Vol. 41 No. SIG 7(TOM 3) IPSJ Transactions on Mathematical Modeling and Its Applications Nov. 2000

Regular Paper

GA Generates New Amino Acid Indices through Comparison between Native and Random Sequences

Satoru Kanai[†,☆] and Hiroyuki Toh[†]

The amino acid sequence of a protein carries its folding information. If the information is encoded by the arrangement of the amino acid residues along the primary structure, the random shuffling of the residues would degrade the information. We developed a new method to compare the native sequence with random sequences generated from the native sequence, in order to extract such information. First, amino acid indices were randomly generated. That is, the initial indices have no significance on the feature of residues. Next, using the indices, the averaged distance between a native sequence and the random sequences was calculated, based on the autoregressive (AR) analysis and the linear predictive coding (LPC) cepstrum analysis. The indices were subjected to the genetic algorithm (GA) using the distance as the fitness, so that the distance between the native sequence and the random sequences becomes larger. We found that the indices converged to hydrophobicity indices by the GA operation. The AR analysis with the converged indices revealed that the autocorrelation in the native sequence is related to the secondary structure.

1. Introduction

Proteins are essential molecules for living organisms, which are involved in a wide variety of biological phenomena. Living organisms can generate various proteins which are composed of 20 kinds of amino acid resides. Amino acid residues are linearly combined by peptide bonds to form a protein. So, proteins are string-like molecules. The amino acid sequences of most protein fold into a globular structure to exert its biological activity. A statement known as Anfinsen's dogma[1] maintains that the information about the folding of a globular protein is carried by the amino acid sequence. If we extracted such information from an amino acid sequence, we could predict the tertiary structure of the sequence. In addition, such information is important to design artificial proteins with desired structures. However, we do not fully understand the relationship between protein sequence and the structure. Techniques for structure prediction or the design of artificial proteins are still far from practical applications. However, this problem has long attracted many people, and various studies have been made thus far.

One of the approaches to tackle this problem is to find the orders or rules held by the amino acid sequences of native proteins. In the analyses, the amino acid sequences are transformed into a series of numerical data. A simple method is the binary transformation, where 0 is assigned to a group of amino acid residues, and 1 is assigned to the residues of another group. Then, an amino acid sequence is expressed as a series of the numbers 1 and 0. For example, hydrophobic residues are expressed by 0, and hydrophilic residues are replaced with 1. A more subtle method for the transformation is the application of amino acid indices. An amino acid index is a set of numerical values, each of which corresponds to an amino acid residue. There are many different amino acid indices, which are roughly classified into six types, hydrophobicity, α and turn propensities, β propensity, composition, physicochemical properties, and other properties[2]. Each residue of a given amino acid sequence is replaced with the corresponding numerical value of a given index. Then, the amino acid sequence of a protein is expressed as one-dimensional numerical value data, like time series data. In this paper, the series of numerical values corresponding to an amino acid sequence is called "profile." The profile has been analyzed by signal processing technique, such as Fourier transformation and autoregressive moving average models, in order to find periodicity or autocorrelation in a given amino acid sequence. There are many example of such approaches. Some people insist that residues are randomly arranged in the amino acid sequences

[†] Department of Bioinformatics, Biomolecular Engineering Research Institute

[☆] Presently with PharmaDesign, Inc.

of native proteins[3)~5)]. However, other people have found periodicity or autocorrelation in the amino acid sequences of native proteins[6)~15)]. In many cases, periodicity or autocorrelation in hydrophobicity are found in amino acid sequences. Thus, the results obtained from the various approaches are still controversial.

In this paper, we propose a new method for the signal processing analysis of amino acid sequences. Our method is regarded as solving an inverse problem against the current signal processing approaches. We did not use any of the known amino acid indices for the study. Instead of assumptions about the features of the amino acid residues, we introduced another assumption for the analysis: if the information about the folding of a protein is carried by the arrangement of the residues along the primary structure, then the information is degraded by the random shuffling of the sequence. Therefore, it is expected to extract information related to protein folding through comparison of sequences of native proteins and the randomly shuffled sequences. First, we connected the amino acid sequences of native proteins to generate a long sequence, which we call "native sequence." For comparison, each sequence constituting a native sequence was randomly shuffled, and then was connected in the same order as in the native sequence. The long sequence composed of the shuffled sequences is called "random sequence." A single random sequence may fortuitously posses folding information. However, it is expected that this information would be lost in most of the random sequences, if many random sequences are generated. Therefore, we used a set of random sequences instead of a single random sequence. First, we prepared a large number of amino acid indices, whose elements were randomly generated. Therefore, the initial amino acid indices did not have any meaning in terms of on the feature of amino acid residues. Using each amino acid index, the native sequence and the random sequences were transformed into profile data. The former is called "native profile," while the latter is called "random profiles." Both profiles were subjected to a univariate autoregressive (AR) analysis[16)]. Then, the distance between the native and random profiles was calculated based on the results of the AR analysis. The distance is known as the linear predictive coding (LPC) cepstrum distance[17)] in the field of speech recognition. Using the distance as the fitness of the index, the population of amino acid indices was subjected to a genetic algorithm (GA)[18)] as follows. The more an amino acid index can distinguish the native sequence from the random sequence, the more descendants the index can reproduce. The amino acid index with the highest fitness in the final generation, therefore, is expected to distinguish the native sequence from the random sequences efficiently. Then, the autocorrelation in the primary structure can be examined by the AR analysis with the obtained indices.

We will show how the GA worked on the evolution of amino acid indices. The relationships between the obtained amino acid indices and the known amino acid indices will be discussed. We will also discuss the relationship between the AR coefficients and the secondary structures of the examined proteins.

2. Materials and Methods

2.1 Preparation of Native Sequences and Generation of Randomly Shuffled Sequences

The proteins used in this study were selected based on the structural classification by CATH[19)]. 20 proteins were selected from the mainly α class. 21 proteins were selected from the mainly β class. 39 proteins were taken from the α-β class. All of the proteins were selected so as to satisfy the following conditions: (a) the sequence length is equal to or greater than 100 amino acid residues, (b) each protein is made of a single domain, and (c) no hetero atoms or ligands are contained in the structure. CATH has one more structural class, few secondary structures. However, none of the proteins included in that class fulfilled the conditions described above. Therefore, no proteins belonging to the last class were used in this study. The proteins used in this study are listed in **Table 1**. The amino acid sequences of the selected proteins were taken from the homepage of the Protein Data Bank (http://www.pdb.bnl.gov/pdb/index.html).

As described below, the amino acid sequences of the selected proteins were subjected to an AR analysis[16)] in this study. However, the sequence length of a single protein was too short to obtain enough samples for the AR analysis. Therefore, all of the amino acid sequences belonging to the same structural class were connected to form a native sequence defined above. Then, three long sequences were constructed, corresponding to the mainly α, the mainly β,

Table 1 The list of proteins used in this study. The column, "ID," indicates the PDB ID code. The column, "CH," indicates the chain when the PDB entry includes several chains. The column, "LN," indicates the length of the amino acid sequence.

(1) Mainly α class

ID	CH	LN	ID	CH	LN
1bip		122	1jvr		137
1hul	A	108	1gdz		151
1rfb	A	119	1lxl		221
2end		138	1col	A	204
153l		185	1pbw	A	216
1lbd		282	1csm	A	256
1sig		339	1cem		363
2spc	A	107	1nfn		191
1aep		161	1lis		136
1rcb		129	1jli		112

(2) Mainly β class

ID	CH	LN	ID	CH	LN
1bw3		125	2eng		210
1clh		166	1whi		122
1ulo		152	1akp		114
1exg		110	1msp	A	126
1stm	A	157	1nfa		178
1knb		196	1thv		207
1cau	B	184	1tnf	A	157
1gff	1	426	1dup	A	152
1tul		108	1gpr		162
1tie		172	1vmo	A	163
1tsp		559			

(3) α-β class

ID	CH	LN	ID	CH	LN
1fus		106	1oun	A	127
1nar		290	1onr	A	316
1fwp		139	1ris		101
2pii		112	2chs	A	127
1dco	A	104	1iba		101
1tbd		134	1msc		129
1vhi	A	142	1kpt	A	105
1mil		104	1svq		114
1chd		203	1ice	A	167
1cex		214	1tib		269
1pdo		135	1tht	A	305
1cfy	A	143	1cby		259
1pvu	A	154	1fvk	A	189
1bhm	A	213	1smn	A	245
1eri	A	276	1rva	A	244
1esc		306	3pte		349
1jon		155	1htm	B	138
1opc		110	1mut		129
1reg	X	122	1lts	A	185
1pbn		289			

and the α-β classes. The lengths of the three sequences were 3677, 3946, and 7050 amino acid residues, respectively. In addition, the three sequences were connected to form one more amino acid sequence, which was referred to as "all data." The four connected sequences were used as native sequences.

Corresponding to each native sequence, a number of shuffled sequences were generated. The amino acid sequence of each constituent protein of the native sequence is randomly shuffled. The shuffled sequences are then connected in the same order as the constituent proteins in the native sequence. The sequence composed of shuffled sequences is used as the random sequence. The random sequence has the same length and the same amino acid composition as the native sequence, but the arrangement of the amino acid residues is randomized. Due to the reason described above, 100 random sequences were generated for each native sequence.

2.2 Genetic Algorithm to Generate Amino Acid Indices

The overall procedure to generate the amino acid indices is committed to the GA[18]. The GA is a heuristic method to solve the combinatorial optimization problems by mimicking the evolution mechanisms of living organisms. In the GA, any possible solution is expressed as a chromosome. Then, a population of chromosomes is subjected to evolutionary operations, such as selection, crossover, and mutation. For the procedure of selection, the fitness of each individual chromosome should be defined, corresponding to the problem under consideration. During the process, the population evolves toward an optimal solution. Note that there is no proof that a solution obtained by the GA is optimal, because the GA is a heuristic method. The following is a standard GA procedure.

P1 Initialization An initial population is randomly generated.

P2 Reproduction and selection The fitnesses of the individuals constituting the population are evaluated. Then, the descendants of an individual are generated, in proportion to the fitness. That is, the greater the fitness is, the larger the number of descendants is. Note that the size of the population is fixed during the whole process. Therefore, an individual with relatively low fitness is expected to become extinct during the procedure.

P3 Crossover Crossover operations between randomly selected pairs of chromosomes are performed.

P4 Mutation Mutation operations against randomly selected chromosomes are performed.

P5 Judgment of termination If a condition of termination is fulfilled, an individual with the highest fitness among the current population is regarded as a quasi-optimal solution. Otherwise, go to **P2** and repeat the procedure until the termination condition is fulfilled.

To solve the problem considered in this manuscript, the standard algorithm was encoded into a program that can perform each operation as follows:

Chromosome representation and initialization

A chromosome indicates an amino acid index. That is, a chromosome is a set of 20 numerical values, each of which corresponds to an unknown feature of an amino acid residue. Note that we did not adopt a binary expression of a chromosome, but used a gray-scale expression. That is, each element of an amino acid index is restricted in a range from 0.0 to 1.0. 500 amino acid indices were created as an initial population, and the size of the population was fixed to 500 during the procedure. Each element of the amino acid index in the initial population was filled with a random number.

Distance between native and random profiles

In this study, we searched for the order that is present in a native sequence, but is degraded in any random sequence generated from the native one. Therefore, a chromosome, which efficiently distinguishes a native sequence from the random sequences, is regarded as effective for selection. To realize this idea, the fitness of a chromosome is defined based on the distance between the native sequence and a random sequence. That is, the greater the distance is, the higher the fitness is. Before entering into the details of the GA operation, we will describe the definition of the distance between the native sequence and a random sequence with an amino acid index.

First, an amino acid sequence, **A**, is converted to a profile, using a given amino acid index, **I**. Let a sequence be $\mathbf{A} = \{A_1, \ldots, A_i, \ldots, A_L\}$, where $A_i \in \{A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V\}$. L indicates the length of the native sequence. Let an amino acid index be $\mathbf{I} = \{I_A, \ldots, I_j, \ldots, I_V\}$, where $0.0 \leq I_j \leq 1.0$, and $j \in \{A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V\}$. The amino acid residue at each i-th position, A_i, is replaced with a numerical value, P_i, where P_i equals I_j if A_i equals j. Then, a profile, $\mathbf{P} = \{P_1, \ldots, P_i, \ldots, P_L\}$, is obtained. The profile obtained from a native sequence is referred to as native profile, while the profile for a random sequence is called random profile. A profile can be treated as time series data if the residue position is regarded as discrete time.

For the time series analysis, the properties of a time series are reduced to be expressed as the AR model. Based on the obtained AR models, the difference between two time series data sets is calculated as the LPC cepstrum distance. This method is a traditional approach in the field of speech recognition. We defined the distance between a native sequence and the random sequence, using the LPC cepstrum distance.

At first, a native profile, **P**, is analyzed as the univariate AR model. A corrected profile, $\mathbf{T} = \{t_1, \ldots, t_i, \ldots, t_L\}$, is calculated from **P**, where $t_i = P_i - M$ and M is the average of **P**. The equation of the AR model of **T** is as follows:

$$t_i = \sum_{m=1}^{k} (b_m \cdot t_{i-m}) + \varepsilon_i \tag{1}$$

where k is the AR order, and ε_i is the white noise at site i. Let $\mathrm{b} = \{b_1, \ldots, b_i, \ldots, b_k\}$ be a set of AR coefficients of the model. The AR coefficients are estimated by minimizing the sum of the squared residual, Q_k.

$$Q_k = \sum_{i=k+1}^{L} \left(t_i - \sum_{m=1}^{k} (b_m \cdot t_{i-m}) \right)^2 \tag{2}$$

Then, the AR coefficients are obtained as the solution of Eq. (3), where S_{rs} and T_r are defined as Eqs. (4) and (5).

$$\begin{bmatrix} S_{11} & S_{12} & \ldots & S_{1k} \\ S_{12} & S_{22} & \ldots & S_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ S_{1k} & S_{2k} & \ldots & S_{kk} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} = \begin{bmatrix} T_1 \\ T_2 \\ \vdots \\ T_k \end{bmatrix}. \tag{3}$$

$$S_{rs} = \sum_{i=k+1}^{L} (t_{i-r} \cdot t_{i-s}), \quad 1 \leq r \leq s \leq k. \tag{4}$$

$$T_r = \sum_{i=k+1}^{L} (t_i \cdot t_{i-r}), \quad 0 \leq r \leq k. \tag{5}$$

The AR order examined in this study ranged from 1 to 8.

Let's consider that two profiles are analyzed by the AR model. Then, two sets of AR coefficients are obtained. In the field of speech recognition, it is known that the Euclidean distance with the sets of AR coefficients is not efficient for practical application, because the AR coefficients do not always reflect the spectral envelope of a given pattern [17]. Instead, the LPC cepstrum distance is widely utilized as the distance between two sets of time series data. Therefore, we adopted the LPC cepstrum distance as the distance between the native profile and a random profile.

The coefficients of the m-th order LPC cepstrum, $\mathbf{c} = \{c_1, \ldots, c_i, \ldots, c_m\}$, for a given profile are calculated based on the AR coefficients:

$$\begin{aligned} c_0 &= \log Q_k \\ c_1 &= b_1 \\ c_i &= b_i + \frac{1}{i} \cdot \sum_{n=1}^{i-1} (n \cdot c_n \cdot b_{i-n}) \end{aligned} \quad (6)$$

where $m \geq k$ and $b_i = 0$ $(i > k)$. The sum of the squared residual, Q_k, is obtained from the AR coefficients by the following equation.

$$Q_k = T_0 - \sum_{m=1}^{k} (b_m \cdot T_m). \quad (7)$$

In this study, the order of the LPC cepstrum, m, was 15.

We can calculate the distance between the profiles, based on the two sets of LPC coefficients corresponding to the profiles. Let $\mathbf{c} = \{c_0, \ldots, c_i, \ldots, c_m\}$ be the set of LPC coefficients for a profile, $\mathbf{P}$. Likewise, let $\mathbf{c}' = \{c'_0, \ldots, c'_i, \ldots, c'_m\}$ be the set of LPC coefficients of another profile, $\mathbf{P}'$. The LPC cepstrum distance between the two profiles, $D(\mathbf{P}, \mathbf{P}')$, is defined as follows:

$$\begin{aligned} D(\mathbf{P}, \mathbf{P}')^2 &= (c_0 - c'_0)^2 \\ &\quad + 2 \cdot \sum_{j=1}^{m} (c_j - c'_j)^2. \end{aligned} \quad (8)$$

Reproduction

The population that is subjected to the GA operation is composed by a set of chromosomes. As described above, each chromosome is an amino acid index. Here, we will define the fitness of a chromosome and describe the operation of reproduction.

Let's consider one chromosome, designated as x, within a population. The raw fitness of chromosome x is obtained as follows. (I) A profile of a native sequence, $\mathbf{P}(x)$, is generated using chromosome x. Then, a set of random profiles, $\{\mathbf{P}'_1(x), \ldots, \mathbf{P}'_i(x), \ldots, \mathbf{P}'_S(x)\}$, is generated by applying chromosome x to the set of random sequences, where S is the number of random sequences generated from a native sequence. (II) The LPC cepstrum distance between the native profile and a random profile i, $D(\mathbf{P}(x), \mathbf{P}'_i(x))$, is calculated. (III) The raw fitness of the chromosome x, $RF(x)$, is then calculated as follows.

$$RF(x) = \frac{100}{S} \cdot \sum_{j=1}^{S} D(\mathbf{P}(x), \mathbf{P}'_i(x)). \quad (9)$$

The procedure is applied to all of the chromosomes included in the current population. Then, the raw fitnesses are assigned to all of the chromosomes within the population. It is often observed that selection pressure is not reflecting the difference in fitness properly and the efficiency of searching for an optimal point becomes worse, as the number of generation in the GA becomes larger. To improve the problem, the raw fitness is re-scaled for selection. There are several methods to re-scale the fitness. Here, we adopted one of the methods which is called sigma truncation [20),21)]. The scaled fitness $SF(x)$ of an amino acid index x by sigma truncation is defined as follows:

$$SF(x) = RF(x) - (RF_{\text{avg}} - C \cdot \sigma) \quad (10)$$

where RF_{avg} and σ are the average and the standard deviation over all of $RF(x)$, respectively. C is a scaling factor, which is obtained as follows:

$$C = (RF_{\text{avg}} - RF_{\text{min}})/\sigma \quad (11)$$

where RF_{min} is the minimum value over all of $RF(x)$.

If the obtained value of C is less than 1.0, then C is set to be 1.0. Likewise, C is set to be 3.0, if the obtained value of C is greater than 3.0. If $SF(x)$ is less than 0.0, then $SF(x)$ is set to be 0.0.

The offspring are generated in proportion to the scaled fitness. The number of offspring of chromosome x is calculated as follows:

$$o_x = \text{int}\left[N \cdot \frac{SF(x) \cdot N_x}{\sum_{i=1}^{N_T} (SF(i) \cdot N_i)}\right] \quad (12)$$

where int$[r]$ is the function that returns an integer that is greater than or equal to r, but less than $r + 1$. N_T is the number of types of chromosomes, and N_i is the number of chromosome i in the current population. Moreover, N is the population size. That is,

$$N = \sum_{i=1}^{N_T} N_i \quad (13)$$

As described above, N was set to 500 in the actual run. However, the total sum of o_x is not always N, due to the dismissal of the decimal fraction by the application of the "int" function in Eq. (12). In order to adjust the population size of the offspring, the number of each chromosome is increased by 1 until the population size of the offspring becomes N. The operation is carried out according the order of the fitness. That is, the operation starts from the chromosome with the highest fitness, and progresses toward the chromosome with the lowest fitness. If the total sum of the offspring becomes N midway through the operation, then the operation is terminated, and the numbers of the offspring of the remaining chromosomes are left as they are. If the total sum is less than N after the operation is applied to all of the chromosomes, then the same procedure is repeated until the condition is satisfied.

Elitism strategy

Some individual chromosomes with high fitness are inherited by the next generation without the operations of crossover and mutation. This strategy is called elitism[22]. In actual runs, the top 1% of the reproduced offspring are regarded as elite when the individuals of the offspring are sorted by scaled fitness. On the other hand, the remaining non-elites are subjected to the following two operations.

Crossover

The uniform crossover operation[23] is adopted for this study. The algorithm allows any number of crossover points at any position. First, a pair of non-elite chromosomes is randomly selected. Then, the two chromosomes are aligned so that the elements for the same amino acid residue in the chromosomes correspond to each other. Each position between two neighboring elements is a crossover point candidate. At each position, a random number ranging from 0.0 to 1.0 is generated. If the random number is greater than a given crossover probability, P_C, a crossover occurs at the position. After the operation, the pair is removed from the non-elite population, and is introduced into a set that will be subjected to the mutation operation. The operation is repeated until the size of the non-elite population becomes 0. When the size of the non-elite population is odd, the chromosome with the highest fitness in the non-elite population is introduced into the set for the mutation operation before the crossover operation described above. P_C is 0.1 in actual runs.

Mutation

In a standard genetic algorithm, each chromosome is expressed by binary codes, and a mutation means exchanging 1 to 0 or *vice versa*. In contrast, the chromosomes in this study were expressed by a set of 20 numerical values ranging from 0.0 to 1.0. In this study, therefore, a mutation means an increase or a decrease in the value of an element of the index by a given constant.

All of the chromosomes included in the set made by the crossover operation are subjected to the mutation operation. Let's consider chromosome x as an example. Chromosome x is an amino acid index with 20 elements. Corresponding to each element, a random number ranging from 0.0 to 1.0 is generated. If the random number is less than a given mutation probability, P_M, the element is not mutated. Otherwise, a random integer is generated to determine whether the change is an increment or a decrement. If the number is odd, then the element is increased. Otherwise, the element is decreased. In this study, P_M is 0.1, and 0.05 is used as the constant value for an increment or a decrement. When the numerical value stored in the element becomes greater than 1.0 by the increment operation, the value is re-set to be 1.0. Likewise, the value is re-set to be 0.0 when the value becomes less than 0.0 by the decrement operation. All of the elements in an index are examined in the same way, and the procedure is applied to all of the chromosomes in the set made by the crossover operation.

Judgment of termination

The chromosomes subjected to crossover and mutation operations were combined with the elite individuals to form the next generation. When the GA operation is repeated by a given number, the program is terminated. Of course, when the number is too small, we cannot obtain a good solution. The number is, therefore, determined by several preliminary trials of the GA operations. In many cases, 50 generations were sufficient for the highest fitness in the population to converge to a constant value. To ensure the convergence, we added 50 more generations. That is, each GA operation was terminated after 100 generations.

2.3 Evaluation of Generated Amino Acid Index

To compare the generated amino acid indices with each other or with known indices, the correlation coefficient between two amino acid indices was calculated.

The cluster analysis was performed using the absolute values of the correlation coefficients as the distance between two indices. In the analysis, the distance $DI(\mathbf{I},\mathbf{I}')$ between two indices $\mathbf{I}$ and $\mathbf{I}'$ is given by:

$$DI(\mathbf{I},\mathbf{I}') = 1 - |CC(\mathbf{I},\mathbf{I}')| \qquad (14)$$

where $CC(\mathbf{I},\mathbf{I}')$ is the correlation coefficient between two amino acid indices. Then, a dendrogram was constructed by the NEIGHBOR program in the PHYLIP 3.5c package[24], by adopting the unweighted pair-group method with the arithmetic mean (UPGMA)[25]. The UPGMA dendrogram was drawn by the TreeView program[26].

3. Results

3.1 Evolution of Amino Acid Indices by the GA Operation

We examined four combined sequences, the mainly α, the mainly β, the α-β structural classes, and all data in this study. For each sequence, the evolution of the amino acid indices by the GA was performed, while changing the AR orders from 1 to 8. Under a given AR order, five GA runs were carried out to check the effect of the initial conditions against the evolution of the amino acid indices. Therefore, the initial seed of the random number was changed for each run.

In the case of the mainly α class, the fitness of the population of amino acid indices converged rapidly despite the difference in initial values, when AR order was set to a value from 2 to 8. The convergence was attaine before the 30th generation in every AR order except for AR order = 1. When AR order = 1, the fitness did not converge during 100 generations in any run with different initial values. Due to the restriction in pages, only the evolutionary process of the amino acid indices with AR order = 4 is shown in **Fig. 1** (1). When the AR orders $\geq$ 2, the converged fitnesses of the five runs for each AR order were similar to each other, despite the differences in the initial conditions. In addition, the obtained amino acid indices from the five runs for each AR order were also similar to each other for the AR orders $\geq$ 2. The absolute values of the correlation coefficients for every pair of the five amino acid indices for each AR order were greater than or equal to 0.990. The GA operation used in this study cannot distinguish the fitness of the amino acid indices that are symmetric to a value 0.5. Therefore, some of the amino acid indices from the five runs showed negative, but high, correlations to the remaining indices (data not shown). We selected one of the five amino acid indices to represent the AR order, so that the distance of the native sequence to the random sequences used for the five runs was the largest among the five amino acid indices. As described above, the fitness did not converge for the AR order = 1. However, a representative amino acid index for the AR order = 1 was formally obtained

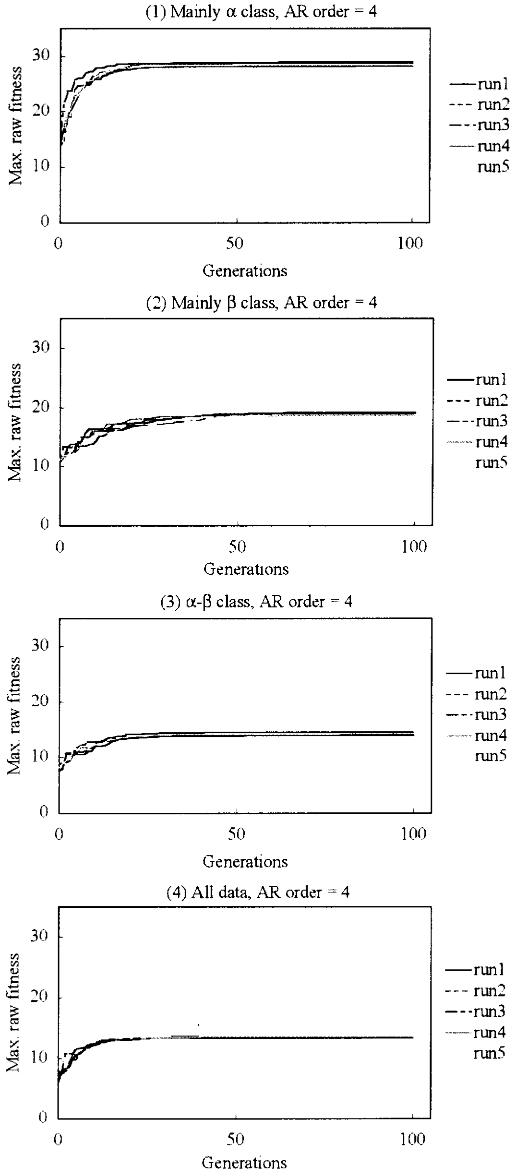

Fig. 1 The plots of the highest raw fitness among the population of amino acid indices at a generation against the generation number.

by the same procedure described above. **Table 2** (1) shows the representative amino acid indices for the AR orders from 1 to 8. Despite the difference in the AR order, the representative amino acid indices for the AR orders from 2 to 5 show high positive correlations to each other (correlation coefficients $\geq$ 0.985). Likewise, the amino acid indices for the AR orders from 6 to 8 showed high positive correlations (correlation coefficients $\geq$ 0.994). However, the amino acid indices for the AR orders from 2 to 5 were negatively correlated with the indices for the AR orders from 6 to 8, and the absolute values of the correlation coefficients were greater than or equal to 0.981. Thus, these two types of amino acid indices were approximately symmetric to an amino acid index value of 0.5. This observation suggests that the amino acid indices obtained from the GA operations for the mainly α class are similar to each other when the AR orders are greater than or equal to 2. In contrast, the amino acid index for the AR order = 1 was different from the other amino acid indices. The absolute values of the correlation coefficients between the representative amino acid indices for the AR orders $\geq$ 2 and that for the AR order = 1 were less than or equal to 0.276.

Like the case of the mainly α class, the amino acid indices for the mainly β class converged in any run with different initial values when the AR orders $\geq$ 3. However, the speed of convergence seemed to be a little slower than that for the mainly α class. Figure 1 (2) shows the evolutionary process of amino acid indices with AR order = 4. The absolute values of the correlation coefficients for every pair of five amino acid indices for each AR order were greater than or equal to 0.982. That is, the amino acid indices, as well as the converged fitnesses, were similar to each other in every run for each AR order, although symmetric indices were often generated in the five runs. In each case of the AR orders = 1 and 2, however, an amino acid index obtained from one run was not similar to those from the other four runs. A representative index of the five indices for each AR order was obtained in the same manner as for the case of the mainly α class, which is listed in Table 2 (2). Except for the case of the AR orders = 1 and 2, the amino acid indices were highly correlated with each other, despite the difference in the AR orders, and the absolute values of the correlation coefficients were greater than or equal to 0.924. The representative amino acid indices for the AR orders = 3, 4, 7, and 8 and those for the AR orders = 5 and 6 were approximately symmetric around an index value of 0.5. Moreover, the absolute values of the correlation coefficients between the representative amino acid index for the AR order = 1 and the indices for the AR orders $\geq$ 3 ranged from 0.740 to 0.832. That is, the representative amino acid index for the AR order = 1 and those for the AR orders $\geq$ 3 were similar to each other. In contrast, the representative amino acid index for the AR order = 2 was different from those for the other AR orders. The absolute values of the correlation coefficients between the representative amino acid index for the AR order = 2 and those for the other AR orders were less than or equal to 0.228.

Like the above two cases, the fitness converged quite rapidly for the α-β class. Figure 1 (3) shows the evolutionary process of amino acid indices with AR order = 4. The converged fitnesses of the five runs for each AR order were similar to each other. Moreover, all of the amino acid indices from the five runs for each AR order were highly correlated. The absolute values of the correlation coefficients of the indices from the five runs for each AR order were greater than or equal to 0.963. Using the method described above, the representative amino acid index was obtained for each AR order (see Table 2 (3)). The absolute values for the correlation coefficients among the representative indices for the AR orders $\geq$ 2 were greater than or equal to 0.918. Therefore, the indices showed high correlation. However, the representative amino acid index for the AR order = 1 was not similar to those for the higher AR orders, and the absolute values of the correlation coefficients were less than or equal to 0.244.

The process of the convergence for all data was basically the same as those for the α-β class. The evolutionary behavior of the amino acid indices with AR order = 4 is shown in Fig. 1 (4). Similar to the case of the mainly α class, the difference in the initial conditions was independent of the convergence in the case of AR orders $\geq$ 2. The absolute values of the correlation coefficients for every pair of amino acid indices from the five runs for each AR order were greater than or equal to 0.990. In the case of the AR order = 1, however, the amino acid index from one run was different from the other

Table 2 The list of representative amino acid indices obtained in this study. The column "AR" indicates the AR order for the examined model. "RF" indicates the averaged raw fitness between a native sequence and five random sequences, using an amino acid index corresponding to the row. The following 20 columns indicated by the one–letter amino acid code show the elements of the converged amino acid indices.

(1) Mainly α class

AR	RF	I	L	M	V	F	Y	W	A	G	P
1	16.731	0.20	0.25	0.15	0.20	0.45	0.30	0.55	0.15	0.25	0.10
2	26.370	1.00	0.90	0.75	1.00	0.90	0.80	1.00	0.50	0.40	0.30
3	26.503	1.00	0.95	0.80	1.00	0.95	0.80	1.00	0.50	0.40	0.35
4	28.341	1.00	0.95	0.75	0.95	1.00	0.80	0.90	0.50	0.40	0.25
5	28.455	1.00	1.00	0.75	0.95	1.00	0.80	0.90	0.45	0.35	0.30
6	28.507	0.00	0.05	0.25	0.05	0.00	0.15	0.10	0.50	0.65	0.70
7	29.092	0.00	0.00	0.20	0.00	0.00	0.15	0.05	0.55	0.60	0.75
8	29.167	0.00	0.00	0.20	0.00	0.00	0.20	0.05	0.55	0.65	0.70
AR		S	T	D	E	N	Q	H	R	K	C
1		0.45	0.10	0.15	0.10	0.35	0.25	1.00	0.15	0.25	0.30
2		0.35	0.30	0.25	0.15	0.20	0.25	0.55	0.15	0.00	0.75
3		0.35	0.30	0.25	0.15	0.20	0.25	0.55	0.15	0.00	0.70
4		0.25	0.40	0.20	0.05	0.15	0.20	0.55	0.15	0.00	0.75
5		0.25	0.35	0.15	0.00	0.15	0.25	0.55	0.10	0.00	0.80
6		0.70	0.65	0.80	0.95	0.85	0.75	0.45	0.85	1.00	0.15
7		0.75	0.65	0.80	1.00	0.80	0.75	0.35	0.85	1.00	0.10
8		0.75	0.70	0.80	0.95	0.80	0.80	0.40	0.90	1.00	0.15

(2) Mainly β class

AR	RF	I	L	M	V	F	Y	W	A	G	P
1	13.271	0.00	0.00	0.80	0.45	0.15	0.65	0.05	0.55	0.40	0.60
2	13.693	0.00	0.45	0.30	0.65	0.90	0.30	0.75	0.50	0.75	0.20
3	16.480	0.75	0.95	0.20	0.55	1.00	0.50	0.90	0.50	0.50	0.35
4	18.899	0.85	1.00	0.05	0.50	1.00	0.50	0.80	0.40	0.35	0.30
5	19.279	0.10	0.00	0.80	0.45	0.00	0.50	0.20	0.55	0.60	0.70
6	21.775	0.15	0.15	0.75	0.40	0.00	0.50	0.25	0.55	0.65	0.65
7	23.319	0.90	0.85	0.15	0.55	1.00	0.50	0.85	0.45	0.30	0.30
8	23.604	0.90	0.90	0.20	0.55	1.00	0.55	0.85	0.45	0.30	0.30
AR		S	T	D	E	N	Q	H	R	K	C
1		0.40	0.80	0.75	0.75	0.75	0.85	0.70	0.95	0.90	0.10
2		0.90	0.65	0.40	0.50	1.00	0.60	0.65	0.00	0.20	0.70
3		0.35	0.20	0.10	0.30	0.00	0.10	0.00	0.25	0.35	0.45
4		0.35	0.30	0.00	0.20	0.00	0.15	0.10	0.25	0.30	0.35
5		0.65	0.65	0.95	0.80	0.95	0.85	0.90	0.70	0.75	0.65
6		0.65	0.60	1.00	0.80	0.85	0.80	0.75	0.65	0.65	0.65
7		0.30	0.40	0.00	0.15	0.05	0.25	0.15	0.35	0.30	0.25
8		0.30	0.40	0.00	0.15	0.10	0.20	0.15	0.35	0.30	0.35

(3) α-β class

AR	RF	I	L	M	V	F	Y	W	A	G	P
1	10.214	0.80	1.00	0.00	0.80	0.00	0.45	0.75	0.30	0.90	0.30
2	13.580	0.90	1.00	0.95	0.55	0.25	0.80	0.95	0.30	0.20	0.40
3	13.771	0.90	1.00	0.90	0.55	0.25	0.75	1.00	0.30	0.10	0.50
4	14.151	0.80	0.90	0.90	0.50	0.25	0.65	1.00	0.30	0.10	0.45
5	16.838	0.95	0.95	0.80	0.75	0.45	0.75	1.00	0.40	0.05	0.40
6	20.019	1.00	0.90	0.95	0.80	0.60	0.75	0.95	0.45	0.10	0.40
7	20.176	0.95	0.90	0.80	0.75	0.55	0.65	1.00	0.45	0.10	0.35
8	20.488	1.00	0.90	0.85	0.80	0.60	0.70	1.00	0.50	0.10	0.40
AR		S	T	D	E	N	Q	H	R	K	C
1		0.25	0.20	0.75	0.70	0.55	0.50	0.95	0.05	0.00	0.55
2		0.15	0.45	0.10	0.05	0.15	0.00	0.25	0.05	0.00	0.70
3		0.20	0.45	0.05	0.05	0.15	0.00	0.25	0.05	0.00	0.75
4		0.15	0.50	0.10	0.10	0.15	0.00	0.20	0.10	0.00	0.65
5		0.15	0.40	0.05	0.00	0.10	0.00	0.25	0.25	0.15	0.55
6		0.10	0.25	0.00	0.10	0.10	0.10	0.30	0.25	0.10	0.65
7		0.05	0.25	0.00	0.05	0.10	0.10	0.30	0.20	0.05	0.60
8		0.10	0.35	0.00	0.15	0.10	0.10	0.30	0.25	0.15	0.65

(4) All data

AR	RF	I	L	M	V	F	Y	W	A	G	P
1	8.979	0.20	0.10	0.05	0.10	0.30	0.35	0.40	0.15	0.35	0.20
2	13.305	0.95	1.00	0.85	0.70	0.60	0.85	1.00	0.45	0.45	0.40
3	13.428	0.05	0.00	0.20	0.25	0.40	0.25	0.05	0.55	0.55	0.65
4	13.521	0.05	0.00	0.20	0.30	0.40	0.20	0.05	0.55	0.60	0.60
5	14.995	0.95	1.00	0.85	0.75	0.65	0.80	0.95	0.40	0.25	0.30
6	16.840	1.00	1.00	0.90	0.80	0.75	0.80	0.90	0.45	0.25	0.35
7	16.882	1.00	1.00	0.90	0.80	0.75	0.80	1.00	0.45	0.25	0.40
8	17.110	1.00	0.95	0.90	0.80	0.75	0.75	0.95	0.45	0.25	0.35
AR		S	T	D	E	N	Q	H	R	K	C
1		0.45	0.25	0.05	0.10	0.25	0.25	1.00	0.15	0.20	0.40
2		0.30	0.40	0.25	0.20	0.15	0.20	0.40	0.10	0.00	0.75
3		0.70	0.60	0.70	0.80	0.85	0.80	0.55	0.90	1.00	0.30
4		0.70	0.60	0.70	0.80	0.85	0.80	0.55	0.90	1.00	0.20
5		0.25	0.30	0.20	0.05	0.05	0.10	0.35	0.05	0.00	0.60
6		0.20	0.30	0.10	0.10	0.05	0.10	0.35	0.10	0.00	0.70
7		0.20	0.30	0.10	0.10	0.05	0.10	0.35	0.15	0.00	0.60
8		0.20	0.35	0.10	0.15	0.05	0.10	0.35	0.10	0.00	0.65

runs, although the fitness of the index was less than those of the remaining four indices. The representative amino acid indices for the AR orders from 1 to 8 were obtained in the same manner as described above, and are also listed in Table 2 (4). The absolute values of the correlation coefficients between the amino acid indices for the AR orders ≥ 2 were greater than or equal to 0.965. Thus, the representative amino acid indices for the AR orders ≥ 2 were highly correlated to each other, despite the difference in the AR order, although the indices for the AR orders $= 2$ and ≥ 5, and those for the AR orders $= 3$ and 4 were approximately symmetric around an index value of 0.5. In contrast, the representative amino acid index for the AR order $= 1$ was different from those for the higher AR orders, and the absolute values of the correlation coefficients between the index for the AR order $= 1$ and those for higher AR orders were less than or equal to 0.049.

3.2 Relationship between the Representative Amino Acid Indices and the Known Amino Acid Indices

The correlation coefficient between every pair of the representative amino acid indices described above was calculated. In addition, the correlation coefficients were calculated between every pair of the indices and 402 known amino acid indices available in an amino acid index database, AAindex1 [2),27)]. All of the representative indices, except for the four indices, showed high correlations not only to each other, but also to the known indices, which are classified into a group of hydrophobicity indices. The four exceptional indices included the one for the AR order $= 1$ from the mainly α class, the one for the AR order $= 2$ from the mainly β class, the one for the AR order $= 1$ from the α-β class, and the one for the AR order $= 1$ from all data. They did not show prominent similarity to any of the known amino acid indices. Likewise, they were not similar to each other (absolute values of correlation coefficients ≤ 0.439), except for the similarity between the index from the mainly α class and all data (correlation coefficient $= 0.906$).

The relationship among the representative indices and the known 149 hydrophobicity indices was examined by a cluster analysis. The four exceptional indices were not included in the study. The result is summarized in a dendrogram shown in **Fig. 2**. As described above, the representative indices on a protein structural class were similar to each other, despite the difference in the AR orders. Reflecting the similarity, such amino acid indices formed a cluster corresponding to each protein structural class in the dendrogram.

The cluster on the mainly α class was distinct from the other clusters, and showed high correlation with four known indices classified into a group of hydrophobicity indices, including an index for the information value for accessibility (average fraction 35%)[28)], an index about the mean polarity [29)], an index about the mean fractional area loss [30)], and an index about the information value for accessibility (average fraction 23%)[28)]. The ID codes of these indices in the AAindex1 are BIOV880101, RADA880108, ROSG850102, and BIOV880102, respectively. The averaged absolute value of the correlation coefficient between the indices derived from the mainly α class and the four hydrophobicity indices described above was 0.952, as indicated at the node X1 in the dendrogram.

The cluster of the mainly β class was also distinct from the other clusters. In the cluster, the index for the AR order $= 1$ was a little distant from the other representative indices from the mainly β class. The 19 known indices belonging to the group of hydrophobicity indices were close to the indices from the mainly β class in the dendrogram, which were different from the indices closely related to the cluster of the mainly α class. In this case, the 19 known indices included an index about the optimal matching hydrophobicity [31)], an index about the transfer energy between an organic solvent and water [32)], and an index about the hydrophobicity [33)]. The ID codes of these indices in AAindex1 are SWER830101, NOZY710101, and ARGP820101, respectively. The 19 indices were relatively distant from the cluster of the mainly β class. However, the averaged absolute value of the correlation coefficients between the indices derived from the mainly β class and the 19 indices described above was 0.700, as indicated at the node Y in the dendrogram.

The cluster of the α-β class occupied a position between those of the mainly α and mainly β classes. This cluster was distinct from the other two clusters, although it was relatively close to the cluster of the mainly α class, rather than that of the mainly β class. The cluster of the α-β class includes that of all data. As shown in the dendrogram, the cluster of all data is closely related to the indices for the AR or-

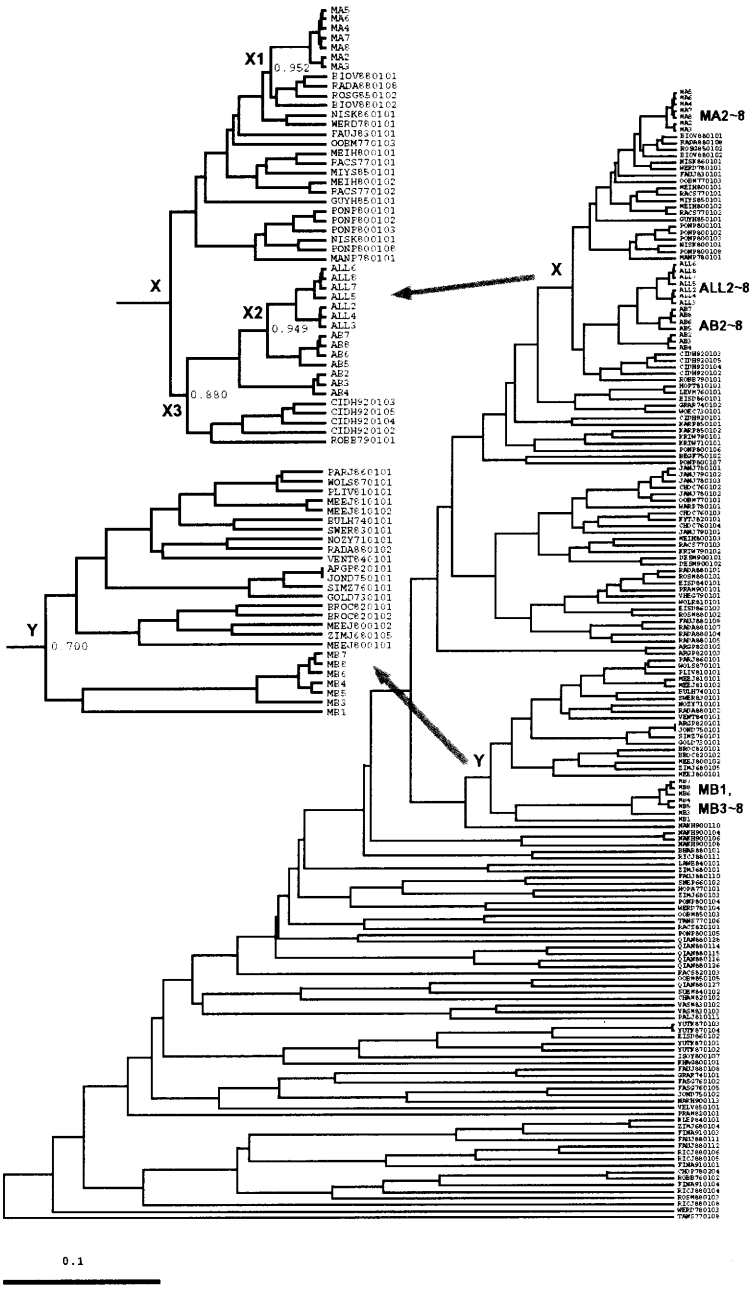

Fig. 2 A dendrogram of the converged indices and the known hydrophobicity indices. The name of the index corresponding to each leaf of the dendrogram is shown at the right side of the dendrogram. The known hydrophobicity indices are indicated by the ID code in the amino acid index database, AAindex1. The structural class and the AR order are shown as a combination of symbols corresponding to the structural class and integers corresponding to the AR order. MA, MB, AB, and ALL are the symbols indicating the mainly α, the mainly β, the α-β classes, and all data, respectively.

ders from 5 to 7 of the α-β class. The averaged absolute value of the correlation coefficient was 0.949, as indicated at the node X2 in the dendrogram. The representative indices included in the clusters of the α-β class and all data showed high correlation with the four amino acid indices about the normalized hydrophobicity scales[34] and an index about the information measured for an α-helix[35]. The ID codes in AAindex1 of the first four indices are CIDH920102, CIDH920103, CIDH920104, and CIDH920105, and the code for the last index is ROBB790101. The averaged absolute value of the correlation coefficient between the indices from the α-β class and all data and the five known indices was 0.880, as indicated at the node X3 in the dendrogram.

In this study, the GA operation was designed to generate amino acid indices that can efficiently distinguish a native sequence from the corresponding random sequences. However, there is no proof that the indices are optimal for distinguishing between them, because GA is just a heuristic approach. Although we can not prove the optimality, we can approximately examine the efficiency for the distinction of a native sequence from the corresponding random sequences by the obtained indices. The efficiency of an amino acid index was evaluated as follows. First, 100 random sequences were generated from a given native sequence. Here, four native sequences used for the GA operation, the mainly α, mainly β, α-β classes, and all data were examined again. Then, the raw fitness of the native sequence against the 100 random sequences was calculated with the given amino acid index by the same manner as described in the section of GA operation. The fitness is here referred to as native raw fitness (NRF). Next, the raw fitness of a random sequence against the remaining 99 random sequences was calculated with the same index by the same manner as described above. The fitness is here referred to as random raw fitness (RRF). The calculation was applied to every random sequence. Then, 100 RRFs were obtained, and the mean (m) and the standard deviation (σ) of the RRFs were calculated. With m and σ, the NRF was normalized to z-score (Z).

$$Z = (\mathrm{NRF} - m)/\sigma \qquad (15)$$

The larger the z-score is, the higher the efficiency of the distinction is considered to be. The calculation was applied not only to the indices obtained by the GA operation, but also to all of the amino acid indices available in AAindex1. The calculation was performed under the different AR orders ranging from 1 to 8. The results are summarized in **Fig. 3**. The

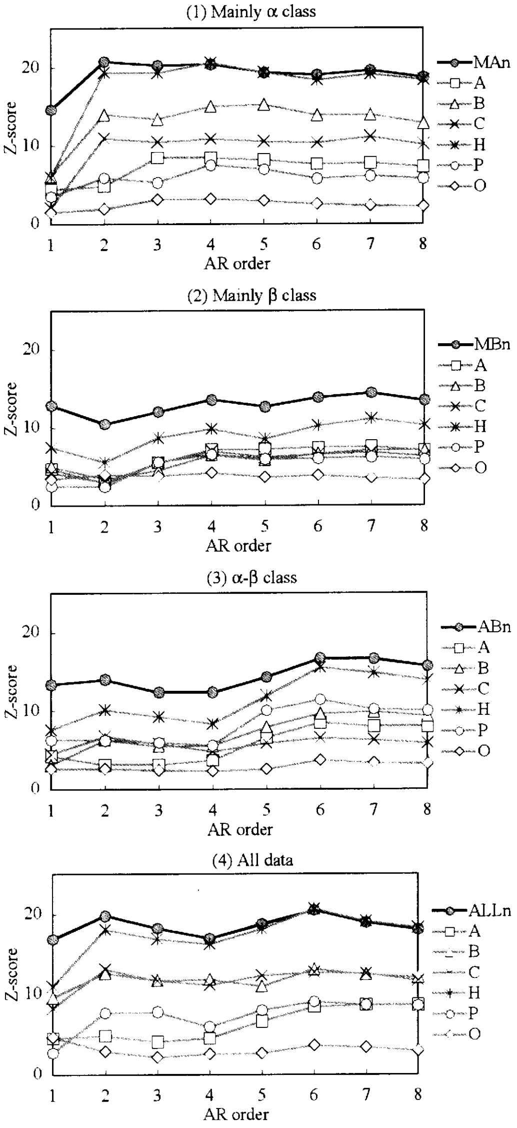

Fig. 3 The plots of the z-scores of the obtained and known indices against the AR orders. MA, MB, AB, and ALL are the symbols indicating the mainly α, the mainly β, the α-β classes, and all data, respectively. The value "n" denotes AR order, which is ranged from 1 to 8. "A", "B", "C", "H", "P", and "O" denote six groups of known indices, α and turn propensities, β propensity, composition, hydrophobicity, physicochemical properties, and other properties, respectively. See details in the text.

ordinate indicates the z-score, and the abscissa indicates the AR order. Closed circles indicate the plot of the z-score of the index obtained by GA with the native sequence corresponding to each graph as the function of the AR order. As shown in the figure, for any native sequence and for any AR order, the z-scores were greater than 10.0 SD. The results suggest high efficiency for the distinction between native and random sequences by the amino acid indices obtained by GA. The results of the calculation with the indices available in AAindex1 were also shown in the figure. To simplify the presentation, the indices were classified into six groups, α and turn propensities, β propensity, composition, hydrophobicity, physicochemical properties, and other properties, according to the study by Tomii and Kanehisa[2]. Then, for each group, only the highest z-score among those of the members in the group at an AR order was plotted against the AR order. As expected from the high correlation between our indices and the known hydrophobicity indices, the differences, the differences in z-score between our indices and the hydrophobicity indices for the mainly α class and all data were quite small, when the AR order ≥ 2. In the case of the α-β class, the differences between the indices obtained in this study and the hydrophobicity indices were also small as the AR order ≥ 6. In the case of the mainly β class, the z-scores of the indices obtained by GA were greater than those of known indices at any AR order. For any native sequence and for any AR order, the highest z-scores for the groups other than the hydrophobicity group were less than those of the indices generated by GA and the hydrophobicity indices, although many of them were greater than 3.0 SD.

4. Discussion

The question addressed in this study was what is the difference between native amino acid sequences and randomly shuffled ones. We considered that such a difference is related to the folding information within the native sequences. As described above, our GA operation, which was designed to maximize the distance between native sequences and shuffled sequences, generated amino acid indices related to the hydrophobicity. The results suggested that the amino acid residues are arranged in a hydrophobic order, and that the order is degraded by the shuffling of the amino acid sequences. As described in the introduction, there are many examples of the data analyses of the primary structures of proteins. Some investigators consider that the design of the native amino acid sequences is random, while others find some periodicity or autocorrelation in the primary structures. Our results strongly support the latter view. That is, the sequences of the native proteins are not random, and the hydrophobicity of amino acid residues is one of the factors related to the design of native proteins. Our hydrophobicity indices were generated in a data-driven manner by the GA. In other words, autocorrelation in hydrophobicity is an intrinsic feature of native amino acid sequences, which was isolated from the sequences by the GA operation.

When the AR order was low, 1 or 2, the calculations sometimes did not converge in a given generation number. For example, the GA operation for the mainly α class did not converge when the AR order = 1. Even if the calculations converged, the obtained indices for the low AR orders were different from each other depending on the difference in the initial populations. For example, in the case of the AR order = 2 for the mainly β class, two different indices were yielded after the GA operations with different initial populations were converged. Similar behavior was observed in the case of the AR order = 1 for all data. Moreover, the obtained indices were not similar to either those for the higher AR orders or the known amino acid indices. On the other hand, the GA calculation for the α-β class with AR order = 1 converged to yield an almost identical index in spite of the difference in initial population. However, the index was different from either those for the higher AR orders or the known amino acid indices. Due to the failure in convergence or the low correlation to all of the known indices, we can't further discuss the indices corresponding to the low AR orders. Such unstable behavior in convergence may suggest that an AR order = 1 or 2 is too low to express the autocorrelation in the native amino acid sequence. Or, the indices may actually characterize the short-range interaction in native amino acid sequences by unknown mechanisms represented by the indices. In this situation, however, further discussion would just make a lot of speculative statements. Therefore, we will exclude the four cases, the AR order = 1 for the mainly α class, the AR order = 2 for the mainly β class, the AR order = 1 for the α-β class, and the AR order = 1 for all data, from the next discussion about the AR analysis.

Our study suggests that hydrophobicity is involved in the design of primary structures of native proteins, and the design in hydrophobicity is related to the secondary structures of the proteins. Then, the next question is how the amino acid sequences are designed in hydrophobicty. In this study, the amino acid sequences were expressed by the AR models (see Eq. (1)), each of which is characterized by the corresponding AR coefficients. The number of AR coefficients for an AR model is equal to the AR order of the model. Let's consider that b_n is an AR coefficient of the n-th order of an AR model with the AR order = m, where $n \leq m$. The absolute value of b_n indicates the strength of the correlation between a pair of residues n sites apart in the primary structure. The sign of b_n indicates the mode of the correlation. If b_n is positive, the residue pairs n sites apart tend to share a similar feature, while the residue pairs tend to have opposite features if b_n is negative. The AR coefficients are, thus, regarded as a measure for autocorrelation in the corresponding primary structures. Therefore, we expressed the native amino acid sequences as AR models using the converged amino acid indices, and examined the relationship between the AR coefficients and the structural class of the sequences.

Figure 4 (1) shows the AR coefficients of the mainly α class. As shown in the figure, the relationships between the AR coefficients and the orders are similar to each other, although the AR orders of the examined models were different from each other. The AR coefficients for the AR order = 1 were not included in the figure, because the fitnesses for the case did not converge within the 100 generations examined in this study. The AR coefficients, b_1 and b_2, had negative values, which means that the residue pairs one or two sites apart tend to have opposite features in hydrophobicity. The AR coefficients, b_3, were approximately zero, which means the residue pair three sites apart does not have any correlation. The AR coefficients, b_4, had positive values. That is, the residue pairs four sites apart share similar hydrophobic features. All of the other AR coefficients, except for those of the seventh order, b_7, were approximately zero. The AR coefficients, b_7, had positive values, although they were relatively small. The pattern of the AR coefficients is consistent with the periodicity of an a helical structure, which is 3.6 residues.

Figure 4 (2) shows the AR coefficients of the mainly β class. Like the cases of the mainly α class, the relationships between the AR coefficients and the orders for the models with different AR orders are similar to each other. The coefficients for the AR order = 2 were not included in the figure, due to the reason described above. As shown in the figure, b_1 had negative values. The AR coefficients, b_2 and b_8, were approximately zero. All of the remaining AR coefficients had negative values, which slowly decreased according to the increase of the order. In contrast, the AR coefficients, b_3, b_5, b_6, and b_8, were approximately zero in the cases of the mainly α class. Therefore, the slow decrease in the AR coefficients is considered to characterize the mainly β class. The negative values of b_1 and b_1 seemed to correspond to the alternative feature for the direction of side-chains on β strands. However, there was a further twist in the obtained results. The AR coefficients for the AR orders = 1, and from 3 to 8, b_2, were approximately zero. This means that the residue pair two sites apart does not have any correlation, although the side-chains of the two residues are expected to extend toward the same direction. The weak coupling of the residues two sites apart may be the reason why the converged indices for the AR order = 2 were not similar to the other indices. In contrast, the AR coefficients, b_4 and b_6, had negative values, although the side chains of the residue pairs four or six sites apart on a β-strand are expected to face the same direction. We cannot explain these results at this stage. It is considered that a long range interaction is involved in the formation of β-strand. However, AR model cannot deal with the long range interaction greater than AR order. This may be the reason that b_4 and b_6 are negative.

Figures 4 (3) and (4) show a similar pattern of AR coefficients. The former was obtained from the analyses of the α-β class, while the latter was from the studies of all data. Like the above two cases, the difference of the AR orders did not affect the relationship between the AR coefficients and the orders. The relationship seemed to be a mixture of that for the mainly α class and that for the mainly β class. The observation is consistent with the result that the amino acid indices for the α-β class and those for all

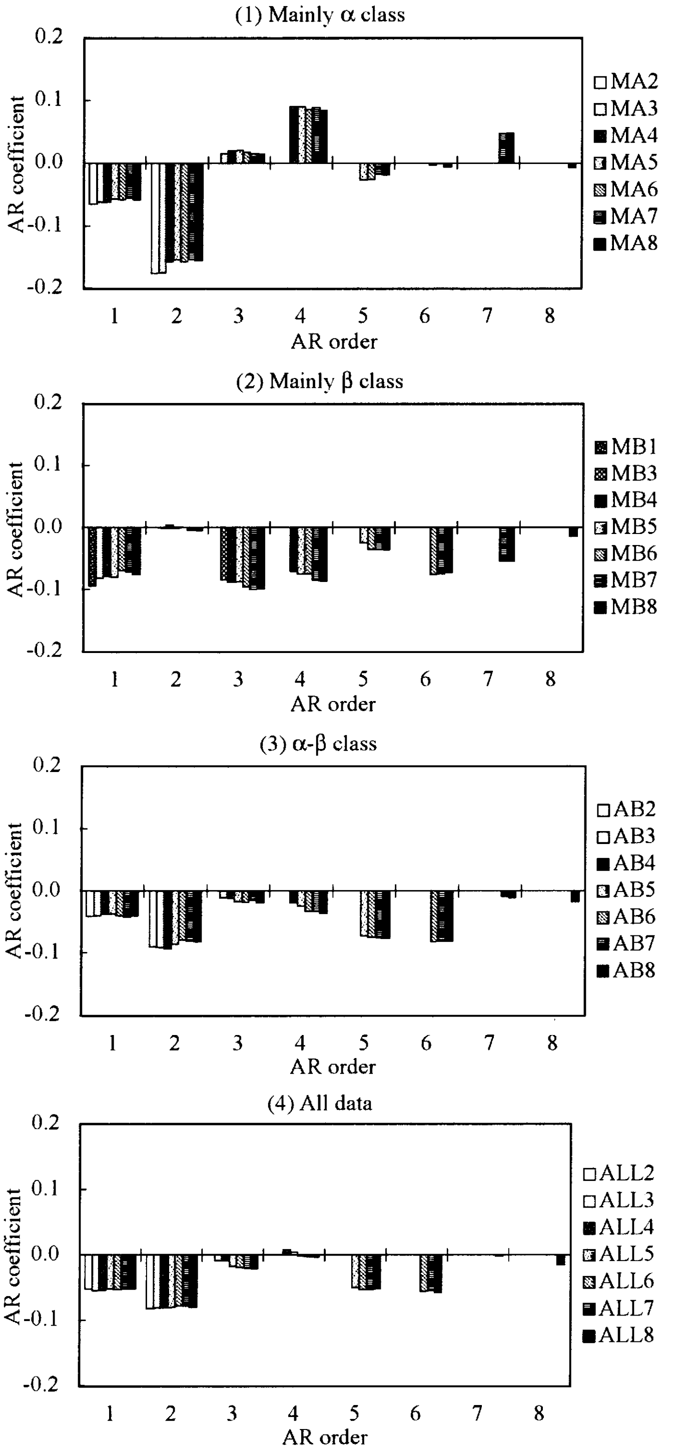

Fig. 4 The relationships between the AR coefficients and the corresponding order of the AR model.

data were similar to each other, and occupied a position between the indices for the mainly α class and those for the mainly β class in the dendrogram.

Our studies suggest that the amino acid sequences of native proteins intrinsicaly contain autocorrelation in hydrophobicity, and that the order in the primary structures is highly related to the secondary structures of the proteins. In this study, we used the LPC cepstrum distance to calculate the fitness. However, the Euclidean distance with a set of the AR coefficients was also efficient to converge the GA operation (data not shown). In that case, the converged indices are similar to those obtained with the LPC cepstrum distance. In this study, a random sequence was constructed by connecting randomly shuffled sequences. However, similar results were obtained when a random sequence with the same amino acid composition as the corresponding native sequence was constructed and shuffled over the entire sequence (data not shown). Here, we adopted a univariate AR analysis, which suggests that hydrophobicity is the main factor constituting the order of the amino acid sequences of native proteins.

Finally, we show a possible application of the obtained indices. The difference among obtained indices corresponding to the three structural classes, mainly α class, mainly β class, and α-β class, suggests a possibility to predict the structural class of a given amino acid sequence by the AR analysis with the indices. However, the sequence length of a protein is too short to be subjected to the AR analysis (see Section 2.1). A possible trick to solve the problem may be to connect the given amino acid sequence with the amino acid sequences of its homologues that can be collected by database searching. Here, we assume that homologous proteins take a similar fold, according to an empirical law of molecular evolution. The long sequence obtained by the connection is regarded as a native sequence, and a large number of random sequences corresponding to the native sequence are generated. Then, the raw fitnesses or z-scores of the native sequence corresponding to the three indices are calculated by the procedure described above. The given amino acid sequence is judged to belong to a structural class, when an index corresponding to the structural class shows the highest fitness or z-score among the three indices. Comparison of a native sequence with the corresponding random sequence is a novel approach to extract the structural information carried by the amino acid sequence. Further development of this approach would provide us a great insight into the relationship between sequences and structures of proteins.

Acknowledgments We thank Dr. Kentaro Tomii, Biomolecular Engineering Research Institute, for useful suggestions about the AAindex1.

References

1) Anfinsen, C.B.: Principles that govern the folding of protein chains, *Science*, Vol.181, pp.223–230 (1973).

2) Tomii, K. and Kanehisa, M.: Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins, *Protein Eng.*, Vol.9, pp.27–36 (1996).

3) White, S.H. and Jacobs, R.E.: Statistical distribution of hydrophobic residues along the length of protein chains: Implications for protein folding and evolution, *Biophys. J.*, Vol.57, pp.911–921 (1990).

4) White, S.H. and Jacobs, R.E.: The evolution of proteins from random amino acid sequences, I: Evidence from the lengthwise distribution of amino acids in modern protein sequences, *J. Mol. Evol.*, Vol.36, pp.79–95 (1993).

5) Rahman, R.S. and Rackovsky, S.: Protein sequence randomness and sequence/structure correlations, *Biophys. J.*, Vol.68, pp.1531–1539 (1995).

6) Black, J.A., Harkins, R.N. and Stenzel, P.: Non-random relationships among amino acids in protein sequences, *Int. J. Pept. Protein. Res.*, Vol.8, pp.125–130 (1976).

7) Eisenberg, D., Weiss, R.M. and Terwilliger, T.C.: The hydrophobic moment detects periodicity in protein hydrophobicity, *Proc.Natl Acad. Sci., USA*, Vol.81, pp.140–144 (1984).

8) Macchiato, M.F., Cuomo, V. and Tramontano, A.: Determination of the autocorrelation orders of proteins, *Eur. J. Biochem.*, Vol.149, pp.375–379 (1985).

9) Liquori, A.M., Sadun, C. and Battisti, A.: Quasi-periodic primary structures of core proteins of human T-lymphotropic leukemia retroviruses, *J. Mol. Evol.*, Vol.26, pp.269–273 (1987).

10) Pande, V.S., Grosberg, A.Y. and Tanaka, T.: Nonrandomness in protein sequences: evidence for a physically driven stage of evolution?, *Proc. Natl Acad. Sci., USA*, Vol.91, pp.12972–12975 (1994).

11) Sun, S. and Parthasarathy, R.: Protein sequence and structure relationship ARMA spectral analysis: application to membrane proteins, *Biophys. J.*, Vol.66, pp.2092–2106 (1994).

12) Irbäck, A., Peterson, C. and Potthast, F.: Evidence for nonrandom hydrophobicity structures in protein chains, *Proc. Natl Acad. Sci., USA*, Vol.93, pp.9533–9538 (1996).

13) Makeev, V.J. and Tumanyan, V.G.: Search of periodicities in primary structure of biopolymers: A general Fourier approach, *Comput. Appl. Biosci.*, Vol.12, pp.49–54 (1996).

14) Rackovsky, S.: "Hidden" sequence periodicities and protein architecture, *Proc. Natl Acad. Sci., USA*, Vol.95, pp.8580–8584 (1998).

15) Weiss, O. and Herzel, H.: Correlations in protein sequences and property codes, *J. Theor. Biol.*, Vol.190, pp.341–353 (1998).

16) Wei, W.W.S.: *Time Series Analysis: Univariate and Multivariate Methods*, Addison-Wesley (1990).

17) Rabiner, L. and Juang, B.H.: *Fundamentals of Speech Recognition*, Prentice Hall (1993).

18) Holland, J.H.: *Adaptation in Natural and Artificial Systems*, University of Michigan Press (1975).

19) Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B. and Thornton, J.M.: CATH – A hierarchic classification of protein domain structures, *Structure*, Vol.5, pp.1093–1108 (1997).

20) Forrest, S.: Documentation for PRISONERS DILEMMA and NORMS programs that use the genetic algorithm, University of Michigan, Ann Arbor (1985).

21) Goldberg, D.E.: *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley (1989).

22) DeJong, K.A.: An analysis of the behavior of a class of genetic adaptive systems, PhD Thesis, University of Michigan (1975).

23) Syswerda, G.: Uniform crossover in genetic algorithms, *Proc. 3rd Int. Joint Conf. on Genetic Algorithms* (ICGA89) (1989).

24) Felsenstein, J.: PHYLIP (Phylogeny Inference Package) version 3.5c, Department of Genetics, University of Washington, Seattle (1993).

25) Sokal, R.R. and Michener, C.D.: A statistical method for evaluating systematic relationship, *Univ. Kansas Sci., Bull.*, Vol.28, pp.1409–1438 (1958).

26) Page, R.D.: TreeView: An application to display phylogenetic trees on personal computers, *Comput. Appl. Biosci.*, Vol.12, pp.357–358 (1996).

27) Nakai, K., Kidera, A. and Kanehisa, M.: Cluster analysis of amino acid indices for prediction of protein structure and function, *Protein Eng.*, Vol.2, pp.93–100 (1988).

28) Biou, V., Gibrat, J.F., Levin, J.M., Robson, B. and Garnier, J.: Secondary structure prediction: combination of three different methods, *Protein Eng.*, Vol.2, pp.185–191 (1988).

29) Radzicka, A., Pedersen, L. and Wolfenden, R.: Influences of solvent water on protein folding: free energies of solvation of cis and trans peptides are nearly identical, *Biochemistry*, Vol.27, pp.4538–4541 (1988).

30) Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H. and Zehfus, M.H.: Hydrophobicity of amino acid residues in globular proteins, *Science*, Vol.229, pp.834–838 (1985).

31) Sweet, R.M. and Eisenberg, D.: Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure, *J. Mol. Biol.*, Vol.171, pp.479–488 (1983).

32) Nozaki, Y. and Tanford, C.: The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. Establishment of a hydrophobicity scale, *J. Biol. Chem.*, Vol.246, pp.2211–2217 (1971).

33) Argos, P., Rao, J.K. and Hargrave, P.A.: Structural prediction of membrane–bound proteins, *Eur. J. Biochem.*, Vol.128, pp.565–575 (1982).

34) Cid, H., Bunster, M., Canales, M. and Gazitua, F.: Hydrophobicity and structural classes in proteins, *Protein Eng.*, Vol.5, pp.373–375 (1992).

35) Robson, B. and Osguthorpe, D.J.: Refined models for computer simulation of protein folding. Applications to the study of conserved secondary structure and flexible hinge points during the folding of pancreatic trypsin inhibitor, *J Mol. Biol.*, Vol.132, pp.19–51 (1979).

(Received August 24, 1999)
(Revised October 2, 1999)
(Accepted December 20, 1999)

Satoru Kanai was born in 1958. He received his M.E. degree from Waseda Univ. in 1984. He had worked in Fujitsu Keiyo Systems Engineering Ltd. (FKY) since 1984. He had been sent on loan to Marine Biotechnology Institute since 1990. He came back to FKY in 1994. He had been sent on loan to Biomolecular Engineering Research Institute since 1996. He had been sent on loan to Fujitsu Ltd. since 1999. He has been working in PharmaDesign, Inc. since 2000. He has been engaging in the research on bioinformatics and molecular phylogenetics.

Hiroyuki Toh was born in 1961. He received his M.S. and Ph.D. degrees from Kyushu Univ. in 1985 and 1989 respectively. He had worked in Protein Engineering Research Institute since 1989. He had worked in Kyushu Institute of Technology as an associate professor since 1993. He has been in Biomolecular Engineering Research Institute since 1996, and now is a research director of the Department of Bioinformatics. He has been engaging in the research on computational molecular biology and molecular evolutionary genetics.