat skin collagen $\alpha 1$ chain (Dayhoff, 1972) reveals that, commencing with glycine-13, every third amino acid is a glycine. The positions between the glycine residues are occupied by a high proportion of glycine complements, these being mainly proline (hydroxyproline) and alanine. This arrangement determines a high degree of informational complementarity between collagen amino acid sequences. The informational-complementarity matrix derived from the *N*-terminal 137-amino acid sequence of rat skin collagen $\alpha 1$ chain demonstrates that a very large number of sequence elements are extensively complementary to each other in terms of our analysis (Fig. 5). The values of P for the matching sequences range down to 4×10^{-8} . A remarkable feature of this collagen seen from the matrix is that the *N*-terminal sequence of five amino acids (Gly-Tyr-Asp-Glu-Lys) cannot be extensively matched with any of the possible 5-amino acid sequences that can be tested out of the remaining 133-amino acid sequence (Fig. 5b). This is largely because the initial *N*-terminal sequence is devoid of glycine complements. Other similar regular discontinuities are also visible in the matrix (Fig. 5b).

Fig. 5. Informational complementarity between collagen amino acid sequences

(a) The codon-to-codon matching matrix for the *N*-terminal 137-amino acid sequence of rat skin $\alpha 1$ collagen (Dayhoff, 1972). The amino acids are numbered from the *N*-terminal. Each point represents the possible codon complementarity of the amino acids concerned; non-complementarities are not registered. (b) Complementarity of amino acid sequences in the rat skin collagen sequence analysed in (a). Lines parallel to the (1,137)-(137,1) diagonal have been drawn through points in the matrix shown in (a). Each line is bounded by two points and includes at least one other point, permitting no more than one space (unmatched amino acid) between points (matched amino acids).

Fig. 6. Informational-complementarity loops for the sheep keratin polypeptide SCMK-B2C

The sequence data are from Dayhoff (1972). The values of P and P' (P' values in parentheses) are: (a) 3.8×10^{-5} (0.001); (b) 9.3×10^{-5} (0.001); (c) 1.1×10^{-5} (0.001); (d) 1.1×10^{-5} (0.001).

The absence of significant complementarity from these regions contrasts dramatically with the extensive complementarities elsewhere in the sequence (Fig. 5).

Sheep keratins were examined as obvious examples of structural proteins with sequences determining an extensive α -helical secondary structure (Lundgren & Ward, 1963). Analysis of the sheep keratin polypeptide SCMK-B2C revealed a multiplicity of low-probability informational-complementarity loops, as shown in Fig. 6. For each of these loops, the probability (P') of obtaining a loop of this size with at least the observed degree of complementarity anywhere in a random sequence of 151 amino acids is approx. 0.001. Note that the L-Q complementarity at the head of the loop in Fig. 6(d) was ignored in this calculation. The most significant symmetries found are derived from the region bounded by serine-22 and glycine-63. This region includes a striking element of repeating sequences occupying positions 30–55 of the polypeptide: Gln-Thr-Ser-Cys-Cys-Gln-Pro-Thr-Ser-Ile-Gln-Thr-Ser-Cys-Cys-Gln-Pro-Thr-Cys-Leu-Gln-Thr-Ser-Gly-Cys-Glu.

Discussion

We have found a multiplicity of low-probability orderings of amino acids in naturally occurring

sequences, based on possible autocomplementarity in the mRNA involving codon-to-codon matching. It must be noted that the orderings demonstrated and the statistical estimates given are independent of the actual codons in the mRNA coding for the amino acid sequences concerned. We believe that the surprising symmetries demonstrated in this paper demand some rationalization.

Before attempting to provide a rationalization of these phenomena it is useful to consider whether the 'successful' sequences described in this paper have any common properties that may be pertinent to the problem. Some major related properties are listed below.

(1) These proteins are major structural proteins with no ascribable catalytic functions. We have been unable to find comparable low-probability amino acid orderings (in terms of P' values) in a number of ligand-binding or catalytic proteins examined.

(2) The general character of the amino acid sequences concerned are peculiar in that the fibroin, keratin and collagen sequences determine homogeneous secondary structures. The other proteins considered are very basic proteins.

(3) In all cases the sequences that can be arranged in informational-complementarity loops involve repeated amino acid sequences. These repeated sequences are obvious for the protamines, fibroin, collagen

and keratin (Dayhoff, 1972, 1973) or have been otherwise detected, as for histone IV (Bauer, 1971) and myelin basic protein (Bauer, 1972).

(4) The sequences examined are stable from an evolutionary point of view, the best example being the histone IV sequences (Dayhoff, 1972). The same observation applies to myelin basic protein (Bauer, 1972; Dunkley & Carnegie, 1974). The protamine, fibroin, collagen and keratin sequences are also examples of evolutionary conservatism in the sense that they involve repeated sequences and a high degree of fidelity in such repeated sequences.

A possible explanation for the phenomena described in this paper is now presented. As discussed above, a codon-to-codon matching strategy yields the best conservation of the general character of an amino acid sequence, in the context of an evolutionary scheme directed towards maximizing coding polynucleotide autocomplementarity (North, 1972). The low-probability orderings described in the present paper may reflect an evolution of the sequences concerned, which has been constrained by both protein and polynucleotide secondary-structure requirements in this extremely conservative fashion.

This explanation is consistent with the non-catalytic (i.e. functionally simple) role and the evolutionary conservatism of these proteins. The cases of fibroin, keratin and collagen, in which the sequences determine homogeneous secondary-structure elements, are consistent with involvement of this evolutionary scheme, which will conserve the general character of amino acid sequences. The success of very basic non-catalytic proteins in this analysis is also consistent with this scheme. In this connexion the discontinuity at the *N*-terminals of the collagen sequence (Fig. 5) can be explained in terms of the hydrogen-bonding of nucleotides at the 5'-end of the mRNA to nucleotides in an untranslated RNA sequence (cf. Min Jou *et al.*, 1972; Brawerman, 1974), although other such discontinuities cannot be obviously rationalized. The marked differences in functional significance of the two halves of the myelin basic protein loop (Fig. 1a) can also be rationalized in terms of our evolutionary scheme.

One can imagine DNA being organized into single-stranded loops, as in the arrangements proposed by Gierer (1966). Alternatively, if genes are organized as continuous duplexes, such single-stranded arrangements could nevertheless arise in autocomplementary regions of DNA during replication. The occurrence of autocomplementarity could have an influence on the positions of possible excision of DNA from a gene, and single-stranded products of such excisions would be rendered more stable if they were extensively autocomplementary. The evolution of clupeine Z by excision and joining of DNA from the clupeine YI and clupeine YII genes (Dayhoff, 1972), and the generation of rat S myelin basic protein from its larger

evolutionary precursor (Dunkley & Carnegie, 1974), can be rationalized in terms of DNA loops similar to those implied by Figs. 3 and 1(a) respectively. Excision and relocation of such DNA loops also provides an evolutionary mechanism for generation of repeated amino acid sequences. In this sense the presence of repeated sequences within the informationally complementary sequences studied here and the peculiar location of the loop in myelin basic protein are consistent with our suggested interpretation of these low-probability orderings.

Our suggested explanation involves an extremely conservative evolutionary mechanism, which can nevertheless be radical in the sense of generating repeated amino acid sequences. However, it is not clear as to whether there is a need for such a restrictive mechanism for conservation of functionally indispensable sequences. Further, if there is a need for autocomplementarity in mRNA, extensive (but not specifically located) base-pairing is easily achieved, without restricting the nature of the amino acids thus encoded. Nevertheless, if our rationalization is incorrect, the multiplicity of low-probability orderings demonstrated in this paper (which are independent of the actual codons in the mRNA) must be explained in terms of some, as yet unknown, relationship between the genetic code and functionally permissible amino acid sequences in these proteins.

This study was supported by grants to G. M. P. and D. R. P. from the Australian Research Grants Committee. We are grateful to Miss Lefi Petrou and Miss Gitta Petersen for technical assistance, and to Dr. M. J. Denton, Dr. C. J. Thompson and Dr. P. Carnegie for suggestions and advice.

References

- Adams, J. M., Jeppesen, P. G. N., Sanger, F. & Barrell, B. G. (1969) *Nature (London)* **223**, 1009-1014
- Ball, L. A. (1972) *J. Theor. Biol.* **36**, 313-320
- Ball, L. A. (1973a) *J. Theor. Biol.* **41**, 243-247
- Ball, L. A. (1973b) *Nature (London)* **242**, 44-45
- Bauer, K. (1971) *Int. J. Protein Res.* **3**, 165-172
- Bauer, K. (1972) *Biochem. J.* **126**, 1245-1248
- Brawerman, G. (1974) *Annu. Rev. Biochem.* **43**, 621-642
- Brezinski, D. P. (1975) *Nature (London)* **253**, 128-130
- Dayhoff, M. O. (ed.) (1972) *Atlas of Protein Sequence and Structure*, vol. 5, National Biomedical Research Foundation, Washington
- Dayhoff, M. O. (ed.) (1973) *Atlas of Protein Sequence and Structure*, vol. 5, Suppl. 1, National Biomedical Research Foundation, Washington
- Dunkley, P. R. & Carnegie, P. R. (1974) *Biochem. J.* **141**, 243-255
- Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Merregaert, J., Min Jou, W., Raeymakers, A., Volckaert, G., Ysebaert, M., Van de Kerckhove, J., Nolf, F. & Van Montague, M. (1975) *Nature (London)* **256**, 273-278

- Fitch, W. M. (1972) *J. Mol. Evol.* **1**, 185-207
- Fitch, W. M. (1974) *J. Mol. Evol.* **3**, 279-291
- Gierer, A. (1966) *Nature (London)* **212**, 1480-1481
- Gralla, J. & De Lisi, C. (1974) *Nature (London)* **248**, 330-332
- Laux, B. E., Dennis, D. & White, H. B., III (1973) *Biochem. Biophys. Res. Commun.* **54**, 894-898
- Levitt, M. & Warshel, A. (1975) *Nature (London)* **253**, 694-698
- Lucas, F. & Rudall, K. M. (1968) in *Comprehensive Biochemistry* (Florkin, M. & Stotz, E. H., eds.), vol. 26B, p. 475, Elsevier, Amsterdam
- Lundgren, H. P. & Ward, W. H. (1963) in *Ultrastructure of Protein Fibers* (Borasky, R., ed.), pp. 39-122, Academic Press, New York
- Mark, A. J. & Petruska, J. A. (1972) *J. Mol. Biol.* **72**, 609-617
- Min Jou, W., Haegeman, G., Ysebaert, M. & Fiers, W. (1972) *Nature (London)* **237**, 82-88
- Nishikawa, K. & Ooi, T. (1974) *J. Theor. Biol.* **43**, 351-374
- North, A. C. T. (1972) *Nature (London)* **239**, 76-77
- Polya, G. M. (1975) *Nature (London)* **255**, 652-653
- Proudfoot, N. J. & Brownlee, G. G. (1974) *Nature (London)* **252**, 359-362
- Rich, A. & Crick, F. H. C. (1961) *J. Mol. Biol.* **3**, 483-506
- Riley, W. T. (1973) *J. Theor. Biol.* **40**, 285-300
- Sanger, F. (1971) *Biochem. J.* **124**, 833-843
- Shotton, D. M. & Watson, H. C. (1970) *Nature (London)* **235**, 811-816
- Suzuki, Y. & Brown, D. D. (1972) *J. Mol. Biol.* **63**, 409-429
- Tinoco, I., Uhlenbeck, O. C. & Levine, M. D. (1971) *Nature (London)* **230**, 362-367
- Urbain, J. (1969) *Biochem. Genet.* **3**, 249-269
- White, H. B., Laux, B. E. & Dennis, D. (1972) *Science* **175**, 1264-1266
- Wu, T. T., Fitch, W. M. & Margoliash, E. (1974) *Annu. Rev. Biochem.* **43**, 539-566