Amino Acid Sequences of α -Helical Segments from S-Carboxymethylkerateine-A

STATISTICAL ANALYSIS

By THOMAS C. ELLEMAN,* W. GORDON CREWETHER* and JOHN VAN DER TOUW†

* Division of Protein Chemistry, CSIRO, Parkville (Melbourne), Vic. 3052, Australia, and

† Division of Mathematics and Statistics, CSIRO, South Melbourne, Vic. 3205, Australia

(Received 2 November 1977)

The distribution of the different types of amino acid residue within two helical segments isolated from the low-sulphur fraction of wool keratin was examined for periodicity and larger sequence repeats. Both were detected, the former corresponding to the geometry of the proposed coiled-coil structure and the latter suggesting a distant gene duplication, though the existence of the repeat might equally well be related to interactions within the microfibril.

The X-ray diffraction patterns of hair, wool and other mammalian keratins contain strong meridional reflections of spacing 14.9 nm (1.49 Å) and 51.5 nm (5.15 Å). Pauling & Corey (1951) suggested that a helical arrangement of amino acid residues (the α -helix) was present in α -keratins, which would account for the 14.9 nm (1.49 Å) arc corresponding to the axial rise per residue. The α -helix, however, would be expected to give a strong reflection on the 54 nm (5.4 Å) layer line, but displaced from the meridian, and Crick (1952) suggested that the coiling of the α -helix into a superhelix (i.e. a coiled-coil) could account for a meridional reflection of spacing 51.5 nm (5.15 Å) rather than an off-meridional reflection of spacing 54 nm (5.4 Å). This distortion would enable the side chains in the interior of adjacent helices (positions a and d; Table 1) to pack more neatly, the side chains of one α -helix fitting into the spaces between the side chains of a neighbouring α -helix, along the line of contact.

Information obtained from the structures of globular proteins indicates that the inner regions of the coiled-coil would be composed of non-polar residues, whereas polar residues would be located on the outer surface in contact with water (North & Phillips, 1969). The geometry of the proposed coiled-coil is such that the non-polar residues occupy positions on the α -helix, alternately three and four residues apart, each residue occupying a position relative to the coiled-coil similar to that of a residue seven residues along the chain. The sequences of the helical regions of keratins might then be expected to show a periodicity in the positions of similar types of residue.

Two types of helical chain segments (I and II) have been isolated from the low-sulphur protein fraction of

wool by methods dependent on the protection against proteinase attack afforded by the structuring of the polypeptide chains (Crewther & Dowling, 1971), and their amino acid sequences have been determined (Hogg *et al.*, 1978; Crewther *et al.*, 1978; Gough *et al.*, 1978). The sequences of these helical segments have been examined by the statistical methods of spectral analysis to detect any periodicity in the sequence and by the chi-squared test to detect regularity in the positions of the proposed seven-residue structural repeat. The sequences have also been examined for gross repetition of sequence at higher levels.

Statistical Procedures and Results

Periodicity within segments

The amino acid sequences were examined for periodicity in the placement of residues by the following procedure. The 0–9 scale of McLachlan (1971) was used for similarity scores between amino acids. Each sequence was aligned with itself for perfect matching and then compared after a series of single-residue displacements. For each frameshift the similarity scores for each comparison of amino acids are summed. Thus each protein of N residues in length produces a score $x(k)$ for each frameshift k , $k = 1, 2, \dots, N - 1$. The frameshift scores are assessed by comparison with the scores that might have been obtained by drawing the appropriate number of similarity scores independently at random, with replacement, from the population of amino acid-similarity scores produced by the frameshifts. By this procedure the variance and the expected value

of $x(k)$ are proportional to $N - k$, the number of similarity scores included in each summation.

To apply a spectral analysis the sequence to be analysed should have the same mean and variance at each position. To ensure that this condition holds the sequence $x(k)$ was transformed to a sequence $z(k)$ by:

$$z(k) = \frac{x(k) - (N - k)\bar{x}}{\sqrt{N - k}}$$

where $\bar{x}$ is the mean similarity score per pair of amino acids averaged over all comparisons at all frameshifts.

The sequence $z(k)$ of the type-I material is shown in Fig. 1.

A spectral analysis was done on the sequence $z(k)$. The technique involves fitting a sine wave of frequency f to the data, for various values of f between 0 and $\frac{1}{2}$ cycle/residue (Jenkins & Watts, 1969). If we let $z(k)$ be the value at the k th position in the sequence, we can define a number, $C(f)$, for each frequency f as follows:

$$C(f) = \frac{1}{M} \left(\left\{ \sum_{k=1}^M [z(k) - \bar{z}] \cos 2\pi f k \right\}^2 + \left\{ \sum_{k=1}^M [z(k) - \bar{z}] \sin 2\pi f k \right\}^2 \right)$$

where $\bar{z}$ is the mean of the sequence $z(k)$ and M , the number of terms in the series, is equal to $N - 1$.

The use of both sines and cosines allows for a possible phase shift. The value $C(f)$ is proportional to

the square of the estimated covariance between the series $z(k)$ and a sine wave of frequency f . The function $C(f)$ is the power spectrum of the series $z(k)$.

Define $Y(f) = 2C(f)/\text{var}(z)$. If the series $z(k)$ comes from a completely random normal process, then each $Y(f)$ is a sample from a chi-squared distribution with 2 degrees of freedom. Moreover, the values corresponding to $f = 1/M, 2/M, 3/M, \dots$, for $f \leq \frac{1}{2}$, are independent. Since we are interested in the largest peaks (Figs. 2 and 3), we have to adjust the probability levels for significance tests accordingly. If the probability that the value of a single random variable exceeds a given value is P , then the probability that at least one of n such random variables chosen independently exceeds that value is $1 - (1 - P)^n$. In the case considered $n = M/2$. Thus the 0.05, 0.01 and 0.001 significance levels for the largest peaks for the type-I and type-II materials correspond to spectral intensities of 14, 17 and 22 respectively. The three largest peaks of Figs. 2 and 3 are consistent with a seven-residue repeat, and are all highly significant; the fourth largest peak is not significant at the 5% level, although an explanation for it has been proposed by Parry *et al.* (1977).

A distribution-free Kolmogoroff-Smirnov test (Jenkins & Watts, 1969) was done on the spectra and both sequences showed a significant departure from randomness (type I, $P < 0.03$; type II, $P < 0.007$). The graph of $Y(f)$ against frequency f , for the type-I material and the type-II material are shown in Fig. 2

Fig. 1. Standardized score at each frameshift alignment of type-I material. Arrows indicate frameshifts that are multiples of seven. See the text for further details.

Fig. 2. Power spectrum of the data of Fig. 1
See the text for details.

Fig. 3. Power spectrum of the standardized score at each frameshift alignment of type-II material
See the text for details.

and Fig. 3 respectively. The maximum value of Y for both type-I and type-II material occurs at a frequency corresponding to one cycle every $3\frac{1}{2}$ residues. Type I also has a high value of Y at a frequency corresponding to one cycle every $\frac{7}{3}$ residues. These are the second and third harmonics of a fundamental period

of 7 and the low intensity of the fundamental is easily explained, since non-polar residues occur more frequently than expected by chance at the third and fourth positions of successive heptads.

To test for any association between the seven positions of the coiled-coil and the nature of the residues at these positions a chi-squared test was used. Of those hydrophobic residues present (phenylalanine, tyrosine, valine, leucine, isoleucine and alanine), the distribution differed from uniformity (type I, $P < 0.00001$; type II, $P < 0.002$ using a 2×7 contingency table), the non-polar residues being more frequently in positions a and d, and the remaining five positions containing relatively more of the polar residues (Table 1). The distribution of the type of polar residue in positions b, c, e, f and g is not uniform, however (Table 2). A chi-squared test on a 3×5 contingency table showed that there is possibly some association between the nature of the charge on a residue and its position (type I, $P < 0.04$; type II, $P < 0.05$), basic residues predominating at positions e and f, whereas acidic residues predominate at positions b and g.

Similarity between segments

The similarity of the functional constraints on the two helical segments necessitates some similarity of the sequences, but the similarity found far exceeds that expected simply on the basis of the seven periodicity. Of the 103 residues compared 32 are identical when the sequences are aligned, without any gaps, from the C-termini. This alignment was assessed by Monte Carlo methods, with the randomization procedure confined such that residues could only be interchanged if their positions differed by a multiple of 7. By the similarity-scoring systems of both McLachlan (1971) and Barker & Dayhoff (1972), the alignment score exceeded by six standard deviations the mean of the optimum frameshift scores of the random sequences.

Similarity within segments

A method was devised for identifying related sequences in two chains, or in different parts of the same protein. This was done by calculating the score obtained for each frameshift and assigning a probability value to that score. This probability was calculated by assuming that the score for each amino acid pair is an independently distributed random number, with a distribution determined by the similarity matrix. A more detailed description of the algorithm for calculating the probabilities is given in the appendix.

When the type-II material was compared with itself, a very high score [$P(60, 244) = 0.000062$; see

Table 1. Contingency table for association between polarity and residue number, modulo 7

		Type I							Type II							
		Residue no., modulo 7							Residue no., modulo 7							
		1	2	3	4	5	6	7	1	2	3	4	5	6	7	
Non-polar		13	3	4	3	10	2	1	4	4	6	10	2	4	12	42
Polar		2	12	11	12	5	12	13	12	12	10	6	13	11	3	67
		15	15	15	15	15	14	14	16	16	16	16	15	15	15	109
Proposed position in coiled-coil		d	e	f	g	a	b	c	e	f	g	a	b	c	d	

Table 2. Contingency table for association between charge and residue number, modulo 7; positions a and d excluded

		Type I					Type II					
		Residue no., modulo 7					Residue no., modulo 7					
		2	3	4	6	7	1	2	3	5	6	
Basic		7	3	2	0	1	4	5	3	0	1	13
Acidic		2	2	5	7	4	3	1	6	8	3	21
Neutral		6	10	8	7	9	9	10	7	7	11	44
		15	15	15	14	14	16	16	16	15	15	78
		e	f	g	b	c	e	f	g	b	c	

the Appendix] occurred at frameshift 49. Even after adjusting for the fact that 108 frameshifts were examined and that the best one was singled out, the score for frameshift 49 is still significant ($P < 0.007$). To test whether the assumptions made were reasonable, Monte Carlo methods were used to check the probabilities calculated. These showed that the calculated probabilities were quite reasonable, with any errors being made on the safe side, since the probabilities of obtaining high scores tended to be overestimated by the algorithm.

Although the periodicity in the sequence of amino acids tends to produce higher scores at frameshifts that are multiples of seven, the score at frameshift 49 was still remarkably high. Using only the score distribution from the 15 frameshifts that are multiples of seven, the probability that at least one of these exceeds $P(60, 244)$, i.e. 0.0035 is 0.05. Monte Carlo methods were again used to verify this with the randomization procedure confined such that residues could only be interchanged if their positions differed by a multiple of 7. Of 1000 sequences generated in this way, only 23 produced scores with probabilities

(as calculated by the algorithm in the Appendix) as low as 0.000062.

In spite of the high degree of similarity between the type-I and type-II material, the 49 frameshift of the type-I material did not diverge significantly from expectation [$P(54, 195) = 0.025$; see the Appendix]. However, the homology between the *N*-terminal portion of the type-II material and the *C*-terminal portion of the type-I material at 49-residue frameshift was considerably better than that between the *N*- and *C*-terminal portions of the type-II material [$P(60, 266) = 0.00000085$, see the Appendix]. This is a significant result, since, of 1000 random comparisons preserving both the seven periodicity and the same degree of homology as the real comparison, none contained a frameshift of greater significance.

Discussion

Both of the helical sequences examined show the periodic nature expected of a coiled-coil in the disposition of the non-polar and polar residues. A periodicity in the charge of the polar residues

also appears to be significant. The sequences of the type-I and type-II helical segments show similarity well beyond the expectation from their periodic nature. They also show similarities between their *N*- and *C*-terminal portions beyond that expected on the basis of their periodic nature alone. The latter similarity might reflect a distant gene duplication of 49 residues in an ancestral precursor gene for both type-I and type-II material. It might equally well indicate similar structural units of the microfibril at a 49-residue spacing, necessitated by the structural requirements for the microfibrils in α -keratin.

References

- Barker, W. C. & Dayhoff, M. O. (1972) in *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed.), pp. 101–110, National Biomedical Research Foundation, Washington
- Crewther, W. G. & Dowling, L. M. (1971) *Appl. Polym. Symp. no. 18*, 1–20
- Crewther, W. G., Inglis, A. S. & McKern, N. M. (1978) *Biochem. J.* **173**, 373–385
- Crick, F. H. C. (1952) *Nature (London)* **170**, 882–883
- Gough, K. H., Inglis, A. S. & Crewther, W. G. (1978) *Biochem. J.* **173**, 365–371
- Hogg, D. M., Dowling, L. M. & Crewther, W. G. (1978) *Biochem. J.* **173**, 353–363
- Jenkins, G. M. & Watts, D. G. (1969) in *Spectral Analysis and its Applications* (Jenkins, G. M. & Parzen, E., eds.), pp. 230–234, Holden-Day, San Francisco, Cambridge, London and Amsterdam
- McLachlan, A. D. (1971) *J. Mol. Biol.* **61**, 409–424
- North, A. C. T. & Phillips, D. C. (1969) *Prog. Biophys. Mol. Biol.* **19**, 5–132
- Parry, D. A. D., Crewther, W. G., Fraser, R. D. B. & MacRae, T. P. (1977) *J. Mol. Biol.* **113**, 449–454
- Pauling, L. & Corey, R. B. (1951) *Proc. Natl. Acad. Sci. U.S.A.* **37**, 261

APPENDIX

Given two proteins of length m and n amino acid residues respectively we can form an $n \times m$ similarity matrix $A = (a_{ij})$, where a_{ij} is the similarity score for residue pair (R_i, R_j) . R_i is the i th residue of one protein and R_j is the j th residue of the second protein. The similarity score between residue pairs, obtained by using, for example, the method of McLachlan (1971), is an integer between 0 and some maximum value r .

For the purpose of calculating the probability levels associated with the total score obtained when comparing part of one protein with part of the other, the following method is used. For each integer s , between 0 and r , we define a number p_s , for the case of two different protein sequences.

p_s = (the number of occurrences
of s in the matrix A)/ mn

For the case of identical sequences, only the triangle above the main diagonal of A is used and:

p_s = (the number of occurrences
of s in the upper triangle of A) $\bigg/ \frac{n(n-1)}{2}$

The number p_s represents the probability that a randomly selected amino acid pair has a similarity score s .

Let $P(k, s)$ be the probability that the total score of a pairwise comparison between two sequences of amino acid residues, each of length k , is equal to s . By definition:

$$P(1, s) = \begin{cases} p_s & \text{for } s = 0, 1, 2, \dots, r \\ 0 & \text{otherwise} \end{cases}$$

If a total score of s for a sequence of k residues is to result and the score for the last pair of residues compared is t , the score for the first $k-1$ pair must be $s-t$. All comparisons are assumed to be independent, hence:

$$P(k, s) = \sum_{t=0}^{\infty} P(k-1, s-t)P(1, t)$$

Omitting terms equal to zero, this simplifies to:

$$P(k, s) = \sum_{t=l_1}^{l_2} P(k-1, s-t)P(1, t), \quad \text{where} \\ l_1 = \max\{0, s-r(k-1)\} \quad \text{and} \quad l_2 = \min\{r, s\}.$$

This recursive relation allows the probabilities to be readily calculated by computer.

Reference

- McLachlan, A. D. (1971) *J. Mol. Biol.* **61**, 409–424