HOMOLOGY BETWEEN THE α1 AND α2 CHAINS OF COLLAGEN

Karl A. Piez

National Institute of Dental Research
National Institutes of Health
Bethesda, Maryland 20014

Gary Balian, Eva Marie Click, Paul Bornstein

Department of Biochemistry
University of Washington
Seattle, Washington 98195

Received July 14,1972

SUMMARY

A sequence of 30 amino acid residues in the α2 chain of rat skin collagen was compared by computer techniques to the sequence of nearly 600 residues in the α1 chain. One region was found with a high degree of similarity that was unlikely to have arisen by chance. It was the same distance from the NH_2-terminus as the α2 sequence and the distribution of charged groups was nearly identical. We conclude that these regions of the α1 and α2 chains are homologous.

Most collagens that have been studied contain two chains of one type (α1) and one of another (α2) each with approximately 1000 amino acid residues. The three chains are parallel and extend the full length of the molecule. Since the α1 and α2 chains are similar in amino acid composition and have the same conformation (see Traub and Piez (1)), it is to be expected that they are closely related in an evolutionary sense and that homology should be demonstrable. Furthermore, electron optical analysis of renatured, collagen-like molecules composed of only one type of chain has demonstrated that the distribution of charged residues along the α1 and α2 chains is similar (2). However, resolution by this technique is at best equivalent to 5-10 amino acid residues and exact chemical identification is not possible. Comparisons of amino acid sequence data are therefore desirable to confirm the expected homology. We present such a comparison here.

The sequence of the first 238 residues in the α1 chain from rat skin or tendon collagen has been reported (3-5). The next 180 residues, which constitute

the COOH-terminal two-thirds of the cyanogen bromide peptide α1-CB8, have now been sequenced; these data will be published separately (6). Together these sequences permit the first 418 residues of the α1 chain to be assembled. The first 112 residues of α1-CB6, the last cyanogen bromide peptide in the α1 chain, have been sequenced (7). The sequence of the remaining 105 residues in α1-CB6 has been completed and made available to us (8); the primary structure of the COOH-terminal 217 residues in α1 is therefore known. The only published sequence from the helical region of the α2 chain is the 30 residues comprising the cyanogen bromide peptide α2-CB2 (9).

Comparison of collagen sequences for homology requires somewhat different criteria, at least in a quantitative sense, than used for other proteins. A high degree of similarity between regions may result by chance from the presence of glycine in every third position and the high concentrations of alanine and proline (or hydroxyproline which is derived from proline). Therefore, it is important to have a quantitative measure of similarity and some indication of statistical significance.

For this purpose we have utilized computer techniques based on concepts similar to those used by other workers including Fitch (10), Sackin (11), Haber and Koshland (12) and McLachlan (13). In the present study, the computer program compared the 30 residues of α2-CB2 of rat skin collagen to all possible 30-residue segments in the known parts of the α1 chain from rat and calf skin collagens. The use of data from two species is justified since species differences are small. The program makes the comparisons residue-by-residue using a score for each amino acid pair that is a measure of similarity and determines the average score for the given segment pair. Only segment pairs where the glycyl residues are aligned were considered since all others would obviously be much less similar. Deletions were not considered since they would have to occur in groups of three residues to maintain the triplet structure and therefore seem unlikely to be a common feature. For simplicity, the known sequences from the NH_2- and COOH-terminal portions were treated as a single sequence without a break. Some of

the 30-residue segments are therefore not real sequences, but their presence does not affect the conclusions. Omitting the sequences of 16 residues at the NH_2-terminus (3) and 25 residues at the COOH-terminus (14) where glycine is not every third residue and which therefore cannot be helical, the known sequence comprises 594 residues or 189 possible 30-residue segments taken in steps of three to maintain alignment of glycyl residues.

Similarity was measured by the use of the score matrix devised by McLachlan (13). This matrix contains a score for all possible amino acid pairs that reflects the frequency with which substitutions involving a given pair have been observed in known homologus proteins. In most cases, amino acids that are frequently substituted for one another are similar in size and chemical character (13). An identical pair (no substitution) has a score of 8 or 9; very similar pairs have scores of 4-6; dissimilar pairs have scores of 0-3. This matrix was devised from published data on globular proteins. The factors that enter into determining the frequency of these substitutions may not be exactly the same as those for collagen, but they are unlikely to be very different.

The output of the computer program showed those segment pairs with scores above a selected value, the average similarity score, and the frequency with which scores were observed within intervals of 0.2. Hydroxyproline and hydroxylysine were treated as if they were proline and lysine, respectively.

To measure statistical significance, the normal distribution of similarity scores was obtained by comparing the sequence of α2-CB2 to "random" sequences generated by scrambling the real sequence with the use of a random number generator in a computer program. The scrambling was done in such a way that glycine was retained in every third position. Since there is evidence that certain other amino acids are restricted with regard to their distribution in the triplet Gly-X-Y (5,6), scrambling was done within positions X and Y of the triplet but not between them.

Comparisons of α2-CB2 to 20 different scrambled sequences for a total of 189 x 20 or 3,780 comparisons were done to give statistically valid data. The

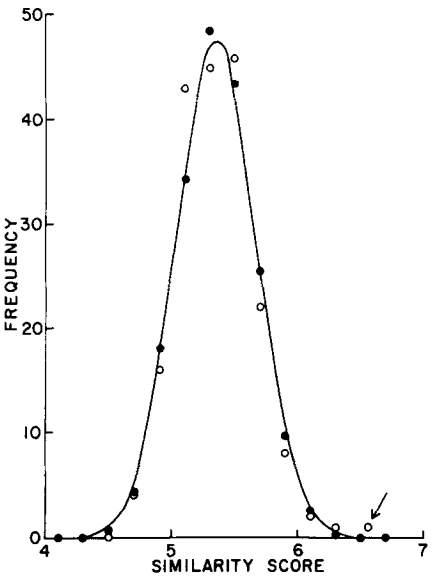

Fig. 1.

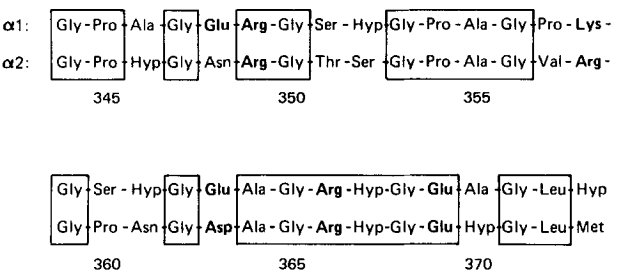

Fig. 2.

Fig. 1. The distribution of similarity scores obtained from a comparison of the sequence of α2-CB2 from the α2 chain of rat skin collagen to all possible 30-residue segments in 594 residues of the α1 chain. The frequency indicates the number of observations in score intervals of 0.2. The open circles show the actual distribution. The closed circles show the normal distribution obtained by comparison of α2-CB2 to 20 scrambled sequences; frequencies are divided by 20. The solid line shows a Gaussian distribution with the same mean and standard deviation. The score for the segment within the α1 chain that is homologous to α2-CB2 is indicated by an arrow.

Fig. 2. A comparison of the sequence of α2-CB2 from the α2 chain of rat skin collagen to residues 344 through 373 in the α1 chain. Residues that are identical in the two chains are enclosed in rectangles. Charged residues are in bold face.

---

average distribution of similarity scores is shown in Fig. 1. The mean score was 5.35 and the standard deviation was 0.315. A Gaussian distribution with these parameters was found to fit the scrambled data reasonably well. The differences are real (Fig. 1) and can be ascribed to the fact that the similarity scores for amino acid pairs show only a rank order and are not statistically related. However, the Gaussian distribution is sufficiently close to be used for an estimation of statistical significance.

The similarity scores obtained from a comparison of α2-CB2 to the 189

possible 30-residue segments in the real α1 sequence were distributed in a manner that was not significantly different from the Gaussian distribution with the exception of the score for one pair (Fig. 1). This pair has a similarity score of 6.56 which lies nearly four standard deviations from the mean and has a probability of chance occurrence of about 5×10^{-5}. It should be noted that this probability is not as low as is usually observed for homologous sequences (see, for example, ref. 13). However, it is probably significant in view of the diluting effect of the glycyl residues which constitute one-third of all segments and are aligned in even the most dissimilar segment pairs.

The sequence of α2-CB2 and the α1 segment that is very similar to it are shown in Fig. 2. Of the 30 residues, 19 are identical (9 out of 20 if the glycines are omitted) and many of the rest are similar in size and/or chemical properties. Particularly striking is the distribution of charged groups. The three basic residues in the α1 chain correspond exactly in position to the three basic residues in the α2 chain. One of the two acidic residues in the α1 chain corresponds to an acidic residue in the α2 chain; the other does not but is next to a basic residue. This nearly identical distribution of charged groups is consistent with the elctron optical analysis of the chains (2) and provides a firm chemical basis for it.

The sequence of α2-CB2 from chick skin collagen is the same as rat skin collagen except residues 3, 8 and 15 which are Ala, Ala and Lys, respectively (9). If these substitutions are made (residues 346, 351 and 358 in Fig. 2), it can be seen that the similarity between the two chains, one from rat and the other from chick, is even closer (21 out of 30 residues are identical for a score of 6.76) than when both chains are from rat.

It can be calculated that the two sequences come from the same region of their respective chains. Preceding α2-CB2 in the α2 chain are α2-CB1, 0 and 4 which together constitute about 345 residues (15). This is well within experimental error of the 343 residues which are known from the sequence data to precede the similar region in the α1 chain. In view of these observations and

the low probability of chance occurrence, we conclude that these regions of the α1 and α2 chains have evolved from a common precursor.

ACKNOWLEDGMENTS

Supported in part by NIH contract 69-2230 and NIH grant AM 11248. Paul Bornstein is the recipient of PHS research career development award K4-AM-42582.

REFERENCES

1. W. Traub and K. A. Piez, Adv. Prot. Chem., 25, 243 (1971).
2. C. Tkocz and K. Kuhn, Eur. J. Biochem., 7, 454 (1969).
3. A. H. Kang, P. Bornstein and K. A. Piez, Biochemistry, 6, 788 (1967).
4. P. Bornstein, Biochemistry, 6, 3082 (1967); W. T. Butler, Biochemistry, 9, 44 (1970); W. T. Butler and S. L. Ponds, Biochemistry, 10, 2076 (1970).
5. G. Balian, E. M. Glick and P. Bornstein, Biochemistry, 10, 4470 (1971).
6. G. Balian, E. M. Glick, M. Hermodsen and P. Bornstein, Biochemistry, in press.
7. K. von der Mark, P. Wendt, F. Rexrodt and K. Kuhn, FEBS Lett., 11, 105 (1970).
8. K. Kuhn, personal communication.
9. J. H. Highberger, A. H. Kang and J. Gross, Biochemistry, 10, 610 (1971).
10. W. M. Fitch, J. Mol. Biol., 16, 9 (1966); 49, 1 (1970).
11. M. J. Sackin, Biochem. Genetics 5, 287 (1971).
12. J. E. Haber and D. E. Koshland, J. Mol. Biol., 50, 617 (1970).
13. A. D. McLachlan, J. Mol. Biol., 61, 409 (1971).
14. J. Rauterberg, P. Fietzek, F. Rexrodt, U. Becker, M. Stark and K. Kuhn, FEBS Lett., 21, 75 (1972).
15. P. P. Fietzek and K. A. Piez, Biochemistry, 8, 2129 (1969).