REPETITIONS IN THE POLYPEPTIDE
SEQUENCE OF CYTOCHROMES

Charles R. Cantor and Thomas H. Jukes

Department of Chemistry and Space Sciences Laboratory
University of California,
Berkeley, California

Received

It seems likely that many proteins have evolved from comparatively short
primordial peptides by processes of duplication, deletions, and amino acid sub-
stitutions due to point mutations (Eck and Dayhoff, 1966; Jukes, 1966a).
In most cases, the evidence for this may have vanished due to the masking
effect of evolutionary changes in the amino acid sequences. In a few cases,
however, remnants of the early peptides are still discernible in the forms
of partially repetitive sequences. Proteins from "primitive" organisms may
be the best source in which to search for such manifestations.

The phenomena of gene duplications and deletions are perceptible as trans-
lations of the genetic message in the polypeptide sequences of certain pro-
teins. Presumably this indicates an evolutionary origin and significance
for the occurrence of these phenomena in these instances. Examples are to be
found in the α and β chains of hemoglobin A, the genes for which occupy
separate chromosomal loci. Their separation is thought to result from dup-
lication and translocation (Ingram, 1963). The two chains also show the
occurrence of deletions, indicated by the presence of gaps in the homology
of the polypeptides, for instance -

*Supported by a National Science Foundation Cooperative Graduate Fellowship
and by grant NsG 479 by the National Aeronautics and Space Administration to
the University of California.
The presence of a gap corresponding to the deletion of 6 consecutive base pairs in the gene is inferred from the comparison of these sequences. The comparison also shows that 7 point mutations have taken place in the gene to produce changes at 5 homologous pairs of loci corresponding to individual amino acid residues indicated by base changes. Such amino-acid changes, when tolerated, may become so numerous as to obscure the original homology, thus erasing the record of the primordial duplication. An example of this is given below for human cytochrome c. The comparison of any two polypeptide chains is aided by enumerating the minimum number of base changes in the coding triplets (Brimacombe, et al., 1965) corresponding to each pair of amino acid residues at two homologous sites. In the comparison above, these are listed underneath the sequences. In this way two amino acids that are related by a single-base change in the code are noted as being more closely related than two which are separated by a two-base or three-base change. Three-base changes at a locus are, of course, difficult to detect because of the extensive synonymity in the third positions of the coding triplets (Brimacombe, et al., 1965).

Partial genetic duplication is evident in the haptoglobins (Smithies, 1964) and in *Clostridium pasteurianum* ferredoxin (Jukes, 1966b).

Fitch (1966) has proposed a method for examining polypeptide chains for regions of partial internal gene duplication, depending upon determining the minimum number of nucleotides which must be altered to permit the conversion of one sequence into the other. The test is carried out by means of a computer. We have devised a similar test, details of which will be published later. Having detected a homology by means of this test, it is often possible to make it visually obvious by means of the procedure used above with the hemoglobin.
An examination of the primary structure of the cytochromes c indicates a region of partial duplication in the cytochrome c of *Neurospora crassa* (Heller and Smith, 1965). Using throughout the numbering system employed for the vertebrate cytochromes c, residues 5 to 19 are homologous with residues 20 to 34 as follows:

<table>
<thead>
<tr>
<th>Residue No.</th>
<th>5</th>
<th>19</th>
</tr>
</thead>
<tbody>
<tr>
<td>lys gly ala asn leu phe lys thr arg cys ala glu cys his gly</td>
<td>20</td>
<td>34</td>
</tr>
<tr>
<td>glu gly glt leu thr gln lys ile gly pro ala leu his gly</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Base Changes: 1 0 1 0 0 2 1 1 1 1 1 1 2 0 0 0

This internal homology is not perceptible in the cytochromes c of other species, including those of certain vertebrates (Margoliash and Smith, 1965), two yeasts (bakers yeast and *Candida krusei*) (Narita, et al, 1963; Narita and Titani, 1965) and the moth *Samia cynthia* (Chan and Margoliash, 1966). As an example, the corresponding sequences for human cytochrome c show no evidence of homology beyond what might be interpreted as being coincidental, as follows:

<table>
<thead>
<tr>
<th>Residue No.</th>
<th>5</th>
<th>19</th>
</tr>
</thead>
<tbody>
<tr>
<td>lys gly lys lys ile phe ile met lys cys ser gln cys his thr</td>
<td>20</td>
<td>34</td>
</tr>
<tr>
<td>val glu lys gly gly lys his lys thr gly pro asn leu his gly</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Base Changes: 2 1 0 2 2 3 2 1 1 1 2 2 0 2 2

The only two identical pairs are the lysines at positions 7 and 22 and the histidines at 18 and 33. In other cytochromes c, this last vestige of a common origin of the two sequences in this region has disappeared; residues 22 and 33 are occupied by asparagine and tryptophan in tuna fish cytochrome c (Kreil, 1963) as follows:

<table>
<thead>
<tr>
<th>Residue No.</th>
<th>5</th>
<th>19</th>
</tr>
</thead>
<tbody>
<tr>
<td>lys gly lys lys thr phe val gln lys cys ala gln cys his thr</td>
<td>20</td>
<td>34</td>
</tr>
<tr>
<td>val glu asn gly gly lys his lys val gly pro asn leu trp gly</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Base Changes: 2 1 1 2 2 3 2 1 2 1 1 2 2 3 2
Evolution can thus in some cases erode the traces of an incident in its early history. In other instances, such as in the often-cited example of the invariant sequence of 11 amino acid residues in loci 70 to 80 of the cytochromes c, an essential portion of the same molecule has remained unchanged during the evolution of a diversity of living species from a common origin (Margoliash and Smith, 1965). It would seem that evolutionary changes in the region of residues 5 to 34 have proceeded more slowly in *Neurospora crassa* cytochrome c than in the other cytochromes c whose primary structure has been described. However, another example of homology in the same region is evident when a typical vertebrate cytochrome c is compared with the diheme peptide from the variant heme protein RHP, of the photoanaerobe *Chromatium* (Dus, Bartsch and Kamen, 1962) as follows, assuming a single amino acid gap between residues 19 and 20 in the vertebrate cytochrome c:

\[
\begin{array}{cccccccccc}
& & & & & & & & & & \\
\text{i Tuna cytochrome c:} & & & & & & & & & & \\
& 10 & & & & & & & & & \\
\text{phe val gln lys cys ala gln cys his thr} & & & & & & & & & & - \\
& 20 & & & & & & & & & & 19 \\
\text{val glu asn gly gly} & & & & & & & & & & \\
\text{ii Chromatium} & & & & & & & & & & \\
\text{phe ala gly lys cys ser gln cys his thr leu} & & & & & & & & & & 10 \\
& 16 & & & & & & & & & & 20 \\
\text{val ala asp glu gly - - - - ser ala lys cys his thr phe} & & & & & & & & & & \\
& 27 & & & & & & & & & & \\
\text{- - asp glu gly ser} & & & & & & & & & & \\
\end{array}
\]

Base changes, i vs. ii: 0 1 1 1 0 0 1 2 0 0 1 0 0 0 0 -

Base changes, internal in ii: - - 0 0 0 1 - - - 1 1 1 0 0 0 1

The primary structure of the remainder of the *Chromatium* heme-carrying protein is not available and would be of great interest as possibly revealing a second incident of internal duplication in the regions preceding and following the known peptide.
REFERENCES


