HOMOLOGY OF PSEUDOMONAS CYTOCHROME c-551 WITH EUKARYOTIC c-CYTOCHROMES

By Saul B. Needleman and Terence T. Blair*

DEPARTMENT OF BIOCHEMISTRY, NORTHWESTERN UNIVERSITY SCHOOL OF MEDICINE

Communicated by Myron L. Bender, May 8, 1969

Abstract.—The homology of *Pseudomonas* cytochrome *c*-551 with eukaryotic cytochromes *c* is examined with a computer-based procedure devised to determine whether similarities exist between these proteins. One method is given by which the more recently evolved cytochromes *c* might have arisen from the *Pseudomonas* protein. This procedure involves only common genetic phenomena and accounts for most of the structural differences between the bacterial and mammalian cytochromes. A time-scale relationship between the *c*-cytochromes from several microorganisms, a mold, yeast, and the eukaryotic organisms is proposed.

Introduction.—The number of amino acid residues in the eukaryotic cytochromes *c* varies between 104 and 111 for the 31 different species thus far examined. Their remarkable similarity apparently is not the result of evolutionary convergence, but rather of homology. Their homology has been used as evidence for the suggestion of a single effective emergence of eukaryotic life on earth.[1] The gene coding for the eukaryotic cytochromes *c* has not changed substantially in form since its initial emergence. However, the problem remains as to what evolutionary steps the gene has undergone before its emergence in the eukaryotic cell, how it developed, and what the ancestral structures might have been. We present evidence here for the consideration of *Pseudomonas* cytochrome *c*-551 as an early ancestor for the eukaryotic cytochromes *c*. It is suggested that the eukaryotic ancestral cytochrome *c* was developed from *P. c*-551 by a series of gene duplications and point mutations. Any attempt at proving homology must account for the 22 amino acid differences between *P. c*-551 and the eukaryotic cytochrome *c* (the model used in the present study is the sequence proposed by Nolan and Margoliash[1]).

The inference of homology is derived from an unusually high correlation (degree of similarity) between the two sequences, obtained by inserting a small number of gaps or deletions to increase their correlatedness. One first looks for identical residues in similar positions in each sequence. Next, one looks for amino acid replacements in which charge, shape, or structural function is retained. Finally, one looks for amino acid residues that might have occurred by the alteration of a single nucleotide base in the triplet codon, even though this might result in an interchange between quite dissimilar amino acids. These latter replacements must be considered to have validity and importance in considering homology as do those involving charge or structural conservatism. As an example, threonine (ACA or ACG) can be interchanged with lysine (AAA or AAG) without regard to charge, shape, or structural similarity of the two amino acids.

Methods.—The complete amino acid sequence of cytochrome *c* from *Pseudomonas fluorescens*[2] and the hypothetical sequence for ancestral cytochrome *c*[1] were used in the test for homology in the present study. The optimum alignment of the sequences was achieved by a computerized comparison of amino acid pairs by the sliding sequence technique. This technique requires that the consecutive amino acids of one protein be aligned opposite those of the other protein to form a series of amino acid pairs. Those segments of the sequence made up of pairs containing the same amino acid are noted (the concordant segments). In addition, one records those pairs of amino acids which are related through similarities in chemical structure (structurally conservative interchanges) and in which the triplet codon representative of one amino acid residue in the pair is convertible to the codon for the other amino acid by a single nucleotide base change. The two amino acid sequences are shifted relative to each other, one amino acid residue at a time, and any new concordant segments are added to the previously recorded segments. The process is repeated until all possible alignments of one sequence opposite the other have been examined for concordant segments, even permitting limited "gaps" between the selected amino acid pairs. In practice, the following conditions are imposed in the program: (1) a lower limit, N_{aa}, on the number of selected amino acids in corresponding positions, (2) a lower limit, P_f, on the per cent of selected amino acids in the total congruent segment, and (3) an upper limit, N_f, on the number of consecutive nonmatching amino acids separating any two selected residue pairs. Data in the present communication were generated with an IBM 1620 model I data processing system using a search program with $N_{aa} = 4$, $P_f = 13$, and $N_f = 20$. Thus, all concordant segments selected by the program contained at least four acceptable amino acid residues which constituted a minimum of 13% of the total of all (acceptable and nonacceptable) residues contained in the concordant segment. No more than 20 consecutive, nonmatching residues were permitted between any two acceptable amino acid residues. The concordant segments were then manually selected and arranged to construct a pairing of the sequences which maximized the correlatedness of the two proteins being compared.

Attempts to derive one sequence from another made use only of accepted genetic phenomena, single nucleotide insertions or replacements, and gene duplication. The total sequence was arbitrarily divided into smaller peptide regions containing 10–20 residues on the basis of the common occurrence of marker amino acids which have been found to be invariant in all of the cytochrome *c* sequences presently known.

Results.—Shown in Figure 1 are the sequences of segments common to the ancestral cytochrome *c* (line *A*) and *P. c*-551 (line *P*).

In the segment containing the first 18 amino acids of hypothetical ancestral cytochrome *c*, 14 amino acids are correlated in the two proteins. Eight are identidal residues, two (Gln/Asn at postition 12 and Arg/Lys at position 13) are conservatively similar by structure, and four other amino acid interchanges can be accounted for by nucleotide conservatism (Gly/Glu-GGe/GAe at position 1, Ala/Pro-GCd/CCd at position 3, Glu/Val-GGd/GUd at position 6, and Ala/Val-GCd/GUd at position 15).

In the next segment of 9 residues, six identical amino acids are similarly oriented in both sequences, four constitute conservative replacements of the first type (structural), and three show nucleotide conservatism (positions 19, 21, and 31). The next segment of 21 residues contains three identically matched, two structurally changed, and nine altered through their nucleotide codons.

In the segment beginning with the tryptophanyl residue at position 59 and ending at the methionyl residue at position 80, a total of five residues of all three types occurs in common sequence in the two proteins. Five identical residues can be aligned in the C-terminal region of the two proteins. In addition,

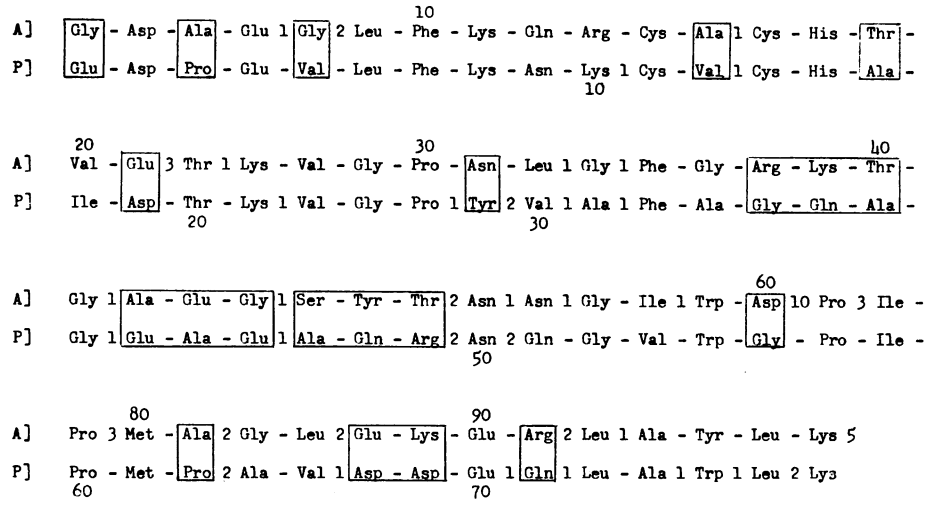

FIG. 1.—Comparison of the amino acid sequences of ancestral cytochrome c^1 and *Pseudomonas* cytochrome c_2-551.[2] The large numerals between residues represent the number of non-acceptable (nonmatching) amino acid residues occurring between concordant segments. The boxed residues are those involved in nucleotide conservatism. The superscripts correspond to the numbering of residues in the ancestral protein sequence. The subscripts correspond to the numbering of residues in the *Pseudomonas* cytochrome.

three amino acids are conservatively changed with structural retention and four show nucleotide conservatism. The last five residues of the hypothetical protein are not matched by any in *P. c*-551.

Comparison of *P. c*-551 with the ancestral protein gives a total of 27 residues (32.9 % of *P. c*-551) which can be identically matched residues. An additional 11 residues (13.6 % of P. *c*-551) can be aligned by considering conservative structure and function retention; 21 more amino acids (25.6 % of P. *c*-551) can be aligned on the basis of nucleotide conservatism. A total of 59 residues (72.3 % of *P. c*-551) are thus common to both sequences.

Figure 2 and 3, respectively, show a possible method of generating segments 1–13 and 14–26 of the eukaryotic ancestral protein from segments 1–10 and 12–20 of *P. c*-551. Line *A* shows the *Pseudomonas* sequence which is used to generate line *B* by gene duplication. Line *C* shows the ancestral sequence derived from line *B*. The letters above the codons represent nucleotide changes necessary to generate the next line.

The segments shown in Figure 4 present two alternative methods of comparing the central segments of both protein sequences. The first comparison (Fig. 4*a*) involves retention of the ancestral tryptophan residue. The second comparison (suggested by Fitch[3]) omits this tryptophan residue as part of a 15-residue deletion. Both comparisons reduce the number of acceptable amino acid pairs over that presented in Figure 1. However, the comparison given in Figure 4*a* necessitates fewer steps to generate the ancestral sequence from *P. c*-551 than does the comparison shown in Figure 1. This tryptophan residue is common to all *c*-cytochromes and probably represents an essential feature of cytochrome *c*.

```
      Pseudomonas c-551   residues 1-10
        G                   AA
A]    GAe - GAb - CCd - GAe - GUe - CUd - UUb - AAe - AAb - AAe

      Glu - Asp - Pro - Glu - Val - Leu - Phe - Lys - Asn - Lys


                  G                 GG    C                             C e   G
B]    GGe - GAb - CCd - GAe - AAe - CCd - GAe - AAe - CUd - UUb - AAe - AAb - AAe

      Gly - Asp - Pro - Glu - Lys - Pro - Glu - Lys - Leu - Phe - Lys - Asn - Lys


      Ancestor cytochrome   residues 1-13
C]    GGd - GAb - GCd - GAe - AAe - GGd - GCe - AAe - CUd - UUb - AAe - CAe - AGe

      Gly - Asp - Ala - Glu - Lys - Gly - Ala - Lys - Leu - Phe - Lys - Gln - Arg
```

FIG. 2.—Generation of segment 1–13 of the ancestor sequence from segment 1–10 of the *Pseudomonas* cytochrome *c*-551 sequence. Line *B* is generated by duplicating the third, fourth, and fifth residues of the sequence given in line *A*.

```
      Pseudomonas c-551   residues 12-20
                      15              G       e    A
A]    UGb - GUd - GCd - UGb - CAb - GCd - AUb - GAb - ACb

      Cys - Val - Ala - Cys - His - Ala - Ile - Asp - Thr


          C   CA                A                 G e   G     G    A     C
B]    UGb - GUd - GCd - UGb - CAb - GCd - GUb - GAe - AAb - GCd - GUb - GAe - AAb

      Cys - Val - Ala - Cys - His - Ala - Val - Glu - Asn - Ala - Val - Glu - Asn


      Ancestor cytochrome   residues 14-26
C]    UGb - GCd - CAe - UGb - CAb - ACd - GUb - GAe - GAe - GGd - GGb - AAe - CAb

      Cys - Ala - Gln - Cys - His - Thr - Val - Glu - Glu - Gly - Gly - Lys - His
```

FIG. 3.—Generation of ancestor sequence segment 14–26 from the *Pseudomonas* sequence segment 12–20. Line *B* is generated by duplicating the four C-terminal residues of the sequence given in line *A*.

It is frequently used as a marker to align sequences, and its inclusion in any sequence comparison probably is preferred to comparisons omitting it.

Discussion.—A definitive technique for tracing the evolution of a protein is not at hand, but the number of amino acid differences between two homologous proteins has been taken as a proximate estimate of the time of divergence of the lines leading to those proteins. The usual criterion by which one protein is said to be homologous to another is a high degree of similarity in their amino acid sequences. With a series of homologous proteins, the possibility exists of deriving the structure for a hypothetical ancestral precursor from which one can postulate lines of evolution representative of the species containing that series of proteins.

The amino acid sequences have been established for a large number of eukaryotic cytochromes *c*. This has led to one suggestion for the sequence of a hypothetical ancestral cytochrome *c* based on the mammalian cytochrome structures. The structure of *Pseudomonas* cytochrome *c*-551, in spite of apparent dissimilarities[1, 2] with the eukaryotic type, is now shown to be homologous with that group. This offers a firmer basis for deducing the structure for the ancestral

(a)

A] Pro 3 [Gly] - Leu 1 Gly - Arg 1 [Thr] - Gly - Gln - Ala - [Glu] - Gly 1 [Ala] 2 Ala 3 [Lys - Asn - Gln] 4 Trp

P] Pro 3 [Asp] - Val 1 Ala - Lys 1 [Ala] - Gly - Gln - Ala - [Gly] - Ala 1 [Ser] 2 Ala 3 [Asn - Lys - Gly] 2 Trp

(b)

A] Gly - Arg - Lys - [Thr] - Gly - Gln - Ala - [Glu] - Gly 15 Glu - [Glu] - Thr - Leu

P] Ala - Lys - Phe - [Ala] - Gly - Gln - Ala - [Gly] - Ala - Glu - [Ala] - Glu - Leu

FIG. 4.—(*a*) An alternate alignment of the internal residues of the two cytochrome sequences beginning with Proline 30. Generation of the ancestral cytochrome *c* from the *Pseudomonas* protein using this alignment requires fewer steps to accomplish than does the alignment presented in Fig. 1, although the number of acceptable matching residues is lower. (*b*) Comparison of the internal sequences of *Pseudomonas* cytochrome *c*-551 and the ancestral cytochrome *c* as suggested by Fitch.[3] The alignment differs significantly from (*a*) in the omission of the tryptophan residue at position 59.

c-cytochrome. This is particularly noteworthy since it has been possible to suggest a pathway through which the amino acid sequence of an eukaryotic cytochrome *c* can be derived from the sequence of the *P. c*-551 protein using only common genetic phenomena. While it is difficult to evaluate the technique on a statistical basis, all attempts at deriving the structures of apparently nonhomologous proteins from that of *P. c*-551 have failed or have required a manyfold increase in the number of manipulations needed to effect that transformation. The relationship between the hypothetical pathway and the pathway which might have been followed during the evolutionary derivation of the earliest form of cytochrome to that as presently constituted is of necessity unknown. This hypothetical pathway is presented only to show that where homology between two proteins can be demonstrated, it is possible to deduce one structure from the other. It is suggested that the *Pseudomonas* protein must have evolved earlier than eukaryotic cytochrome *c*.

Dus *et al.*[4] recently have suggested that the cytochrome c_2 of *Rhodospirillum rubrum* (*R.* c_2) holds a position between the cytochromes of prokaryotic and eukaryotic organisms. They found some degree of homology between *P. c*-551 and other cytochromes *c* and suggest that *Rhodospirillum rubrum* is an immediate precursor for the eukaryotic group; *P. c*-551 is placed on a side pathway from cytochrome c_2. However, we feel that *R.* c_2 occupies a position intermediate between *P. c*-551 and eukaryotic types on the same pathway. As might be expected, *Pseudomonas* cytochrome *c* is considered to be more closely related to cytochrome c_2 than to the eukaryotic type. Cantor and Jukes[5] already have suggested a probable low degree of homology between the cytochromes of *Neurospora crassa* and *Pseudomonas fluorescens* but propose that the *Pseudomonas* protein did not evolve from the same prototype as did the other *c*-cytochromes.

A total of 59 amino acid residues (more than 72% based on the *P. c*-551 sequence) are found in the same sequential array in *Pseudomonas fluorescens* cytochrome *c*-551 and the ancestral cytochrome *c* as representatives of the eukaryotic

cytochromes *c*. This includes 27 cases in which the amino acids in the two proteins are identical, 11 are related by structural conservatism, and 21 more by nucleotide conservatism without regard to retention of structure or change. Of the 27 identical amino acid residues, 14 are identical with those amino acids which have remained invariant on all the mammalian sequences thus far elucidated.

The region covering residues 70–80 (invariant) in the eukaryotic sequence is entirely lacking in *P. c*-551. This region has been proposed as the site for binding cytochrome *c* to the mitochondrial membrane. If one considers the reasonable assumption that the eukaryotic organisms developed from prokaryotic types and that the two proteins being compared are not the result of convergent evolution, it is evident that there would be no reason to expect this segment in the sequence of *P. c*-551. Even including these 11 residues, Fitch and Margoliash[6] have calculated that 27–29 residues must remain invariant in the eukaryotic sequence. Twenty-six residues are coincident between *P. c*-551 and the ancestral protein, though the invariant residues are not the same in both sequences. If *P. c*-551 is to be used in calculating an ancestral structure for the cytochromes, the list of absolutely invariant residues will have to be revised. For one group of proteins of identical function, the hemoglobins, Zuckerkandl and Pauling[7] have suggested that no more than seven amino acid residues will remain completely invariant.

For a few other groups of proteins, each containing molecules of similar chain length, e.g., the insulins, fibrinopeptides, and the hemoglobins, homology has been suggested on the basis of their nearly identical amino acid sequence. A common ancestor for the hemoglobins also has been postulated.[7] The various globin proteins are thought to have arisen from a common ancestor which underwent gene duplication; the two resulting proteins followed separate evolutionary paths. This process has led to five individual new proteins of similar length and function, each homologous to the other. The suggestion has been made that the sequences of the ferredoxins as presently constituted may have been derived from a smaller common ancestor rather than from an ancestor of similar size.[8]

The degree of similarity in two protease zymogens (identical and structurally conservative residues) is 53 per cent. In the hemoglobins the degree of similarity is 51 per cent. Even without the inclusion of nucleotide conservatism, the present comparison shows 47 per cent similarity. It is likely that more mutations will occur at sites not directly associated with the function of the protein, and it is evident that a clearer understanding of evolutionary derivation of one protein from another can be had only by considering nucleotide exchanges as well as identical and structurally related amino acids. Taken to an extreme, it is perhaps wise to weigh the different types of possible exchanges, i.e., zero, one, two, or three nucleotide codon changes.

The heme-binding region in *Pseudomonas* cytochrome *c*-551 is much the same as that in the eukaryotic cytochromes, i.e., two thioether-bonded cysteine residues separated by two variable amino acids; these are followed by a histidine residue in both examples. The distribution of hydrophobic residues in *P. c*-551 correlates well with the hydrophobic regions in the mammalian cytochromes. A total of 14 residues (66.6%) lies within the commonly accepted hydrophobic regions while 3 more (14.3%) are coincident with isolated hydrophobic residues or

extend a hydrophobic region. Only four residues in the *Pseudomonas* protein lie outside the usual hydrophobic regions of the mammalian sequences. Although there is more variability with the basic residues, these in general tend to locate within the usually accepted basic regions.

It is evident, then, from the above considerations, as well as from the high degree of coincidence of amino acids in the sequence array and the ability to uniquely arrive at the mammalian structure from that of the *Pseudomonas* protein, that *Pseudomonas* cyrochrome *c*-551 must be considered as an early ancestor of the so-called mammalian group of *c*-cytochromes, enabling us to extend back in time data from which can be deduced the structure of the ancestral cytochrome molecule from which all others evolved.

Summary.—*Pseudomonas* cytochrome *c*-551 differs from the eukaryotic cytochromes *c* in the total number of amino acid residues, in composition, and in the order of residues in large segments of the amino acid sequence. Only in the mode of binding of the heme to the protein and in the general, visible absorption pattern is there any obvious similarity between the bacterial and eukaryotic types. Nevertheless, we found a high degree of similarity between *Pseudomonas* *c*-551 and an ancestral cytochrome *c* sequence representative of the modern mammalian proteins. On this basis, *Pseudomonas* cytochrome *c*-551 is suggested as being homologous with and ancestral to the eukaryotic type of *c*-cytochromes.

* This work was supported in part by the U. S. Public Health Service general research support grant 1-501-FR-05370-02 to S. B. N.

[1] Nolan, C., and E. Margoliash, in *Annual Review of Biochemistry*, ed. P. D. Boyer (Palo Alto: Annual Reviews, Inc., 1968), vol. 37, p. 727.

[2] Ambler, R. P., *Biochem. J.*, **89**, 349 (1963).

[3] Fitch, W., personal communication.

[4] Dus, K., K. Sletten, and M. D. Kamen, *J. Biol. Chem.*, **243**, 5492, 5507 (1968).

[5] Cantor, C. R., and T. H. Jukes, these PROCEEDINGS, **56**, 177 (1966).

[6] Fitch, W., and E. Margoliash, *Biochem. Genet.*, **1**, 65 (1967).

[7] Zuckerkandl, E., and L. Pauling, in *Evolving Genes and Proteins*, ed. V. Bryson and H. J. Vogel (New York: Academic Press, 1965), p. 97.

[8] Eck, R. V., and M. O. Dayhoff, *Science*, **152**, 363 (1966).