CONFORMATIONAL IMPLICATIONS OF AMINO ACID SEQUENCE REGULARITIES IN COLLAGEN

Guido SALEM and Wolfie TRAUB

Department of Structural Chemistry, Weizmann Institute of Science, Rehovot, Israel

Received 9 December 1974

1. Introduction

It is well established that the amino acid sequence of a protein determines the folded conformation appropriate to its biological function. However, investigations of such relationships have shown them to be very complex and to involve sequences of widely varying specificity [1]. We have therefore chosen to investigate the skeletal protein collagen, which has a fairly uniform conformation in accordance with its primarily structural function, in the expectation that it may indicate relatively simple relationships between sequence and conformation.

Collagen has long stiff rod-like molecules consisting of three polypeptide chains wound about a common axis in a triple helical structure [2]. Each chain has a little over one thousand amino acid residues, including glycine in every third position, except near the ends of the chains, and unusually large amounts of the imino acids proline and hydroxyproline. In many types of collagen all three chains have identical sequences, but the predominant type of collagen in most vertebrate tissues has two identical chains, designated α_1, and a third, α_2, of somewhat different sequence. The three chains all extend over the full length of the molecule and have their N-terminal ends on the same side.

The polypeptide chains are themselves coiled as well as twisted about a common axis in a rope-like structure, and the left-handed helical symmetry of the molecule relates the structurally equivalent Gly X.Y. tripeptides by a translation of 2.9Å and a rotation of some 108°. Because every third residue lies near the central axis of the molecule there is room in this position for only the smallest amino acid, glycine. Residues in both the X and Y positions lie on the surface of the molecule, but differ in their backbone conformations and intramolecular stereochemical environments. The three chains are joined by systematic hydrogen bonding between NH groups of glycine residues and CO groups of residues in X positions (fig.1).

In recent years there have been extensive determinations of amino acid sequences of several vertebrate collagens covering the complete length of the α_1 chains [3–12], and it has emerged that several residues are non-randomly distributed between positions X and Y [7,9,13]. We have now analysed the α_1 sequence for preferred distributions of the various amino acid residues, not only between X and Y positions, but also in terms of possible side-chain pair interactions. We have discovered several of these, and, by examining models of the molecular conformation, we have found some possible stereochemical explanations for them.

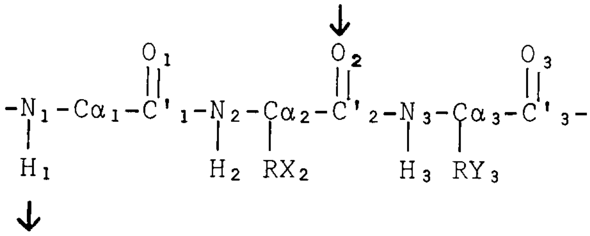

Fig.1. Tripeptide sequence in collagen indicating notation used intext and location of interchain hydrogen bonds.

2. Technical details

In our analysis, we have assumed that the amino acid sequences of the helical portions of all three

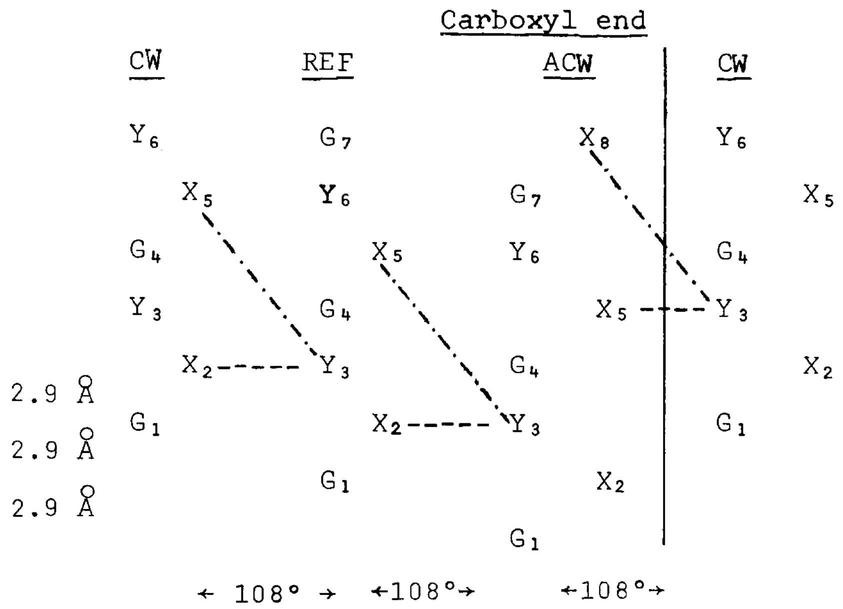

Fig.2. Schematic illustration of spatial relationships between amino acid residues in the three polypeptide chains of the collagen molecule. REF, CW and ACW indicate reference, clockwise and anti-clockwise chains, as viewed from the carboxyl end of the molecule.

chains are identical, and we have used a combination of α_1-chain sequences found for positions 1–402 in rat skin collagen and 403–1011 in calf skin collagen [14]. Partial sequence determinations indicate that differences between rat and calf skin collagen are very small [15], but that α_1 and α_2 chains may differ in about half of the X and Y positions [16,17].

Equivalent residues on the three chains are related by 2.9Å translation and 108° rotation, as illustrated in fig.2. It can be seen from this figure that, for example, X_2–Y_3 interactions relating the clockwise to reference, and reference to anticlockwise, chains are equivalent to X_5–Y_3 interactions between anticlockwise and clockwise chains.

The molecular conformation used in our model studies is that derived from X-ray structure analyses of $(\text{Gly Pro Pro})_n$ and several other collagen-like polytripeptides [18–20], which is evidently a very good approximation to the collagen structure [2]. This has recently been borne out by a comparison of the very detailed collagen X-ray pattern shown by stretched rat tail tendon with patterns calculated for various triple helix conformations, which showed much better agreement for the $(\text{Gly Pro Pro})_n$ structure than for any of the alternative models [21,22].

A CPK [23] space-filling molecular model was built with about ten residues on each of the three chains. With the aid of angular dials [24] on the amide nitrogen and carbonyl carbon atoms, the dihedral angles ϕ and ψ were adjusted so as to fix the polypeptide backbone in the $(\text{Gly Pro Pro})_n$ conformation, and the chains were then assembled into the triple helix structure by connecting the interchain hydrogen bonds. Various residues were attached in positions X and Y and their steric relationships investigated by rotating about side-chain single bonds without, however, altering the backbone dihedral angles.

3. Results and discussion

Table 1 shows the distribution of the various amino acid residues between X and Y positions, and it can be seen that several of them are far from evenly divided. It is known that proline residues are hydroxylated after polypeptide chain synthesis by an enzyme which acts specifically at the Y position [25]; hence the uneven distribution of proline and hydroxyproline residues. Enzymatic conversion of lysine to hydroxylysine also occurs only at the Y position, but, in skin-type collagen, to a much smaller degree than hydroxylation of proline [26,27]. The restriction of all 12 phenylalanines, and all but one of the 19 leucines, to X-positions has been explained in terms of severe steric hindrance for these residues in position Y [28].

There are also several amino acids, including arginine, glutamic acid, lysine, glutamine and threonine, which though not 'forbidden' in either position evidently favour one of them. We have considered whether in such situations the preferred sequence may allow interactions which serve to stabilise the molecular conformation. These could be either interactions between side chains and backbone or between side chains and side chains. The latter situation would require that the amino acid sequence cause specific pairs of side chains to occur in steric proximity to each other, and table 2 shows an analysis of the sequence data for such possible pair preferences, several of which evidently exist.

We have used the molecular model to search for possible steric interactions which might account for the various sequence regularities indicated by tables 1 and 2, and have found several plausible explanations.

One of the simpler situations concerns glutamine, which shows a preference for the Y position, but no specific mode of pairing with any other side chain. When glutamine is located in the Y_3 position the amino group at the end of its side chain can form a hydrogen bond to the backbone C_3O_3 on the clockwise chain, as shown in fig.3a. This figure also shows that asparagine, which shows no preference between X and Y positions, has too short a side chain to make this type of interchain hydrogen bond.

Arginine and lysine, on the other hand, do have sufficiently long side chains for their amino groups to bind in this way, and their preference for position Y is also consistent with such hydrogen bonding. However, table 2 strongly indicates that these positively charged residues interact specifically with negatively charged glutamic and aspartic acids. An interchain interaction between arginine at position Y_3 and glutamic acid on the clockwise chain at X_2 could explain the observed X_2-Y_3 and X_5-Y_3 preferences, given equivalent interactions between all three polypeptide chains (see Section 2). Lysine and glutamic acid show X_5-Y_3 and X_8-Y_3, but no X_2-Y_3 preferences, indicating an interaction between lysine at position Y_3 on the reference chain and glutamic acid at X_5 on the clockwise chain. The pair preferences in table 2 also suggest a similar interaction between lysine at Y_3 and aspartic acid at X_5 on the clockwise chain, and possibly also an interchain interaction between arginine at Y_3 and aspartic acid at X_2 on the clockwise chain.

Table 1
Distribution of amino acid residues between X and Y positions

Amino acid	Position X	Postition Y
Pro	116	4
Hyp	1	112
Phe	12	0
Leu	18	1
Arg	9	42
Lys	12	20
Gln	8	18
Asn	7	4
Glu	38	6
Thr	3	13
Ser	17	18
Asp	14	15
Ala	60	61
Val	9	8
Ile	3	4
Met	2	5
His	2	0
Hyl	0	4

We have been able to build satisfactory models incorporating hydrogen bonding between terminal amino and carboxyl groups for all four of these interactions, but we have also found it possible to devise hydrogen bonding schemes with the basic and acidic residues arranged in other sequences. However, the different modes of pairing are not sterically equivalent and it appears that only the ones we have described would allow the lysine and arginine residues to make hydrogen bonds simultaneously to an acidic side chain and to C_3O_3 on the clockwise polypeptide chain (fig.3b). This could account for the preference of arginine and lysine for position Y and indirectly for the glutamic acid bound to these residues occurring in position X.

These pairings of acidic and basic side chains clearly involve charge as well as hydrogen bonding interactions, and there seems to be some tendency for oppositely charged side chains to occur in proxi-

Table 2
Distribution of neighbouring amino acid residues

Y_3	$X_{2,5,8}$	Pro (116)	Ala (60)	Glu (38)	Leu (18)	Ser (17)	Asp (14)	Lys (12)	Phe (12)	Arg (9)	Val (9)	Gln (8)	Asn (7)
Hyp (112)	2	39	20	11	11	10	0	1	7	5	2	2	1
	5	46 [38.6]	21 [19.9]	13 [12.6]	3 [6.0]	7 [5.6]	4 [4.7]	4 [4.0]	3 [4.0]	1 [3.0]	0 [3.0]	3 [2.7]	2 [2.3]
	8	31	23	12	6	6	5	4	4	5	4	3	3
Ala (61)	2	31	6	7	1	3	5	1	1	0	1	1	1
	5	24 [21.0]	11 [10.9]	6 [6.9]	3 [3.3]	2 [3.1]	0 [2.5]	5 [2.2]	1 [2.2]	3 [1.6]	2 [1.6]	2 [1.4]	1 [1.3]
	8	20	11	10	0	3	4	3	3	2	2	0	0
Arg (42)	2	8	9	10	0	1	4	1	0	0	2	3	0
	5	6 [14.5]	4 [7.5]	9 [4.7]	4 [2.2]	3 [2.1]	2 [1.7]	0 [1.5]	5 [1.5]	2 [1.1]	3 [1.1]	1 [1.0]	0 [0.9]
	8	21	3	3	5	2	2	0	2	0	0	0	0
Lys (20)	2	7	9	1	0	1	1	0	0	0	0	0	1
	5	3 [6.9]	4 [3.6]	5 [2.3]	1 [1.1]	1 [1.0]	5 [0.8]	0 [0.7]	0 [0.7]	0 [0.5]	0 [0.5]	0 [0.5]	1 [0.4]
	8	5	5	6	0	1	2	0	0	0	0	0	0
Gln (18)	2	7	1	3	2	1	0	1	2	0	1	0	0
	5	9 [6.2]	3 [3.2]	0 [2.0]	1 [1.0]	0 [0.9]	0 [0.7]	0 [0.6]	0 [0.6]	0 [0.5]	2 [0.5]	0 [0.4]	0 [0.4]
	8	5	3	3	2	2	1	0	0	0	0	1	0
Ser (18)	2	10	3	1	0	0	0	1	2	0	0	0	0
	5	5 [6.2]	4 [3.2]	3 [2.0]	3 [1.0]	0 [0.9]	0 [0.7]	0 [0.6]	0 [0.6]	1 [0.5]	0 [0.5]	0 [0.4]	0 [0.4]
	8	9	3	3	1	0	0	1	0	0	0	0	0
Asp (15)	2	1	5	0	1	0	2	3	0	1	0	1	0
	5	1 [5.2]	4 [2.7]	2 [1.7]	1 [0.8]	1 [0.8]	0 [0.6]	1 [0.5]	0 [0.5]	1 [0.4]	2 [0.4]	1 [0.4]	0 [0.3]
	8	3	6	0	2	0	1	2	0	0	0	0	0
Thr (13)	2	2	1	3	2	0	1	1	0	1	0	0	0
	5	8 [4.5]	1 [2.3]	0 [1.5]	0 [0.7]	1 [0.7]	0 [0.5]	0 [0.5]	1 [0.5]	0 [0.3]	0 [0.3]	0 [0.3]	0 [0.3]
	8	6	2	1	0	1	0	0	1	1	0	0	0
Val (8)	2	2	2	1	0	0	0	0	0	1	1	0	0
	5	5 [2.8]	2 [1.4]	0 [0.9]	0 [0.4]	0 [0.4]	0 [0.3]	0 [0.3]	0 [0.3]	0 [0.2]	0 [0.2]	0 [0.2]	0 [0.2]
	8	4	1	0		0	0	2	0	0	0	0	0
Gln (6)	2	0	1	1	0	1	0	1	0	1	1	0	0
	5	2 [2.1]	1 [1.1]	0 [0.7]	0 [0.3]	2 [0.3]	0 [0.2]	0 [0.2]	0 [0.2]	0 [0.2]	0 [0.2]	0 [0.1]	0 [0.1]
	8	2	0	0	0	0	0	0	0	1	0	0	0

For each amino acid residue at Y_3, shown in the left-hand column, the corresponding three rows show that the numbers of the various residues occurring at positions X_2,X_5 and X_8 in the polypeptide chain sequence (cf. fig.2). Figures in ( ) show total numbers at position X or Y, and figures in [ ] the number of pairs, N, expected for a random distribution of residues among the 337 tripeptides (eg. for ProHyp N=(116) (112)/337=38.6). Large deviations from N are underlined.

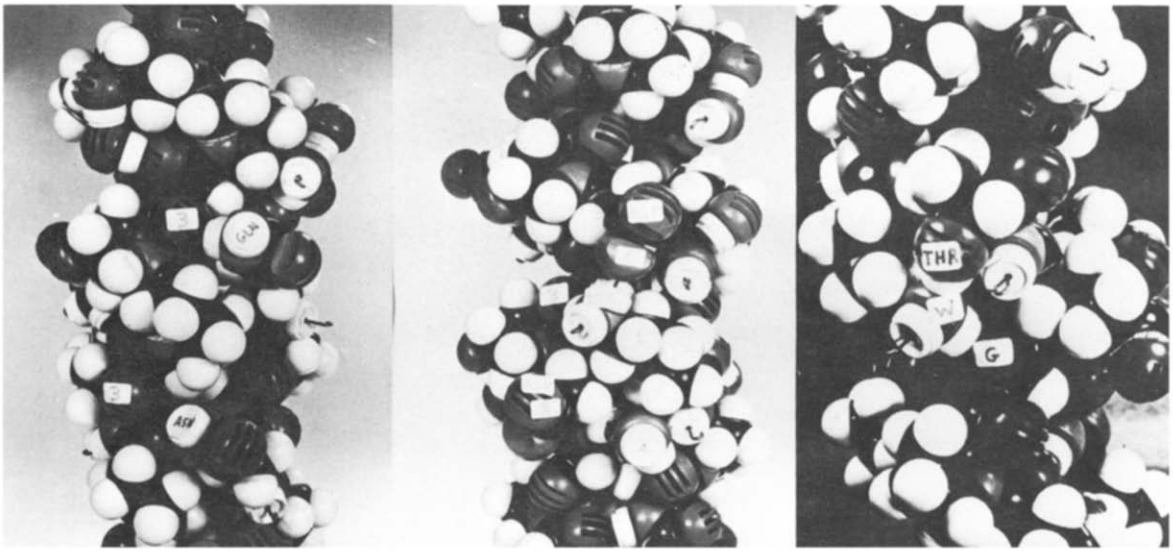

Fig.3. Three views of space-filling model of collagen illustrating suggested intramolecular interactions. White atoms are hydrogen, black atoms carbon, rounded corrugated grey atoms oxygen and wedge-shaped grey atoms nitrogen. The three photographs show: (a) The terminal amino group of glutamine at Y_3 hydrogen bonded to the backbone C_3O_3 on the clockwise chain. (b) The terminal amino group of lysine at Y_3 hydrogen bonded simultaneously to the terminal carboxyl group of aspartic acid at X_5 and the backbone C_3O_3, both on the clockwise chain. This view also illustrates the impossibility of simultaneous hydrogen bonding to C_3O_3 and aspartic acid at X_2. (c) A water bridge (W) linking the OH of threonine at Y_3 to the glycyl C_1O_1 on the same chain.

mity in other sequences as well (e.g. aspartic acid at Y_3 with lysine at nearby X positions), whereas side chains with the same charge occur close together less frequently than might be expected on a random basis. (see table 2).

We also found a mode of interchain hydrogen bonding for aspartic acid in the Y position connecting its side-chain carboxyl group and the backbone NH at X_2 on the clockwise chain, provided this residue is not a proline. This could account for the infrequent occurrence of X_2-Y_3 and X_5-Y_3 proline–aspartic–acid pairs (table 2).

The preference of threonine for the Y position can be explained by the formation of a hydrogen-bonded water bridge linking the threonine OH group at Y_3 to C_1O_1 on the same chain (fig. 3c). The threonine side chain is fixed in the orientation appropriate to this bond by the methyl group, which sits in a hydrophobic pocket formed by C_3' of the threonine, C_β of X_2 on the clockwise chain, and C_γ and C_δ of a proline residue at X_5 on the clockwise chain. As can be seen from table 2, there is indeed a tendency for proline to pair with threonine in this way. A similar intrachain water bridge involving the OH of hydroxyproline has recently been suggested to account for this residue's role in stabilising the triplex helix conformation [28,29].

We have also investigated frequencies of side-chain pairs not appearing in table 2, including X_2-Y_5, X_2-Y_6 and Y_3-Y_6, and have found some indications of additional regularities, notably between charged side chains.

We cannot claim to have completely explained the regularities in the amino acid sequence. The model-building approach can indicate possible, and impossible, interactions, but can not assess the energy differences between alternative possibilities nor provide proof of the existence of any interactions, which may have to await X-ray analyses of crystalline fragments of collagen.

We have considered the possibility of correlating the amino acid sequence with functional aspects of collagen structure other than stability of the molecular conformation. There are some indications of preferred sequences near sites of enzymic activity, including the absence of aspartic acid immediately preceding hydroxyproline (cf. table 2), the common occurrence of arginine at the Y position following glycosylated hydroxylysine [30,31] and of hydrophobic residues at the two X positions preceding it. However, it appears that the sequence regularities we have discussed above are independent of hydroxyproline or hydroxylysine positions. There are also long range sequence regularities which direct the selfassembly of collagen molecules to form fibrils [14,32], presumably through intermolecular contacts at interacting edges [33]. Consequently, we have analysed the distribution of the various amino acid residues among the ten different X and ten different Y orientations implicit in the helical symmetry of the molecule. This shows that the very uneven distributions between X and Y positions are independent of azimuthal direction, so that concentrations of particular residues at interacting edges could at most make only a rather small contribution to this effect.

Therefore, to summarise our conclusions, it appears that the amino acid sequence of collagen is far from random in the X and Y positions, that many of the sequence regularities arise from side-chain specific intramolecular interactions, and that, estimating from the excess of residues in the X or Y positions, a third or more of the tripeptide units may achieve additional stabilisation in the tryplet helix conformation through the kinds of interactions we have described.

Acknowledgements

We are indebted to Drs Paul Bornstein, Peter Fietzek, Klaus Kühn and Bernard Rigby for stimulating discussions and to Dr Ada Yonath for help in computing the sequence analysis. This investigation was supported by research grant Az 11-1790 from Stiftung Volkswagenwerk.

References

[1] Dickerson, R. E. (1972) Scientific American 226, 58–72.

[2] Traub, W. and Piez, K. A. (1971) Advan. Protein Chem. 25, 243–352.

[3] Kang, A. H., Bornstein, P. and Piez, K. A. (1967) Biochemistry 7, 788–795.

[4] Bornstein, P. (1967) Biochemistry 6, 3082–3093.

[5] Butler, W. T. (1970) Biochemistry 9, 44–50.

[6] Butler, W. T. and Ponds, S. L. (1971) Biochemistry 10, 2076–2081.

[7] Balian, G., Click, E. M. and Bornstein, P. (1971) Biochemistry 10, 4470–4478.

[8] Balian, G., Click, E. M., Hermodsen, M. and Bornstein, P. (1972) Biochemistry 11, 3798–3806.

[9] Fietzek, P. P., Wendt, P., Kell, I. and Kühn, K. (1972) FEBS Lett. 26, 74–76.

[10] Fietzek, P. P., Rexrodt, F., Hopper, K. and Kühn, K. (1973) Eur. J. Biochem. 38, 396–400.

[11] Wendt, P., von der Mark, K., Rexrodt, F. and Kühn, K. (1972) Eur. J. Biochem. 30, 169–183.

[12] Rauterberg, J., Fietzek, P. P., Rexrodt, F., Becker, U., Stark, M. and Kühn, K. (1972) FEBS Lett. 21, 75–79.

[13] Fietzek, P. P., Rexrodt, F., Wendt, P., Stark, M. and Kühn, K. (1972) Eur. J. Biochem. 30, 163–168.

[14] Hulmes, D. J. S., Miller, A., Parry, D. A. D., Piez, K. A. and Woodhead-Galloway, J. (1973) J. Mol. Biol. 79, 137–148.

[15] Fietzek, P. P. and Kühn, K. Eur. J. Biochem. (In press).

[16] Fietzek, P. P., Kell, I. and Kühn, K. (1972) FEBS Lett. 26, 66–68.

[17] Fietzek, P. P. and Kühn, K. (1973) FEBS Lett. 36, 289–291 and private communication.

[18] Yonath, A. and Traub, W. (1969) J. Mol. Biol. 43, 461–477.

[19] Segal, D. M., Traub, W. and Yonath, A. (1969) J. Mol. Biol. 43, 519–527.

[20] Traub, W., Yonath, A. and Segal, D. M. (1969) Nature 221, 914–917.

[21] Traub, W. and Salem, G. (1972) Acta Cryst. A28, S38.

[22] Salem, G. (1973) M.Sc. Thesis, Weizmann Institute, Rehovot.

[23] Koltun, W. L. (1965) Biopolymers 3, 665–679.

[24] Hauschka, P. V. and Segal, D. M. (1966) Biopolymers 4, 1051–1052.

[25] Hutton, J. J., Kaplan, A. and Udenfriend, S. (1967) Arch. Biochem. Biophs. 121, 384–391.

[26] Butler, W. T. (1968) Science 161, 796–798.

[27] Miller, R. L. (1971) Arch. Biochem. Biophys. 147, 339–342.

[28] Traub, W. (1974) Isr. J. Chem. 12, 435–439.

[29] Ramachandran, G. N., Bansal, M. and Bhatnager, R. S. (1973) Biochim. Biophys. Acta 322, 166–171.

[30] Morgan, P. H., Jacobs, H. G., Segrest, J. P. and Cunningham, L. W. (1970) J. Biol. Chem. 245, 5042–5048.

[31] Isemura, M., Ikenaka, T. and Matsushima, Y. (1973) J. Biochem. (Tokyo) 74, 11–21.

[32] Doyle, B. B., Hukins, D. W. L., Hulmes, D. J. S., Miller, A., Rattew, C. J. and Woodhead-Galloway, J. (1974) Biochem. Biophys. Res. Commun. 60, 858–864.

[33] Segrest, J. P. and Cunningham, L. W. (1973) Biopolymers 12, 825–834.