Amino Acids (2008) 35:541–549
DOI 10.1007/s00726-008-0082-1

ORIGINAL ARTICLE

Analysis of 3D structural differences in the IgG-binding domains based on the interresidue average-distance statistics

Takeshi Kikuchi

Received: 14 February 2008 / Accepted: 28 March 2008 / Published online: 23 April 2008
© Springer-Verlag 2008

Abstract It is well-known that the IgG-binding domain from staphylococcal protein A folds into a 3α helix bundle structure, while the IgG-binding domain of streptococcal protein G forms an ($\alpha + \beta$) structure. Recently, He et al. (Biochemistry 44:14055–14061, 2005) made mutants of these proteins from the wild types of protein A and protein G strains. These mutants are referred to as protein A219 and protein G311, and it was showed that these two mutants have different 3D structures, i.e., the 3α helix bundle structure and the ($\alpha + \beta$) structure, respectively, despite the high sequence identity (59%). The purpose of our study was to clarify how such 3D structural differences are coded in the sequences with high homology. To address this problem, we introduce a predicted contact map constructed based on the interresidue average-distance statistics for prediction of folding properties of a protein. We refer to this map as an average distance map (ADM). Furthermore, the statistics of interresidue distances can be converted to an effective interresidue potential. We calculated the contact frequency of each residue of a protein in random conformations with this effective interresidue potential, and then we obtained values similar to ϕ values. We refer to this contact frequency of each residue as a $p(\mu)$ value. The comparison of the $p(\mu)$ values to the ϕ values for a protein suggests that $p(\mu)$ values reveal the information on the folding initiation site. Using these techniques, we try to extract the information on the difference in the 3D structures of protein A219 and protein G311 coded in their amino acid sequences in the present work. The results show that the ADM analyses and the $p(\mu)$ value analyses predict the information of folding initiation sites, which can be used to detect the 3D difference in both proteins.

Keywords Average distance map · IgG-binding domain · Protein structure · Folding initiation site

Introduction

Many proteins are the targets of drug design against pandemic diseases (see, e.g., Chou et al. 2006; Du et al. 2007; Gao et al. 2007; Li et al. 2007; Oxenoid and Chou 2005; Schnell and Chou 2008; Wang et al. 2007a, b; Wei et al. 2006a, b; Wu and Yan 2008; Ye et al. 2007; Zhang et al. 2006), and hence it is vitally important to acquire their three-dimensional (3D) structural information via various computational approaches (see, e.g., Chou 2004; Lubec et al. 2005). The information on the 3D structure and function of a protein must be coded on its amino acid sequence, and how we decode a protein sequence is the main problem for predicting 3D structural information in bioinformatics and molecular biophysics. We usually use sequence alignment of sequentially homologous proteins to solve such problems (see, e.g., Chou 2004). Except for some special cases (e.g., Chou et al. 1999), usually for 3D-structure prediction of a protein with sequence homology, more than 30% sequence homology is required, and we can make a relatively accurate model of the 3D structure of a protein if we can use a template structure with more than 50% sequence homology.

Recently, however, a pair of proteins sharing 59% sequence homology but possessing different 3D structures

T. Kikuchi (✉)
Department of Bioscience and Bioinformatics,
College of Information Science and Engineering,
Ritsumeikan University, 1-1-1 Nojihigashi,
Kusatsu, Shiga 525-8577, Japan
e-mail: tkikuchi@is.ritsumei.ac.jp

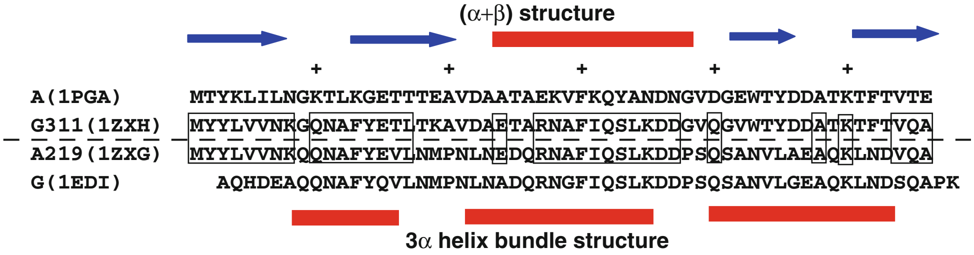
(α+β) structure
+ + + + +
A(1PGA) MTYKLILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDGEWTYDDATKTFTVTE
G311(1ZXH) MYYLVVNKGQNAFYETLTKAVDAETARNAFIQSLKDDGVQGVWTYDDATKTFTVQA
A219(1ZXG) MYYLVVNKQQNAFYEVLNMPNLNEDQRNAFIQSLKDDPSQSANVLAEAQKLNDVQA
G(1EDI) AQHDEAQQNAFYQVLNMPNLNADQRNGFIQSLKDDPSQSANVLGEAQKLNDSQAPK
3α helix bundle structure

Fig. 1 Multiple sequence alignment for protein A, protein A219, protein G and protein G311. The PDB code of each protein is indicated in *parenthesis*. The location of a β strand is indicated by a *blue arrow*, and that of an α helix is shown by a *red rectangle*

was reported (He et al. 2005; Alexander et al. 2005). It has been observed that the IgG-binding domain from staphylococcal protein A folds into the 3α helix bundle structure, and the IgG-binding domain of streptococcal protein G forms the ($\alpha + \beta$) structure. He et al. (2005) made mutants from the wild types of protein A and G strains with the phage display technique. These mutants are referred to as protein A219 and protein G311, respectively, and the authors showed that these two mutants have different 3D structures, i.e., the 3α helix bundle structure and the ($\alpha + \beta$) structure, respectively, despite the high sequence identity (59%). It should be noted that there is just 11% of the sequence identity between the sequences of protein A and protein G. The difference in the 3D structure formation of these proteins has been investigated recently with molecular dynamics technique starting with the native structures (Scott and Daggett 2007).

The purpose of our paper is to clarify how the 3D structural difference is coded in the sequences with high homology. However, in the present case, the sequence homology is too high, and it is difficult to solve the present problem with the usual sequence alignment or other standard bioinformatical techniques.

For our purpose, we introduce a predicted contact map constructed based on interresidue average-distance statistics for prediction of folding properties of a protein. We refer to this map as an average distance map (ADM) (Kikuchi et al. 1988; Kikuchi 2002). ADM was originally developed to predict compact areas such as domains in the sequence of a protein. We can show that the information of folding processes of a protein is reflected in its ADM, i.e., the locations of subdomains predicted by its ADM are related to early folding segments (Ichimaru and Kikuchi 2003; Nakajima et al. 2005). In this paper, we present the results of analyses of ADMs for IgG-binding proteins and their mutants.

At the same time, the statistics on interresidue distances can be converted to an effective interresidue potential (Kikuchi 1996, 1999). We calculated the contact frequency of each residue of a protein with random conformations using the interresidue potential based on the interresidue average-distance statistics, and we obtained values similar to ϕ values (Daggett and Fersht 2000). Of course, the underlying concept of these values, which we refer to as $p(\mu)$ values, is not exactly the same as ϕ values. However, the comparison of the $p(\mu)$ values with the ϕ values for a protein suggests that a profile of $p(\mu)$ values reveals the information on the folding initiation site. In this paper, we try to analyze how the difference in 3D structures of the mutants of IgG-binding protein A and protein G can be detected from their sequences with the $p(\mu)$ value analyses in combination with the ADM analyses.

Methods

Proteins used in the present work

We treat protein A, protein A219, protein G, and protein G311 in this study. The sequences of these proteins are shown in Fig. 1. The identical residues between protein A219 and protein G311 are indicated, and the sequence homology between these two proteins is 59%. The 3D scaffold of protein A and protein A219 is 3α helix bundle structure, and that of protein G and protein G311 is ($\alpha + \beta$) structure. We present the 3D structures of protein A and protein G in Fig. 2.

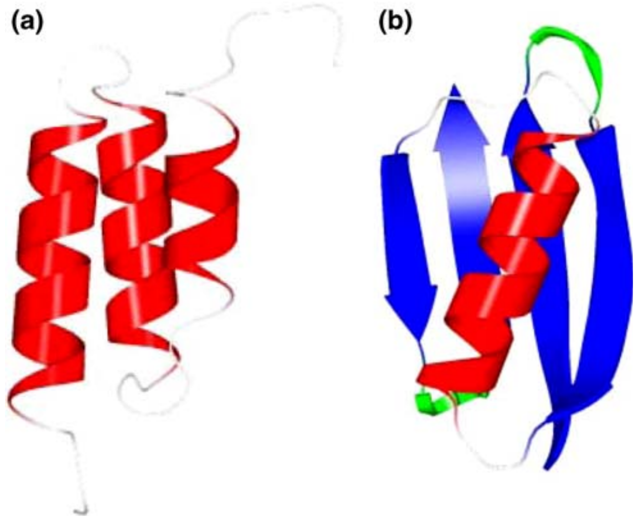

Fig. 2 3D structures of **a** protein A (3α helix bundle structure) and **b** protein G [($\alpha + \beta$) structure]. The α helices are shown in *red* and the β strands in *blue*

Brief survey of the average distance map (ADM) method

The ADM method is described in Kikuchi et al. (1988) and Kikuchi (2002) in detail. Here, let us briefly present a survey of the method.

Calculation of the average distances between residues in each range defined as separation between residues along the sequence of a protein

Interresidue average distances were calculated according to ranges defined in advance. When i and j are residue numbers along a given sequence, a range is defined as the length between two residues along the sequence. That is, a range is defined as $M = 1$ when $1 \leq k \leq 8$; $9 \leq k \leq 20$, $21 \leq k \leq 30$, $31 \leq k \leq 40$ and so on define ranges $M = 2, 3, 4, \ldots$, respectively, where $k = |i - j|$. An average distance, $d(A, B, M)$, where A and B denote amino acid types, was calculated for every residue pair in the range M. Then, a contact map for a protein with unknown 3D structure based on the average distances was constructed. If the average distance between i-th and j-th residues in range M is less than a cutoff distance defined in advance, i.e., $d(A, B, M) \leq d_c(A, B, M)$, we make a plot on the map. Here, $d_c(M)$ denotes a cutoff distance of the range M. The cutoff value was determined in the following way.

Definition of cutoff distances used to construct ADMs

The procedure to define a cutoff distance value in each range for the construction of the ADM of a given sequence is as follows. A cutoff distance for each range is determined in order that the contact density of the whole ADM is close to that of the real distance map (RDM) of a given protein, where the RDM represents a contact map constructed based on the actual 3D structure of a protein. (A contact on RDM is defined when the interresidue Cα atomic distance is less than 15 Å in the present study). The contact density on RDM can be approximated by the formula $\rho_{\rm av} = {\rm C/N}$, where $\rho_{\rm av}$ is the average value of contact density for the entire region of a map, N is the total number of residues, and C is an adjustable constant (Kikuchi et al. 1988). $C = 36.12$ is used in the present work. To reproduce a value of $\rho_{\rm av} = {\rm C/N}$, cutoff distances for construction of ADM of a given protein are determined. Thus, a different cutoff distance is used for a different range in the construction of ADM. Furthermore, the following formula is assumed, i.e., the number of residue pairs that make contacts obeys the following equation for range M (Kikuchi et al. 1988).

$$P(M)_C = \left(\frac{D}{M}\right)P(M)_t, \tag{1}$$

$P(M)_c$ is the number of amino acid pairs whose average distances in the range M are less than the cutoff distance, i.e., residue pairs to be plotted on the ADM, and $P(M)_t$ is the total number of residue pairs in the range M, i.e., 210 pairs of residues minus the number of pairs with statistically insufficient occurrence (Kikuchi et al. 1988). D is an adjustable parameter that gives the overall average density $\rho_{\rm av}$ of ADM close to the predicted value of $\rho_{\rm av}$ on RDM.

Contact density difference

A map is divided into two parts by a line parallel to the ordinate at the i-th residue or by a line parallel to the abscissa at the i-th residue as shown in Fig. 3a and b. We define a contact density difference as $\Delta\rho_i = \rho_i - \tilde{\rho}_i$ where ρ_i and $\tilde{\rho}_i$ denote the contact density of the triangular and trapezoidal parts, respectively (see Fig. 2).

$\Delta\rho_i = \rho_i - \tilde{\rho}_i$ is scanned from residues 1 to N, and then a scanning plot of contact density differences is obtained. We call the scanning plot produced by the division using the line parallel to the ordinate horizontal scanning, and the plot produced by the line parallel to the abscissa vertical scanning. The h of $\Delta\rho_i^{\rm h}$ and v of $\Delta\rho_i^{\rm v}$ denote the horizontal and vertical divisions of a map, respectively.

Definition of compact regions

A scanning plot usually shows some peaks and valleys, which denote a large change in contact density values on a map. Schematic drawing of a horizontal scanning plot of $\Delta\rho_i^{\rm h}$ from 1 to N is shown in Fig. 3c, and the peak and valley appear at a and b in the figure at the large change in the contact density values. The same situation is observed in the vertical scanning in Fig. 3d. Thus, we can detect the boundary of a compact region on a map by a peak and a valley appearing in horizontal and vertical scanning plots of contact density differences.

Prediction of locations of subdomains

The positions of the peaks of scanning plots also define the locations of subdomains as indicated in Fig. 4. A fictitious contact map with two compact areas near the diagonal of a map is presented in this figure, showing the horizontal and vertical scanning plots. The peaks at the residues A and B in the horizontal scanning plot and the residues C and D in the vertical scanning plot indicate the existence of two domains. Thus, we predict the regions A–C and B–D on the map as possible compact regions or domains in the given protein.

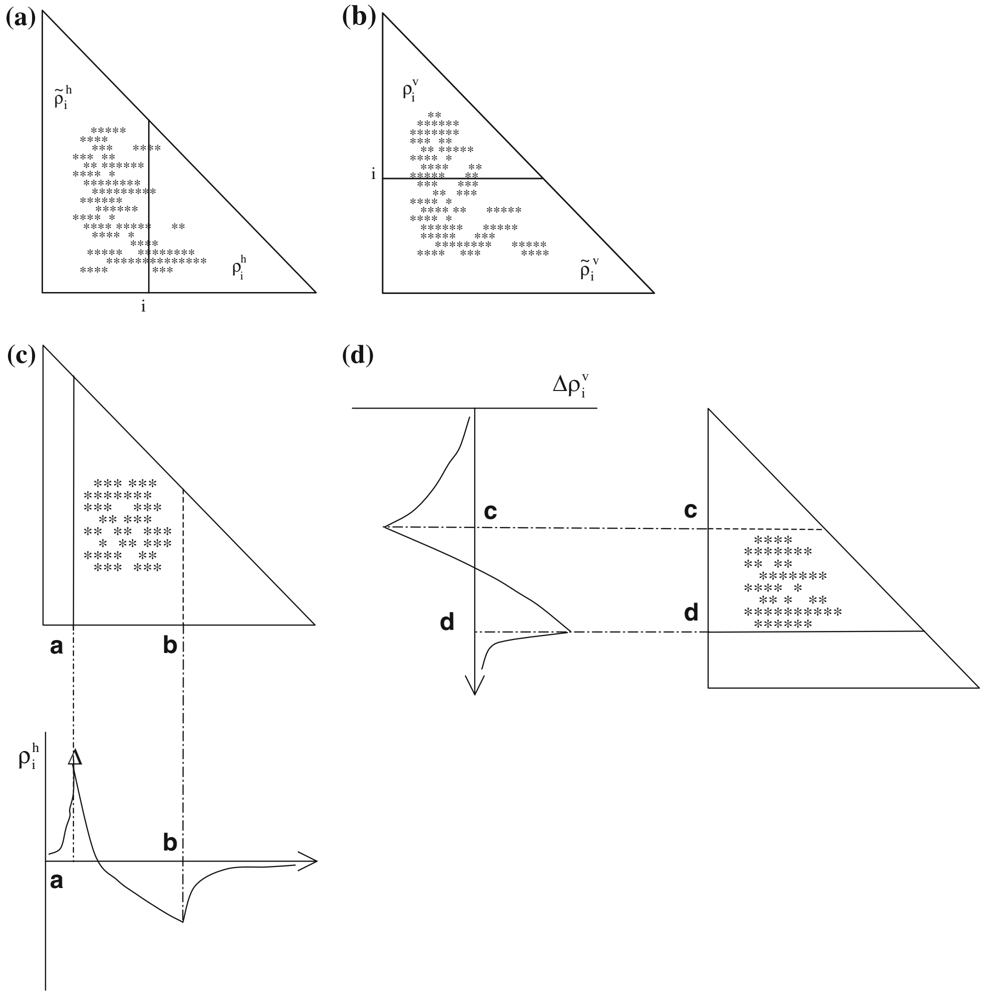

Fig. 3 **a** Contact map for a fictitious protein divided by a line parallel to the ordinate at the residue i. The density of contacts (plots) on the triangular part of the divided map is defined as ρ_i^h and that of trapezoidal part as $\tilde{\rho}_i^h$. **b** Contact map for a fictitious protein divided by a line parallel to the abscissa at the residue i. The density of contacts (plots) on a triangular part of the divided map is defined as ρ_i^v and that of trapezoidal part as $\tilde{\rho}_i^v$. **c** Schematic drawing of a scanning plot of the differences in contact density, $\Delta\rho_i^h = \rho_i^h - \tilde{\rho}_i^h$, from 1 to N on a fictitous map. A peak and a valley appear at a and b in the figure at a large change in contact density values. We refer to the scanning plot produced by the division using the line parallel the ordinate as horizontal scanning. **d** Schematic drawing of a scanning plot of the differences in contact density, $\Delta\rho_i^v = \rho_i^v - \tilde{\rho}_i^v$, from 1 to N on the fictitous map. A peak and a valley appear at c and d in the figure at a large change in contact density values. We refer to the scanning plot produced by the division using the line parallel the abscissa as vertical scanning

Interresidue potential calculated from the average-distance statistics

We derive an interresidue effective potential from the average-distance statistics, which reproduces the average distances and standard deviations used for the construction of ADMs using the Gaussian function as a distribution function (Kikuchi 1996).

Model of a protein

We employ a simple bead model for a protein, i.e., each amino acid is represented by its Cα atom, and the detailed structure of each amino acid is ignored. A peptide bond is represented as a virtual bond with a length of 3.8 Å. In this model, each bond angle θ and each dihedral angle ϕ are variables as shown in Fig. 5.

Interresidue effective potential

When $\varepsilon_{ij}^M(r_{ij})$ is an interresidue effective potential between residue i and j, $\varepsilon_{ij}^M(r_{ij})$ is expressed as Eq. 2 where $\bar{r}_{\mathrm{AB}}^M$ is the average distance between Cα atoms of residue types A and B in range M and σ_{AB}^M is the standard deviation.

$$\varepsilon_{ij}^M(r_{ij}) = kT\frac{\left(r_{ij} - \bar{r}_{\mathrm{AB}}^M\right)^2}{\sigma_{\mathrm{AB}}^M} + kT\ln Z + \frac{1}{2}\ln 2\pi\sigma_{\mathrm{AB}}^M, \qquad (2)$$

r_{ij} is the distance between Cα atoms of residues i and j. Z is the partition function. Residue types A and B correspond to the residue types of i and j. k and T are the Boltzmann constant and temperature, respectively. The constant terms in Eq. 2 can be regarded as the zero point. In the present study, we put $\varepsilon_{ij}^M(r_{ij}) = \varepsilon_{\mathrm{HC}}$ when $\bar{r}_{ij}^M \le r_{\mathrm{cut}}$. We set $r_{\mathrm{cut}} = 1.9$ Å and $\varepsilon_{\mathrm{HC}} = 50$ kcal/mol. These values were obtained empirically (Kikuchi 1996).

Monte Carlo simulation

A simulation was performed from a totally randomized conformation, i.e., we performed sampling of random structures with the present potential using the standard Metropolis Monte Carlo technique. In a Monte Carlo simulation, each dihedral angle, ϕ, and bond angle, θ, of a residue was varied within $-\gamma\pi \le \phi,\ \theta \le \gamma\pi$ followed by

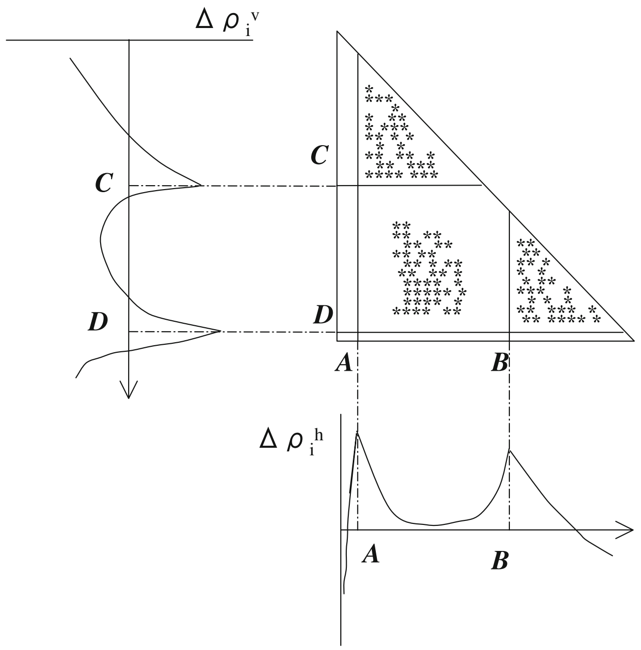

Fig. 4 A fictitious contact map with two compact areas near the diagonal of a map with horizontal and vertical scanning plots. The peaks at the residues *A* and *B* in the horizontal scanning plot and the residues *C* and *D* in the vertical scanning plot indicate the existence of two domains. The regions *A*–*C* and *B*–*D* on the map are the possible compact regions or domains in the protein

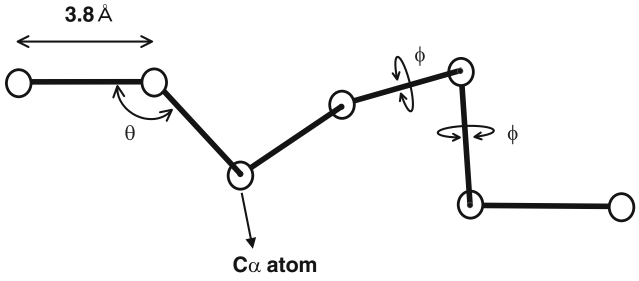

Fig. 5 The model of a protein in this study. Each amino acid is represented by its Cα atom and the detailed structure of each amino acid is ignored. A peptide bond is represented as a virtual bond with the length of 3.8 Å. In this model, only bond angle θ and dihedral angle ϕ are variable

the Metropolis judgment. γ and temperature parameter T were adjusted in order that the acceptance ratio in the Monte Carlo routine was around 0.5. This procedure was iterated for all residues. For a whole simulation, this routine was iterated 60,000 times.

Calculations of the contact frequency during the simulations

The contact frequency $g(i,\ j)$ for a residue pair of the sequence, in other words, the contact probability, was calculated. Contact is defined as r_{ij} of less than 10 Å in this study. A measure of high contact frequency, $Q(\mu,\ \nu)$, is defined according to the following equation, where μ and ν are the μ-th and ν-th residues.

$$Q(\mu,\nu) = \frac{\left(g(\mu,\nu) - \left(\frac{\sum_{|i-j|=m} g(i,j)}{\sum_{|i-j|=m}}\right)\right)}{D(m)}. \tag{3}$$

Here, $D(m)$ is defined as in Eqs. 4 and 5.

$$D(m) = \sqrt{\frac{\sum_{|i-j|=m}\left(g(m) - g(i,\ j)_{|i-j|=m}\right)^2}{\sum_{|i-j|=m}}} \tag{4}$$

$$g(m) = \frac{\sum_{|i-j|=m} g(i,\ j)}{\sum_{|i-j|}} \tag{5}$$

$p(\mu) = \sum_{\nu} Q(\mu,\nu)$ expresses a residue showing high contact frequency with other residues and this value corresponds to a ϕ value (Daggett and Fersht 2000). We ran 10 simulations for each protein, and took average values of the 10 simulations.

A ϕ value is an experimentally observed value defined for each residue, and this value denotes the measure of the involvement of each residue in the native structure formation during its folding transition state (Daggett and Fersht 2000).

Results

ADMs for protein A, protein A219, protein G, and protein G311

The ADMs for protein A, protein A219, protein G, and protein G311 are presented in Fig. 6a–d. For protein A, ADM predicts the region 9–54 as a domain including the 3α helices, and for protein A219, the region 16–56 is predicted as a domain that contains the second and third helices and a part of the first helix. The ADMs of these 3α bundle proteins basically show that the predicted compact regions or subdomains contain 3α helices.

On the other hand, the ADM for protein G predicts the region 20–54, located in the C-terminal side, as a subdomain, and for protein G311 the region 29–45, as presented in Fig. 6c and d, respectively. That is, the compact regions predicted by the ADMs of the $(\alpha + \beta)$ proteins contain mainly the central α helix and β strand(s) located in the C terminal region of the sequence. From these ADMs, the following predictions are made. For proteins with the 3α helix bundle structure, the threee helices form compact regions during the folding, and for the $(\alpha + \beta)$ structure proteins, the C terminal part is mainly involved in the folding. Thus, some differences can be observed on the ADMs.

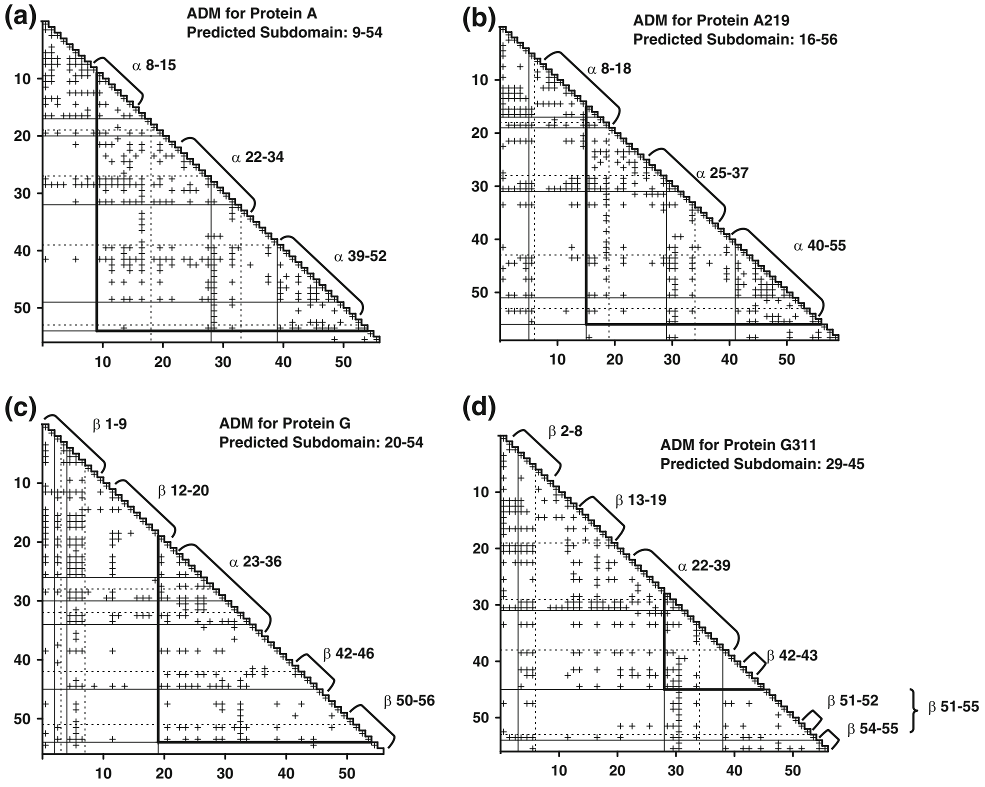

Fig. 6 Average distance maps (ADMs) for the proteins treated in this paper. The positions of secondary structures are indicated near the *diagonal lines* of the maps. **a** ADM for protein A. The predicted region of the subdomain is 9–54, enclosed by *thick black lines* on a map. **b** ADM for protein A219. The predicted region of the subdomain is 16–56, enclosed by *thick black lines* on a map. **c** ADM for protein G. The predicted region of the subdomain is 20–54, enclosed by *thick black lines* on a map. **d** ADM for protein G311. The predicted region of the subdomain is 29–45, enclosed by *thick black lines* on a map

Analysis with the interresidue effective potential and $p(\mu)$ value analysis

Figure 7a shows the $p(\mu)$ values for protein A and protein A219. The positions of the peaks of the $p(\mu)$ values for protein A are 16-V, 30-F, and 44-V, and the peaks appear in the $p(\mu)$ values for protein A219 at the same position. These residues at the peaks are totally conserved between protein A and protein A219. The residues are located in the first, second, and third helices, respectively. Furthermore, in the native structures of these proteins, the residues at the peaks of the $p(\mu)$ values form a hydrophobic packing as shown in Fig. 8a. It is recognized that the $p(\mu)$ values of some residues around 16-V in the N-terminal helix are higher than the $p(\mu)$ values of the residues in the second and third helices, which suggests the deep involvement of the N-terminal helix in the folding. The experimentally observed ϕ values for protein A by Sato et al. (2004) are also presented in Fig. 7a (see also Sato et al. 2006; Sato and Fersht 2007). We notice from Fig. 7a that the locations of peaks of $p(\mu)$ values correspond well to the ϕ values.

We also show in Fig. 7b the $p(\mu)$ values for protein G and protein G311, whose structures exhibit $(\alpha + \beta)$ topology. This figure indicates that the $p(\mu)$ values of the central α helix and the third β strand are higher compared with those of the 3α helix bundle structures. Among the residues on and near the peaks of the $p(\mu)$ curve for protein G, 34-A and 43-W, and 23-A and 45-Y form hydrophobic packing in the native structure as shown in Fig. 8b. These residues correspond to 34-L and 43-W, and 23-A and 45-Y in protein G311, and they also form the hydrophobic contacts in the native structure of protein G311 (Fig. 8b). The ϕ value obtained by McCallister et al. (2000) also indicates the role of the C-terminal region and the α helix in the folding, although their detailed analysis suggests the involvement of a β turn between β3 and β4.

Discussion

The predictions by the ADMs for protein A and protein A219 suggest almost the same folding properties, i.e., the ADMs predict a folding process in which 3α helices form a compact region and grow to the native structure. This observation is reflected in the $p(\mu)$ profiles—the predicted contact frequency of each residue with the other residues—for protein A and protein A219, i.e., the three peaks are observed in the $p(\mu)$ profile and each peak corresponds to each helix. These results correspond to the fact that the three residues at the peaks form the hydrophobic contact in the native structures of both protein A and protein A219 as presented in Fig. 8a. These results also correspond to the ϕ value profile for protein A. Thus, the analyses with average-distance statistics predict the folding properties of protein A and protein A219.

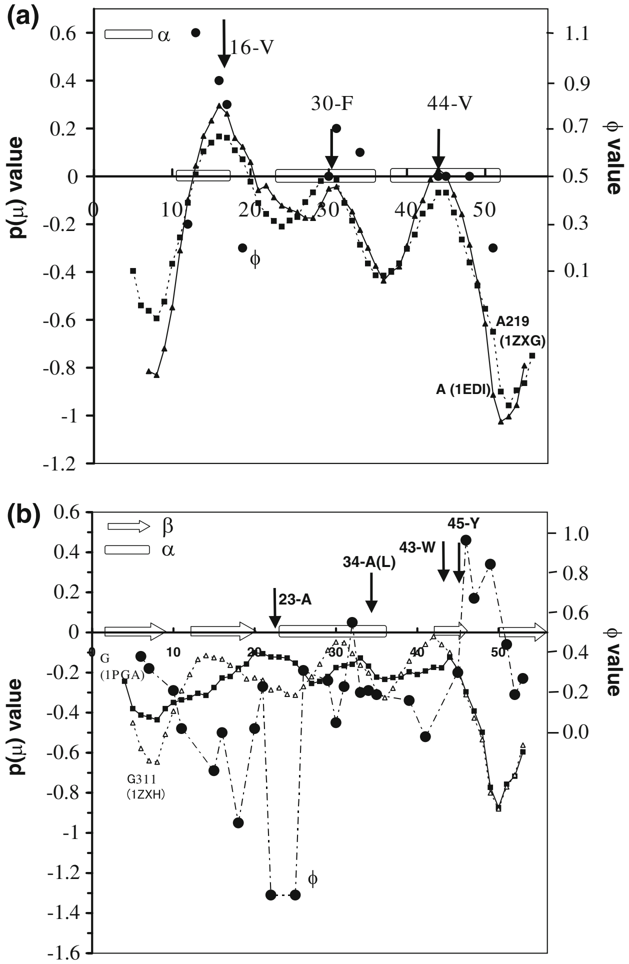

Fig. 7 **a** Plots of $p(\mu)$ values for protein A and protein A219. The plot for protein A is denoted by *triangles with a solid line* and that for protein A219 by *rectangles with a broken line*. The plot of ϕ values for protein A (Sato et al. 2004) is also indicated with *filled circles*. The *black arrows* indicate the location of the peaks of the $p(\mu)$ values for protein A and protein A219. The positions of α helices are indicated by *open rectangles*. **b** Plots of $p(\mu)$ values for protein G and protein G311. The plot for protein G is denoted by *rectangles with a solid line*, and that for protein G311 by *triangles with a broken line*. The plot of ϕ values for protein G is also indicated with *circles with a chain line*. The *black arrows* indicate the location of the peaks of the $p(\mu)$ values for protein G and protein G311. The positions of an α helix and β strands are indicated by an *open rectangle* and *open arrows*, respectively

For protein G and protein G311, the ADM profiles predict that each protein folding starts at the C-terminal region, i.e., folding occurs at the central α helix and the third (and fourth) β strand(s) in each protein. The $p(\mu)$ profiles also indicate the activation of the central helix and the third strand for the contact formation. The tendency in the location of peaks of the $p(\mu)$ profile for protein G somewhat resembles the location of the peaks in the ϕ value profile obtained by McCallinger et al. (2000) as seen in Fig. 7b. It is observed that some of the residues on and near the peaks of the $p(\mu)$ profiles for protein G and protein G311 form hydrophobic contacts in the native structures as presented in Fig. 8b, although the correspondence between the residues forming the hydrophobic contacts and the residues on the peaks of the $p(\mu)$ profile is more remarkable in the cases of protein A and protein A219. However, a remarkably high peak is observed at the β turn between the third and fourth β strands in the ϕ value profile (McCallinger et al. 2000) but not in the $p(\mu)$ profile. The involvement of the C-terminal β turn is also indicated by the results of unfolding simulations (Scott and Daggett 2007). Thus, we need further detailed simulations to predict the fine structure of the ϕ value profile. We are currently working on this problem by incorporating the transition structure information into the contact-frequency calculations.

The problem in this paper is whether we can detect the 3D structural difference between protein A219 and protein G311 from the sequences with the high 59% sequence homology. As an attempt, let us compare the local sequences around the highest peaks of the $p(\mu)$ profiles for protein A219 and protein G311. The highest peak of the $p(\mu)$ profile for protein A219 is located at the residue 16-V, and that for G311 at the residue 42-V. Tentatively, let us take the sequence of $i \pm 5$ residues, where i stands for the residue number of the highest peak of a $p(\mu)$ profile. First of all, we compare the residues 11–21 of protein A219 with the corresponding part of protein A (see Figs. 1 and 9). We note that only one residue differs between the local sequences of protein A219 and protein A, i.e., the amino acid identity of this part is 90.9%. In contrast, when we compare this part of the sequence of protein A219 with the corresponding local sequence of protein G, the amino acid identity is 9%. It should be recalled that the 3D scaffold of protein A219 is the same as that of protein A, that is, the 3α helical bundle structure. In other words, protein A219 folds to the 3D structure of a protein whose corresponding local sequence is highly homologous to the sequence around the highest peak of the $p(\mu)$ profile for protein A219.

Next, the local sequence 37–48 of protein G311 is compared with the corresponding part of protein G (see Fig. 9). The amino acid identity of these local sequences is 81.8%. When we compare 37–48 of protein G311 with the corresponding part of protein A, the amino acid identity is 18% (Fig. 9). The 3D structure of protein G311 is the ($\alpha + \beta$) structure, i.e., the 3D structure of protein G311 is the same as that of protein G, whose corresponding local sequence shows high homology with the local sequence around the highest peak of the $p(\mu)$ profile for protein G311. Thus, as far as the IgG-binding domains are concerned, we can detect the 3D structural difference between protein A219 and protein G311 using the $p(\mu)$ value

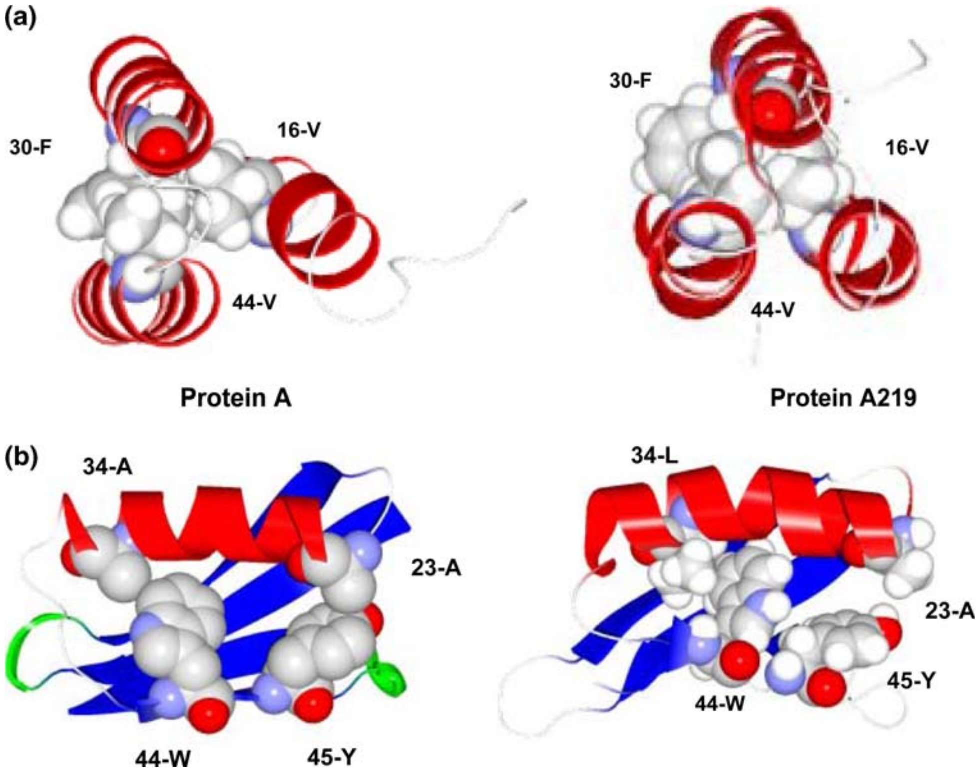

Fig. 8 **a** The 3D structures of protein A and protein A219 with the residues showing the peaks in the profiles of $p(\mu)$ plots. These residues form the hydrophobic packing. **b** The 3D structures of protein G and protein G311 with the residues located near the peaks in the profiles of the $p(\mu)$ plots, which form hydrophobic packing

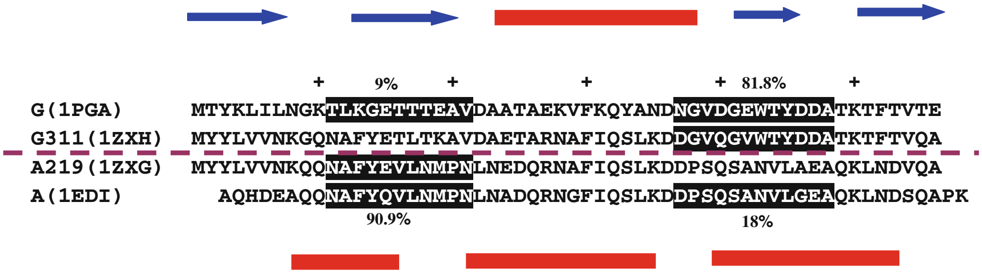
+ 9% + + + 81.8% +
G(1PGA) MTYKLILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDGEWTYDDATKTFTVTE
G311(1ZXH) MYYLVVNKGQNAFYETLTKAVDAETARNAFIQSLKDDGVQGVWTYDDATKTFTVQA
A219(1ZXG) MYYLVVNKQQNAFYEVLNMPNLNEDQRNAFIQSLKDDPSQSANVLAEAQKLNDVQA
A(1EDI) AQHDEAQQNAFYQVLNMPNLNADQRNGFIQSLKDDPSQSANVLGEAQKLNDSQAPK
90.9% 18%

Fig. 9 Comparison of the local sequences near the residue at the peaks of the $p(\mu)$ profiles. We take the sequence of $i \pm 5$ residues, where i stands for the residue number of the highest peak of a $p(\mu)$ profile. The local sequence used in the protein A219 comparisons is the residues 11–21. This part is compared with the corresponding regions of the sequences of protein A and protein G. The local sequences used for the sequence comparisons are indicated by *black bars*. The sequence identity between these regions of protein A219 and protein A is 90.9%, whereas it is 9% when the local sequence of protein A219 is compared with the corresponding part of protein G. The highest peak of the $p(\mu)$ profile for protein G311 is at 42-V. The region 37–48 of protein G311 is compared with the corresponding part of protein G. The amino acid identity of these local sequences is 81.8%, but the amino acid identity of 37–48 for protein G311 with the corresponding part of protein A is 18%

profiles calculated based on the average-distance statistics and local sequence homology. That is, even though two IgG-binding domain sequences show very high homology, the difference in local sequence homology around the peaks of the $p(\mu)$ profiles determines the 3D structural difference. We consider that a region containing the highest $p(\mu)$ values promotes the folding of a protein and therefore the difference in the folding can be detected in the present manner. We are considering that the present strategy can be extended to protein 3D-structure prediction. We are currently working in this direction.

Acknowledgments This work was supported by Grant-in-Aids for Scientific Research (C) (no. 19510202) from the Ministry of Education, Culture, Science, Sports, and Technology (MEXT) of Japan.

References

Alexander PA, Rozak DA, Orban J, Bryan PN (2005) Directed evolution of highly homologous proteins with different folds by phage display: implications for the protein folding code. Biochemistry 44:14045–14054

Chou KC (2004) Review: structural bioinformatics and its impact to biomedical science. Curr Med Chem 11:2105–2134

Chou KC, Watenpaugh KD, Heinrikson RL (1999) A model of the complex between cyclin-dependent kinase 5(Cdk5) and the activation domain of neuronal Cdk5 activator. Biochem Biophys Res Commun 259:420–428

Chou KC, Wei DQ, Du QS, Sirois S, Zhong WZ (2006) Review: progress in computational approach to drug development against SARS. Curr Med Chem 13:3263–3270

Daggett V, Fersht AR (2000) Transition states in protein folding. In: Pain RH (ed) Mechanisms of protein folding, 2nd edn. Oxford University Press, Oxford

Du QS, Wang SQ, Chou KC (2007) Analogue inhibitors by modifying oseltamivir based on the crystal neuraminidase structure for treating drug-resistant H5N1 virus. Biochem Biophys Res Commun 362:525–531

Gao WN, Wei DQ, Li Y, Gao H, Xu WR, Li AX, Chou KC (2007) Agaritine and its derivatives are potential inhibitors against HIV proteases. Med Chem 3:221–226

He Y, Yeh DC, Alexander P, Bryan PN (2005) Solution NMR structures of IgG binding domains with artificially evolved high levels of sequence identity but different folds. Biochemistry 44:14055–14061

Ichimaru T, Kikuchi T (2003) Analysis of the differences in the folding kinetics of structurally homologous proteins based on predictions of the gross features of residue contacts. Proteins 51:515–530

Kikuchi T (1996) Inter-Cα atomic potentials derived from the statistics of average interresidue distances in proteins: application to bovine pancreatic trypsin inhibitor. J Comput Chem 17:226–237

Kikuchi T (1999) Study of protein fluctuation with an effective inter-Cα atomic potential derived from average distances between amino acids in proteins. J Comput Chem 20:713–719

Kikuchi T (2002) Application to the prediction of structures and active sites of proteins and peptides. In: Pandalai SG (ed) Recent research developments in protein engineering. Research Signpost, Kerala

Kikuchi T, Némethy G, Scheraga HA (1988) Prediction of the location of structural domains in globular proteins. J Protein Chem 7:427–471

Li Y, Wei DQ, Gao WN, Gao H, Liu BN, Huang CJ, Xu WR, Liu DK, Chen HF, Chou KC (2007) Computational approach to drug design for oxazolidinones as antibacterial agents. Med Chem 3:576–582

Lubec G, Afjehi-Sadat L, Yang JW, John JP (2005) Searching for hypothetical proteins: theory and practice based upon original data and literature. Prog Neurobiol 77:90–127

McCallinger EL, Alm E, Baker D (2000) Critical role of b-hairpin formation in protein G folding. Nat Struct Biol 7:669–673

Nakajima S, Emma A-S, Kikuchi T, Arredondo-Peter R (2005) Prediction of folding pathway and kinetics among plant hemoglobins using an average distance map method. Proteins 61:500–506

Oxenoid K, Chou JJ (2005) The structure of phospholamban pentamer reveals a channel-like architecture in membranes. Proc Natl Acad Sci USA 102:10870–10875

Sato S, Fersht AR (2007) Searching for multiple folding pathways of a nearly symmetrical protein: temperature dependent ϕ-value analysis of the b domain of protein A. J Mol Biol 372:254–267

Sato S, Religa TL, Daggett V, Fersht AR (2004) Testing protein-folding simulations by experiment: B domain of protein A. Proc Nat Acad Sci USA 101:6952–6956

Sato S, Religa TL, Fersht AR (2006) ϕ-analysis of the folding of the B domain of protein a using multiple optical probes. J Mol Biol 360:850–864

Schnell JR, Chou JJ (2008) Structure and mechanism of the M2 proton channel of influenza A virus. Nature 451:591–595

Scott KA, Dagget V (2007) Folding mechanisms of proteins with high sequence identity but different folds. Biochemistry 46:1545–1556

Wang SQ, Du QS, Chou KC (2007a) Study of drug resistance of chicken influenza A virus (H5N1) from homology-modeled 3D structures of neuraminidases. Biochem Biophys Res Commun 354:634–640

Wang SQ, Du QS, Zhao K, Li AX, Wei DQ, Chou KC (2007b) Virtual screening for finding natural inhibitor against cathepsin-L for SARS therapy. Amino Acids 33:129–135

Wei DQ, Du QS, Sun H, Chou KC (2006a) Insights from modeling the 3D structure of H5N1 influenza virus neuraminidase and its binding interactions with ligands. Biochem Biophys Res Comm 344:1048–1055

Wei DQ, Zhang R, Du QS, Gao WN, Li Y, Gao H, Wang SQ, Zhang X, Li AX, Sirois S, Chou KC (2006b) Anti-SARS drug screening by molecular docking. Amino Acids 31:73–80

Wu G, Yan S (2008) Prediction of mutations engineered by randomness in H5N1 neuraminidases from influenza A virus. Amino Acids 34:81–90

Ye Y, Wei J, Dai X, Gao Q (2007) Computational studies of the binding modes of A2A adenosine receptor antagonists. Amino Acids. doi:10.1007/s00726-007-0604-2

Zhang R, Wei DQ, Du QS, Chou KC (2006) Molecular modeling studies of peptide drug candidates against SARS. Med Chem 2:309–314