MOLECULAR ASSEMBLY IN COLLAGEN

Wolfie TRAUB

Laboratory of Biochemistry, N.I.D.R. National Institutes of Health, Bethesda, MD 20014, USA and **Department of Structural Chemistry, The Weizmann Institute of Science, Rehovot, Israel*

Received 11 April 1978
Revised version received 17 May 1978

1. Introduction

Although the triple helical molecular conformation of collagen has been known for many years, the manner in which the molecules are packed together in native fibrils has continued to be the subject of much investigation. Collagen molecules consist of three polypeptide chains, each having a little over one thousand residues, including glycine in every third position except near the telopeptide ends of the chains. In many types of collagen all three chains have identical sequences, but the predominant type of collagen in most vertebrate tissues (type I) has two identical chains designated α1, and a third, α2, of somewhat different sequence [1]. Extensive work on collagens from several animal species has established essentially the entire amino acid sequence of type I collagen and large portions of the sequences of types II and III [2–5]. The three α chains form right-handed helices in coiling about a common molecular axis in a rope-like conformation. However, the structurally equivalent Gly–X–Y tripeptides in different chains are related by left-handed helical symmetry with a translation of 2.9Å and a rotation of 110±2° [6]. The experimental uncertainty in this last parameter allows a range of 30 to 45 residues per turn for the pitch of the α chains.

In native fibrils, the approximately 3000Å long collagen molecules are known to be staggered by multiples of a distance, D, equal to 670Å and corresponding to about 234 residues. Various lines of evidence further suggest that the molecules are first assembled, with a D stagger into thin structural units, termed microfibrils, which are probably supercoiled and are further associated to form fibrils [7]. Several analyses have been made of the amino acid sequence of the α1(I) chains with the aim of identifying the residues whose interactions determine the ordered molecular assembly in fibrils, and these showed periodicities of D for large hydrophobic and for charged residues as well as some shorter periodicities including $D/6$ and $2D/11$ [8–10]. However, attempts to define the three-dimensional packing geometry in terms of intermolecular interactions between various residues [11,12,13] have not, up to now, elucidated the origin of these shorter sequence regularities.

This paper describes a further sequence analysis based on data for α2 as well as the α1 chains of collagen types I, II and III. Attention was particularly focused on those potentially interacting residues which are conserved [3] in all α1-type chains and whether or not they are also conserved in α2. This analysis has thrown new light on the structural significance of the sequence regularities and has led to a detailed model for the ordered assembly of collagen molecules, which can account for these regularities as well as a wide range of chemical, X-ray and electron micrograph data.

2. Amino acid sequence analysis

Amino acid sequence data for α1(I), α1(II), α1(III) and α2 collagen chains from several animal species [2,3,4,5] have been compared to determine features of different degrees of variability.

* Present address

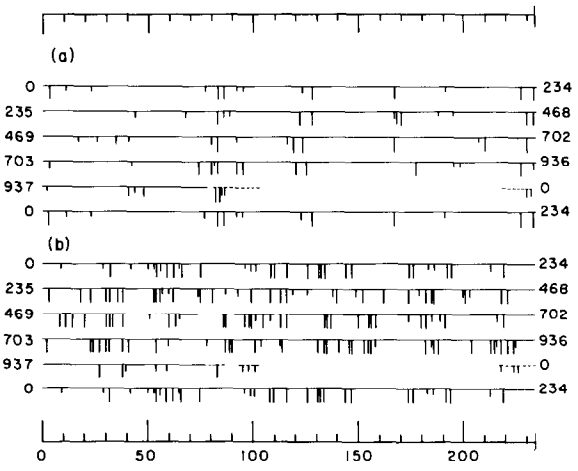


Fig.1. Sequence positions of amino acid residues which are conserved to various degrees in different collagen chains. (i) Occur in all known α1-type and α2 sequences. (ii) Occur in all α1-type chains, but not α2. (iii) Occur in α1 (I) chains. (a) shows positions of large hydrophobic residues including Phe, Met, Ile, Leu and Val; (b) shows positions of charged residues including Arg, Lys, Glu and Asp. Dashed lines indicate terminal non-helical regions of the collagen molecule. Residues are numbered from the beginning of the helical region.

Figure 1a shows the positions of large hydrophobic residues in the α1(I) sequence and, in particular, those which are conserved in all known sequences of α1-type chains or in both α1-type and α2 chains. Figure 1b shows analogous data for charged amino acid residues.

It is clear from fig.1a that the positions of large hydrophobic residues are far from random. They fill less than 10% of the non-glycine positions in α1(I), but half of these positions are conserved in all known sequences, though more often than not these include substitutions of one large hydrophobic residue for another. There is a concentration of large hydrophobics conserved in α1 and α2 into four groups with positions around 83+n234, 125+n234, 167+n234 and 233+n234 (where n=1,2,3 or 4). In other words, potentially interacting hydrophobic groups are separated not only by D, but also by $2D$, $3D$ and even $4D$ when the hydrophobic regions of the telopeptides are included. This is suggestive of nearest- and second-nearest-neighbour interactions between molecules, as might occur if these regions corresponded to nuclei of strong hydrophobic interactions between collagen molecules in a five-stranded microfibril. The high degree of conservation of large hydrophobic residues in these regions is quite remarkable given that the α1(I) and α2 sequences differ in about 50% of non-glycine positions and α1(I) and α1(III) in some 63% [14].

These residues also show a shorter periodicity of about 42 residues, equal to $2D/11$, (see fig.2), but this is not a regular repeat throughout the sequence, as has previously been reported from analyses of the α1(I) sequence alone [8–10,13].

There are also indications in fig.1a of additional groups of large hydrophobic residues in α1(I) around 41+n234 and 191+n234, which are not always conserved in other sequences. These residues are also consistent with the intermittent $2D/11$ periodicity.

Charged amino acid residues are also highly conserved. They comprise almost 25% of non-glycine residues in α1(I) and 70% of these are conserved in all known sequences (fig.1b). Many of these conserved charged residues are separated from each other by D or multiples of D, but they are much more broadly distributed than the conserved hydrophobics. In fact, conserved and unconserved charged residues do not appear to be distributed differently. Both tend to fall in charged regions which alternate with uncharged regions with a strong periodicity of about 39 residues equal to $D/6$ (fig.2), as has previously been reported for α1(I) chains [9,10]. The conserved large hydrophobics fall in the uncharged regions.

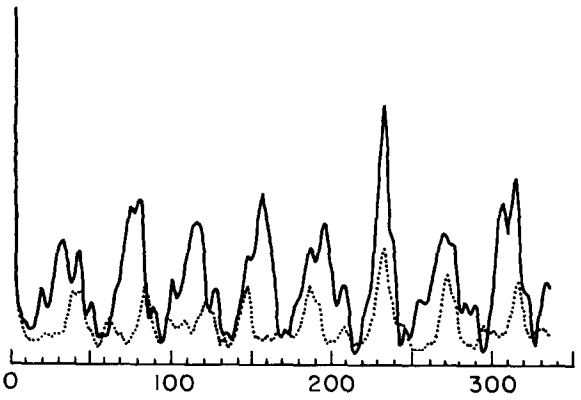


Fig.2. Distribution of inter-residue separations for residues conserved in α1-type (and generally also α2) chains. Data for large hydrophobic residues and charged residues are shown by dotted and full lines respectively.

Figure 1b also shows a class of residues which are charged in α1(I) but not in α2. These tend to be clustered in several groups showing an approximate $n234$ separation with positions around $60+n234$, $100+n234$, $180+n234$ and $215+n234$. These residues correspond quite well to the $D/6$ periodicity.

There are only 15 residues which are charged in α2 but not in α1(I) and they show no strong periodicity. There are about 70 large hydrophobic residues which occur only in α2 and these show a weak tendency to be separated by multiples of 14 and 21. Analyses of periodicities of various groups of residues were made with the aid of a programme, written by Dr Benes Trus, which can be used to sort out frequently occuring residue separations (fig.2).

3. Geometry of molecular packing

It is possible to postulate a detailed model of molecular assembly that can account for the sequence regularities described above as well as a great many experimental results concerning the structure of collagen.

The data shown in fig.1a strongly suggest that the conserved large hydrophobic residues play a role in stabilizing a D stagger between collagen molecules, and that both α1 and α2 chains are involved in such stabilizing interactions. Equivalent residues on α1 and α2 chains (eg. α1Phe317 and α2Leu317) presumably lie on complementary interacting edges which, because of the molecular conformation of collagen [6], would have an azimuthal separation of either about 110° or 140° (fig.3).

Equivalent intermolecular contacts at intervals of $D/6$ along each α chain would cause conserved hydrophobic residues separated by D (e.g. 83, 317, 551, 785, 1018) to be lined up along a single interacting edge in accordance with the arrangement of fig.3. Alternatively, one might consider other submultiples of D as intercontact intervals or even divisions of D by $n\pm1/3$, which would correspond to large hydrophobics from all three chains being lined up along each interacting edge. However, only one of these other intervals, $D/5.67 = 41.3$ residues, is close to an observed periodicity in the amino acid sequence and this fits poorly with the distribution of charged residues. On the other hand, the distribution of large

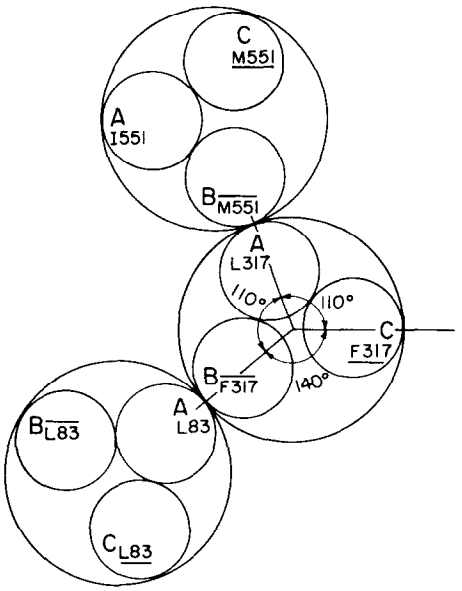


Fig.3. Cross-sectional view, from N-terminal end, of D-staggered collagen molecules associated through hydrophobic interactions involving both α2 chains (at A edges) and α1 chains (at B edges). Residues are identified by the one-letter amino acid code, and lines below or above the sequence number distinguish the N- and C-terminal α1 chains, respectively.

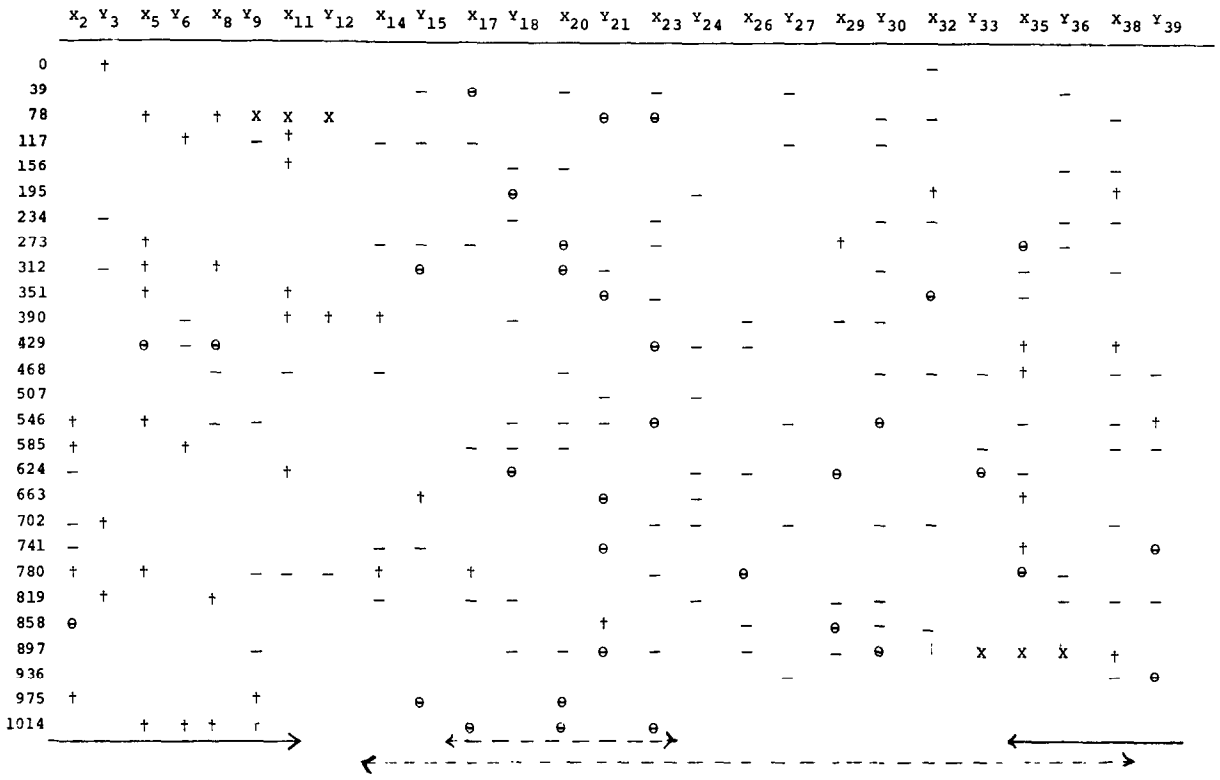
hydrophobic residues fits in well with a $D/6$ periodicity, even though their 'local' separation is $2D/11$. As is shown in table 1, in an α chain with a 'pitch'* of 39 residues per turn all the conserved hydrophobics would be confined to about a third of the azimuth and the conserved charges would be largely confined to the remaining two thirds. I have therefore chosen $D/6$ as the intercontact interval. This conclusion is also supported by the results of a recent comprehensive statistical analysis of potential intermolecular interactions in collagen [15].

Figure 4 illustrates the azimuthal distribution of large hydrophobic residues and potential intermolecular crosslink sites in a collagen molecule composed of three identical α chains with a 'pitch'* of 39

* The 'pitch' corresponds to the number of residues along an α chain between geometrically equivalent intermolecular contacts. This would equal the pitch of an α chain for straight collagen molecules, but would be modified by any supercoiling of the molecules, as would occur in twisted microfibrils

Table 1

Azimuthal distribution, in a collagen α chain with a 'pitch' of 39 residues per turn, of large hydrophobics conserved in α1-type chains (†) and charged residues conserved in α1-type chains (-, or θ for those which do not occur in α2). Full, dashed and double dashed lines, respectively, indicate regions of heavy concentrations of hydrophobics, charged residues and α1-specific charged residues. XXX indicates crosslink sites



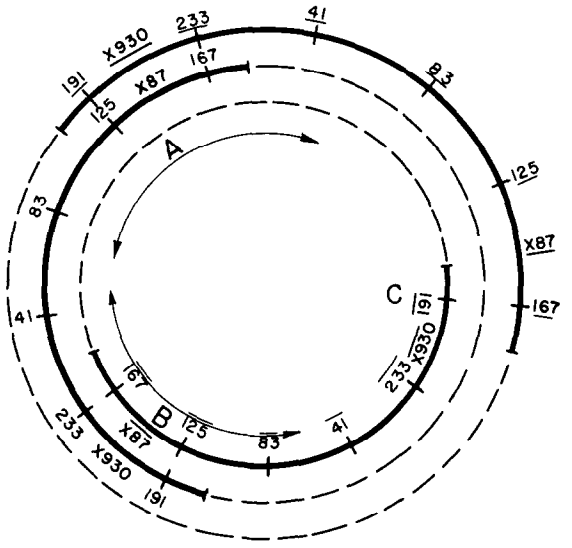


Fig.4. Azimuthal distribution of regions of hydrophobic residues (full lines) in a collagen molecule composed of three identical α chains with a 'pitch' of 39 residues per turn. The molecule is viewed, in projection, from its N-terminal end with the α chains represented by circles. Equivalent residues on the three α chains are related by anti-clockwise rotations of 111°, 111° and 138°, respectively, starting from the first α chain from the N-terminal end of the molecule (outermost circle, lines below residue sequence numbers) and proceeding through the middle chain (middle circle) and C-terminal chain (innermost circle, lines above sequence numbers). Locations of groups of large hydrophobic residues (e.g. 83+n234) are indicated by the first number of the series, and potential crosslink sites are indicated by X87 and X930. The arrows indicate the approximate widths of interacting edges at A and B which include complete sets of conserved hydrophobic residues; an analogous edge at C would be almost 30° wider.

residues per turn. This should have two complementary interacting edges both of which include all the conserved large hydrophobic residues (cf. fig.3) and a crosslink site. Of the three possible edges, A and B are narrower than C and are 111° apart, as would be appropriate to a five-stranded microfibril. These considerations would apply to collagen molecules of types I, II or III. However, for type I collagen this choice of interacting edges would also imply that α2 is the middle chain, otherwise there would be no need for both the 167+*n*234 and 233+*n*234 hydrophobic series to be conserved in α2. This position for α2 has also been proposed from considerations of intramolecular stability [16].

Because the crosslink sites are known to be on α1 chains [2,17], this choice of edges also implies that the site at residue 930 occurs on the first (i.e. N-terminal) α1 chain at interacting edge A and the site at 87 on the third chain at edge B. Residues Hyl 87 and Hyl 930 are cross-linked to the C- and N-terminal telopeptides respectively on molecules staggered by 4*D* [2,17], so microfibrils joined at the A and B edges would have a right-handed helical arrangement of *D*-staggered collagen molecules, as shown in fig.3.

The three α chains must, of course, occur in the same order along the length of the molecule on any interesting edge. Therefore for maximal interactions between large hydrophobics, the first, second and third chains at edge A must interact primarily with the second, third and first chains, respectively, at edge B on an adjacent molecule. The last of these three sets of interactions, that is between the α1 (C-terminal) and α1 (N-terminal) chains, would not involve the large hydrophobic residues, but would, in fact, involve regions of the α1 sequence where there is a concentration of residues which are charged in α1 but not in α2 (cf. fig.4, and table 1).

CPK packing models were used to study possible steric interactions between complementary regions of collagen molecules and indicated several 'good fits' including Phe317(α1C) between Leu83(α2) and Phe 86(α2), Leu317(α2) between Leu548(α1C) and Met551(α1C), and Arg333(α1N) between Asp96(α2) and Asp101(α1C), (where α1N and α1C indicate N- and C-terminal α1 chains respectively). In terms of the right-handed helical arrangement of molecules, residues at edge A interact with residues at edge B on the adjacent molecule, which are about 234 positions further from the N-terminus (fig.3). The exact stagger has been estimated by maximising interactions between large hydrophobics and between oppositely charged residues along the A and B edges, and this appears to be between 233 and 234 residues. This is the absolute stagger, that is between identical residues on N-terminal α1 chains on adjacent molecules. The average sequence-number separation between interacting residues in different chains should be 232½, 232½ and 235½, respectively, for α1N(A)–α2(B), α2(A)–α1C(B) and α1C(A)–α1N(B) interactions (where A and B indicate interacting edges). These different numbers are a consequence of the one-residue stagger between α chains in each molecule. The higher figure for the third type of interaction is consistent with McLachlan's [9] observation that in the α1 chain groups of residues with complementary charges tend to be separated by slightly more than 234 sequence positions.

The approximately 100° width for the interacting edges indicated in fig.4 is also consistent with model-building studies, though individual interactions depend on the length and flexibility of the side-chains involved and the different azimuthal orientations of residues in X and Y positions [18]. With these reservations, we can roughly define edge A as including residues (5 to 18)+*n*39(α2), (17 to 30)+*n*39 (α1C) and (32 to 45)+*n*39 (α1N), and edge B as including residues (5 to 18)+*n*39 (α1C), (20 to 33)+*n*39 (α1N) and (32 to 45)+*n*39 (α2).

4. Discussion

The derivation of this model for collagen assembly has been based largely on the search for complementary interacting edges along which are concentrated the non-polar regions of the amino acid sequence and, in particular, the conserved large hydrophobic residues. The pronounced *D* periodicity of these residues implies that the assembly model is consistent with a maximisation of strong hydrophic interactions. Though an overall maximisation of charge interactions has not been sought, these would predominate in the model along about a third of the interacting edges and account for most of the charged residues which are conserved in α1-type but not α2 chains. My approach has been influenced by the fact that in protein struc-

tures hydrophobic residues are almost invariably buried, but charged residues generally in contact with solvent, as well as by examples of assembly of fibrous proteins which are dominated by hydrophobic interactions [19,20]. Other authors, who have made statistical analyses of the collagen assembly problem [11,12,15,21], have given relatively greater emphasis to charge interactions. Calculations in these studies of optimum helical pitch, interacting edges and α2-chain position, corresponding to estimated minimum-energy assembly configurations, have led to various results and appear to have a delicate dependence on the geometrical and energy approximations used in the computations. However I have been encouraged by the recent statistical analysis of Piez and Trus [15] which, in spite of the different approach, has led to an assembly model similar to that presented here.

Two features of the above analysis, the grouping of invariant large hydrophobic residues and the 111° separation of interacting edges, provide support for a five-stranded microfibril as compared with other suggested models for collagen assembly. Miller and Parry [22] have shown that such a microfibril can conform to the 4-fold screw axis indicated by the medium-angle X-ray pattern of rat tail tendon, if the collagen molecules are twisted about the microfibril axis with a pitch equal to $20D/(10n\pm1)$ (where n is an integer). In fact, it can also be shown that a left-handed twist with a pitch of $20D/11$ would bring about equivalent intermolecular contacts at $D/6$ intervals in the microfibril if the untwisted α chains have a pitch of $2D/11$. This is the most common interval between large hydrophobics, as described above (see fig.2), and it seems possible that they might serve in the initial alignment of straight collagen molecules before the microfibrils are twisted up.

It would appear from the above analysis that microfibril formation should not be greatly affected by the α2 chain. The two complementary interacting edges would be substantially the same if all three chains were of the α1 type, and indeed, judging from the observation of the characteristic 670Å periodicity in electron micrographs, stable microfibrils can be formed from only α1(I) [23], α1(II) [24] or α1(III) [25] chains. The α2 chain increases the difference between edge C, on the one hand, and edges A and B, on the other. Its main function may, therefore, be to modify the outside of the microfibril so as to enhance intermicrofibrillar association. This might help explain the occurrence in vivo of thicker fibrils in type I collagen than in types II and III.

Acknowledgements

Much of the study reported here was carried out while I was enjoying a sabbatical leave at the National Institutes of Health. This gave me the opportunity of many stimulating and critical discussions with Drs Karl Piez and Benes Trus, who were working on the same problem using a more statistical approach [15]. These discussions served to modify our originally quite different viewpoints, and our final conclusions substantially reinforce each other. I am also particularly grateful to Drs William Butler, Peter Fietzek, Andrew Kang and Klaus Kühn for making available to me amino acid sequence data on collagen prior to publication.

References

[1] Miller, E. J. (1976) Mol. Cell. Biochem. 13, 165–192.
[2] Fietzek, P. P. and Kuhn, K. (1976) Int. Rev. Connect. Tissue Res. 7, 1–60.
[3] Butler, W. T., Finch, J. E. and Miller, E. J. (1977) Biochemistry 16, 4981–4990.
[4] Dixit, S. N., Seyer, J. M. and Kang, A. H. (1977) Eur. J. Biochem. 73, 213–221.
[5] Dixit, S. N., Seyer, J. M. and Kang, A. H. (1977) Eur. J. Biochem. 81, 599–607.
[6] Traub, W. and Piez, K. A. (1971) Adv. Protein Chem. 25, 243–352.
[7] Piez, K. A. and Miller, A. (1974) J. Supramolec. Struct. 2, 121–137.
[8] Hulmes, D. J. S., Miller, A., Parry, D. A. D., Piez, K. A. and Woodhead-Galloway, J. (1973) J. Mol. Biol. 79, 137–148.
[9] McLachlan, A. D. (1977) Biopolymers 16, 1271–1297.
[10] Hulmes, D. J. S., Miller, A., Parry, D. A. D. and Woodhead-Galloway, J. (1977) Biochem. Biophys. Res. Commun. 77, 574–580.
[11] Cunningham, L. W., Davies, H. H. and Hammonds, R. G. (1976) Biopolymers 15, 485–502.
[12] Trus, B. L. and Piez, K. A. (1976) J. Mol. Biol. 108, 705–732.
[13] Piez, K. A. and Trus, B. L. (1977) J. Mol. Biol. 110, 701–704.

[14] Bornstein, P. and Traub, W. (1978) in: The Proteins, (Neurath, H. and Hill, R eds) vol. IV, Academic Press, New York.
[15] Piez, K. A. and Trus, B. L. (1978) J. Mol. Biol. in press.
[16] Traub, W. and Fietzek, P. P. (1976) FEBS Lett. 68, 245–249.
[17] Henkel, W., Rauterberg, J. and Stirtz, T. (1976) Eur. J. Biochem. 69, 223–231.
[18] Yonath, A. and Traub, W. (1969) J. Mol. Biol. 43, 461–477.
[19] McLachlan, A. D. and Stewart, M. (1975) J. Mol. Biol. 98, 293–304.
[20] Nakashima, Y., Wiseman, R. L., Konigsberg, W. and Marvin, D. A. (1974) Nature 253, 68–71.
[21] Hofmann, H., Fietzek, P. P. and Kühn, K. (1978) J. Mol. Biol. in press.
[22] Miller, A. and Parry, D. A. D. (1973) J. Mol. Biol. 75, 441–447.
[23] Tkocz, C. and Kühn, K. (1969) Eur. J. Biochem. 7, 454–462.
[24] Stark, M., Miller, E. J. and Kuhn, K. (1972) Eur. J. Biochem. 27, 192–196,
[25] Wiedemann, H., Chung, E., Fujii, T., Miller, E. J. and Kühn, K. (1975) Eur. J. Biochem. 51, 363–368.