Secondary structure breakers and hairpin structures in myoglobin and hemoglobin

Kenichiro Imai*, Naoyuki Asakawa, Toshiyuki Tsuji, Masashi Sonoyama and Shigeki Mitaku

Department of Applied Physics, Graduate School of Engineering, Nagoya University, Furocho, Chikusa-ku, Nagoya, 464-8603, Japan

*E-mail: imai@bp.nuap.nagoya-u.ac.jp

Abstract

Secondary structure breakers, particularly hairpin structures, impose strong limitations on the global structure of a protein. Three kinds of secondary structure breakers (proline, glycine and amphiphilic residues) were studied in myoglobin and hemoglobin, which are typical all-α type proteins. Secondary structure breakers were located as predicted in about two thirds of the interhelical loops. Charge symmetry analysis provided evidence that high charge symmetry drives the formation of hairpin structures. Based on this information about breakers and hairpin structures, the possibility of folding a protein *in silico* is discussed.

1. Introduction

An amino acid sequence contains all the information required to define the three-dimensional structure of a protein. Although a protein is represented in a database by a one-dimensional sequence composed of 20 types of amino acids, the true protein is a polymer that forms a complex three-dimensional molecular structure through the interplay of various physical interactions. A major goal in protein science is to sufficiently understand the physicochemical processes controlling protein folding so that the *in silico* regeneration of a protein's structure from its sequence will be possible. Given the current status of computational approaches, structure prediction based solely on a physicochemical approach is extremely difficult. Therefore, most research groups participating in CASPs (Critical Assessment of Techniques for Protein Structure Prediction) rely on the rapidly growing number and size of sequence and structure databases.

Some classes of proteins exhibit rather unique physicochemical properties. For example, membrane proteins have significant hydrophobic segments that penetrate membranes. This property, together with the existence of amphiphilic residue clusters in the neighborhood of the hydrophobic regions, was instrumental in the development of a software system (SOSUI) for predicting membrane proteins from sequences [1][2][3]. In addition, very extended proteins such as calmodulin and troponin C show very simple physical characteristics such as large net negative charges, resulting in electric repulsion between the amino- and carboxyl-terminal domains. Extended proteins can be predicted accurately only by considering the net charge distribution in an amino acid sequence [4]. However, most proteins are water soluble and globular, and for these proteins it is very difficult to predict local structural elements based on the physicochemical properties of the amino acid sequence.

As a first step towards developing analyses for local structures in soluble globular proteins, we have focused on studying the 'break points' in myoglobin and hemoglobin, which are typical examples of all-α type proteins [5]. Sequence components that break α-helices, together with hairpin structures, may determine the fold of a protein if the helices can be assumed to be rigid bodies. Recently we analyzed in detail the conditions required for making breaks in the local order of α-helices and β-sheets [6]. In addition to the well-known secondary structure breakers proline and glycine, we revealed that amphiphilicity peaks can also be potential breakers. Furthermore, detailed analysis of amino acid sequences near the secondary structure breakers showed that other environmental factors, such as hydrophobicity and the existence of small polar residues, are necessary in order to determine the final local structure adjacent to the potential breakers. As discussed in the present paper, when this method was applied to the amino acid sequences of myoglobin and hemoglobin, break points could be predicted, except for two exceptions. Detailed investigation of the charge distribution around the break points revealed significant charge symmetry that physically stabilizes hairpin structures. We here present a simple model for the folding of globin-type proteins based on information about the breaks and hairpins.

2. Methods

2.1 Data sets for analysis

Seventeen proteins with globin-type folds were selected from the SCOP database [7]. Sequence identity amongst these proteins was less than 30 %. However, investigation of the amino acid sequences by breaker prediction and charge symmetry analysis lead to the conclusion that the 17 proteins shared substantial physicochemical properties. The results for two proteins, myoglobin from sperm whale (PDB ID: 1mbo) and hemoglobin (domain 1) from pig roundworm (PDB ID: 1ash), are presented in this paper. We selected this pair of proteins, since the results of the analyses for them showed the most significant difference.

2.2 Prediction of secondary structure breakers

Our recently developed method for predicting secondary structure breakers is based on three main factors and four environmental parameters [6]. The approach has two principle steps. First, candidates are enumerated by the main factors, then the breakers are predicted, discriminating the candidates by the environmental factors.

The three main factors impacting secondary structure breakers were defined by the existence of proline and glycine, and peaks in the protein's amphiphilicity index. The parameters for the existence of proline and glycine were represented by $P(i)$ and $G(i)$, respectively. A value of 1 for $P(i)$ represents the existence, and a value of zero represents the absence, of proline at the i-th residue. $G(i)$ represents the corresponding parameter for glycine. The amphiphilicity index was previously defined as the transfer energy for the hydrophobic stem group based on accessible surface area, and was first used in developing the membrane protein prediction system, SOSUI [3]. Finite amphiphilicity index values were calculated for lysine (3.67), arginine (2.45), histidine (1.45), glutamic acid (1.27), glutamine (1.25), tryptophan (6.93), and tyrosine (5.06). The first five residues bear very polar side chains, whereas the last two residues are only weakly polar. The amphiphilicity index for strongly polar residues was named the A-index, and that for weakly polar residues was named the A'-index. The A-index was used for the enumeration process, and A'-index was used for the discrimination process.

The averages of the three main parameters, $\langle P(j) \rangle$, $\langle G(j) \rangle$, and $\langle\langle A(k) \rangle\rangle$, were calculated by the following equation:

$$\langle X(j) \rangle = \left[\sum_{i=j-m}^{j+m} X(i) \right] \Big/ (2m+1) \tag{1}$$

$$\langle\langle X(k)\rangle\rangle = \left[\sum_{j=k-l}^{k+l}\langle X(j)\rangle\right]\Big/(2l+1) \tag{2}$$

The parameter $X(j)$ represents one of $P(j)$, $G(j)$ or $A(j)$. The window size $(2m+1)$ of the moving average was 3 for proline and glycine, and 7 for the A-index. The window size $(2l+1)$ of the double average for $\langle X(j)\rangle$ was 9. We calculated the double average of the amphiphilicity index because the plot of the moving average for $\langle A(j)\rangle$ is still very notchy. The peak values obtained for the three parameters, $\langle P(j)\rangle$, $\langle G(j)\rangle$, and $\langle\langle A(k)\rangle\rangle$, were identified as potential secondary structure breakers. The thresholds for $\langle P(j)\rangle$ and $\langle G(j)\rangle$ were zero, and that for $\langle\langle A(k)\rangle\rangle$ was 0.4. A single residue of proline and glycine can break the secondary structure, but their cluster does not form many breaking points but a single break. In this meaning, we set the threshold for proline and glycine as zero. The peaks for $\langle P(j)\rangle$ were good secondary structure breakers in themselves. However, glycine residues and clusters of amphiphilic residues were located both at the termini of secondary structure elements, and inside the secondary structures. The branching of these two potential breakers depends on the physicochemical properties of the surrounding regions. Therefore, the second step of the prediction process was applied only to candidates containing clusters of glycine and amphiphilic residues.

The four environmental parameters were the average of the Kyte and Doolittle hydrophobicity index $\langle H(j)\rangle$ [8], the helical periodicity score $\langle HPS(j)\rangle$, the density of small polar residues $\langle ST(j)\rangle$, and the average A'-index $\langle A'(j)\rangle$. The average was calculated by Eqs. (1) using a window size of 7, except for the helical periodicity score $\langle HPS(j)\rangle$.

$$\langle HPS(j)\rangle = \max\{|\langle HP(j)\rangle|, |\langle HP(j-1)\rangle|\} \tag{3}$$

where $\langle HP(j)\rangle = \left[H(j+5)-H(j+3)+H(j+1)-H(j)+H(j-2)-H(j-4)\right]\Big/5$. This score is a very simple index, but effectively represents the α-helical periodicity.

We calculated the deviation of the parameters from the corresponding values for the dataset of false data, namely the breaker candidates in the secondary structure core regions. We assumed that the environment of a break point could be represented by a segment 15 residues long. The details of the calculation have been described previously [6]. The discrimination function obtained from the primary component analysis is defined by Eqs. (4):

$$Score(l) = a_0 + a_1\Delta\langle H(l)\rangle + a_2\Delta\langle HPS(l)\rangle + a_3\Delta\langle ST(l)\rangle + a_4\Delta\langle A'(l)\rangle \tag{4}$$

in which l represents the position of a candidate secondary structure breaker. The parameters, $\Delta\langle H(l)\rangle$, $\Delta\langle HPS(l)\rangle$, $\Delta\langle ST(l)\rangle$ and $\Delta\langle A'(l)\rangle$, are the deviation from the average values for the false data set. The coefficients of the discrimination score were as follows: $a_1 = 2.01$, $a_2 = 1.90$,

$a_3 = 32.9$, $a_4 = 5.30$, and $a_0 = -0.68$ for breaks by glycine; and $a_1 = 3.15$, $a_2 = 1.36$, $a_3 = 42.5$, $a_4 = 2.40$, and $a_0 = -1.71$ for breaks by amphiphilicity peaks [6].

Accuracies as high as 90 % were achieved for predicting the loop regions, including the loop core and the breaking regions [6]. Almost two-thirds of the loops were covered by the breakers predicted by this method [6]. Namely, the specificity and the sensitivity of the prediction were about 90 % and 66 %, respectively.

2.3 Analysis of charge symmetry centers

We analyzed the charge distribution of amino acid sequences by two methods: the charge density (CD) map and the charge symmetry analysis. In both methods, the charge densities of segments are calculated, as defined by the following equation.

$$\langle C(j,k) \rangle = \left[\sum_{i=j}^{k} C(i) \right] \bigg/ (k-j+1) \tag{5}$$

The CD map is a triangle map in which the charge density $\langle C(j,k) \rangle$ of a segment from j-th to k-th position of an amino acid sequence is plotted at the point of (j, k) in the map by a pseudo-color [9]. Positive charge is expressed by blue color and negative charge by red. The values of $C(i)$ are 1 and -1 for positive and negative charges, respectively, and 0 for neutral residues. In the analysis of charge density map, we compared it with another kind of the triangle map, the contact map, in which the distance between C_α atoms is shown by four colors: black, orange, purple and pink represent contacts within 7Å, 10 Å, 13 Å and 18 Å, respectively.

The other method analyzing the charge distribution is the charge symmetry analysis. When the distribution of electric charges changes sign, the inverse charge symmetry is formed at which the polypeptide is often folded into a hairpin structure. Charges with opposite sign tend to attract each other, and the inverse symmetry of the electric charges in the amino acid sequence can cause hairpin structures. To analyze the charge symmetry, the charge density $\langle C(j-3,j+3) \rangle$ of a window of seven residues was used instead of the electric charge $C(i)$ on each amino acid.

$$\langle C(j-3,j+3) \rangle = \left[\sum_{i=j-3}^{j+3} C(i) \right] \bigg/ 7 \tag{6}$$

When we evaluate the charge symmetry at position j in an amino acid sequence, two values of charge density at the symmetric position with respect to the j-th residue, $\langle C(j-k-3,j-k+3) \rangle$ and $\langle C(j+k-3,j+k+3) \rangle$, are compared. If the symmetry is perfect, the two values have the same absolute value and the opposite sign. We define a value, $D(j,k)$, dividing the difference between $\langle C(j-k-3,j-k+3) \rangle$ and $\langle C(j+k-3,j+k+3) \rangle$ by 2, as shown in the following equation.

$$D(j,k) = \frac{1}{2} \left(\langle C(j+k-3,j+k+3) \rangle - \langle C(j-k-3,j-k+3) \rangle \right) \tag{7}$$

Then, the value of $D(j,k)$ is equal to $\langle C(j-k-3, j-k+3)\rangle$ and $\langle C(j+k-3, j+k+3)\rangle$ for the case of perfect charge symmetry. Therefore, the sum of the squares of deviation can be used for scoring the symmetry.

$$S(j,m) = \frac{1}{2m}\sum_{k=1}^{m}\left\{\begin{array}{c}\left(\langle C(j+k-3, j+k+3)\rangle - D(j,k)\right)^2 \\ +\left(\langle C(j-k-3, j-k+3)\rangle + D(j,k)\right)^2\end{array}\right\} \tag{8}$$

Herein, m represents the scan length. There are symmetries not only in long twin segments, but also in short twin segments. Therefore, we varied the scan length from 5 to 15, and the maximum values of $S(j,m)$ among the eleven numbers were used to estimate the charge symmetry score, CS score.

$$\langle CS\ score(j)\rangle = \max\left\{S(j,m)\right\}_{m=5\sim15} \tag{9}$$

When these scores are plotted as a function of the amino acid sequence, the position of the inverse charge symmetry can be identified from the valley in the plot.

3. Results

We previously investigated secondary structure breaker mechanisms and elucidated that at least three factors cause breaks in secondary structure: clusters of proline, clusters of glycine, and clusters of amphiphilic residues. The predicted breakers covered about two-thirds of all loop segments, and the predicted breakers were located in loop region with accuracies as high as 90 % [6]. Figure 1(a) and 1(b) show the predicted breakers plotted as a function of amino acid sequence for hemoglobin from pig roundworm, and sperm whale myoglobin, respectively. The colored bars at the top of the graphs represent the eight helical regions shown in the crystal structures depicted to the right of the graphs. The peaks in the upper three traces in each graph indicate the positions of predicted secondary structure breakers based on the existence of proline, glycine and amphiphilic residues. These results are combined at the bottom of each graph. Herein, we define the breaking region by a segment of six residues around the terminus of a secondary structure [6]. According to this definition, the predicted breakers agreed well with the known breaking regions in both proteins, allowing the shift of about ± 3 residues. The reason for the shift of about three residues is the coarse graining of the physical properties of amino acid sequences used in the present method of the breaker prediction. Because three residues is less than one turn of the α-helix, the shift of this level seems reasonable as the migration of the end of an α-helix. Comparing Figures 1(a) and 1(b), it is apparent that the types of breakers are not necessarily the same, but the breakers are located at the same breaking region. The sequence homology of these two proteins is only 15%, indicating that they are a remote homolog pair of proteins. The fact that the same structure (or in this case, the
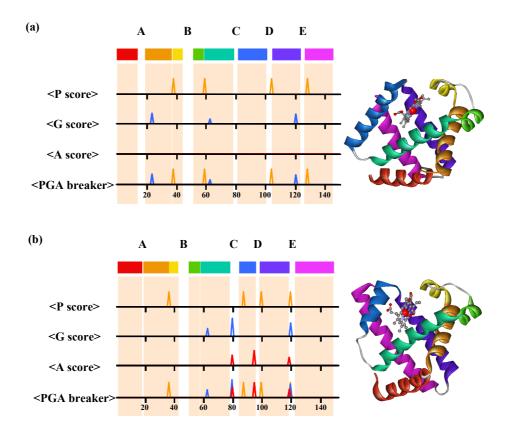
Figure 1. Results of secondary structure breaker prediction for hemoglobin (a) and myoglobin (b).

The orange peaks, blue peaks and the red peaks show the positions of proline breakers, glycine breakers and amphiphilic breakers, respectively. The values of peaks represent the prediction score of each breaker. The bars above the plots of the scores represent the α-helices and the breaks of helices are designated from A to E

break in the secondary structure) can be formed by several different mechanisms is considered the basis for the robustness of protein secondary structures.

The fourth mechanism by which secondary structure can be broken was elucidated from the charge density (CD) map previously developed [9]. Figures 2(a)-2(d) show the CD maps, in which the charge density of clusters is marked by pseudo color: positive by blue, and negative by red. The lower right triangles in Figure 2(a) and 2(b) show the charge density maps of hemoglobin and myoglobin, respectively. The contact maps are shown in the upper left triangle. The orange bars at the bottom of the maps delineate the helical regions of the proteins. The contact maps of hemoglobin and myoglobin are very similar, but the CD maps are considerably different. The CD map of myoglobin is primarily blue, indicating that clusters longer than about 20 residues are positively charged. In contrast, both blue and red areas are apparent in the CD map for hemoglobin. This fact indicates that the electric charges in a hemoglobin molecule are almost neutralized. The CD map indicates that not only two myoglobin molecules but also all segments longer than 20
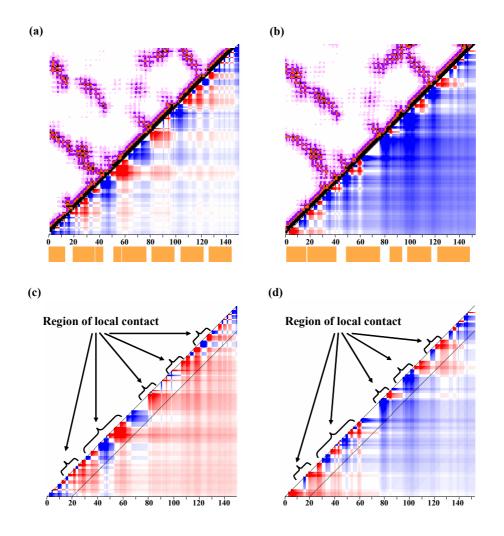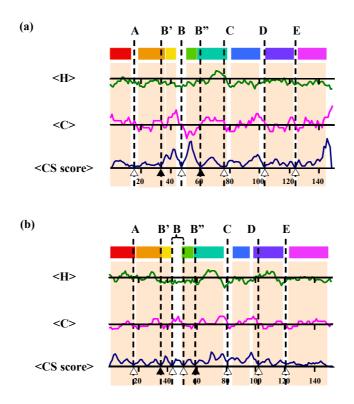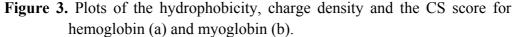
Figure 2. Charge density (CD) map with the contact maps of hemoglobin (a) and myoglobin (b). The CD maps for hemoglobin (c) and myoglobin (d) when the electric charge for histidine is set to zero.

The upper left triangles in (a) and (b) show contact maps plotting pairs of C_α atoms close in space. The distance between C_α atoms is shown by four colors; black, orange, purple and pink represent contacts within $7\,\text{Å}$, $10\,\text{Å}$, $13\,\text{Å}$ and $18\,\text{Å}$, respectively. The light orange bars at the bottom of the maps designate the helical regions of the proteins.

residues will overall be electrically repulsed, while electric repulsion beyond 20 residues for hemoglobin will be small.

In contrast to Figures 2(a) and 2(b), in which a histidine residue is assumed to have a positive charge, Figures 2(c) and 2(d) show the CD maps of hemoglobin and myoglobin, respectively, on the assumption of the neutral histidine. It is well known that the pK_a value of histidine is near pH 7.0 and that histidine is generally neutral when there are large positive charges in its vicinity. The color of the off-diagonal areas in Figure 2(c) changed significantly to light red, while the color in Figure

Figure 3. Plots of the hydrophobicity, charge density and the CS score for hemoglobin (a) and myoglobin (b).

The breaks of α-helices from A to E are in good accordance with the significant minima of the CS scores which are marked by triangles and the dotted lines. The dotted lines are drawn for convenience to identify the correlation between <H>, <C> and <CS score>. The breaks B' and B" seem to be the satellite breaks of the break of B, which is discussed in Figure 5 in detail. The electric charge for histidine is set to zero in the plots for myoglobin.

2(d) remained blue. It is reported that the globin domain of the hemoglobin of pig roundworm is monomeric at pH 8.5, at which histidine is mostly neutralized [10]. The correlation between the degree of dissociation of histidine and the polymerization of proteins is an interesting problem.

The regions of local structure in the CD maps show patchy patterns, and the charge boundary appears to be correlated with the hairpin structures, as seen from the contact map. The boundary between patches of charge density in Figure 2 becomes a type of symmetry center, and the segments of amino and carboxyl sides of the symmetry center interact with each other by electrostatic attraction. Therefore, the charge symmetry centers in amino acid sequences tend to form hairpin structures. This physical intuition can be confirmed by the charge symmetry (CS) score by Eqs. (9) which becomes smaller for points having clearer charge symmetry. Figure 3 shows plots of the hydrophobicity, the charge density, and the CS score for hemoglobin (Figure. 3(a)) and myoglobin (Figure. 3(b)). The minimum of the CS score is in general agreement with the point of charge inversion. Since this score is not calculated from short segments around the inversion point, but

rather from segments longer than eleven residues, the electrostatic attraction between sufficiently long charged clusters gives the minimum of the CS score. There are five large breaks from A to E, and two small breaks (B' and B") in hemoglobin and myoglobin. The positions of the breaks are consistent with the significant minima of the CS score. Breaks B and C in hemoglobin, and breaks A and B in myoglobin, were not predicted from the analysis of proline, glycine and amphiphilic residues. The most plausible mechanism for the fourth type of breaker is electrostatic attraction due to the inverse charge symmetry.

Figure 4 schematically shows the folding of hemoglobin, assuming hairpin structures at CS centers. Open triangles indicate the positions of the CS centers. The two closed triangles indicate the satellite CS centers near break B. Each triangle corresponds to the triangles in Figure 3. Arrows represent the secondary structure breakers, the colors of which correspond to the three kinds of breakers. The purpose of the diagram is to help explain the forces influencing folding, and not for the prediction of the three dimensional structure. Therefore, the angles of the hairpins are adjusted to the actual structure. However, the fact that the globin fold can be regenerated by the very simple rule of hairpin formation at the CS centers strongly suggests that the physical force mediating folding in these types of proteins is electrostatic attraction at the inverse symmetry of charge distribution.

Figure 5 shows the structure around break B, viewed from the top of the break. The three breaks form a S-shaped structure, caused by the combination of hairpins as shown schematically in Figure 5(b). Breaks B, B' and B" probably function as a unit and form a hinge.
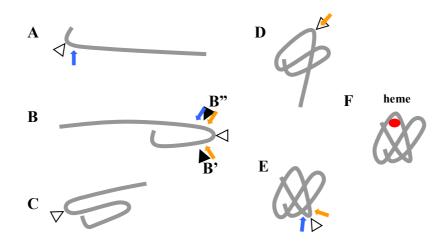

Figure 4. A simple mechanism for the folding of hemoglobin, assuming hairpin structures at the designated positions on the CS centers.

Arrows represent the secondary structure breakers; the color code is the same as in Figure 1. Open triangles show the positions of the CS centers. The two closed triangles indicate the satellite CS centers near break B.
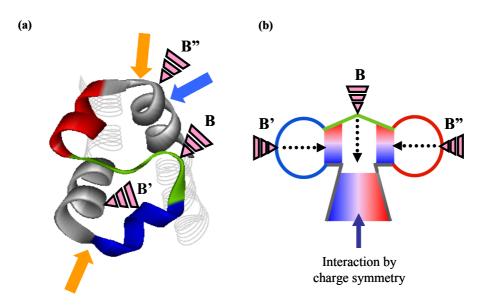
Figure 5. The structure around break B, viewed from the top of the break B (a), and a schematic diagram of three breaks that form a S-shaped structure resulting from the combined hairpins (b).

The triangles indicate the CS centers near break B, B' and B".

4. Discussion

The physicochemical properties of the amino acid sequences of hemoglobin and myoglobin were analyzed from two aspects: the secondary structure breaks caused by proline, glycine and amphiphilic residues, and charge symmetry at the hairpin structures.

It is well established that proline is a secondary structure breaker due to the rigidity of this cyclized residue. It is also known that glycine is abundant in the vicinity of break regions [11][12]. Recently, we found that clusters of amphiphilic residues can be secondary structure breakers depending on the environment around the sequences [6]. Combining the effects of primary factors (proline, glycine and amphiphilic residues) and several environmental factors permitted us to develop a method for accurately predicting secondary structure breakers [6]. In the present paper, the application of this method to the amino acid sequences of hemoglobin and myoglobin revealed that five out of seven loops in these two globins are caused by proline, glycine or amphiphilic residues.

In the remaining two loops, significant charge symmetry was found in the coarse-grained charge distribution. Simple consideration of coulombic interactions between segments on both sides of a charge symmetry center suggests that electrostatic interactions may lead to the formation of hairpin structures. Consequently, all the breaks in the structures of hemoglobin and myoglobin are

understandable from a physical viewpoint. Furthermore, by forming hairpin structures at the charge symmetry centers, the shape of the globin fold could be regenerated. This finding strongly suggests that the secondary structure breakers and the charge symmetries discussed in this paper can be substantial forces driving protein folding. Furthermore, the inverse charge symmetry is probably the fourth mechanism by which secondary structures are broken.

Secondary structure prediction has historically focused on the ordered, structurally regular regions of proteins (i.e., α-helices and β-sheets), rather than on breaks in the secondary structures. Early prediction methods were based on the propensities of single amino acids to adopt a given type of secondary structure; these methods had accuracies in the range of 50-55 % [13][14]. Divergent evolutionary profiling by multiple alignments, combined with other information technology approaches, has recently improved the accuracy of predictions [15][16][17][18][19][20][21] [22][23]. However, accuracies of 80 % have yet to be attained by these methods. The importance of disordered structures has recently been noted, and several methods have been reported for predicting disordered structures [24][25][26]. However, these methods were not designed to predict many short breaks between secondary structure elements. In contrast, the method described in this paper focuses on the breaks in the structure of proteins with globin folds, and reveals a novel mechanism for the formation of breaks: inverse charge symmetry. Adding this mechanism to those previously reported (i.e., clusters of proline, glycine and amphiphilic residues), all the breaks in the structures of hemoglobin and myoglobin could be predicted.

An important problem that remains to be solved is that of secondary structure formation. In this work, the physical factors driving the formation of secondary structure were not discussed. The minima of the CS score are located not only at the breaking regions but also at the secondary structure cores around 30-th, 90-th and 110-th residues (Figure 3). It is physically reasonable that the electric dipoles of an α-helical structures can be stabilized by the inverse charge symmetry, when the negatively and the positively charged residues are positioned at the N- and the C-terminal regions of α-helix, respectively. Therefore, the factors for the branching of a CS center to the hair-pin and the α-helical structures seem to be subtle problem. The physical factors driving secondary structure formation is presently under investigation in our laboratory.

References

[1]   T. Hirokawa, S. Boon-Chieng, S. Mitaku, *Bioinformatics*, **14**, 378-379 (1998).

[2]   T. Tsuji, S. Mitaku, *CBIJ*, **4**, 110-120 (2004).

[3]   S. Mitaku, T. Hirokawa, T. Tsuji, *Bioinformatics*, **18**, 608-616 (2002).

[4]   N. Uchikoga, S. Takahashi, R. Ke, M. Sonoyama, S. Mitaku, *Protein Sci.*, **14**, 74-80 (2005).

[5] C. Branden, J. Tooze, Introduction to Protein Structure, Garland Publishing, (1999).

[6] K. Imai, S. Mitaku, *BIOPHYSICS*, **1**, 55-65 (2005).

[7] A. G. Murzin, S. E. Brenner, T. Hubbard, C. Chothia, *J. Mol. Biol.*, **247**, 536-540 (1995).

[8] J. Kyte and R. F. Doolittle, *J.Mol.Biol.*, **157**, 105-132 (1982).

[9] K. Imai, S. Mitaku, *CBIJ*, **3**, 194-200 (2003).

[10] J. Yang, A. P. Kloek, D. E. Goldberg and F. S.Mathews, *Proc. Natl. Acad. Sci. USA*, **92**, 4224-4228 (1995).

[11] P. Y. Chou, G. D. Fasman, *Adv. Enzymol.Relat. Areas. Mol. Biol.*, **47**, 45-148 (1978).

[12] M. Levitt, *Biochemistory*, **17**, 4277-4285 (1978).

[13] P. Y. Chou, G. D. Fasman, *Biochemistory*, **13**, 222-245 (1974).

[14] J. Garnier, D. J. Osguthorpe, B. Robson, *J. Mol. Biol.*, **120**, 97-120 (1978).

[15] B. Rost, C. Sander, *J. Mol. Biol.*, **232**, 584-599 (1993).

[16] D. T. Jones, *J. Mol. Biol.*, **292**, 195-202 (1999).

[17] A. A. Salamov, V. V. Solovyev, *J. Mol. Biol.*, **247**, 11-15 (1995).

[18] R. D. King, M. J. Sternberg, *Protein Sci.*, **4**, 2298-2310 (1996).

[19] D. Frishman, P. Argos, *Proteins*, **27**, 329-335 (1997).

[20] J. A. Cuff, M. E. Clamp, A. S. Siddiqui, M. Finlay, G. J. Barton, *Bioinformatics*, **14**, 892-893 (1998).

[21] H. Kim, H. Park, *Protein Eng.*, **16**, 553-560 (2003).

[22] J. J. Ward, L. J. McGuffin, B. F. Buxton, D. T. Jones, *Bioinformatics*, **19**, 1650-1655 (2003).

[23] B. Rost, *Methods. Biochem. Anal.*, **44**, 559-587 (2003).

[24] P. Romero, Z. Obradovic, C. R. Kissinger, J. E. Villafranca, A. K. Dunker, *Proc. IEEE Intl. Conf. Neural Networks*, **1**, 90-95 (1997).

[25] R. Linding, L. J. Jensen, F. Diella, P. Bork, T. J. Gibson, R. B. Russell, *Structure*, **11**, 1453-1459 (2003).

[26] J. J. Ward, L. J. McGuffin, K. Bryson, B. F. Buxton, D. T. Jones, *Bioinformatics*, **20**, 2138-2139 (2004).