Bull. Inst. Chem. Res., Kyoto Univ., Vol. 60, No. 5～6, 1982

Analysis of Protein Sequence Homology by Correlation Coefficients

Yasushi KUBOTA*

Received August 28, 1982

A method for detection of protein sequence homology by correlation coefficients was developed. In this way, any sequence repetition in a protein and the extent of the sequence homology among proteins can be estimated quantitatively. First of all, the collected 34 parameters inherent in amino acids, such as hydrophobicity, were classified into a smaller number of groups by the factor analysis, and based on the classification, proper six parameters for the computation of the arithmetic average of correlation coefficients were selected so as to give rise to a good correlation for the CD and EF hand regions (Ca^{2+}-binding sites), while poor ones for other segments in carp parvalbumin. Then, the method was examined on the proteins; calf collagen α1-chain, wool keratin and light meromyosin as the examples for proteins having some internal repetitions of a short segment, calmodulin, cardiac troponin C and myosin A1 light chain as the examples for those having repetitions of a long segment. The above three proteins were also compared with the two Ca^{2+}-binding sites of carp parvalbumin. The method was extended to two dimensional way, in expression by a comparison matrix. The procedure extended to the comparison matrix is applied to the following homologous proteins: calmodulin and skeletal muscle troponin C; vitamin D-dependent Ca^{2+}-binding protein and parvalbumin; bovine trypsinogen and porcine elastase; hemoglobin α-chain and myoglobin. The results show that the procedure may be a useful method for searching homologous segments and identifying important sites which take similar native conformation in amino acid sequences.

KEY WORDS: Protein/ Sequence homology/ Primary structure/ Tertiary structure/ Correlation analysis/ Amino acid parameters/

I. INTRODUCTION

Proteins, which are principal substances in living organism, are made up of α-amino acids in linear sequences in which the units are linked together by peptide bonds. There are twenty different amino acids used in the construction of proteins, and an amino acid sequence of a protein is genetically provided for. A number of amino acid sequences of various proteins have been elucidated recently,[1] and sequence information given by DNA plays an essential role in structural and functional aspects of the proteins. Thus, with the accumulation of data on protein sequences, the sequence homology becomes one of the most important concepts as a problem of comparative biochemistry and molecular evolution. The detection of homology, however, has usually been performed qualitatively by comparing chemical similarity of corresponding amino acid residues.

Correlation coefficients have been known to be one of the most important statistical quantities which are utilized for analysis of irregular phenomena.[2] A correlation

* 窪田 綏: Laboratory of Physical Chemistry of Enzyme, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611, JAPAN

coefficient is adequate to estimate quantitatively the extent of mutual relationship between two variables X and Y; that is, X and Y are identical, the value is 1, and if random, the value is 0, *i.e.,* no correlation at all. Because proteins are linear copolymers of amino acids, numerical expression of a protein sequence obtained by replacing amino acid residues by some parameter will give a basis to apply the method of correlation coefficients. The purpose of this study is to describe a method to estimate the sequence homology quantitatively in terms of correlation coefficients computed from parameters inherent in amino acids, and also reveal correlations of homologous sequences to three-dimensional structures.

Many physical parameters inherent in amino acid, *e.g.,* hydrophobicity, propensity to form α-helix and β-structure, etc. are available. These parameters, however, are not always independent, *i.e.,* correlated more or less with each other. First of all, the collected 34 parameters of amino acids are classified into a smaller number of groups by the factor analysis, which is one of the methods to analyze variables,[3,4] and proper parameters will be selected so that they can be used for detection of sequence homology by correlation coefficients. As a reference sequence for the selection of proper parameters, the sequence of parvalbumin which contains the two Ca^{2+}-binding regions called CD hand and EF hand sequences, is adopted, since the two segments have been known to be homologous and to have similar conformations to bind Ca^{2+} specifically according to X-ray crystallography.[5]

In the proposed method, some internal sequence repetitions of proteins are exhibited by the periodic patterns of the autocorrelation coefficients, and the candidates for the Ca^{2+}-binding regions are revealed as high correlative regions to the CD or EF hand region by the dominant peaks of the correlation coefficients. The method can be extended to two dimensional analysis, in expression by a comparison matrix and this extension makes more effective in identification of homologous segments which have the structural similarity.

II. METHODS

Homology in protein sequences can be represented by correlation coefficients which are calculated, as described in detail elsewhere,[6] from the sequences after replacing residues by numerical values of various parameters of the amino acids such as hydrophobicity. Brief description is as follows. Let us consider an amino acid sequence of a protein, X, n residues long. Autocorrelation coefficient $A(\tau)$, as a function of τ residues lag, can be calculated by comparing quantities $x(i)$ at the position i with $x(i+\tau)$ at the position $i+\tau$;

$$A(\tau)=\frac{\sum_{i=1}^{n-\tau}(x(i)-\langle x\rangle)(x(i+\tau)-\langle\!\langle x\rangle\!\rangle)}{[\{\sum_{i=1}^{n-\tau}(x(i)-\langle x\rangle)^2\}\{\sum_{i=1}^{n-\tau}(x(i+\tau)-\langle\!\langle x\rangle\!\rangle)^2\}]^{1/2}}, \qquad (1)$$

$$\langle x\rangle=\frac{1}{n-\tau}(\sum_{i=1}^{n-\tau}x(i)), \quad \langle\!\langle x\rangle\!\rangle=\frac{1}{n-\tau}(\sum_{i=1}^{n-\tau}x(i+\tau)).$$

If the sequence has any repetition of τ_0 residues long, $A(m\tau_0)$ ($m=0, 1, 2, \ldots$) must exhibit a high value.

Similarly, as a measure of the extent of homology between two amino acid sequences (or part of the sequences), X and Y, the correlation coefficient $C(j)$ at the position j of the sequence Y is defined as

$$C(j)=\frac{\sum_{i=1}^{n}(\mathrm{x}(u+i-1)-\langle \mathrm{x}\rangle)(\mathrm{y}(j+i-1)-\langle \mathrm{y}\rangle)}{[\{\sum_{i=1}^{n}(\mathrm{x}(u+i-1)-\langle \mathrm{x}\rangle)^2\}\{\sum_{i=1}^{n}(\mathrm{y}(j+i-1)-\langle \mathrm{y}\rangle)^2\}]^{1/2}}, \tag{2}$$

$$\langle \mathrm{x}\rangle=\frac{1}{n}(\sum_{i=1}^{n}\mathrm{x}(u+i-1)),\quad \langle \mathrm{y}\rangle=\frac{1}{n}(\sum_{i=1}^{n}\mathrm{y}(j+i-1)).$$

This procedure provides a way to evaluate quantitatively homology in amino acid sequences of proteins. To visualize such homology, the method can be effectively extended as described previously.[7] Let X and Y be two different (or same) protein sequences, then, a correlation coefficient, $C_p(i,j)$, of parameter p at the position (i,j) in a square array is defined as

$$C_p(i,j)=\frac{\sum_{l=-k}^{k}(\mathrm{x}_p(i+l)-\langle \mathrm{c}\rangle)(\mathrm{y}_p(j+l)-\langle \mathrm{c}\rangle)}{[\{\sum_{l=-k}^{k}(\mathrm{x}_p(i+l)-\langle \mathrm{c}\rangle)^2\}\{\sum_{l=-k}^{k}(\mathrm{y}_p(j+l)-\langle \mathrm{c}\rangle)^2\}]^{1/2}}, \tag{3}$$

where $(2k+1)$ is equal to the length of segments to be compared and examined from 5 to 15 to choose a proper length. In order to reduce the signal-noise ratio the average correlation coefficients $\langle A(\tau)\rangle$, $\langle C(j)\rangle$ and $\langle C(i,j)\rangle$ are introduced;

$$\langle A(\tau)\rangle=\frac{1}{n}\sum_{p=1}^{n}A_p(\tau), \tag{4}$$

$$\langle C(j)\rangle=\frac{1}{n}\sum_{p=1}^{n}C_p(j), \tag{5}$$

$$\langle C(i,j)\rangle=\frac{1}{n}\sum_{p=1}^{n}C_p(i,j), \tag{6}$$

where n is a number of parameters of amino acids. Since Eqs. (4), (5) and (6) are the arithmetic average, the n kinds of parameters should be selected so as to be independent of each other as much as possible for a reason to avoid artificial weighting on those of the parameters which are correlated.

The factor analysis, one of the methods to analyze variables, provides a basis for classification of a set of variables into a smaller number of groups or categories. In this way a variable is represented in a form of a linear combination of several factors which are independent of each other. This is accompanied by the factor model;

$$\mathbf{X}_j=\sum_{i=1}^{m}\mathrm{a}_{ji}\mathbf{F}_i+\mathrm{a}_j\mathbf{u}_j, \tag{7}$$

where $\mathbf{X}_j=(\mathrm{x}_{j1},\ \ldots\ldots,\ \mathrm{x}_{js})$ is a s-dimensional vector (in the present case x_{ji} corresponds to a quantity of amino acid i of parameter j, and $s=20$). The vectors $\mathbf{F}_1, \ldots\ldots, \mathbf{F}_m$ are called common factors, since they are common to the set of variables $\mathbf{X}_j$, and the vectors $\mathbf{u}_1, \ldots\ldots, \mathbf{u}_n$ are called unique factors, since they are unique to each variable. Based on the correlations among the n variables of $\mathbf{X}_j$, the analysis is then to determine the coefficients of the common factors (factor loading a_{ji}) and those of the unique factors (uniqueness a_j^2) so that $\mathbf{F}_i$ and $\mathbf{u}_j$ are uncorrelated, $i=1, 2, \ldots\ldots, m$, $j=1, 2, \ldots\ldots, n$. After the factor loadings have been determined, the next step in the analysis is to search new proper common factors so that one variable may be described by one factor as much as possible, *i.e.*, to find new factor loadings of a variable having a high value for a certain factor and low values for the other factors. This involves the technique of factor rotation, *i.e.*, calculation of the variance of the squares of factor loadings;

$$\sigma_p{}^2=\frac{1}{n}\Big(\sum_{j=1}^{n}(a_{jp}{}^2)^2\Big)-\frac{1}{n^2}\Big(\sum_{j=1}^{n}a_{jp}{}^2\Big)^2,\quad \sigma^2=\sum_{p=1}^{m}\sigma_p{}^2, \tag{8}$$

where *n* is the number of variables (in the present case, n=34). The new factor loadings are determined so as to make σ^2 maximum with the orthogonal rotation of factor axis. That is so called varimax method which is mostly used in application. In this way, a minimum number of factors of practical significance, *m*, can be found

Table I. List of the 34 collected parameters of amino acids

#	Parameter	Reference
P1	Propensity to form β-structure (Levitt)	[8]
P2	β-structure-coil equilibrium constant, s_β (Ptitsyn & Finkelstein)	[9]
P3	Preference for parallel β-strands (Lifson & Sander)	[10]
P4	Preference for β-strands (Lifson & Sander)	[10]
P5	Propensity to form β-structure (Chou & Fasman)	[11]
P6	Preference for antiparallel β-strand (Lifson & Sander)	[10]
P7	Average surrounding hydrophobicity, $\langle H\rangle$ (Manavalan & Ponnuswamy)	[12]
P8	Partial specific volume	[13]
P9	Transfer energy (Bull & Breese)	[14]
P10	Molecular weight	[15]
P11	Average volume of buried residue (Lesk & Chothia)	[16]
P12	Average non-boned energy per residue (Oobatake & Ooi)	[17]
P13	Average percent in proteins (Dayhoff)	[1]
P14	Dihedral angle between four successive C^α atoms (Levitt)	[18]
P15	Short range non-bonded energy per atom (Oobatake & Ooi)	[17]
P16	Bulkiness (Zimmerman *et al.*)	[19]
P17	Hydrophobicity (Jones)	[20]
P18	Propensity to form α-helix (Levitt)	[8]
P19	Propensity to form α-helix (Chou & Fasman)	[11]
P20	Helix-coil equilibrium constant, s_α (Ptitsyn & Finkelstein)	[9]
P21	Propensity to form reverse turn (Levitt)	[8]
P22	Propensity to form β-turn (Chou & Fasman)	[11]
P23	pK value of amino group (pK-N)	[15]
P24	Transfer energy (Janin)	[21]
P25	Average non-bonded energy per atom (Oobatake & Ooi)	[17]
P26	Side chain interaction parameter, ξ_s (Krigbaum & Rubin)	[22]
P27	Transfer energy (Levitt)	[18]
P28	Polarity (Grantham)	[23]
P29	Contact number of a residue (Nishikawa & Ooi)	[24]
P30	Propensity to bury inside of a molecule (Wertz & Scheraga)	[25]
P31	Polarity (Zimmerman *et al.*)	[19]
P32	Long range non-bonded energy per atom (Oobatake & Ooi)	[17]
P33	Relative mutability (Dayhoff)	[1]
P34	pK value of carboxyl group (pK-C)	[15]

from the collected parameters of amino acids. After all, each variable has large factor loadings on one or at most a few factors and small factor loadings in the remaining factors. Thus, variables which belong to the same group are more correlative with each other and those which belong to different group are less correlative with each other. This procedure will be applied to the classification of the collected parameters of the amino acids[1,8~25] listed in Table I.

Based on the classification by the factor analysis, physical parameters used for Eqs. (4), (5) and (6) are selected so that the EF hand sequence has a good correlation with the CD hand sequence, and poor correlations with the other segments in carp parvalbumin. In addition to this, the number of parameters are preferably reduced for a reason of saving computing time.

With proper parameters determined, in the first place, the method of the average autocorrelation coefficient, $\langle A(\tau)\rangle$, is applied to the proteins which have been known to have some repetitions of a short segment (*e.g.*, seven residues lag); calf collagen α1-chain,[26] wool keratin[27] and light meromyosin from rabbit skeletal muscle (L-MM),[28] and a long segment which have been assumed to be caused by gene duplications; calmodulin from bovine brain,[29] cardiac troponin C (TN-C)[30] and myosin A1 light chain.[31] The last three proteins are also compared with the two Ca^{2+}-binding sites, the CD and EF hand regions in carp parvalbumin by using Eq. (5), since these proteins seem to contain similar three-dimensional structures to the Ca^{2+}-binding sites. Secondly, the method extended to the comparison matrix (Eq. (6)) is applied to the following homologous proteins: calmodulin and skeletal muscle troponin C,[32] and vitamin D-dependent calcium-binding protein from bovine intestine (ICaBP)[33] and carp parvalbumin[34] for the calcium binding proteins; bovine trypsinogen[35] and porcine elastase[36] for the serine proteases from eukaryotic organisms; horse hemoglobin α-chain[37] and sperm whale myoglobin[38] for the globin family. Since globins and serine proteases are known to have similar three-dimensional structures, respectively, as determined by the X-ray crystallography,[39~44] the quantitative comparison of homology of the sequences and tertiary structures of these proteins will be done. The comparisons of the sequences are also performed on globin family; hemoglobin α-chain, myoglobin, lamprey hemoglobin,[45] erythrocruorin from *Chironomus*,[46] and leghemoglobin,[47] which have been considered to be distantly related with each other in globin family.

The comparison in three-dimensional structures of proteins can be performed independently of sequence homology by the superposition technique as described by Nishikawa and Ooi.[48] The extent of coincidence of the two conformations, A and B, is estimated by the distance between the corresponding atoms, $l_{ii}=|r_i^B-r_i^A|$, for the i-th residues after the best superposition of the two conformations obtained by translation and rotation of the coordinates of C^α atoms so as to minimize the r.m.s. deviations between corresponding atoms.

All computations were performed with FACOM M-160 AD at the computing center of the Institute for Chemical Research, Kyoto University.

III. RESULTS

Classification of the parameters Table II shows that the 34 parameters are represented by seven factors, or classified into seven groups by the factor analysis. The parameters in the table are arranged in the decreasing order of values of factor loading, a_{ji}s, which correspond to a component of every parameters on the orthogonal axes of seven factors. Each of factors is interpreted as follows:

FI: This factor may correspond to a tendency to form β-structure, since the para-

Table II. Final pattern matrix of the 34 parameters (Table I), and average correlation coefficient *C* between the CD and EF hand regions of parvalbumin. Factor loadings less than 0.25 are omitted.

	FI	*FII*	*FIII*	*FIV*	*FV*	*FVI*	*FVII*	*C*
P1	0.89							0.27
P2	0.86				0.42			0.32
P3	0.83			0.25	0.35			0.32
P4	0.83	0.29			0.39			0.45
P5	0.82	0.29			0.39			0.41
P6	0.70	0.45			0.34			0.49
P7	0.65		0.26		0.64			0.48
P8	0.64		0.26	0.62				0.45
P9	−0.54	−0.43		−0.43	−0.43			0.40
P10		0.95						0.10
P11	0.26	0.91						0.18
P12		0.88						0.33
P13		−0.80		0.26	−0.35			0.32
P14		−0.79	−0.33			0.28		0.07
P15		−0.73			0.61			−0.06
P16	0.54	0.62		0.41				0.34
P17	0.33	0.56		0.54	0.38	0.28		0.38
P18			0.96					0.13
P19			0.94					0.18
P20	0.46	0.26	0.72	−0.30				0.52
P21	−0.58	−0.25	−0.69		−0.28			0.52
P22	−0.56	−0.25	−0.63		−0.37			0.49
P23				0.89				0.44
P24	0.29				0.90			0.22
P25	0.35				0.87			0.26
P26	−0.30		−0.29		−0.82			0.43
P27	−0.31	−0.28			−0.76		0.37	0.43
P28	−0.50				−0.74			0.54
P29	0.55				0.73	0.32		0.49
P30	0.44	0.35			0.71	0.25		0.50
P31	−0.25		0.28		−0.65		0.55	0.26
P32	0.58	0.32			0.64	0.30		0.43
P33		−0.33			−0.32	−0.81		0.44
P34	0.41			0.28			−0.74	0.40

meters related to β-structure are included in this factor. To some extent, the parameters related β-structure correlate with hydrophobic property, since their factor loadings on the factor FV, which reflects hydrophobicity as explained below, are relatively high (*e.g.*, a_{ji}s on FV for P2, P3, P4 and P5 are 0.42, 0.35, 0.39 and 0.39, respectively). Interestingly the parameter of preference for antiparallel β-strands (P6) are also related to FII (a_{ji}=0.45).

FII: This factor may correspond to a size of an amino acid, where short range non-bonded energy (P15) and hydrophobicity (P17) are related to the factor of FV (a_{ji} on FV for P15 and P17 are 0.61 and 0.38, respectively).

FIII: This factor reflects the propensity to form α-helix, where P20, P21 and P22 have relatively high components on FI.

FIV: The only one parameter, pK-N is extracted. This parameter seems to be independent of the other parameters, because of relatively low factor loadings of pK-N on the other factors (at most −0.24 on FII).

FV: The parameters on hydrophobicity or polarity are mainly involved in this group. Therefore, FV can be interpreted as a factor of hydrophobic (or polar) nature. Some of the parameters reflect a propensity to form β-structure (*e.g.*, P28, P29, P30, etc).

FVI and FVII: Each of these factors has only one parameters, relative mutability and pK-C, respectively.

Selection of proper parameters When the average correlation coefficients, $\langle C(i,$

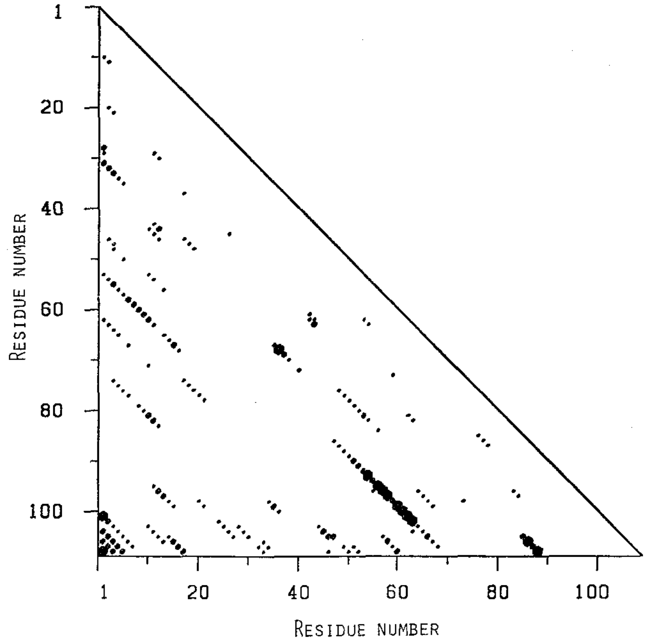

Fig. 1. Comparison matrix for carp parvalbumin with itself obtained by the use of proper six parameters giving the best result for the detection of sequence homology (P8, P21, P23, P28, P33, P34 in Table I). Small, medium, and large dots indicate values with $0.4 \leqslant \langle C(i,j) \rangle < 0.5$, $0.5 \leqslant \langle C(i,j) \rangle < 0.6$, and $0.6 \leqslant \langle C(i,j) \rangle$, respectively.

$j)\rangle$, are plotted against the residue numbers, i and j, for the sequence of parvalbumin a comparison matrix is obtained as shown in Fig. 1. As described previously,[7] all points of $\langle C(i,j)\rangle$s less than 0.4 were omitted and the appropriate length of segments in Eq. (3), $(2k+1)$, to compute $C_p(i,j)$, was 11 (or $k=5$). Furthermore, at least several (*e.g.*, 10) successive $\langle C(i,j)\rangle$s of greater than 0.4 are meaningful, because of the appearance by chance of short segments less than a few residues (*e.g.*, 3). Thus, existence of some homologous segments in proteins could be represented by successive dots parallel to the diagonal.

The criteria for the selection of proper parameters for detection of sequence homology were based on the sequence of parvalbumin which contains the two Ca^{2+}-binding regions, called CD hand (residues from 38 to 69) and EF hand (residues from 77 to 108); that is, if a combination of parameters gives a high correlation for the CD and EF hand regions, and low correlations for the other regions, this combination is adopted as a good one. The following set of six parameters seems to be the best combination satisfying the above criteria; partial specific volume (P8), propensity to form reverse turn (P21), pK-N (P23), polarity (P28), relative mutability (P33) and pK-C (P34). These parameters were selected one by one from the seven groups in a trial and error manner with regard to the correlation coefficients, C, between the CD and EF hand sequences listed in Table II. Fig. 1 shows the comparison matrix computed from Eq. (6) with the above set. The alignment of dots

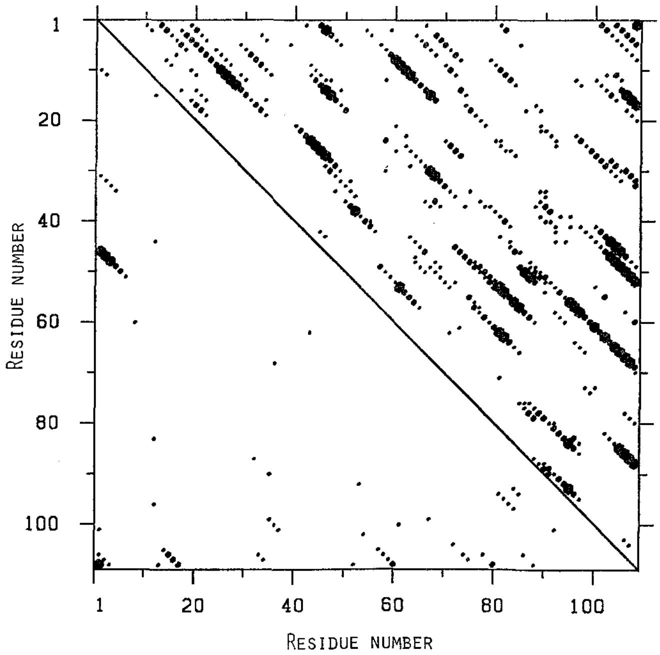

Fig. 2. Comparison matrix for carp parvalbumin with itself. The upper right triangle separated by the diagonal; the use of the parameters with a high correlation coefficient, C (P6, P20, P21, P22, P28, P29, P30). The lower left half; the use of the 7 kinds of parameters nearest to the factor axes in Table II (P1, P10, P18, P23, P24, P33, P34).

greater than 0.4 is obtained over 22 residues, from 47 to 68, and 86 to 107, for the Ca^{2+}-binding regions.

Another two ways to select the parameters from the seven groups might be possible; one is the use of the seven parameters nearest to the factor axes in Table II such as propensity to form β-structure (P1), molecular weight (P10), etc. The result, however, was not good as shown in the lower half of Fig. 2, because of poor homology on the Ca^{2+}-binding regions. The other is the use of the parameters with a high correlation coefficient, C, of the order of 0.4, as the extreme case for the selection. However, as shown in the upper right half of Fig. 2, many correlative segments appeared not only on the Ca^{2+}-binding regions, but also on the other segments which were not related to the tertiary structures. Therefore, the above two ways are not adequate for the selection of proper parameters for Eq. (6). Consequently, for the further comparison of sequences and structures of homologous proteins, the six parameters selected at the first step of this section will be employed, as the proper parameters to compute $\langle C(i,j)\rangle$ in Eq. (6). This combination is also available to the computation of $\langle A(\tau)\rangle$ and $\langle C(j)\rangle$ in Eq. (4) and (5).

Homologous sequences in proteins The first application of the procedure by correlation coefficients is to proteins which have some internal sequence repetitions, *e.g.*, calf collagen α1-chain, calmodulin, etc. The use of the average autocorrelation, $\langle A(\tau)\rangle$, as a function of the residue lag, τ, effectively indicates the periodicity in a sequence. The collagen molecule, which is the main constituent of skin, bone and tendon, has been known to be a triple chain molecule formed by two α1-chain and

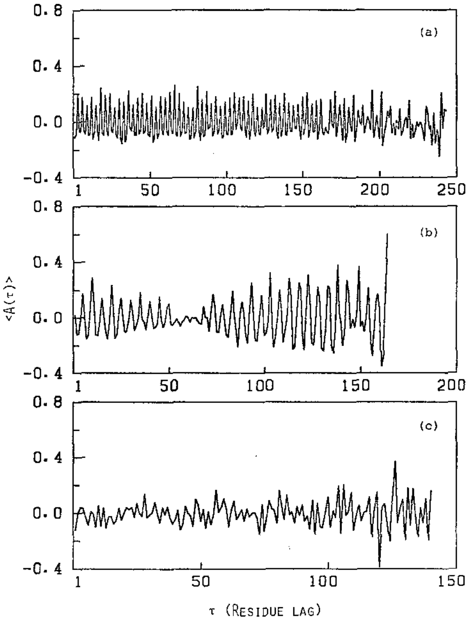

Fig. 3. (a) The periodic pattern in the sequence of calf collagen α1-chain expressed by the autocorrelation coefficients $\langle A(\tau)\rangle$. A periodicity of the three-residues is extensively observable. (b) The periodic pattern in the sequence of sheep keratin B2A. (c) The pattern for light meromyosin from rabbit skeletal muscle.

one α2-chain.[26] Fig. 3a illustrates the periodic pattern in the sequence of calf collagen α1-chain, as a typical example of internal sequence repetition of the short segment. This periodicity of the three-residues covers all the sequence without any interruption of the repetition along the sequence, and reflects that the Gly-X-Y triplet sequence is minimal for triple-chain helix formation, where proline (Pro) can usually occur in the position X, and Hydroxyproline (Hyp) in the position Y.

As the next example for a protein with short repeating segments, sheep keratin B2A was taken. As shown in Fig. 3b, the molecule reveals a clear periodicity of the five residues intervals. Interestingly, the periodicity interrupts at the 50th residue lag, and appear again from the 73th residue lag after the disturbance of a lag of 23 residues. It should be noted that the periodic pattern has relatively higher peaks at ten residue intervals than that five residue intervals.

The pattern for light meromyosin, a typical rod-shaped molecule[28] is of interest, since the molecule seems to be similar to α-tropomyosin (α-TM) which shows a marked periodicity of seven residue intervals throughout the whole sequence as reported before.[6] In contrast to α-TM, however, the periodicity at seven residue intervals was not clearly indicated as seen in Fig. 3c, suggesting that the molecule, as a coiled-coil molecule which favors hydrophobic interactions between two subunits, is less

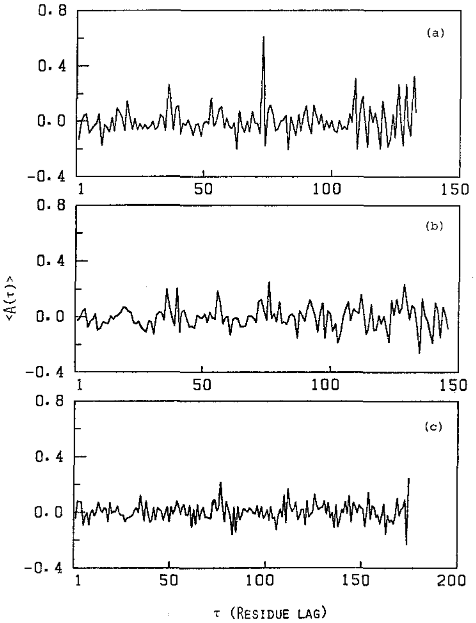

Fig. 4. (a) Plots of autocorrelation coefficients ⟨A(τ)⟩ *vs.* the residue lag τ for calmodulin from bovine brain. The observed major peak (0.61) at the residue lag of 73 clearly shows that the molecule is composed of two homologous segments. (b) Plots of autocorrelation coefficients ⟨A(τ)⟩ *vs.* the residue lag τ for cardiac TN-C. The disturbance of internal repetition is recognizable, judging from the peak height near the middle of the sequence. (c) Plots of autocorrelation coefficients ⟨A(τ)⟩ *vs.* the residue lag τ for myosin A1 light chain. Internal repetition is so weak.

stabilized than α-TM.

Calmodulin has been considered to be a calcium binding protein composed of two homologous sequences each of which contains two Ca^{2+}-binding sites.[29] Therefore, it is of interest to test another proteins, which seem to be related to calmodulin, in terms of autocorrelation. Fig. 4a shows the result on calmodulin, as a typical example for a calcium binding protein. One major peak (0.61) at the residue lag of 73 residues and two minor peaks (0.27 and 0.31) at lags of 36 and 109 residues are observed respectively, clearly showing that the molecule is composed of two homologous segments (probably produced by a gene duplication). Fig. 4b and 4c show the results on cardiac TN-C and myosin A1 light chain, respectively. The peaks near the middle of the sequences are much smaller than that for calmodulin, indicating that internal sequence repetitions in these molecules are so weak. Such disturbances of internal repetitions in cardiac TN-C and myosin A1 light chain may be related to the reduction of the number of Ca^{2+}-binding sites.

The average correlation coefficient, $\langle C(j) \rangle$, is applicable to identification of sequences in one protein homologous to a specified sequence, especially biologically important sites, in another protein. For instance, Ca^{2+}-binding sites in parvalbumin, the CD and EF hand sequences, which are known to have the specific structures for Ca^{2+}-binding, may be adopted as the reference sequences. Calmodulin was tested to identify the candidates for the Ca^{2+}-binding sites in the sequence, using

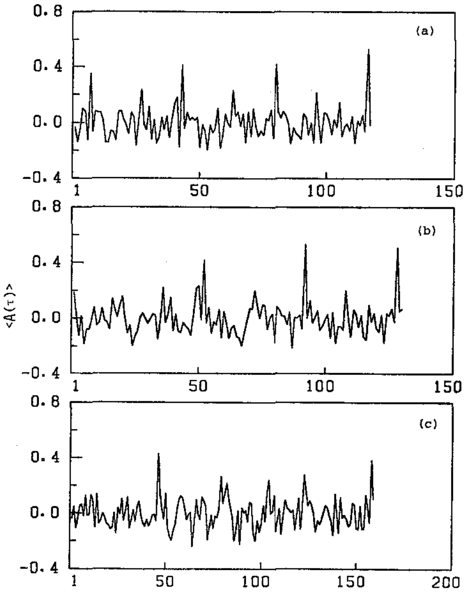

Fig. 5. (a) Comparison of calmodulin with the EF hand sequence of carp parvalbumin by plotting correlation coefficients $\langle C(j) \rangle$ against the residue number. (b) Comparison of cardiac TN-C with the EF hand sequence by plotting correlation coefficients $\langle C(j) \rangle$ against the residue number. (c) Comparison of myosin A 1 light chain with the EF hand sequence by plotting correlation coefficients $\langle C(j) \rangle$ against the residue number.

the EF hand sequence. As shown in Fig. 5a, four peaks were clearly observed, indicating the starting residues of a homologue to the EF hand sequence as the residues 7, 43, 80 and 116, respectively. In this way, the candidates for the Ca^{2+}-binding sites may be easily detected by the peak position of correlation coefficients. When the CD hand sequence was used for the comparison, a similar result was obtained, although peak IV was as low as 0.2. Fig. 5b and 5c were obtained for cardiac TN-C and myosin A 1 light chain, respectively, by scanning with the EF hand sequence, illustrating smaller numbers of the candidates for the Ca^{2+}-binding sites in the sequences. Cardiac TN-C lacks the peak I, and the number of bound Ca^{2+}-ions seems to be reduced from four to three (Fig. 5b). For myosin A 1 light chain, only two peaks were observed for homologous regions to the EF hand sequence, suggesting that the candidates for the Ca^{2+}-binding site were at most two regions, starting from residues 46 and 158, respectively. (Fig. 5c) When scanned with the CD hand sequence, similar results were obtained for these proteins.

The procedure can be extended to the search of homologous sequences with deletions or insertions in any two different (or same) proteins, by plotting correlation coefficients, $\langle C(i,j)\rangle$s, in a square array. Fig. 6 shows the comparison matrix for skeletal TN-C and calmodulin which are regarded as homologous proteins. Successive large dots which represent $\langle C(i,j)\rangle$s greater than 0.6 appear near along the diagonal, indicating that these proteins are closely related to each other. As seen from the correlation values, the homology of half of the molecules in the N-termnal side is good. Two alignments of large dots other than those near the diagonal, one 90th

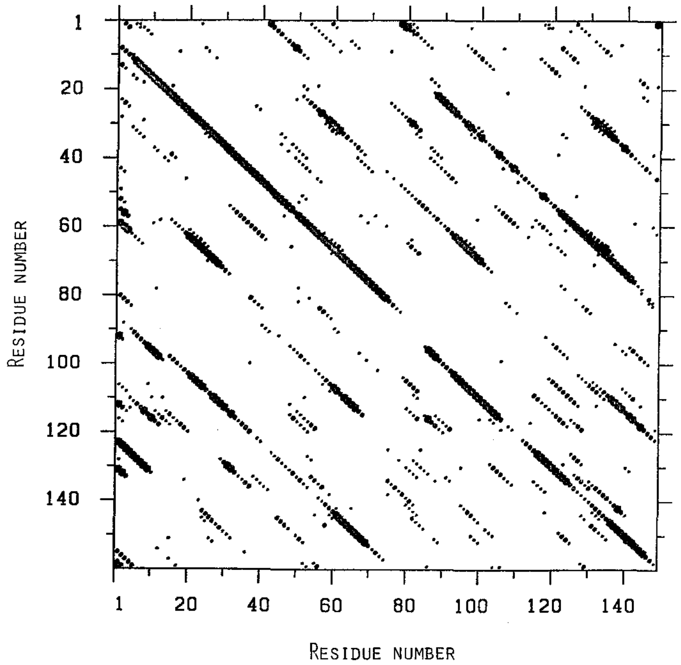

Fig. 6. Comparison matrix for skeletal TN-C (ordinate) and calmodulin (abscissa). Small, medium, and large dots are the same values as in Fig. 1.

to 157th for TN-C and 4th to 74th for calmodulin (in the lower half of Fig. 6), and the other 21st to 78th for TN-C and 87th to 144th for calmodulin (in the upper half of Fig. 6) correspond to the homologous Ca^{2+}-binding regions, respectively, which contain two Ca^{2+}-binding sites. This implies that each of the molecules is divided into two similar sequences (see Fig. 4a).

According to the recent study on vitamin D-dependent calcium-binding protein from bovine intestine (ICaBP) by X-ray crystallography,[49] the protein contains two Ca^{2+}-binding sites similar to those in parvalbumin, but one domain, the I-II domain (residues from 3 to 36), is rather different; it has a larger and rearranged Ca^{2+}-binding loop. When the sequences of these proteins were compared, the interesting features could be obtained (Fig. 7); first, the III-IV domain (residues from 46 to 74) is homologous to the Ca^{2+}-binding sites in parvalbumin, although the extent of homology (*i.e.*, correlation values) of the EF hand sequence is relatively higher than that of the CD hand sequence. Secondly, the I-II domain is less homologous to the Ca^{2+}-binding sites in parvalbumin, especially to the EF hand sequence. Some insertions in the I-II domain against the CD hand sequence can be recognized, although the precise locations are not identified. This is consistent with the result of X-ray crystallographic study, and it is remarkable that this was obtained by the knowledge of the amino acio sequence only.

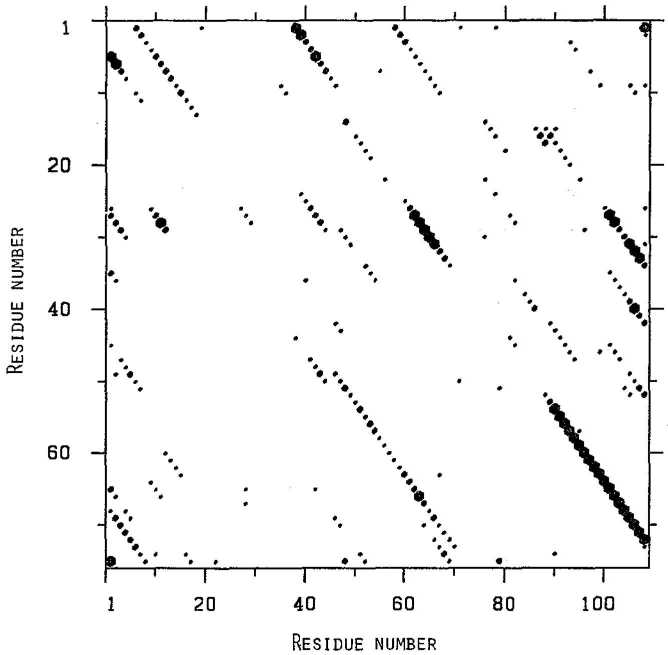

Fig. 7. Comparison matrix for ICaBP (ordinate) and parvalbumin (abscissa). Small, medium, and large dots are the same values as in Fig. 1.

The quantitative estimate of relationship between sequence homology and structural homology will be examined in the next section.

Correspondence to tertiary structure In oider to examine correspondence of the sequence homology by the present method to the structural homology, bovine trypsinogen and porcine elastase were compared as shown at the top of Fig. 8. Deviation of the corresponding C^{α} atom, l_{ii}, of i-th residue obtained by the superposition technique was plotted against the residue number of elastase, at the bottom of Fig. 8. In the comparison matrix large dots of correlation values greater than 0.6 appear near along the diagonal, suggesting that these proteins are closely related to each other. A good homology is present for the segment containing the active sites of serine protease, His 46 and Ser 183 of trypsinogen, but the sites near Asp 90 for trypsinogen and Asp 93 for elastase exhibit a poor homology, suggesting that the conformation may vary in these regions. In fact the region near Asp 93 for elastase seems to coincide poorly with the region near Asp 90 for trypsinogen (over 3A). The regions of disulfide bonds connecting residues 31–47, 154–168, 179–203, 122–189, 13–143, and 115–216, of trypsinogen are located in the poorly homologous regions, at most the ends of good homologous regions. Interestingly, a region of poor correlations to connect the homologous segments is located at the surface of the molecules. The results show that the homology of sequence gives rise to homology of tertiary structure; the deviation for the homologous regions is less than 2–3A. This is reasonable in a sense, because an effort to select the

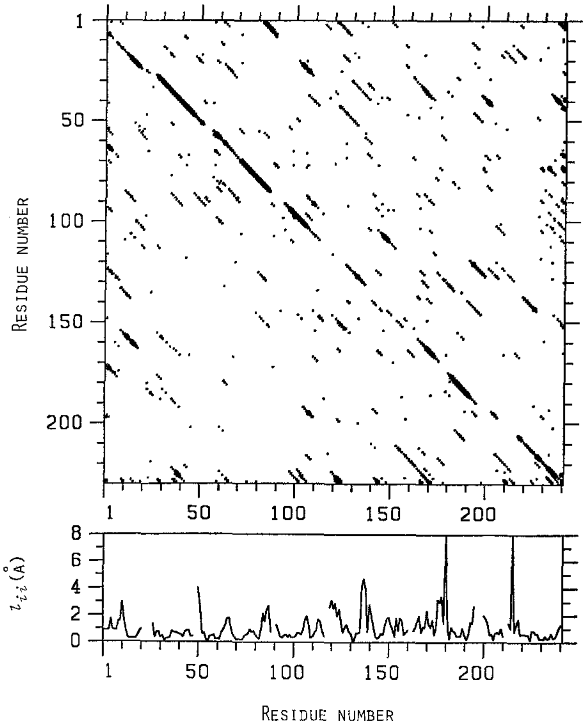

Fig. 8. Top: comparison matrix for bovine trypsinogen (ordinate) and porcine elastase (abscissa). Bottom: plots of deviation in position of C^{α} atom of i-th residue, l_{ii}, against the residue number of elastase. The alignment of the sequence was according to the sequence homology. Small, medium, and large dots are the same values as in Fig. 1.

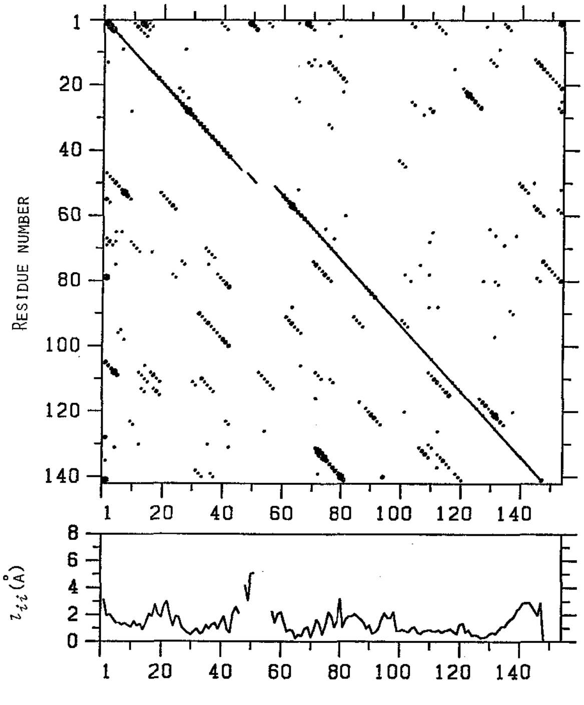

Fig. 9. Top: comparison matrix for hemoglobin α-chain (ordinate) and myoglobin (abscissa). Bottom: plots of l_{ii} against the residue number of myoglobin. The alignment of the sequence was according to the sequence homology. Small, medium, and large dots are the same values as in Fig. 1. The alignment of the sequences is shown by a thin line.

parameters used for detection of homology was made so as to give a good correspondence between the homology of sequence and structure as much as possible.

When hemoglobin α-chain and β-chain, which have been regarded as closely related to each other in globin family, were compared, a good correspondence of the same degree as for the case of serine proteases was obtained as reported previously.[7] Here, hemoglobin α-chain and myoglobin were compared, as an another interesting pair for comparison in globin family. As shown in Fig. 9, the deletion of D-helix in α-chain is recognizable as a blank region after the 42nd residue of myoglobin near the diagonal. In contrast to the good homology of three quarter of α- and β-chain in the C-terminal side, it was found that the homology of half of α-chain and myoglobin in the N-terminal side was good, judging from the distribution of large dots lying near the diagonal. These homologous regions are corresponding to A (C-terminal side of the helix), B, C and E-helix, and give rise to homologous three-dimensional structure with the deviations less than 2–4A.

IV. DISCUSSION

The statistical technique of correlation coefficient calculated from the digitalized amino acid sequence was successfully applied to quantitative estimation of sequence homology. In the first place the present method provides a way of detecting the sequence repetitions by using the average autocorrelation coefficient, $\langle A(\tau)\rangle$. There are two major types in the sequence repetitions: one is the repetition of a short segment as shown in Fig. 3, and the other is that of a long segment (Fig. 4). It has been proposed that the repeating sequence arose from gene duplication or multiplication during evolution.[50] The result on calmodulin supports this possibility, since it has the strong internal sequence repetition as seen from the peak height in the middle (Fig. 4a). On the other hand, the repetition of sequences of collagen and light meromyosin may be interpreted in terms of structural requisite rather than gene duplication; for example, LMM has to satisfy a multiple seven residues lag conditions like α-tropomyosin, because its coiled-coil structure requests the periodic location of hydrophobic residues to associate two α-helical subunits (Fig. 3c). Similar inference seems to be valid for collagen as the triple stranded, coiled-coil molecule (Fig. 3a). In the case of keratin molecule, however, the prominent repeating units are found to be five residues long by the present study, rather than ten residues long as reported by Elleman,[27] and in the absence of data at present it is not possible to say whether this repetition was caused by structural requisite or gene duplications.

Secondly, the use of the average correlation, $\langle C(j)\rangle$, helps to search a functional region such as Ca^{2+}-binding region. Calmodulin gives a typical example in assessing the effectiveness of the attempts to detect the candidates of Ca^{2+}-binding sites as seen in Fig. 5a. Because of integrity of calmodulin as a calcium binding protein having the four Ca^{2+}-binding sites, the average autocorrelation coefficient, $\langle A(\tau)\rangle$, gave the dominant peak in the middle of the sequence, indicating that the molecule is composed of two strong homologous sequences. On the other hand, in the case of cardiac TN-C and myosin A 1 light chain the peak near the middle is very weak (Fig. 4b and 4c) and the homology within the sequence may have been obscured by the reduction of the number of Ca^{2+}-binding sites in the proteins (Fig. 5b and 5c).

When the method is extended to a square array, the average correlation coefficient, $\langle C(i,j)\rangle$, can be used successfully for the detection of homologous segments of any two proteins, and also the comparison matrix is convenient to visualize the overall and/or partial sequence homology. The use of comparison matrix has further advantage for detection of homologous sequence with a deletion or insertion in one sequence, because high correlative values will be successively observed on a neighbouring diagonal after a break at the position, but the use of the correlation coefficient, $\langle A(\tau)\rangle$ and $\langle C(j)\rangle$ is not effective for this case.

Comparison of ICaBP and parvalbumin gave the interesting feature for the examination of relation between sequence and tertiary structure (Fig. 7). The sequence homology estimated by the present method suggests that the III-IV domain in ICaBP is more homologous to the EF hand sequence than the CD hand sequence in parvalbumin, and this result is consistent with that of the X-ray crystallography;[49] that is, the r.m.s. deviations for the EF and CD hand regions are 2.40A and 3.25A, respectively. However, the I-II domain is rather different from the III-IV domain; there are two insertions, Ala 15 and Asn 21 in the domain and the r.m.s. deviations for the EF and CD hand regions are comparable to each other, i.e., 2.75A and 2.86A, respectively, as reported. Nevertheless, the sequence of the domain is more homologous to the CD hand sequence than the EF hand sequence, although the homology is relatively weaker than that for the III-IV domain and the EF hand sequence. Thus, the sequences of the CD and EF hand regions seem to be two extreme amino acid sequences despite the similarity of their tertiary structures. Because of the absence of homology between Ca^{2+}-binding sites of the I-II domain and the CD hand region, it is difficult to assign the domain as a candidate of a Ca^{2+}-binding site from the present result. As the whole molecule, it is found that ICaBP resembles to the two-third of parvalbumin, from the residues of 38 to 108, although some absence of homology in the II-III connecting region and helix III is recognizable, and this is supported from the X-ray crystallography.

The present method extended to a square array is applicable to estimating

```
                                                       10                  20                  30
Horse (α-chain)                     V - L S A A D K T N V K A A W S K V G G H A G E Y G A E A L E

Myoglobin                           V - L S E G E W Q L V L H V W A K V E A D V A G H G Q D I L I
Lamprey             P I V D S G S V A P L S A A E K T K I R S A W A P V Y S N Y E T S G V D I L V
Erythrocruorin                          L S A D Q I S X V Z A S F D K V K - - - - G D P V G I L Y
Leghemoglobin                       V A F T E K Q D A L V S S S F E A F K A N I P Q Y S V V F Y T

                                     40                    50                            60
                    R M F L G F P T T K T Y F P H F - D L S H - - - - - G S A Q V K A H G K K V G

                    R L F K S H P E T L E K F D R F K H L K T E A E M K A S E D L K K H G V T V L
                    K F F T S T P A A Q E F F P K F K G M T S A D E L K K S A D V R W H A E R I I
                    A V F K A D P S I M A K F T Q F A G - K D L E S I K G T A P F H T E A N R I V
                    S I L E K A P A A K D L F S F L A N G V D - - - - P T N P K L T G H A E K L F

                                 70                    80                  90                 100
                    - D A L T L A V G H L D D - L P G A L S D L S N L H A H K L R V D P V N F K L

                    - T A L G A I L K K K G H - H E A E L K P L A Q S H A T K H K I P I K Y L E F
                    - N A V N D A V A S M D D T E K M S M K D L S G K H A K S F Q V D P Q Y F K V
                    - G F F S K L I G E I P E N I A D V N T F P A S K H V - - R G V T H B Z L B B
                    A L V R D S A G Q L K A S G T V V A D A A L G S V H A Q - K A V T N P E F - V

                                    110                 120                 130
                    L S H C L L S T L A V H L P N D F T P A V H A S L D K F L S S V S T V L T S K

                    I S E A I I H V L H S R H P G N F G A D A Q G A M N K A L E L F R K D I A A K
                    L A - V I - - - - - - - - - A D T V A A G D A G F E K L S M I C I L M L R S A
                    F R A G F V S Y M K A H - - T D F G A E A - A A W G T L A D T F F G M I F S K
                    V K E A L L K T I K A A V G D K W S D E L S R A W E V A Y D E L A A A I K A K

                   140
                    Y R

                    Y K E L G Y Q G
                    Y
                    M
```

Fig. 10. The alignments of the sequences of proteins in globin family against hemoglobin α-chain. The regions of good homologous segments ($\langle C(i,j)\rangle$ greater than 0.4) are underlined. The alignments are referred to Ref. (1).

numerically the extent of homology along the alignment of protein sequences such as globin family as shown in Fig. 10. The comparison was performed on the proteins in globin family which seem to be distantly related to each other. The alignments of homologous sequences against α-chain are shown, where the regions of good homologous sequences ($\langle C(i,j)\rangle$ greater than 0.4) are underlined. There were two segments, which had greater correlation coefficients than 0.4, and the length of the aligned dots greater than 10, according to the sequence homology of hemoglobin α-chain and myoglobin. Interestingly, the regions of good homologous sequences were obtained on the N-terminal sides for vertebrate proteins. On the other hand, the sequence of leghemoglobin has a good homology on the C-terminal side.

The present method using correlation coefficients could be effectively applied to the detection of repetitions and homologous regions in amino acid sequences, and also provides a way for estimation of relationship between the sequence homology and the structural homology in proteins.

ACKNOWLEDGMENTS

The author would like to acknowledge the continuing guidance and encouragement of Prof. Dr. Tatsuo Ooi. He is also grateful to Dr. Sho Takahashi and Dr. Ken Nishikawa for helpful suggestions.

REFERENCES

(1) M. O. Dayhoff Ed., "Atlas of Protein Sequence and Structure," Vol. 5, Suppl. 3, National Biomedical Research Foundation, Washington, D. C. 1978.
(2) D. R. Cox and P. A. W. Lewis, "The Statistical Analysis of Series of Events," Chapman and Hall, London, 1978.
(3) H. H. Harman, "Mathematical Methods for Digital Computers," A. Ralston and H. S. Wilf Ed., Vol. 1, John Wiley & Sons Inc., New York, N.Y. 1960, pp. 204–220.
(4) A. A. Afifi and S. P. Azen, "Statistical Analysis, A Computer Oriented Approach," 2nd ed., Academic Press, New York, N.Y. 1979, pp. 324–341.
(5) R. H. Kretsinger and C. E. Nockolds, *J. Biol. Chem.*, **248**, 3313 (1973).
(6) Y. Kubota, S. Takahashi, K. Nishikawa and T. Ooi, *J. Theor. Biol.*, **91**, 347 (1981).
(7) Y. Kubota, K. Nishikawa, S. Takahashi and T. Ooi, *Biochim. Biophys. Acta*, **701**, 242 (1982).
(8) M. Levitt, *Biochemistry*, **17**, 4277 (1978).
(9) O. B. Ptitsyn and A. V. Finkelstein, *Quart. Rev. Biophys.*, **13**, 339 (1980).
(10) S. Lifson and C. Sander, *Nature*, **282**, 109 (1979).
(11) P. Y. Chou and G. D. Fasman, *Adv. Enzymol.*, **47**, 45 (1978).
(12) P. Manavalan and P. K. Ponnuswamy, *Nature*, **275**, 673 (1978).
(13) E. J. Cohn and J. T. Edsall, "Proteins, Amino Acids, and Peptides," Van Nostrand-Reinhold, Princeton, New Jersey, 1943.
(14) H. B. Bull and K. Breese, *Arch. Biochem. Biophys.*, **161**, 665 (1974).
(15) H. A. Sober Ed., "Handbook of Biochemistry, Selected Data for Molecular Biology," 2nd ed, The Chemical Rubber Co., Cleveland, Ohio, 1970.
(16) A. M. Lesk and C. Chothia, *J. Mol. Biol.*, **136**, 225 (1980).
(17) M. Oobatake and T. Ooi, *J. Theor. Biol.*, **67**, 567 (1977).
(18) M. Levitt, *J. Mol. Biol.*, **104**, 59 (1976).
(19) J. M. Zimmerman, N. Eliezer and R. Simha, *J. Theor. Biol.*, **21**, 170 (1968).
(20) D. D. Jones, *J. Theor. Biol.*, **50**, 167 (1975).
(21) J. Janin, *Nature*, **277**, 491 (1979).
(22) W. R. Krigbaum and B. H. Rubin, *Biochim. Biophys. Acta*, **229**, 368 (1971).

(23) R. Grantham, *Science,* **185**, 862 (1974).
(24) K. Nishikawa and T. Ooi, *Int. J. Peptide Protein Res.,* **16**, 19 (1980).
(25) D. H. Wertz and H. A. Scheraga, *Macromolecules,* **11**, 9 (1978).
(26) P. P. Fietzek and K. Kühn, *Eur. J. Biochem.,* **52**, 77 (1975).
(27) T. C. Elleman, *Biochem. J.,* **130**, 833 (1972).
(28) M. Elzinga and B. Trus, in "Methods in Peptides and Protein Sequence Analysis," C. Birr Ed., Elsevier, Amsterdam, 1980, pp. 213–224.
(29) D. M. Watterson, F. Sharief and T. C. Vanaman, *J. Biol. Chem.,* **255**, 962 (1980).
(30) J.-P. van Eerd and K. Takahashi, *Biochemistry,* **15**, 1171 (1976).
(31) G. Frank and A. G. Weeds, *Eur. J. Biochem.,* **44**, 317 (1974).
(32) J. H. Collins, M. L. Greaser, J. D. Potter and M. J. Horn, *J. Biol. Chem.,* **252**, 6356 (1977).
(33) C. S. Fullmer and R. H. Wasserman, *J. Biol. Chem.,* **256**, 5669 (1981).
(34) C. J. Coffee and R. A. Bradshaw, *J. Biol. Chem.,* **248**, 3305 (1973).
(35) O. Mikeš, V. Holeyšovský, V. Tomášek and F. Šorm, *Biochem. Biophys. Res. Comm.,* **24**, 346 (1966).
(36) D. M. Shotton and B. S. Hartley, *Nature,* **225**, 802 (1970).
(37) G. Braunitzer and G. Matsuda, *J. Biochem.,* (Tokyo) **53**, 262 (1963).
(38) A. B. Edmundson, *Nature,* **205**, 883 (1965).
(39) F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi and M. Tasumi, *J. Mol. Biol.,* **112**, 535 (1977).
(40) R. C. Ladner, E. J. Heidner and M. F. Perutz, *J. Mol. Biol.,* **114**, 385 (1977).
(41) T. Takano, *J. Mol. Biol.,* **110**, 569 (1977).
(42) H. Fehlhammer and W. Bode, *J. Mol. Biol.,* **98**, 683 (1975).
(43) W. Bode and P. Schwager, *J. Mol. Biol.,* **98**, 693 (1975).
(44) L. Sawyer, D. M. Shotton, J. W. Campbell, P. L. Wendell, H. Muirhead, H. C. Watson, R. Diamond and R. C. Ladner, *J. Mol. Biol.,* **118**, 137 (1978).
(45) G. Braunitzer and H. Fujiki, *Naturwissenschaften,* **56**, 322 (1969)
(46) G. Buse, S. Braig and G. Braunitzer, *Hoppe-Seyler's Z. Physiol. Chem.,* **350**, 1686 (1969).
(47) N. Ellfolk and G. Sievers, *Acta Chem. Scand.,* **23**, 2994 (1969).
(48) K. Nishikawa and T. Ooi, *J. Theor. Biol.,* **43**, 351 (1974).
(49) D. M. E. Szebenyi, S. K. Obendorf and K. Moffat, *Nature,* **294**, 327 (1981).
(50) W. C. Barker, L. K. Ketcham and M. O. Dayhoff, *J. Mol. Evol.,* **10**, 265 (1978).