Bull. Inst. Chem. Res., Kyoto Univ., Vol. 63, No. 2, 1985

Optimization of Amino Acid Parameters for Correspondence of Sequence to Tertiary Structures of Proteins

Motohisa OOBATAKE*, Yasushi KUBOTA** and Tatsuo OOI*

Received May 4, 1985

New five parameters, numerical values of which represent some aspects of amino acid residues, were derived to complement previous six parameters[1,2] and were used successfully for detection of sequence homology of proteins and that of structural correspondence of homologous segments. First, 34 parameters from known physical quantities were used as initial values of optimization and 18 convergent parameters were found by optimizing correlation coefficient for four pairs of homologous segment (Set I). Then, these optimum parameters were classified into five groups by a factor analysis and final five parameters near the factor axes were selected. Using these five parameters for 12 homologous proteins (Set II), structural homologous segments were obtained with a high probability (above 90%) when two segments of greater than ten residues long have successive correlation coefficients C_{seq} values of greater than 0.5. In addition to homologous proteins, structural correspondence for all 15 non-homologous proteins (Set III) were investigated. The extent of coincidence becomes worse than the homologous case under the same condition; none the less, it becomes better under the conditions that the length of segments to be compared is more than 13 residues with successive C_{seq} of greater than 0.5. Furthermore, it is shown that structural correlation coefficient C_{str} defined by a contact number of C^{α} atoms calculated only by X-ray C^{α} coordinates' data correctly selects the homologous segments with r.m.s. deviation 1.42Å on average without information of amino acid sequences.

KEY WORDS: Protein/ Homology/ Sequence/ Tertiary structure/

INTRODUCTION

Many physical parameters inherent in amino acids (e.g., hydrophobicity, propensity to form α-helix and β-structure, etc.) have been obtained by various investigators from such view points as experiments on amino acids and the statistical analysis of protein structures. Some of the parameters such as polarity have been used for detection of sequence homology of proteins by the recently developed method of correlation coefficients,[1,2] since these parameters seem to reflect the tertiary structure of proteins. In this method, a protein sequence is expressed numerically by replacing amino acid residues by a physical index property, and the six parameters including partial specific volume, etc., selected from many inherent values have been used as the best set of parameters for detection of protein homology. In order to complement these parameters, it is necessary to search another best set of parameters started from the initial physical values by the minimization method, and also to reduce the number of parameters used (at the present time, six parameters) to save computing time.

* 大島玄久，大井龍夫: Laboratory of Physical Chemistry of Enzyme, Institute for Chemical Research, Kyoto University, Uji, Kyoto

** 窪田 綏: Biological Information Division, CSK Research Institute, 4–39–5 Tamagawa, Setagaya, Tokyo 158

In this study, we construct a function with 20 variables (i.e., the number of type of amino acids) and minimize the function with the 34 collected parameters as the initial values for each amino acid residue. The optimized parameters obtained by minimization will be calssified into groups according to a factor analysis[3] and we will select the best set of parameters used for the calculation of average correlation coefficient. We will also discuss the efficiency of the parameters obtained by this way for detection of sequence homology of proteins and check the extent of structural correspondence of homologous segments.

METHODS

Optimization of physical parameters for amino acids

For two protein sequences, X and Y, a correlation coefficient, $C_p(i, j)$, already used,[1,2] is defined as follows:

$$C_p(i,j)=\frac{\sum_{l=-k}^{k}(x_p(i+l)-\langle x_p\rangle)(y_p(j+l)-\langle x_p\rangle)}{\{[\sum_{l=-k}^{k}(x_p(i+l)-\langle x_p\rangle)^2]\cdot[\sum_{l=-k}^{k}(y_p(j+l)-\langle x_p\rangle)^2]\}^{1/2}} \tag{1}$$

where $x_p(i)$ is an index value of a parameter p specific to an amino acid residue at the position i in X, $y_p(j)$ at the position j in Y, $2k+1$ (e.g., 21 for $k=10$) is the lengths of segment to be compared, and $\langle x_p\rangle$ is an average value over twenty kinds of amino acids. If two segments of the length n are homologous in tertiary structure (e.g., coincidence of the two conformations within 2 Å for C^α atoms on average), the correlation coefficient is desired to be as high as possible, and if not homologous (e.g., greater than 5 Å in r.m.s. deviation for C^α atoms) the two segments should have little correlation, that is, the values of correlation coefficient, $C_p(i, j)$, should be near zero. Our problem is to find such optimized parameters which reflect homology of tertiary structure of proteins in terms of correlation coefficient. In order to obtain such optimized parameters which give high correlation coefficients for the segments having similar conformations and correlation coefficients near zero for those of different conformations, we have to solve the optimum problem, that is, we minimized a following objective function with 20 variables:

$$f=\frac{1}{l'}\sum_{a'=1}^{l'}\left|\frac{1}{N_{a'}}\sum_{i,j}^{N_{a'}}C_{p'}(i,j)\right|-\frac{1}{l}\sum_{a=1}^{l}\frac{1}{N_a}\sum_{i,j}^{N_a}C_p(i,j) \tag{2}$$

where $C_p(i, j)$ is the correlation coefficient of a parameter p between two segments of the lengths N_a for reference proteins having similar conformations, and $C_{p'}(i, j)$ is that between two segments of the lengths $N_{a'}$ having different conformations (e.g., greater than 5 Å in r.m.s. deviation). As the reference proteins (or segments) employed in the second term of Eq. (2) we selected four pairs (Set I) of following homologous proteins (i.e., $l=4$): two Ca^{2+} binding regions of carp parvalbumin (residues 42–66 and 81–105), hemoglobin α-chain (1–42, 54–93, and 95–134) and myoglobin (1–42, 60–99, and 101–140) for α-proteins, and regions I (1–23) and III (88–110), II (28–84) and IV

(117–173) of γ-crystallin for β-proteins, respectively. As the non-homologous region employed in the first term of Eq. (2) (i.e., $l'=1$), one segment from each of hemoglobin α-chain (79–88) and myoglobin (39–48) with initially very high correlation values calculated from parameters such as hydrophobicity despite poor homology in tertiary structure (greater than 5Å in r.m.s. deviation) is taken. As the initial values for optimization of parameters through Eq. (2), the 34 collected parameters listed in Table I were used.[4-22] The optimized parameters obtained by Eq. (2) in this way will be classified into groups according to the factor analysis,[3] which is available to select the best set of n parameters used for the average sequence correlation coefficient, $C_{seq}(i,j)$;

$$C_{seq}(i,j)=\frac{1}{n}\sum_{p=1}^{n} C_p(i,j) \tag{3}$$

This arithmetic average was made by reason of reduction of signal-noise ratio. In order to test how well the extent of the optimized parameters determined by the above procedure reflects the tertiary structures of proteins, we compare the following sets (Set II) of homologous proteins which have similar tertiary structures but different amino acid sequences (abbreviated name in protein data bank): Ca-binding sites of parvalbumin (1CPV); hemoglobin α-chain, β-chain (2MHB), and myoglobin (2MBN) for the globin family; elastase (1EST), beta-trypsin (2PTN), subtilisin (1SBT), and proteinase b (2SGB) for the serine protease; hen egg-white lysozyme (2LYZ) and bacteriophage T4 lysozyme (1LZM); trypsin inhibitor (3PTI) and ovomucoid (1OVO); cytochrome C_2 (1C2C) and cytochrome C (3CYT) for the cytochrome family; ferredoxin having 2-fold structural repeat (1FDX); azulin (1AZU) and plastocyanin (1PCY); four regions of γ-crystallin (1GCR). A similar estimation was made for the following 15 proteins which have non-homologous tertiary structures each other but different amino acid sequences (Set III); parvalbumin, cytochrome C and myoglobin for α-proteins; prealbumin (2PAB), superoxide dismutase (2SOD), elastase, concanavalin A (3CNA), and γ-crystallin for β-proteins; trypsin inhibitor, ribonuclease A (1RN3), hen egg-white lysozyme, thermolysin (3TLN), and phospholipase A2 (1BP2) for $\alpha+\beta$ proteins; lactate dehydrogenase (4LDA), and triose phosphate isomerase (1TIM) for α/β proteins. A survey of comparison was made for all combinations of these proteins. Since the three-dimensional structures of the above proteins are known,[23] we can estimate quantitatively the extent of correspondence between homology of the sequence and tertiary structure.

Extent of the coincidence in tertiary structure

We have two methods to estimate the coincidence of tertiary structures. One is the superposition method[24] which is done by superposing a three-dimensional structure of one protein or segments on another so that the sum of the squares of deviations between corresponding C^α atmos is minimized. Complementary, the r.m.s. deviation by a method of distance between residues, $[\sum_{i,j}^{N}(r_{ij}-r'_{ij})^2/N]^{1/2}$, is also calculated where r_{ij} and r'_{ij} is the distance between C^α atoms of i-th and j-th residue of protein X and X′, respectively.

The other is a method to use correlation coefficients between two conformations of proteins (or segments). Let us consider two protein structures, X and X′. The structure correlation coefficient, $C_{str}(i, j)$, between a partial tertiary structure centered at the i-th C^{α} atom of X, and that at the j-th C^{α} atom of X' is defined as follows:

$$C_{str}(i, j) = \frac{\sum_{l=-k}^{k} (n(i+l) - \langle n(i) \rangle)(n'(j+l) - \langle n'(j) \rangle)}{\{[\sum_{l=-k}^{k} (n(i+l) - \langle n(i) \rangle)^2] \cdot [\sum_{l=-k}^{k} (n'(j+l) - \langle n'(j) \rangle)^2]\}^{1/2}} \quad (4)$$

where

$$\langle n(i) \rangle = \frac{1}{2k+1} \sum_{l=-k}^{k} n(i+l),$$

$$\langle n'(j) \rangle = \frac{1}{2k+1} \sum_{l=-k}^{k} n'(j+l),$$

$n(i)$ is a contact number which indicates a number of C^{α} atoms within 8 Å centered at the i-th C^{α} atom in X, $n'(j)$ at the j-th C^{α} atom in X′, and $k=10$. This method is especially advantageous when we focus on the correspondence between two primary structures from C^{α} coordinates of proteins having nearly the same tertiary structures. For example, when we compare two α-helices having structural repetition with the periodicity of 3–4 residues by the superposition method, one amino acid residue corresponds to any amino acid residue in the other helix because of superposition on each other by rotating one helix, but for the method of structure correlation coefficient, only the amino acid residue with the same environment in space corresponds if the contact number is taken into account. Thus both methods should be used together when deriving the correspondence between two amino acid sequences by comparing those tertiary structures from C^{α} coordinates. All computations were performed with FACOM M180 II AD at the computing center of the Institute for Chemical Research, Kyoto University.

RESULTS

Optimization of physical parameters

After the calculation of Eq. (2) using the initial values of the 34 parameters listed in Table I, of these the 18 parameters such as propensity to form reverse turn (p21) were found to converge to the optimum values (marked by an asterisk in Table II). That is, correlation values, $C_p(i, j)$, from homologous regions using their convergent values were greater than 0.5 and those for non-homologous regions were near zero (the values listed in Table II). Since the mutual correlations of these optimum parameters are not independent but correlate more or less with each other, we classified these parameters into groups according to the factor analysis. As listed in Table II on the factor analysis, the 18 optimum parameters were classified into five factors (FI to FV), and then we took five parameters near the factor axes taking one from the every group as represented by bold letters and an additive one (op25) from the group

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

Names	Parameters	References	Convergence value of $C_p(i,j)$ Homologous regions(**)	Non-homologous region(***)
P1	Propensity to form β-structure (Levitt)	(4)	0.25, 0.43, 0.52	0.38
P2*	β-structure-coil equilibrium constant, s_β (Ptitsyn & Finkelstein)	(5)	0.66, 0.48, 0.52	0.00
P3*	Preference for parallel β-strands (Lifson & Sander)	(6)	0.69, 0.55, 0.50	0.00
P4*	Preference for β-strands (Lifson & Sander)	(6)	0.50, 0.51, 0.49	0.00
P5*	Propensity to form β-structure (Chou & Fasman)	(7)	0.79, 0.65, 0.53	0.00
P6*	Preference for antiparallel β-strand (Lifson & Sander)	(6)	0.56, 0.46, 0.55	0.00
P7	Average surrounding hydrophobicity, $\langle H \rangle$ (Manavalan & Ponnuswamy)	(8)	0.35, 0.45, 0.35	0.00
P8	Partial specific volume	(9)	0.45, 0.51, 0.37	0.26
P9*	Transfer energy (Bull & Breese)	(10)	0.69, 0.59, 0.54	0.00
P10	Molecular weight	(11)	0.08, 0.44, 0.40	−0.09
P11	Average volume of buried residue (Lesk & Chothia)	(12)	0.16, 0.48, 0.40	0.23
P12	Average non-bonded energy per residue (Oobatake & Ooi)	(13)	0.31, 0.49, 0.49	0.28
P13	Average percent in proteins (Dayhoff)	(14)	0.29, 0.32, 0.38	−0.16
P14	Dihedral angle between four successive C^α atoms (Levitt)	(15)	0.12, 0.35, 0.43	0.29
P15	Short range non-bonded energy per atom (Oobatake & Ooi)	(13)	−0.11, 0.44, 0.28	0.11
P16	Bulkiness (Zimmerman *et al.*)	(16)	0.23, 0.36, 0.54	0.00
P17*	Hydrophobicity (Jones)	(17)	0.63, 0.56, 0.57	0.00
P18	Propensity to form α-helix (Levitt)	(4)	0.10, 0.04, 0.34	0.13
P19	Propensity to form α-helix (Chou & Fasman)	(7)	0.20, −0.05, 0.34	−0.11
P20*	Helix-coil equilibrium constant, s_α (Ptitsyn & Finkelstein)	(5)	0.67, 0.60, 0.55	0.00
P21*	Propensity to form reverse turn (Levitt)	(4)	0.73, 0.68, 0.50	0.00
P22*	Propensity to form β-turn (Chou & Fasman)	(7)	0.66, 0.64, 0.54	0.00
P23	pK value of amino group (pK–N)	(11)	0.48, 0.37, 0.21	0.23
P24*	Transfer energy (Janin)	(18)	0.61, 0.41, 0.56	0.00
P25*	Average non-bonded energy per atom (Oobatake & Ooi)	(13)	0.64, 0.62, 0.52	0.00
P26*	Side chain interaction parameter ξ_s (Krigbaum & Rubin)	(19)	0.48, 0.50, 0.44	0.00
P27*	Transfer energy (Levitt)	(15)	0.73, 0.54, 0.54	0.00
P28*	Polarity (Grantham)	(20)	0.57, 0.47, 0.41	0.00
P29*	Contact number of a residue (Nishikawa & Ooi)	(21)	0.73, 0.57, 0.53	0.00
P30*	Propensity to bury inside of a molecule (Wertz & Scheraga)	(22)	0.66, 0.48, 0.46	0.00
P31	Polarity (Zimmerman *et al.*)	(16)	0.30, 0.33, 0.18	0.20

P32*	Long range non-bonded energy per atom (Oobatake & Ooi)	(13)	0.75,	0.61, 0.56	0.00
P33	Relative mutability (Dayhoff)	(14)	0.52,	0.24, 0.26	0.00
P34	pK value of carboxyl group (pK-C)	(11)	0.50,	0.39, 0.25	−0.08

*) 18 parameters of these converged to optimum values.

**) The values correspond to two Ca^{2+} binding regions of carp parvalbumin, hemoglobin α-chain and myoglobin, and regions I and III, II and IV of γ-crystallin, respectively.

***) This corresponds to hemoglobin α-chain (residues 79–88) and myoglobin (39–48).

Table II. Final pattern matrix of the 18 optimized parameters and selected parameters (*). Facotr loadings less than 0.25 are omitted and large values are represented by bold letters

	FI	FII	FIII	FIV	FV
op5	**0.95**				
op21*	**−0.92**			0.28	
op32	**−0.83**			0.34	
op29	**0.75**			−0.60	
op22	**−0.69**	0.31		0.58	
op27	**−0.68**			0.64	
op30	**0.59**		−0.50		0.51
op24*		**−0.91**	−0.30		
op25*	−0.45	**0.63**		0.55	
op26*			**0.94**		
op2*				**−0.91**	
op28	−0.33		0.32	**0.77**	
op3		−0.59		**−0.74**	
op20	0.62			**−0.74**	
op17	0.60			**−0.73**	
op9	−0.65			**0.73**	
op4	0.35			**−0.69**	0.56
op6	0.53			−0.54	**0.59**

FII were took. The values for amino acids of the five parameters, op2, op21, op24, op25, and op26 are shown in Table III. Most of these are correlative to the 34 physical parameters listed in Table I: the highest values of correlation coefficients between five selected parameters and 34 physical parameters is 0.79 for op2 and preference for all β-strands (p4), 0.76 for op21 and average non-bonded energy per residue (p12), 0.61 for op24 and transfer energy (p24), −0.63 for op25 and pK-C (p34), 0.78 for op26 and side-chain interaction parameters, ξ_s (p26), respectively. The set of five optimum parameters obtained in this way will be used in Eq. (3) (i.e., $n=5$), and homologous segments in the sequences can be identified by calculating comparison matrices obtained by plotting $C_{seq}(i,j)$ against the residue number, i, of one protein and j, of the other protein.[1)]

Table III. The values for 20 amino acids of the optimized five parameters

	op2	op21	op24	op25	op26
D	−2.05	3.32	−0.33	8.86	2.85
N	0.03	2.49	1.31	2.27	5.56
T	2.39	1.09	−1.52	−4.75	8.60
S	1.47	0.94	−0.83	−1.60	6.78
E	0.93	2.20	0.48	4.04	5.16
Q	1.02	1.49	−1.12	1.79	4.15
P	0.41	2.12	−0.58	5.19	5.14
G	0.12	2.07	0.64	−0.56	9.14
A	2.01	1.34	0.46	−2.49	4.55
C	1.98	1.07	0.20	−3.13	−0.78
V	3.50	1.32	0.54	−3.97	3.81
M	1.75	0.70	0.15	−4.96	2.18
I	3.70	0.66	3.28	−10.87	2.10
L	2.73	0.54	0.43	−7.16	3.24
Y	2.23	−0.17	−2.21	9.25	2.40
F	2.68	0.80	0.52	−6.64	4.37
K	2.55	0.61	−1.71	−9.97	10.68
H	−0.14	1.27	−1.31	4.22	4.48
R	0.84	0.95	−1.54	2.55	5.97
W	2.49	−4.65	1.25	−17.84	1.97

Selection of homologous sequences by structure correlation coefficients

Comparing two homologous proteins of Set II, homologous segments are selected from the C^{α} coordinates by picking up C_{str} successively greater than 0.6 for more than ten residues. Then these selected segments shown in Table IV are superposed with each other, so that all these corresponding segments have the good structure correspondences with r.m.s. deviation, 1.42 Å, on average. This correspondence of each amino acid between two proteins is the same as the conventional alignment.[14] Thus, this method is convenient to select automatically the homologous segments only from C^{α} coordinates' data and does not require the information on amino acid sequence.

Estimation of efficiency of the five optimized parameters

Structural correspondence depends in general on the length of homologous segments in protein sequences and their sequence correlation values. Therefore, we checked the extent of coincidence in three-dimensional structures of homologous segments of various length, i.e., comparing homologous proteins, segments having both the same tertiary structures and the same environment are picked up. Fig. 1 shows the extent of structural correspondence of homologous segments in which C_{seq} for all the residues is greater than 0.5 for various lengths of the homologous proteins of Set II, i.e., the numbers of segments selected from successive $C_{seq}(i, j)$ values of greater than 0.5 are plotted against the length of homologous segments as ratios to the total number (a function of the lengths, i.e., 27 to 9 for 5 to 15 of the lengths) of homologous

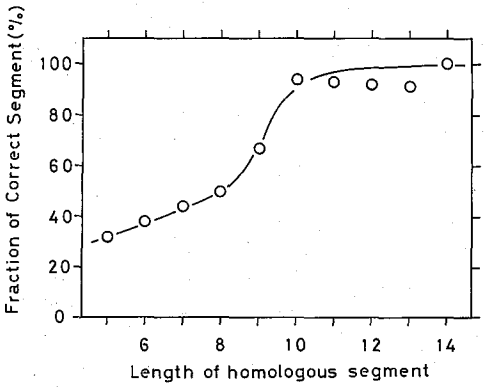

Fig. 1. Ratio of a number of correctly selected segments to a total number of homologous segments in Set II. Region of those homologous segments is defined by the successive C_{str} values of greater than 0.6 and r.m.s. deviations within 2 Å by the superposition method. From a pair of those proteins, the segments are selected by the condition that C_{seq} values greater than 0.5 succeed above the length of abscissa.

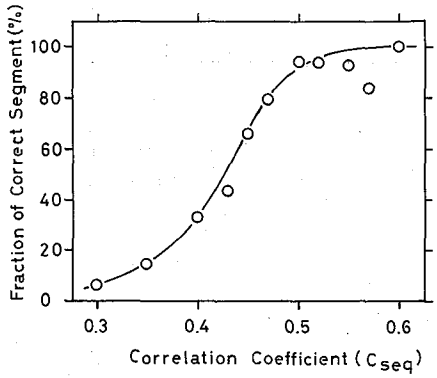

Fig. 2. Plots as in Fig. 1 but the segments are selected by the condition that C_{seq} values greater than those of abscissa succeed above ten residues.

segments which should be detected from successive $C_{str}(i, j)$ values of greater than 0.6 and r.m.s. deviations within 2 Å by the superposition method. A sharp increase with length and good correspondence with high probability above 10 residues long are observed. The number of correctly selected homologous segments decreased with the length of the homologous segment. This indicates that the probability of occurrence of homologous segments gradually decreases with increasing length, and means that the increase in restriction (i.e., the increase in the length of homologous segment) improves the fit to the homologous segments but cannot select many homologous segments. A similar estimation was made for segments with ten residues long having various cut off values of C_{seq} as shown in Fig. 2, in which good correspondences are derived when C_{seq} is above 0.5. The total number of homologous segments is a function of the cut off values, i.e., 27 to 9 for 0.3 to 0.6 of those values. This result indicates that when two segments of greater than ten residues long have successive correlation $C_{seq}(i, j)$ values of greater than 0.5, these are structurally homologous at least with a probability of 90%.

Since effective length and correlation value are determined, homologous segments for homologous proteins are selected by taking the segments of greater than ten residues long with successive C_{seq} above 0.5. This is done by using the five optimized parameters of this research shown in the middle part of Table IV and by using six parameters of the previous research[1] in the right part of Table IV. Only one fragment between elastase and proteinase b is selected by using the five parameters unfavorably. 13 fragments were selected simultaneously from two parameter sets favorably but neither of several fragments were selected unfavorably.

Table IV. Selection of homologous segments by structure correlation

Protein X	Protein X′	Fragment X	Fragment X′	N	C_{str}
parvalbumin	parvalbumin	42– 66	81– 105	25	0.81
hemoglobin	myoglobin	1– 42	1– 42	42	0.89
(α-cahin)		54– 93	60– 99	40	0.83
		95–134	101–140	40	0.85
hemoglobin	hemoglobin	1– 13	2– 14	13	0.88
(α-cahin)	(β-chain)	18– 43	17– 42	26	0.80
		55–125	60–130	71	0.99
		131–141	136–146	11	0.83
γ-crystallin	γ-crystallin	5– 16	44– 55	12	0.76
		29– 38	72– 81	10	0.70
		41– 51	89– 99	11	0.78
		1– 23	88–110	23	0.81
		28– 84	117–173	57	0.89
		107–124	149–166	18	0.75
		1– 17	1– 17	17	0.89
elastase	beta-trypsin	20– 29	17– 26	10	0.75
		29– 55	24– 50	27	0.87
		42– 62	36– 56	21	0.83
		45– 91	38– 84	47	0.88
		91–116	82–107	26	0.88
		132–158	122–148	27	0.81
		122–146	111–135	25	0.88
		178–193	167–182	16	0.84
		212–240	195–223	29	0.92
		113–123	88– 98	11	0.73
elastase	proteinase b	14– 24	166–176	11	0.74
		193–202	6– 15	10	0.79
					(−0.05
elastase	subtilisin	18– 30	206–218	13	0.75
lysozyme	lysozyme	4– 14	58– 68	11	0.75
(egg-white)	(T4)	79– 88	148–157	10	0.68
cytochrome c_2	cytochrome c	3– 12	2– 11	10	0.81
		13– 30	13– 30	18	0.89
		40– 52	40– 52	13	0.75
		62– 75	59– 72	14	0.75
		91–112	82–103	22	0.79
ferredoxin	ferredoxin	1– 13	88–100	13	0.89
azulin	plastocianin	1– 27	28– 54	27	0.85
		29– 38	79– 88	10	0.67
average					0.81

C_{seq} and C_{seq}' is calculated by using the parameters of this research (five parameters) and of the is a root mean square deviation between C^{α} atoms by using the method of superposition and distance, between hemoglobin and myoglobin were omitted from this table. This segment (*) is not selected

coefficient C_{str} and by sequence correlation coefficient C_{seq}

R. m. s. sup (Å)	R. m. s. dist (Å)	C_{seq}	Fragment X	Fragment X′	C_{seq}'	Fragment X	Fragment X′
1.41	0.96	0.69	53– 67	92–106	0.60	51– 63	90–102
1.39	1.08	0.66	12– 24	12– 24	0.62	13– 30	13– 30
1.25	0.88						
0.71	0.46						
1.26	0.81	0.63	2– 19	3– 20			
1.32	0.74				0.61	24– 38	23– 37
0.99	0.80	0.67	54– 65	59– 70	0.59	73– 83	78– 88
		0.76	85–102	90–107	0.71	87–102	92–107
1.96	1.28	0.61	113–127	118–132	0.63	110–141	115–146
0.36	0.29						
1.52	1.34						
0.26	0.20	0.63	44– 53	92–101	0.62	45– 54	93–102
0.77	0.50						
1.68	0.95	0.70	37– 70	126–159	0.70	44– 60	133–149
2.20	1.69						
0.84	0.54						
1.93	1.49						
2.34	1.92	0.81	32– 42	27– 37	0.75	27– 50	22– 45
2.79	2.30						
2.25	2.01	0.67	56– 65	49– 58	0.60	57– 68	50– 61
					0.70	70– 86	63– 79
0.62	0.46	0.59	91–106	82– 97	0.66	94–106	85– 97
1.80	1.41						
2.07	1.69	0.72	124–137	113–126	0.72	181–192	170–181
0.86	0.91	0.71	182–191	171–180	0.68	227–240	210–223
1.18	1.28	0.68	224–240	207–223			
1.95	1.64						
1.78	1.27						
1.14	0.77						
4.43	3.78)*	0.64	85– 97	76– 88			
1.90	1.74						
1.38	0.57						
1.78	1.45						
0.48	0.44				0.76	13– 39	13– 39
0.81	0.71	0.69	14– 37	14– 37			
1.06	0.51	0.58	39– 50	39– 50	0.64	61– 76	58– 73
0.63	0.60				0.72	103–112	93–102
1.63	1.13						
1.87	1.11				0.63	33– 44	6– 17
2.10	1.63						
1.69	1.46						
1.42	1.08						

previous research[1] (six parameters), respectively. N is the length of fragments. R. m. s. sup and dist respectively. Short fragments which overlap with this table between hemoglobin α and β-chain and from C_{str} but from C_{seq}.

In addition to homologous proteins we also tried an estimation of the efficiency of the parameters with all combinations for the 15 non-homologous proteins of Set III. Fig. 3 shows the average r.m.s. values of the spacial coincidence plotted against the length of segments. For the correlation values of 0.3 and 0.4, the extent of coincidence becomes worse, even though the length of segment is more than 12 residues long. However, at the values greater than 0.5, the extent drastically becomes better when the length is more than 12 residues, i.e., the average r.m.s. values are 3.5 Å, 2.4 Å and 0.7 Å at 11, 12 and 13 residue lengths of the segments, respectively. The dotted line indicates the average r.m.s. values for the segments randomly compared in the 15 non-homologous proteins. It should be noted that the number of detected homologous segments in non-homologous proteins diminishes significantly in comparison to that in homologous proteins.

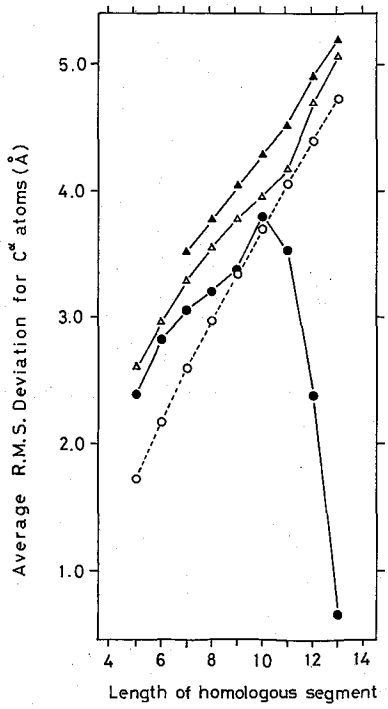

Fig. 3. Plot of r. m. s. deviation for C^{α} atoms by superposition method averaged over selected segments vs. lengths of segments which define selection condition for non-homologous proteins of Set III. The values of C_{seq} above which the segments are selected are 0.3 (▲), 0.4 (△) and 0.5 (●). The dotted line indicates random level of average r. m. s. deviation, that is, selection was made only by the length regardless of C_{seq} values for all pair of segments in Set III.

DISCUSSION

The minimization of the physical parameters by Eq. (2) gave us the optimum parameters reflecting the tertiary structure of proteins. Since the optimum parameters obtained depend on the reference proteins (or regions) taken in Eq. (2), it is important to assess the suitability of selection of these proteins from the practical point of view. Although the primary structure contains the entire information, the secondary structure gives a structural basis for protein folding. Therefore, we selected the reference proteins for the second term of Eq. (2) from both α- and β-proteins. In addition to

this, the reference proteins for the second term were required to have some complicated and specific native conformations (which are homologous to each other) as well as the adequate sequence homology. The proteins (or regions) used in this study as the reference are typical proteins which satisfy the above requisite, and this is the least requisite for the minimization of the physical parameters by which the optimum parameters are to be calculated. The major problem is selection of the regions which should be taken for the first term of Eq. (2). There are several regions (e.g., regions 1–20 and 53–72 of parvalbumin) other than the regions finally selected having different conformations which exhibit high correlation values whenever we use parameters such as hydrophobicity which are considered to characterize tertiary structure of proteins. Since it was difficult to automatically select the regions for the first term of Eq. (2), we selected the most effective regions in a trial and error manner. Thus the five sets of parameters obtained by Eq. (2) may be optimum ones to identify the structurally homologous segments.

It was interesting that the minimization of Eq. (2) has a tendency to converge to good parameters (i.e., high correlation for homologous regions but poor correlation for the other regions), when we used the physical parameters which reflect a hydrophobic (or polar) nature and a tendency to form β-structure as the initial values for minimization (see the parameters with asterisks in Table I). Furthermore, all of the parameters listed in Table I could not move from their own values, when we minimized the Eq. (2) without the first term. This means that the physical parameters themselves are optimum in this sense. Therefore, it was possible to find the best combination from the physical parameters to identify homologous sequences as described previously.[1,2] However, the set of parameters had to include one or two parameters which only served to diminish correlation values for regions having different conformations. This is the reason why the inclusion of hydrophobic parameters which characterize the tertiary structure of proteins is important but not enough to identify the homologous sequence, and we had to include the parameters pK-C and pK-N in addition to hydrophobic parameters as reported previously.[1,2]

The efficiency of the five set parameters obtained by the present procedure was quantitatively estimated on a wide variety of homologous proteins as well as non-homologous proteins (Fig. 1 to 3). Moreover, from the results of Table IV the segments simultaneously selected by two parameter sets are homologous with high probability. These results suggest that we may identify homologous segments having similar native conformations with a probability of 90% and 70% for homologous and non-homologous proteins, respectively, if all the successive residues in the segments of the length greater than ten residues long have sequence correlation values greater than 0.5. This leads us to approach the prediction of the three-dimensional structure of proteins from their primary structures as follows. A protein of unknown three-dimensional structure is compared to various proteins of known three-dimensional structure by the present procedure, and constructed by combining detected homologous segments according to their corresponding regions in the protein. The structure of the protein derived in such a way may be subjected to a procedure for energy minimization of the protein. Thus, this method of sequence correlation coefficient may

be a powerful way to predict the three-dimensional structure of proteins.

ACKNOWLEDGEMENT

We express our sincere thanks to Dr. John Shepherd for reading the manuscript and valuable comments.

REFERENCES

(1) Y. Kubota, K. Nishikawa, S. Takahashi and T. Ooi, *Biochim. Biophys. Acta*, **701**, 242 (1982).
(2) Y. Kubota, *Bull. Inst. Chem. Res., Kyoto Univ.* **60**, 309 (1982).
(3) A. A. Afifi and S. P. Azen, "Statistical Analysis, A Computer Oriented Approach," 2nd ed., Academic Press, New York, 1979, pp. 324–341.
(4) M. Levitt, *Biochemistry*, **17**, 4277 (1978).
(5) O. B. Ptitsyn and A. V. Finkelstein, *Quart. Rev. Biophys.*, **13**, 339 (1980).
(6) S. Lifson and C. Sander, *Nature*, **282**, 109 (1979).
(7) P. Y. Chou and G. D. Fasman, *Adv. Enzymol.*, **47**, 45 (1978).
(8) P. Manavalan and P. K. Ponnuswamy, *Nature*, **275**, 673 (1978).
(9) E. J. Cohn and J. T. Edsall, "Proteins, Amino Acids, and Peptides," Van Nostrand-Reinhold, Princeton, New Jersey, 1943.
(10) H. B. Bull and K. Breese, *Arch. Biochem. Biophys.*, **161**, 665 (1974).
(11) H. A. Sober Ed., "Handbook of Biochemistry, Selected Data for Molecular Biology," 2nd ed., The Chemical Rubber Co., Cleveland, Ohio, 1970.
(12) A. M. Lesk and C. Chothia, *J. Mol. Biol.*, **136**, 225 (1980).
(13) M. Oobatake and T. Ooi, *J. Theor. Biol.*, **67**, 567 (1977).
(14) M. O. Dayhoff Ed., "Atlas of Protein Sequence and Structure," Vol. 5, Suppl. 3, National Biomedical Research Foundation, Washington D. C. 1978.
(15) M. Levitt, *J. Mol. Biol.*, **104**, 59 (1976).
(16) J. M. Zimmerman, N. Eliezer and R. Simha, *J. Theor. Biol.*, **21**, 170 (1968).
(17) D. D. Jones, *J. Theor. Biol.*, **50**, 167 (1975).
(18) J. Janin, *Nature*, **277**, 491 (1979).
(19) W. R. Krigbaum and B. H. Rubin, *Biochim. Biophys. Acta*, **229**, 368 (1971).
(20) R. Grantham, *Science*, **185**, 862 (1974).
(21) K. Nishikawa and T. Ooi, *Int. J. Peptide Protein Res.*, **16**, 19 (1980).
(22) D. H. Wertz and H. A. Scheraga, *Macromolecules*, **11**, 9 (1978).
(23) F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, Jr., M. D. Rodgers, O. Kennard, T. Shimanouchi and M. Tasumi, *J. Mol. Biol.*, **112**, 535 (1977).
(24) S. C. Nyberg, *Acta Crystallogr., Ser. B*, **30**, 251 (1974).
(25) P. Manavalan and P. K. Ponnuswamy, *Arch. Biochem. Biophys.*, **184**, 476 (1977).