ACTA CHEMICA SCANDINAVICA 17 (1963) S 250 – S 256

A Comparison of the Amino Acid Sequences of the Cytochromes *c* of Several Vertebrates

E. MARGOLIASH, S. B. NEEDLEMAN and J. W. STEWART

Biochemical Research Department, Abbott Laboratories, North Chicago, Illinois, USA

The amino acid sequences of four vertebrate cytochromes *c* (horse, human, pig and chicken) show a remarkable degree of similarity. All are 104 residues long, are acylated at the amino-terminal residue, the two cysteinyl residues binding the heme are in positions 14 and 17, and two long sequences (residues 16 to 43 and 63 to 82) are identical in all four proteins. Comparing each cytochrome *c* in turn with the others, 28 amino acid exchanges are obtained. These occur at only 15 positions which are grouped into three areas along the chain. The variations do not disrupt the groupings of basic and hydrophobic residues, even though such residues represent a large proportion of the entire protein. Some positions are particularly prone to change. At residue 89 four different amino acids are observed. Residues in certain positions appear to be characteristic of particular species and known phylogenetic relations are in conformity with the observed variations. The relations of the amino acid sequence of cytochrome *c* to the possible secondary and tertiary structures of the protein are discussed. The consistence of the observed amino acid exchanges with known triplet nucleotide sequences in ribonucleic acid, coding for particular amino acids, is examined.

Complete amino acid sequences of the cytochromes *c* prepared from horse[1–4], man[5], chicken[6] and pig[6] have been determined. The purpose of the present article is to compare these structures and to examine the implications of the observed amino acid replacements.

Table 1 gives the amino acid sequence of horse heart cytochrome *c*, while Table 2 lists the differences between the amino acid sequences of the horse protein and the corresponding sequences of the other proteins examined.

COMMON FEATURES IN THE STRUCTURE OF VERTEBRATE CYTOCHROMES *c*

The most striking single fact revealed by an examination of Tables 1 and 2 is the remarkable degree of similarity of the amino acid sequences of the four cytochromes *c* listed, even though the species examined derive from phylogenetic origins relatively as divergent as chicken and man. This fundamental homology

Table 1. Amino acid sequence of horse heart cytochrome c[1].

Acetyl – Gly – Asp – Val* – Glu – ***Lys*** – Gly – ***Lys*** – ***Lys*** – **Ileu*** – **Phe** – Val* – $GluNH_2$ – ***Lys*** – CyS* – Ala – $GluNH_2$ – CyS*
(10; HEME)

– ***His*** – Thr* – Val* – [Glu – ***Lys*** – Gly – Gly – ***Lys*** – ***His*** – ***Lys***] – Thr* – Gly – Pro* – [$AspNH_2$ – **Leu** – ***His*** – Gly –
(20; 30)

Leu – **Phe** – Gly – ***Arg*** – ***Lys***] – Thr* – Gly – $GluNH_2$ – **Ala** – **Pro*** – Gly – **Phe** – **Thr*** – **Tyr** – Thr* – [Asp – Ala – $AspNH_2$
(40; 50)

– ***Lys*** – $AspNH_2$ – ***Lys*** – Gly] – **Ileu*** – Thr* – **Try** – ***Lys*** – Glu – Glu – Thr* – [**Leu** – **Met** – Glu – **Tyr** – **Leu** – Glu –
(60)

$AspNH_2$] – Pro* – ***Lys*** – ***Lys*** – **Tyr** – **Ileu*** – Pro* – Gly – Thr* – ***Lys*** – **Met** – **Ileu*** – **Phe** – Ala – Gly – **Ileu*** – ***Lys*** – ***Lys*** –
(70; 80)

Lys – Thr* – Glu – ***Arg*** – Glu – Asp – **Leu** – **Ileu*** – [Ala – **Tyr** – **Leu** – ***Lys*** – ***Lys*** – Ala] – Thr* – $AspNH_2$ – GluCOOH
(90; 100; 104)

Sequences of six or more helix forming residues are **boxed**. Non-helix forming residues are marked **with an** asterisk. Basic residues are in *italics* and hydrophobic residues **in bold type.**

Table 2. Differences in the amino acid sequences of horse[1], pig[6], human[5] and chicken[6] cytochromes *c*. A blank space indicates that the amino acid is identical to the one found in the horse protein.

	Residue position in peptide chain														
Species	3	11	12	15	44	46	47	50	58	60	62	83	89	92	104
Horse	Val	Val	$GluNH_2$	Ala	Pro	Phe	Thr	Asp	Thr	Lys	Glu	Ala	Thr	Glu	Glu
Pig							Ser			Gly			Gly		
Human		Ileu	Met	Ser		Tyr	Ser	Ala	Ileu	Gly	Asp	Val	Glu	Ala	
Chicken	Ileu			Ser	Glu		Ser			Gly	Asp		Ser	Val	Ser

presumably stems from the genetic derivation of these proteins from a common primordial cytochrome *c*, amino acid variations being limited to those consistent with structures able to carry out the necessary specific electron transfer and energy conserving functions.

All four vertebrate cytochromes *c* under consideration are single chains 104 residues long, have an amino-terminal residue acylated at the α-amino group and the heme attached through its two vinyl side chains by thio-ether linkages to cysteinyl residues in positions 14 and 17. The structural similarities around the attachment of the prosthetic group to the peptide chain, first observed by Tuppy and collaborators[7,8] for a number of different cytochromes of the *c* group, have been fully borne out by the amino acid sequences discussed here. These similarities consist of two cysteinyl residues, separated by two other residues, preceded by a basic amino acid, lysine or arginine, and followed by the sequence His-Thr. This pattern has, however, also been found in cytochromes functionally entirely different from vertebrate cytochrome *c*, such as the protein from *Rhodospirillum rubrum*[9] and more recently in the cytochromoid of *Chromatium*[10]. These features therefore most probably represent requirements for the attachment of the heme moiety and the provision of at least part of its immediate environment, irrespective of the specificity of the overall protein for particular enzyme systems.

The 28 amino acid substitutions observed by comparing, in turn, each of the cytochromes *c* with the other three, occur at only 15 positions along the peptide chain, leaving two long sequences, those from residue 16 to 43 and residue 63 to 82, completely unchanged. It is probable that the first of these sequences, which contains the three histidyl residues in the molecule as well as the unreactive lysyl residue[11], is part of the "crevice"[12] structure which enfolds the prosthetic group and may contain one[13] or both[14,15] of the hemochrome-forming groups. The lack of variation in this area could thus be ascribed to specific amino acid sequence requirements for the formation of a "crevice" which is functionally efficient.

The second large constant part of the sequence (residues 63 to 82) contains one of the major groupings of basic residues and three of the eight groupings of hydrophobic residues[3]. It is tempting to speculate that it represents the area of the protein responsible for the inhibition of cytochrome oxidase[16] and cytochrome *c* peroxidase[17]. Both basic[16–18] and non-cationic groups[19] have been implicated in the protein-enzyme bond.

Basic amino acids (lysine, arginine and histidine) constitute 22 to 23 % of the

entire protein. Except for one variation, that of residue 60 in the horse protein which is exchanged for a glycine in all the other proteins, none of the basic residues vary. Similarly, except for an isoleucine in position 3 in chicken cytochrome *c*, where the other proteins carry a valyl residue, the eight groups of hydrophobic residues[3] are not disturbed, even though such residues represent 22 to 25 % of the peptide chain. The only other increase in the number of such residues is found in human cytochrome *c* (residues 11, 12 and 58, isoleucine, methionine and isoleucine, respectively). The effect of this change is to increase the size or continuity of two hydrophobic areas already present in the other proteins. This remarkable stability of the basic residues, which are largely grouped together, and of the eight clusters of hydrophobic amino acids, increases the probability that these groupings do indeed represent, as previously noted[3], definite functional or structural necessities.

COMMENTS ON THE SPATIAL CONFIGURATION OF THE CYTOCHROME *c* CHAIN

A recent electron microscopic study of horse heart cytochrome *c* by the negative contrasting technique[20] gives an estimate of the dimensions of the molecule of 38 to 40 Å × 28 Å. The protein is visualized as consisting of three segments of about equal length, running parallel to each other and connected by two turns[20]. The thickness of the straight portions identify them as α-helices. The model looks like a letter e collapsed in the vertical direction[20]. This structure might be considered to be consistent with the amino acid sequence of the protein. The chain carries four prolyl residues at position 30, 44, 71 and 76, respectively. Since helices are interrupted at such residues, this situation may provide for a wide 180° turn in the vicinity of the two first prolines and a similar tight turn at the last two prolines. The three straight sections left would contain 29, 26 and 28 residues, respectively, which could provide α-helices of about 43.5, 39 and 42 Å in length, in reasonable agreement with the dimensions of the straight portions of the above model[20].

Such a model, however, implies a very high helical content, calculated by Levin[20] to be about 80 %. This does not appear to be in accord with the distribution of helix-forming and non-helix forming residues as identified by Blout[21] from the conformation of synthetic polyamino acids. The non-helix forming residues are marked in Table 1 by an asterisk. The sequences of six or more helix-forming amino acids are boxed, since the very smallest polypeptides which show properties indicative of a tendency towards a helical conformation are six residues long[22]. There are only five sections which, on this basis, could form helices and these encompass no more than 34.6 % of the entire chain. In particular it is notable that the area near the attachment of the heme contains 6 non-helix forming residues in three of the cytochromes and 7 such amino acids in the human protein, out of a total of 12 residues, making it improbable that this region is helical as has been assumed[14].

With regard to the proline residues, all four are not essential since in chicken cytochrome *c* the proline in position 44 is replaced by a glutamyl residue. Such a substitution does not necessarily change the number of non-helical portions of the molecule. Indeed as was shown by Kendrew and collaborators[23] for myoglobin, non-helical sections can occur without the presence of a proline.

DIFFERENCES IN THE AMINO ACID SEQUENCES OF VERTEBRATE CYTOCHROMES *c*

As noted above the amino acid exchanges occur at only 15 sites along the chain and are grouped in three regions located between residues 3 and 15, 44 and 62, 83 and 104. Some sites appear to be particularly prone to change. Among the four proteins examined four different residues appear in position 89, while three are found in position 92. This indicates loci at which mutations appear to be favored, similar to the genetic "hot-spots" observed by Benzer[24] in bacteriophage.

Residues in certain positions are probably characteristic of particular species or groups of closely related species. Thus horse cytochrome *c* has a threonine in position 47 and a lysine in position 60, where the other proteins carry seryl and glycyl residues, respectively, in these locations. The horse protein is the only one of the present set containing 19 lysyl residues as compared to the 18 present in the other cytochromes *c*. This extra basic group must account for the relatively easy separation of the horse protein from other cytochromes *c* by cation exchanger chromatography[11]. Horse cytochrome *c* also differs from the other proteins by its total lack of serine. Similarly chicken cytochrome *c* is unique in having an isoleucine in position 3 and a serine at the carboxyl-terminal end of the chain, positions in which all the other proteins have valine and glutamic acid respectively. Human cytochrome *c* is distinguished by an Ileu-Met sequence in place of the Val-GluNH_2 found in the other proteins in positions 11 and 12.

As expected from known phylogenetic relations there are only three amino acid differences between the cytochromes of horse and pig. Chicken cytochrome *c*, the only bird protein represented in the present series, differs by as many as 9 residues from horse, 10 from human and 7 from pig cytochrome *c*. Differences of the same magnitude occur among the mammalian proteins. For example, the human protein differs from horse cytochrome *c* by 11 residues and by 9 residues from the pig protein. Clearly, more precise phylogenetic correlations will be possible only when the amino acid sequences of many more cytochromes *c* from suitably chosen species have been worked out.

The finding that triplets of nucleotides in ribonucleic acid "code" for individual residues in proteins, makes it possible to examine whether the exchanges of amino acids in a series of related proteins result from changes of single nucleotides in the corresponding triplets, or a more extensive variation. For this purpose the composition of the coding triplets, obtained by experimental observations with synthetic polynucleotides[25–27], is not sufficient and it is necessary to consider the actual sequence of nucleotides in each triplet. From known amino acid exchanges in various proteins Smith[28] has proposed six alternative sets of coding triplets and Jukes[29] has proposed a single such set. Table 3 lists the amino acid exchanges found in the four vertebrate cytochromes *c* together with their corresponding nucleotide triplets according to Smith's [28] code 1. Of the 18 types of amino acid exchanges observed 8 involve the change of one nucleotide in the coding triplet and of these, four result from changes of a purine for a pyrimidine base while four are due to exchanges of like bases only. The remaining 10 residue variations appear to be due to more complex changes, presumably through intermediate stages of evolution involving other amino acids. The four residues observed in

Table 3. Nucleotide triplet coding sequences corresponding to amino acid exchanges observed in four vertebrate cytochromes *c*. The nucleotide sequences are those proposed by Smith[28] (code No. 1).

Position of residues in peptide chain	Amino acid interchange	Corresponding nucleotide triplet sequences
3, 11	**Val/Ileu**	**UUG/UUA**
12	**$GluNH_2$/Met**	**UCG/GUA**
15	**Ala/Ser**	**CUG/CUU**
44	**Pro/Glu**	**CCU/UAG**
46	**Phe/Tyr**	**UUU/UAU**
47, 89	**Thr/Ser**	**CUA/CUU**
50	**Asp/Ala**	**UGA/CUG**
58	**Thr/Ileu**	**CUA/UUA**
60	**Lys/Gly**	**UAA/UGG**
62	**Glu/Asp**	**UAG/UGA**
83, 92	**Ala/Val**	**CUG/UUG**
89, 104	**Glu/Ser**	**UAG/CUU**
89	**Thr/Gly**	**CUA/UGG**
89	**Thr/Glu**	**CUA/UAG**
89	**Glu/Gly**	**UAG/UGG**
89	**Ser/Gly**	**CUU/UGG**
92	**Glu/Val**	**UAG/UUG**
92	**Glu/Ala**	**UAG/CUG**

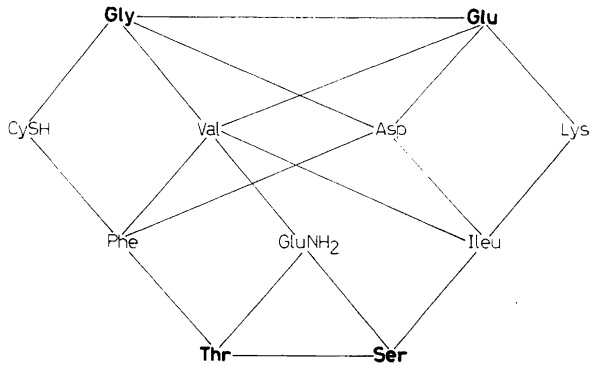

Fig. 1. Possible intermediates in the amino acid exchanges at position 89 of the peptide chain of vertebrate cytochromes *c*. The nucleotide triplet coding sequences used were those proposed by Smith[28] (code No. 1). The residues in **bold type** are the ones so far observed in four vertebrate cytochromes *c*, the others represent possible intermediates. Residues joined by a line can transform one into the other as a result of a single nucleotide change in the coding triplet.

position 89, threonine, glycine, glutamic acid and serine, furnish a good example of such a possibility. These residues can be arranged in two pairs, glycine and glutamic acid on the one hand, and threonine and serine on the other. Transformation within each pair can be obtained by a single base change, while to

proceed from one pair to the other several alternative pathways, involving cysteine, valine, aspartic acid, lysine, phenylalanine, glutamine and isoleucine as intermediates, may be envisaged. The pathways are summarized in Fig. 1. Similar though less complex pathways may be considered for the exchanges occurring at positions 12, 44, 50, 60, 62 and 92. It should however be strongly emphasized that since the above discussion is based on a triplet nucleotide code which is incomplete, as evidenced by the recent discovery of coding triplets not containing uridylic acid[25,26], it is impossible to decide at present whether the apparent complexities of some residue exchanges will not be simplified by a fuller knowledge of coding triplets, or whether the protein did in fact evolve through numerous intermediate forms. The finding of cytochromes *c* carrying the postulated intermediate residues may not always be possible since some may have been eliminated in evolution. Nevertheless the establishment of a large series of amino acid sequences for a set of homologous proteins will help, as already indicated[28,29], in determining the sequence of nucleotides in coding triplets.

REFERENCES

1. Margoliash, E., Smith, E. L., Kreil, G. and Tuppy, H. Nature **192** (1960) 1125.
2. Margoliash, E. and Smith, E. L. *J. Biol. Chem.* **237** (1962) 2151.
3. Margoliash, E. *J. Biol. Chem.* **237** (1962) 2161.
4. Tuppy, H. and Kreil, G. *Monatsh. Chem.* **92** (1962) 780.
5. Matsubara, H. and Smith, E. L. *J. Biol. Chem.* **237** (1962) PC 3575.
6. Chan, S. K., Needleman, S. B., Stewart, J. W., Walasek, O. F. and Margoliash, E. *Federation Proc.* **22** (1963) 658.
7. Tuppy, H. and Bodo, G. *Monatsh. Chem.* **85** (1954) 1024.
8. Tuppy, H. and Paléus, S. *Acta Chem. Scand.* **9** (1955) 353.
9. Paléus, S. and Tuppy, H. *Acta Chem. Scand.* **13** (1959) 641.
10. Dus, K., Bartsch, R. G. and Kamen, M. *J. Biol. Chem.* **237** (1962) 3083.
11. Margoliash, E. *Brookhaven Symp. Biol.* **15** (1962) 266.
12. George, P. and Lyster, R. L. J. *Proc. Natl. Acad. Sci. U. S.* **44** (1958) 1013.
13. Margoliash, E., Frohwirt, N. and Wiener, E. *Biochem. J.* **71** (1959) 559.
14. Ehrenberg, A. and Theorell, H. *Acta Chem. Scand.* **9** (1955) 1193.
15. Paléus, S., Ehrenberg, A. and Tuppy, H. *Acta Chem. Scand.* **9** (1955) 365.
16. Smith, L. and Conrad, H. in Falk, J. E., Lemberg, R. and Morton, R. K. (Eds.) *Haematin Enzymes,* Pergamon Press, London, 1961, p. 260.
17. Beetlestone, J. *Arch. Biochem. Biophys.* **89** (1960) 35.
18. Estabrook, R. W. in Falk, J. E., Lemberg, R. and Morton, R. K. (Eds.) *Haematin Enzymes,* Pergamon Press, London, 1961, p. 276.
19. Conrad, H. and Wasserman, A. R. *Federation Proc.* **20** (1961) 42.
20. Levin, Ö. *Arch. Biochem. Biophys.* **Suppl. 1** (1962) 301.
21. Blout, E. R. in Stahmann, M. A. (Ed.) *Polyamino Acids, Polypeptides and Proteins,* The University of Wisconsin Press, Madison, 1962, p. 275.
22. Mitchell, J. C., Woodward, A. E. and Doty, P. *J. Am. Chem. Soc.* **79** (1957) 3955.
23. Kendrew, J. C., Watson, H. C., Strandberg, B. E., Dickerson, R. E., Phillips, D. C. and Shore, V. C. *Nature* **190** (1961) 666.
24. Benzer, S. *Proc. Natl. Acad. Sci. U. S.* **47** (1961) 403.
25. Jones, O. W., Jr., and Nirenberg, M. W. *Proc. Natl. Acad. Sci. U. S.* **48** (1962) 2115.
26. Wahba, A. G., Gardner, R. S., Basilio, C., Miller, R. S., Speyer, J. F. and Lengyel, P. *Proc. Natl. Acad. Sci. U. S.* **49** (1963) 116.
27. Speyer, J. F., Lengyel, P., Basilio, C. and Ochoa, S. *Proc. Natl. Acad. Sci. U. S.* **48** (1962) 63.
28. Smith, E. L. *Proc. Natl. Acad. Sci. U. S.* **48** (1962) 677, 859.
29. Jukes, T. H. *Proc. Natl. Acad. Sci. U. S.* **48** (1962) 1809.

Received April 2, 1963.