THE OCCURRENCE OF GAPS IN PROTEIN SEQUENCES

Charles R. Cantor

Department of Chemistry
Columbia University
New York, New York 10027*

Received March 29, 1968

Recently a large number of attempts have been made to search for homology among protein sequences. (Eck and Dayhoff, 1966; Jukes, 1966; Fitch and Margoliash, 1967; Ingram, 1966; Fitch, 1966; Hartley, et. al., 1965; Brew, et. al., 1967; Manwell, 1967). It has rarely been possible, except in the case of cytochrome c, to place uninterrupted polypeptide sequences side by side and find regions of extensive homology dispersed throughout them. Instead, it seems to be necessary to introduce interruptions in one or both of the sequences at various points in order to extend the apparent homology across most of the protein chains. The breaks in the sequences usually called gaps may result from additions of residues to one chain or deletions from the other. They may originate from recombination in DNA. A simple procedure has been developed to test whether a gap does in fact increase the apparent homology between two sequences. The existence of gaps in the sequence of amino acids of some proteins can be fairly convincingly demonstrated. In many other cases however, gaps which have been postulated may prove to be uncertain if they are subjected to statistical tests.

In some cases the presence of a gap seems to improve the

* This work was supported by Grant GM 14825 from the National Institutes of Health.

apparent homology between two sequences so extensively that no justification is necessary to account for its presence in the protein chain. Part of the qualitative evidence for such a gap which has long been thought to occur between residues 18 and 19 of the hemoglobin β chain (Braunitzer et. al., 1961) is displayed in Figure 1. The minimum base differences (mbd) shown below each pair of amino acids are found by examining the set of degenerate codons for each amino acid. With the gap there are a total of 11 mbd for the 20 residues compared. If the gap is closed there are three possible ways of comparing the sequences. These are shown in Figure 1. They result in alignments with either 19 mbd, or 26 mbd. The placement of a gap in the above comparison results in a vastly improved apparent homology. In many other cases, however, the results are not nearly so obvious. For example Dunnill used 8 gaps to develop the possibility of homology between egg white and T4 lysozymes

	SEQUENCES	TOTAL MBD
Hbα(11-33)	Lys-Ala-Ala-Trp-Gly-Lys-Val-Gly-Ala-His-Ala-Gly-Glu-Tyr-Gly-Ala-Glu-Ala-Leu-Glu-Arg-Met	
Hbβ(12-31)	Thr-Ala-Leu-Trp-Gly-Lys-Val Asn-Val-Asp-Glu-Val-Gly-Gly-Glu-Ala-Leu-Gly-Arg-Leu	
MBD	1 0 2 0 0 0 0 1 1 1 0 2 0 1 0 0 0 1 0 1	11
Hbα(11-33)	Lys-Ala-Ala-Trp-Gly-Lys-Val-Gly-Ala-His-Ala-Gly-Glu-Tyr-Gly-Ala-Glu-Ala-Leu-Glu-Arg-Met	
Hbβ(12-31)	Thr-Ala-Leu-Trp-Gly-Lys-Val-Asn-Val-Asp-Glu-Val-Gly-Gly-Glu-Ala-Leu-Gly-Arg-Leu	
MBD	1 2 2 2 1 2 1 1 1 1 0 2 0 1 0 0 0 1 0 1	19
Hbα(11-33)	Lys-Ala-Ala-Trp-Gly-Lys-Val-Gly-Ala-His-Ala-Gly-Glu-Tyr-Gly-Ala-Glu-Ala-Leu-Glu-Arg-Met	
Hbβ(12-31)	Thr-Ala-Leu-Trp-Gly-Lys-Val-Asn-Val-Asp-Glu-Val-Gly-Gly-Glu-Ala-Leu-Gly-Arg-Leu	
MBD	1 0 2 0 0 0 0 2 1 1 1 1 1 2 1 0 2 1 1 2	19
Hbα(11-33)	Lys-Ala-Ala-Trp-Gly-Lys-Val-Gly-Ala-His-Ala-Gly-Glu-Tyr-Gly-Ala-Glu-Ala-Leu-Glu-Arg-Met	
Hbβ(12-31)	Thr-Ala-Leu-Trp-Gly-Lys-Val-Asn-Val-Asp-Glu-Val-Gly-Gly-Glu-Ala-Leu-Gly-Arg-Leu	
MBD	1 0 1 1 2 2 1 2 2 1 1 1 2 0 1 1 2 2 2 1	26

Figure 1. Comparisons of sequences chosen from the α and β human hemoglobin chains. The top comparison contains a gap of 2 residues on the β chain. The other 3 comparisons are uninterrupted sequences.

(1967). Large numbers of gaps were used by Hill to align immunoglobulin sequences (1966). Many other examples could be cited. A difficulty is that, when two relatively dissimilar sequences are compared, it is almost always possible to improve the apparent homology by the insertion of one or more gaps. Tests have been made on the effects of putting gaps at random into unrelated sequences using a digital computer. These show that only when sequences which are already very similar are compared will the best placement of a gap be unlikely to lead to increased homology. Thus every gap postulated in a protein sequence should be treated with suspicion until its existence can be justified.

Consider, first, the problem of comparing two fixed arbitrary short sequences such as the hemoglobin fragments discussed above. There are 3 ways in which 20 amino acids from the two sequences may be compared without a gap of 2 in the β chain. With the gap, 21 different alignments are possible. If random sequences are compared in N different ways the number of times a given degree of homology will be found is equal to NP, where P is the a priori probability of choosing sequences which result in the homology. Let Pg be the a priori probability for the comparison containing the gap, and Po, the probability without the gap. Then, for the above hemoglobin comparison, we can calculate the relative frequencies of chance occurrences of an homology with or without the gap: $fg/fo = 21\ Pg/3\ Po$. Unless Pg is much less than Po the placement of a gap will increase fg/fo and thus make it appear more likely that the homology with the gap is simply due to a random event.

We have used the procedure suggested by Fitch (1966) for calculating the probability, P(L,J), of finding J minimum base differences when comparing L codons. For α and β human hemo-

globin we find $P(20,19) = 3{\cdot}7 \times 10^{-4}$ and $P(20,11) = 2{\cdot}6 \times 10^{-9}$. Thus $fg/fo = 4{\cdot}9 \times 10^{-5}$. Since this is much less than one it indicates that the gap significantly improves the apparent homology. Values of fg/fo near or greater than 1 clearly argue against the existence of the gap.

What happens when one wants to compare two entire protein chains instead of preselected short sequences? One approach, introduced by Fitch (1966), involves the comparison of all sets of chains of length L chosen from the two intact protein chains. The result, is a distribution of numbers of comparisons as a function of the minimum base changes for the amino acid residues compared. This distribution can be tested against distributions generated from random sequences to see if any homologies exist. The effect of adding gaps into the sequences is to increase the number of possible sequence comparisons. Thus the extra homology introduced by a gap must more than compensate for the increased number of comparisons. The numbers of possible comparisons of L amino acid pairs chosen from sequences of length N_1 and N_2 with various numbers and sizes of gaps on sequence 2 are given below.

Conditions	Number of Comparisons
no gaps	$(N_2-L+1)\ (N_1-L+1)$
1 gap of G residues	$(N_2-L+1)\left[(N_1-L+1-G)L+G\right]$
2 gaps of G residues each	$(N_2-L+1)\left[(N_1-L+1-2G)(L(L+1)/2)+G(L+1)\right]$
one gap of 1 and one gap of 2	$(N_2-L+1)\left[(N_1-L-2)L^2+3L\right]$

These equations are valid only as long as $N_2+G<N_1$

All of the results shown were calculated for gaps occurring in only one chain. The equations can easily be modified to include the effects of gaps in both protein sequences. In general, adding

gaps increases the number of comparisons by about L^{N_g}, where N_g is the number of gaps. Thus it can be seen that comparisons involving large numbers of gaps must be looked on with suspicion unless the existence of the gaps can be justified. This can be done either statistically, or by analogy to other similar proteins where the relative homologies are greater, or the gaps already proven.

The use of digital computers has greatly simplified the task of searching for similarities in protein sequences. We have modified the procedures developed by Fitch (1966) to include the placement of a gap at every possible position in a protein sequence. The results of comparing the first 40 residues of human hemoglobin α and β chains are shown in Figure 1. All possible sets of 16 consecutive amino acids were chosen. In each of these a single gap of either 1, 2 or 3 residues was placed at all possible positions and the resulting sequence was compared with all ungapped sequences on the other chain. From these calculations six distribution curves were produced. These are plotted on probability paper in Figure 2. This type of a plot would yield a straight line for a Gaussian distribution. All of our results show marked curvature which is indicative of a substantial homology between the α and β chains (Fitch, 1966). There is a significant difference between the distributions produced when a gap is located on the α chain as opposed to the β chain only if a gap of 2 is used. This gap, on β, shifts the distribution towards smaller numbers of base changes. These results provide strong confirmation for the existence of a gap of 2 amino acids in the β chain of human hemoglobin. They demonstrate the usefulness of computer methods in searching for the existence of gaps.

From the above it is evident that the existence of a gap

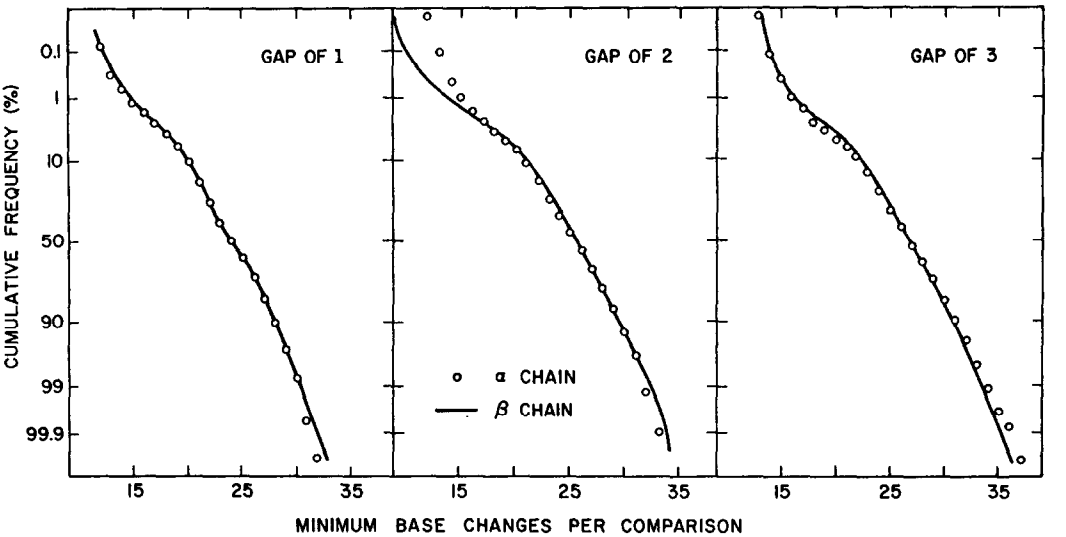

Figure 2. Results of comparing the first 40 residues of the α and β chains of human hemoglobin. Details are given in the text.

is most easily seen when the homology on both sides of the gap is very extensive. This means that gaps postulated to occur very near the ends of protein sequences seem destined to remain highly spectulative. Here the homology on one side of the gap can almost never be good enough to justify the gap.

In conclusion, it is recommended that considerable caution be used in postulating the existence of gaps in protein sequences.

Acknowledgment: The author has been greatly assisted by many informative discussions with Thomas Jukes.

REFERENCES

Braunitzer, G., Hilschmann, N., Rudloff, V., Hilse, K., Liebold, B. and Muller, R., Nature, 190, 480 (1961).
Brew, K., Vanamann, T.C. and Hill, R.L., J. Biol. Chem., 242, 3747 (1967).
Dunnill, P. Nature, 215, 621 (1967).

Eck, R.V. and Dayhoff, M.P., Atlas of Protein Sequence and Structure, Natl. Biomedical Res. Foundation, Silver Spring, Maryland, 1966.
Fitch, W.M., J. Mol. Biol., 16, 8 17 (1966).
Fitch, W.M. and Margoliash, E., Science, 155, 279 (1967).
Hartley, B.S., Brown, J.R., Kauffman, D.L. and Smillie, L.B., Nature, 207, 1157 (1965).
Hill, R.L., Delaney, R., Fellows, Jr., R.E. and Lebovitz, H.E., Proc. Natl. Acad. Sci., 56, 1762 (1966).
Ingram, V.M., The Hemoglobins in Genetics and Evolution, Columbia University Press, New York, 1963.
Jukes, T.H., Molecules and Evolution, Columbia University Press, New York, 1966.
Manwell, C., Comp. Biochem. Physio, 383 (1967).