A high performance prediction system of coiled coil domains containing heptad breaks: SOSUIcoil

Hideki Tanizawa^{1*}, Ganga D. Ghimire^{1, 2}, and Shigeki Mitaku¹

¹ Nagoya University, School of Engineering, Department of Applied Physics,
Furocho, Chikusa-ku, Nagoya 464-8603, JAPAN.

² Venture Business Laboratory, Nagoya University,
Furocho, Chikusa-ku, Nagoya 464-8603, JAPAN

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

(Received October 8, 2008 ; accepted December 12, 2008 ; published online December 29, 2008)

Abstract

The coiled coil structure in proteins is characterized by heptad repeats of hydrophobic amino acids, but many breaks of the heptad repeats are observed within coiled coil regions which are the main cause of the errors in the currently used prediction systems for coiled coils. For understanding the characteristics of coiled coils including heptad breaks, the features of coiled coils were studied focusing on three problems: (1) the determination of appropriate register for the breaks in heptad repeat regions, (2) the discriminations of coiled coil regions using physical properties of amino acid segments which have heptad repeats and (3) the elucidation of the structural difference among several types of heptad breaks. Appropriate registers of heptad repeats and breaks were determined by two steps: first, the typical template of heptad repeats was applied to amino acid sequences and then several types of template for heptad breaks were applied to the segments around inconsistent points of the heptad repeat regions, leading to the most appropriate registers. Then, the coiled coil regions were discriminated from other types of regions by the canonical discriminant analysis, using ten parameters (three physicochemical properties and seven number densities of amino acids). The novel coiled coil prediction system SOSUIcoil showed better performances of coiled coil prediction than other prediction systems. Furthermore, the structures of segments around the heptad breaks were analyzed, indicating that some types of heptad breaks tend to form coiled coil structure whereas the other types are at the end of coiled coils.

Key Words: Coiled coil prediction, physicochemical properties, SOSUIcoil, heptad breaks regions

Area of Interest: Bioinformatics and Bio Computing

1. Introduction

Coiled coil domains are found in many proteins such as fibrous structural proteins, motor proteins, transcription factors and membrane fusion proteins [1]. The coiled coil domains are known to be structurally stable and contribute to the stability of the tertiary and quaternary structure of the proteins [2]. Amino acid sequences of the coiled coil show the characteristic repeats of hydrophobic amino acids called the heptad repeat, in which the hydrophobic residues occur at every three and four residues [3]. Using the periodic occurrence of amino acid residues in their sequences, several software systems have already been developed for the prediction of coiled coil regions: COILS [4], PAIRCOIL [5], MULTICOIL [6], Learncoil [7], Paircoil2 [8] and Marcoil [9].

Although these prediction system are successful in predicting the coiled coil for typical heptad repeat, all prediction system has common weakness in the prediction of coiled coils around the breaks of the heptad repeats. There are four types of breaks of the heptad repeats [1] [10]: (1) The stutter is the break in which three residues are deleted from the heptad repeats [11]. The stutter has the local tendency to loosen the supercoiling. (2) The stammer is the break in which four residues are deleted [10]. The stammer has the tendency to tighten the supercoiling as the pitch of the helix locally decreased at this type of breaks. (3,4) There are two types of skips in which one residue is inserted or deleted, skip(+) and skip(-), respectively. [12]. The structure of a skip(+) type of the break seems well to form π -turn in the α -helix (ex. DNA polymerase I (Poll)), and skip(+) also tended to become a kink of the helix (ex. GreA).

The functions of heptad breaks are not well understood yet, but heptad breaks in the real proteins are highly conserved, suggesting their structural and functional importance [13]. It was also reported that skips in the coiled coil of myosin rod could have important effects on the way myosin molecules pack in the cross-section of the thick filaments [12]. In addition, Jerry H. Brown suggested that the local underwinding or overwinding caused by specific breaks in the heptad repeat has an effect on the global structure and can modify both the assembly of the protein and its interaction properties [13]. Therefore, heptad breaks are very important units in amino acid sequences and may provide us new insights into the complex relationship between amino acid sequence and the higher levels of structures.

One of the difficulties in predicting the coiled coil structures from amino acid sequences lies in the ambiguity of the register of the hydrophobic heptad repeat around the heptad breaks. It is easy to identify the typical heptad repeats and the heptad breaks can be easily observed when they are sandwiched by typical heptad repeat regions. However, the heptad repeats often become obscure around the heptad breaks, and the identification of the heptad break register of the hydrophobic residues around the heptad breaks is not easy. In fact, the coiled coil prediction system of COILS [4] and Paircoil2 [8] does not take the heptad breaks into consideration. Marcoil which takes the heptad breaks into account seems to be the best prediction system in its accuracy so far [9]. Therefore, the development of a method to determine the heptad break registers with high sensitivity than other methods is required for the improvement of the coiled coil prediction.

Previously, we developed a software system based on the physicochemical parameters of amino acids for the accurate prediction of membrane proteins [14]. The important concept in this method is that the stability of transmembrane regions is determined by the average hydrophobicity at the central region of membrane and the average amphiphilicity at the interface between the membrane interior and water. The similar concept can be applied to the prediction of coiled coils. In the case of the coiled coil proteins, hydrophobic residues occur at the core region, and the outfield region contain various polar residues. Therefore, the average values of physical properties of the core and

the outfield regions of coiled coil will be the parameters for the prediction of the coiled coil structure. In fact, many of the previously developed prediction systems for coiled coils use the information about not only the core but also the outfield region for the high performance prediction. However, those systems use only the probability of the occurrence of amino acids and neglect the fact that an amino acid has various kinds of physicochemical properties. Therefore, the use of the average values of various physiochemical properties as the discrimination parameters at the core and the outfield regions leads to completely novel method for the coiled coil prediction.

In the current work, we developed a novel method to determine the register of the hydrophobic residues at the heptad repeats and breaks. We also developed a method to discriminate the regions with the heptad repeat registers between the coiled coil and other structures by using the physicochemical parameters. Combining those methods, a novel prediction system SOSUicoil for the coiled coil was obtained in which the accuracy was better than other systems and the contribution of various physical properties to the coiled coil structure could be discussed in detail.

2. Material and method

2.1 Datasets for analysis

The datasets of amino acid sequences of proteins with and without coiled coil domains were prepared from the Uniprot and PDB-REPRDB [15], respectively. Table 1 shows the numbers of proteins with and without coiled coil domains used for developing the prediction system of coiled coils. We omitted redundant proteins with the homology above 30%. The seven residues on each side of coiled coil regions were eliminated from the data of coiled coil domains to refrain from the potential errors of non-coiled coil residues. Proteins without coiled coil domains were checked by SOCKET [2] to remove potential coiled coil proteins. The coiled coil regions from Uniprot have several ambiguous data which are not supported by crystallographic or NMR data. However, we used them as the training dataset for the prediction of coiled coil domain, since many data are necessary for the prediction by two stages: the determination of the register and the discrimination of coiled coil regions using the register. Furthermore, the influence of random errors in training data can be statistically masked by the averaging of many data.

Test dataset was obtained from SCOP database [16] (version 1.73) (all the proteins were listed in Table 2). The redundant data with homologies above 30% were omitted. The data used in training dataset were also removed from the test dataset. We compared the performance of the current method with other systems by using two kinds of definition of the coiled coil. One was the definition by SCOP and another was the definition by SOCKET. Since SOCKET does not predict skip regions in default setting, the option “-e 1” was applied (see documentation of SOCKET [2]). The two methods of the definition of the coiled coil have complementary advantages and disadvantages. The definition by SOCKET uses the exact distance between two helices and the coiled coils at the heptad breaks which show different distance between the interacting helices are sometimes judged not to be the coiled coil. On the other hand, the coiled coils in SCOP are judged by the observation of researchers and more coiled coil is given by SCOP than SOCKET. The disadvantage of SCOP is the arbitrariness of the definition of coiled coil. Therefore, we used two kinds of coiled coil regions for the evaluation of the prediction systems.

Table 1. Training dataset of coiled coil and non-coiled coil domains

	Number of proteins	Number of residues	Source of data
Coiled coil domains	1,350	36,639	Uniprot
Protein without coiled coil domain	1,011	285,269	PDB-REPRDB
(total)	2,361	321,908	

Table 2. PDB codes of 54 test dataset of coiled coil proteins

Each data were derived from Protein Data Bank.

1fif	2fxo	1hbw	1ha0	1jun	2ocy	1aa0
2np0	1d7m	1aq5	1mg1	1jad	2b9c	1ebo
1flk	1wpa	1ajy	1mv4	1qu7	1qu1	2e7s
2ezp	1ezj	1m7l	1vsg	1ik9	2d3e	1ox3
2cpb	1r48	1qey	2vsg	1av1	1wyy	1lj2
1fdm	1hf9	1dip	1g5g	1mlj	1joc	1m5i
2b5u	2a93	1zta	2ch7	1fu1	1env	1l8d
1c1g	1pjf	1cii	1qoy	1dkg		

2.2 Determination of registers of hydrophobic heptad repeats and breaks

The coiled coil structure is typically characterized by the hydrophobic heptad repeats, in which hydrophobic residues occur every seven residues. However, there are breaks of the heptad repeats in amino acid sequences even within the coiled coil structure, and the existence of the breaks of the heptad repeats decrease the accuracy of the coiled coil prediction. The heptad breaks are conserved in many coiled coils, suggesting the importance of such breaks. Therefore, we developed a novel method to determine the register of hydrophobic heptad repeats and their breaks for two purposes: the improvement of the accuracy of the coiled coil prediction and the study of the functional meaning of the heptad breaks.

Figure 1 shows the relationship between the amino acid sequences and the helical structure in the coiled coil. It is known that the pitch of the helix in the coiled coil is a little smaller than that of the typical α -helix, 3.6 residues. As shown in Figure 1A every seventh residues overlap in the coiled coil corresponding to the pitch of 3.5 residues. The reason for the heptad repeats of hydrophobic residues comes from this sequence-structure relationship. Figure 1A also shows that seven residues correspond to two turns, and the residue at the third position “d” from the first hydrophobic residue at “a” usually become hydrophobic.

Figure 1. Helical structure of the coiled coil

(A) Helical wheel diagram of the coiled coil based on the heptad register “a,b,c,d,e,f,g”. (B) We categorized the registers into three groups: the core, the medium and the outfield regions.

The physical properties of residues have to be closely related to the environment and we divided the seven positions into three groups: the core region (“a” and “d”), the outfield region (“b”, “c” and “f”) and the medium region (“e” and “g”). For determining the register of the hydrophobic heptad repeats, we have used only the hydrophobicity of the core region because it is the most important factor for the supercoiling. The evaluation function $S(i)$ for the typical heptad repeats was, therefore, represented by Eq. (1)

$$S(i) = H(i-14) + H(i-11) + H(i-7) + H(i-4) + H(i) + H(i+3) + H(i+7) + H(i+10) + H(i+14) \quad (1)$$

in which $H(i)$ represents hydropathy index by Kyte and Doolittle [17] at the sequence position i . This evaluation function is devised for determining the phase of the heptad repeat of hydrophobic residues, using the window of 29 residues from $(i-14)$ to $(i+14)$. When the center of the window is at the position of the core, “a” or “d”, in Figure 1A, the evaluation function shows a large positive value. Whereas, the evaluation function shows a large negative value when the center of the 29 residue window is positioned at the outfield region. The interval between the neighboring hydrophobic core residues is typically three (a - d) or four (d - a), and this fact was used for determining the register of the heptad repeats.

Then, the difference between the evaluation function $S(i)$ and the maximum value of $S(i)$ in the five residue window around i -th position was calculated by Eq. (2) for clarifying the position of the core region.

$$D(i) = S(i) - \text{Max}\{S(i+k) : (-2 \leq k \leq 2, k \neq 0)\} \quad (2)$$

Figure 2 shows the schematic plot of $D(i)$ as a function of the sequence number for explaining the algorithm of the determination of the heptad repeat register. When we take the highest and the next highest peaks of $D(i)$ in the window of 28 residues, the interval between two peaks is closely related to the phase of the heptad repeats. As shown in Figure 2A, the interval between two peaks is 3, 10, 17 or 24 ($7k-4$) residues, only when the peaks correspond to the “a” and “d” positions. Whereas, the interval of 4, 11, 18 or 25 ($7k-3$) residues is obtained only when the positions of the peaks are inversed (Figure 2B). However, when the interval is 7, 14, 21, or 28 ($7k$), both of the two peaks may be “a” or “d”, and the register cannot be determined uniquely. Then, we take the third highest peak and determine the register according to the same procedure. Other intervals are not

consistent with the heptad repeats and the evaluation functions for the heptad breaks have to be applied to those regions.

Figure 2. The schematic diagram of the processes for determining the heptad repeat register. The highest and the next highest peaks of $D(i)$ are selected in the window of 28 residues. (A) When intervals are 10 residues for example, the highest and the second peaks were assigned to “a” and “d”, respectively. (B) When intervals are 18 residues for example, the highest and the second peaks were assigned to “d” and “a”, respectively.

There are four kinds of heptad breaks: skip(+), skip(-), stutter and stammer. The heptad breaks of skip(+) and skip(-) are the insertion and deletion of a single residues in the region of “a”-“d” and “d”-“a”, respectively. The heptad breaks of stutter and stammer are the deletion of three residues of “a”, “b” and “c” (stutter) and the deletion of four residue of “d”, “e”, “f”, and “g” (stammer), respectively. There are eight positions for each kind of heptad breaks in the window of 29 residues leading to 32 evaluation functions of heptad breaks (four kinds of heptad breaks times eight positions) described as follows

$$\begin{aligned}
 S_{j,s}(i) = & H(i \mp 14 + \alpha_1(j,s)) + H(i \mp 11 + \alpha_2(j,s)) + H(i \mp 7 + \alpha_3(j,s)) + \\
 & H(i \mp 4 + \alpha_4(j,s)) + H(i) + H(i \pm 3 + \alpha_5(j,s)) + H(i \pm 7 + \alpha_6(j,s)) + \\
 & H(i \pm 10 + \alpha_7(j,s)) + H(i \pm 14 + \alpha_8(j,s)) \\
 & \dots\dots\dots (3)
 \end{aligned}$$

in which $H(i)$ indicated the hydropathy index at the i -th position of the sequence. The parameter j represents the number of heptad breaks variations; 1-8 correspond to skip, 9-12 correspond to stutter, 13-16 correspond to stammer, respectively. The upper signs in Eq.(3) correspond to $s=1$ and the lower signs correspond to $s=2$. The parameters from $\alpha_1(j,s)$ to $\alpha_8(j,s)$ represent the correction terms for shifted position because of heptad breaks. The each value of the correlation terms are shown in Table 3. The best register of the heptad break is determined by the highest $S_{j,s}(i)$ values template. After determining the heptad repeats and breaks, all the residues are assigned to one of three categories as shown in Figure 1: the core, the medium and the outfield regions for the

discrimination between the coiled coil and other structures.

Table 3. Modulation factor $\alpha_k(j,s)$ for 32 heptad break patterns.

Upper sign correspond to $s=1$. Lower values correspond to $s=2$.

		$\alpha_1(j,s)$	$\alpha_2(j,s)$	$\alpha_3(j,s)$	$\alpha_4(j,s)$	$\alpha_5(j,s)$	$\alpha_6(j,s)$	$\alpha_7(j,s)$	$\alpha_8(j,s)$
skip(+)	$j = 1$	∓ 1	0	0	0	0	0	0	0
	$j = 2$	∓ 1	∓ 1	∓ 1	0	0	0	0	0
	$j = 3$	0	0	0	0	± 1	± 1	± 1	± 1
	$j = 4$	0	0	0	0	0	0	± 1	± 1
skip(-)	$j = 5$	± 1	± 1	0	0	0	0	0	0
	$j = 6$	± 1	± 1	± 1	± 1	0	0	0	0
	$j = 7$	0	0	0	0	0	∓ 1	∓ 1	∓ 1
	$j = 8$	0	0	0	0	0	0	0	∓ 1
stutter	$j = 9$	∓ 1	0	0	0	0	0	0	0
	$j = 10$	∓ 1	0	∓ 1	0	0	0	0	0
	$j = 11$	0	0	0	0	± 1	0	± 1	0
	$j = 12$	0	0	0	0	0	0	± 1	0
stammer	$j = 13$	0	± 1	0	0	0	0	0	0
	$j = 14$	0	± 1	0	± 1	0	0	0	0
	$j = 15$	0	0	0	0	0	∓ 1	0	∓ 1
	$j = 16$	0	0	0	0	0	0	0	∓ 1

Table 4. Determination of the registers of the 1st and 2nd peaks of $D(i)$ from the intervals between the peaks

If intervals between the peaks of $D(i)$ are “3,10, 17, 24” or “4, 11, 18, 25”, we could determined the one register. In the case of “7, 14, 21, 28”, two patterns of the registers are allowed. In other cases, amino acid sequences are analyzed by the registers of heptad breaks.

Intervals	Heptad repeat register of 1 st and 2 nd peaks of $D(i)$	Next procedure
3, 10, 17, 24 $(7k - 4)^{*1}$	“a” - “d”	
4, 11, 18, 25 $(7k - 3)^{*1}$	“d” - “a”	
7, 14, 21, 28 $(7k)^{*1}$	“a” - “a” or “d” - “d”	Analysis using the next highest peak
1, 2, 5, 6, 8, 9, 12, 13, 15, 16, 19, 20, 22, 23, 26, 27 (otherwise)		Analysis by the heptad break registers.

^{*1} k is integer ($k \geq 1$)

2.3 Discrimination of coiled coil by Canonical Discriminant Analysis

We calculated the discrimination score for each 28 residues window between the coiled coil and other structures, using 10 physicochemical properties in the three regions of the core, the medium and the outfield. Namely, the discrimination analysis was carried out by 30 parameters which are the simple average of the physicochemical properties for the register in Eq. (4).

$$X_p^{(r)}(k) = \frac{1}{m^{(r)}(k)} \sum_{\substack{k \leq i \leq k+28, \\ R(i)=r}} x_p(i) \cdots \cdots \cdots (4)$$

in which the suffices, p and r , are the types of properties and the three region of the core, the medium and the outfield, respectively. The value $x_p(i)$ is the property p of i -th residue, the value $m^{(r)}(k)$ is the number of residues of region r from the k -th to $k+28$ -th residues, and the value $R(i)$ is the region of i -th residue.

The physicochemical properties used in the present analysis were selected by try-and-error approach from various physicochemical properties, including densities of amino acids. Finally, we selected 10 properties for the discrimination of the coiled coil domains: the hydropathy index [17], the total charge density, the amphiphilicity [18], and the density of aromatic residues, “Pro”, “Gly”, “Glu”, “Gln”, “Leu” and “Val”. The distribution of 10 properties in three registers of the core, the medium and the outfield were calculated, using 1,350 proteins with the coiled coil domain and 1,011 proteins without coiled coil, respectively. Figure 3 shows the averages of 10 properties at seven positions of the heptad repeats. The order of the seven positions is according to Figure 1B. The average values at the core region are very different from those at the outfield region. The averages at the medium region were similar to those in the outfield region, but slight difference was found for several properties. Since the values within the same region are very similar, we averaged those values for the discrimination analysis.

Figure 3. Distribution of the average values of 10 physicochemical properties in seven registers
 Profile of physicochemical properties: amphiphilicity(Amp), total charge density(Chg),
 hydrophathy(Hyd), aromatic(Aro) and amino acid composition, of positive and negative data were
 averaged and compared.

To discriminate the positive and negative data, a canonical discriminant analysis was carried out using 30 parameters. Among the eigenvectors of the correlation matrix of the 30 parameters, the vector provided the largest ratio between the two groups was chosen as the coefficients $A_p^{(r)}$ shown in Table 5. Finally, the discrimination score of k -th residues were calculating as following equation.

$$Z(k) = \max_{k-28 \leq j \leq k} \left[\sum_{r=1}^3 \sum_{p=1}^{10} A_p^{(r)} \cdot X_p^{(r)}(j) \right] \dots\dots\dots (5)$$

The threshold value of 1.6 was chosen for discriminating between two groups of learning data most clearly. The final column of Table 5 shows the contribution of ten properties to the discrimination between the coiled coil and the non-coiled coil which is the sum of the contribution of the three regions, $\sum_{r=1}^3 A_p^{(r)} \cdot \overline{X_p^{(r)}}$, in which $\overline{X_p^{(r)}}$ is the average of the parameter $X_p^{(r)}$.

Table 5. Coefficients and contributions of 10 parameters in the canonical discriminant analysis

10 physicochemical properties were sorted by the values of contribution. To compare the importance between the properties, we calculated contributions using following process. First, multiply average values by coefficient. Absolute of each resulted values were summarized and used as contribution.

	Coefficient $A_p^{(r)}$			Contribution
	core	medium	outfield	
Amphiphilicity index (Amp)	0.640	0.450	0.360	0.930
Number density of glutamic acid (Glu)	1.048	1.941	1.983	0.592
Charge density (Chg)	-1.040	-0.440	-0.390	0.500
Hydropathy index (Hyd)	0.100	-0.080	-0.160	0.400
Number density of leucine (Leu)	1.361	0.583	-0.751	0.380
Density of aromatic residues (Aro)	-0.676	-1.363	-2.057	0.362
Number density of glutamine (Gln)	0.117	1.005	1.164	0.186
Number density of glycine (Gly)	-1.178	-0.828	-1.172	0.139
Number density of valine (Val)	-0.365	-0.431	-0.785	0.074
Number density of proline (Pro)	-0.332	0.238	0.240	0.031

3. Result

3.1 Performance of the coiled coil prediction system

The accuracy of the prediction system SOSUIcoil was evaluated by the dataset from PDB. Table 6 shows the performance of the present predictor, comparing with other coiled coil predictors, Marcoil, Paircoil2 and COILS which are available through internet. Although the precision of our system is almost the same as other systems, there is significant difference between other systems and the present method in the values of the recall. One of the reasons for significant advantage of the large recall in our system is probably due to the accurate determination of the registers of the heptad repeats and breaks. In fact, the poor detection of coiled coil around irregular repeats is the main reason of low recall in the systems COILS and Paircoil2. COILS does not consider the register for the heptad breaks, and Paircoil2 remove the irregular region from the analyzing sequences. Therefore, the recall for those systems is much lower than the present method. In contrast to COILS and Paircoil2, Marcoil takes the probability for the heptad breaks into account and the recall is greatly improved. However, the present method which calculates all possible registers of heptad breaks showed larger recall than the Marcoil system.

Looking at the structure of the predicted coiled coil regions which were in fact the helical structures but not the coiled coil, we realized that many of them had partner domains with which the hydrophobic core was formed similar to the coiled coil. The prediction of those regions as the coiled coil is correct in the meaning that the complex structure is stabilized by the hydrophobic core due to the heptad repeats. When we counted those helical domains as the members of coiled coil, the precision became almost 100% and the recall by all predictors also increased, as shown in the parenthesis of Table 6. The precision and the recall by our predictor exceeded 90%.

One of the advantages of the present method is that the stabilizing mechanism of the coiled coil can be discussed by the contribution of the physicochemical parameters. The coefficients of 30 parameters in Eq. (5) to the discrimination scores are shown in Table 5. The largest contribution to the discrimination score were the amphiphilicity and the densities of glutamic residues. The molecular complexes in water are usually stabilized by amphiphilic groups and, therefore, the large contribution to the coiled coil discrimination indicates that the coiled coil structure is stabilized not only by the hydrophobic residues in the core region but also by the amphiphilic residues in the outfield region. The large contribution of glutamic residues is probably due to the amphiphilic nature of this residue [19]. The densities of leucine and aromatic residues showed also large contribution to the discrimination score. Leucine contributes to the leucine zipper, and the positive coefficient of the leucine density in the core region is reasonable, while the negative coefficient in the outfield region is because of the unfavorable hydrophobic interactions of leucine in water. The aromatic residues have the effect for decreasing the stability of coiled coils by the steric hindrance of large side chains on the unfavorable hydrophobic interaction.

Table 6. Performance of four coiled-coil predictors for the dataset of coiled coils longer than 28 residues.

Recall is defined by $TP/(TP+FN)$. Precision is defined by $TP/(TP+FP)$. F-measure is defined by $2Recall*Precision/(Recall + Precision)$

The data in parentheses are the result for the dataset of alpha helices longer than 14 residues .

		SOSUIcoil	Marcoil	Paircoil2	COILS
SOCKET	Recall	88.8% (90.6%)	84.4% (86.3%)	56.2% (60.0%)	69.8% (73.3%)
	Precision	83.3% (99.1%)	85.2% (100%)	85.7% (100%)	83.7% (100%)
	F-measure	86.0% (94.7%)	84.8% (92.6%)	67.9% (75.0%)	76.1% (84.6%)
SCOP	Recall	68.8%	64.4%	44.9%	50.0%
	Precision	96.9%	96.7%	96.9%	97.1%
	F-measure	80.5%	77.3%	61.4%	66.0%

3.2 Structural features around heptad breaks

We analyzed 54 proteins with coiled coil domain for the test of our predictor SOSUIcoil, and identified 43 heptad breaks: 17 skip(+), 3 skip(-), 12 stammer and 11 stutter. Table 7 shows the structural features of the segments around the heptad breaks and the results of the coiled coil prediction by the present system, Marcoil, Paircoil2 and COILS. It should be noted that the regions around the heptad breaks of stammer and stutter show mostly straight helices, whereas many skips show the breaks of the helical structure. Among the four types of heptad breaks, the stammer and the stutter does not show significant influence on the coiled coil structure. Since the predictors, COILS and Paircoil2, do not consider the continuity of the coiled coil structure at the heptad breaks, the performance of the predictor decrease by the neglect of the properties of the heptad breaks. It seems that the combination of the sensitive determination of heptad breaks and the discrimination by physicochemical parameters in this work well predicts the structure at the heptad breaks. The heptad breaks are not abundant in coiled coil domains, but the position of the heptad breaks are conserved, suggesting the functional importance [13]. Therefore, the correct identification of the heptad breaks and the correct structure prediction is important for understanding the behaviors of the proteins in relation to their functions.

Table 7. Form of protein around skip(+), skip(-), stammer and stutter

(A) Structural features of each heptad breaks were categorized into 3 structural features.

(B) The numbers of heptad breaks predicted as coiled coil by each software systems.

A				
	Structural features around heptad breaks			Total
	Straight	Bend	End	
Skip(+)	7	8	2	17
Skip(-)	2	0	1	3
Stammer	10	1	1	12
Stutter	8	2	1	11
Total	27	11	5	43

B					
	Total	Numbers of coiled coils around heptad breaks			
		SOSUIcoil	Marcoil	Paircoil2	COILS
Skip(+)	17	15	13	5	9
Skip(-)	3	2	2	1	2
Stammer	12	11	11	10	9
Stutter	11	8	8	5	8
Total	43	36	34	21	28
Ratio		88%	83%	51%	68%

Figure 4. The analysis of a coiled coil protein, transcript cleavage factor GreA

(A) The crystal structure transcript cleavage factor GreA (PDB:1GRJ) includes a skip(+) in the helix1. The structure were rendered with Molscript [20]

(B) The first α -helix which has the skip(+) type heptad break in the middle of the helix could not predicted except for SOSUIcoil. Furthermore, SOSUIcoil showed good correlation with the region which defined by SOCKET.

Figure 5. The analysis of a coiled coil protein, glycoprotein

(A) The crystal structure of glycoprotein (PDB: 1VSG). There is a skip(+) at middle region of the second helix.

(B) The first helix could be predicted as coiled coil only by SOSUIcoil. The coiled coil regions predicted by SOSUIcoil well agreed with the regions defined by SOCKET.

Figure 4 and Figure 5 shows the example of prediction results of coiled coil between our system SOSUIcoil and other three prediction system: COILS, Paircoil2 and Marcoil. The coiled coil regions from PDB were identified by a software system SOCKET, which is well known as a coiled coil identification system [16]. The crystal structure of the 158-residue protein GreA [21](PDB:1GRJ) has antiparallel α -helical coiled-coil dimer in their N-terminal domain as shown in Figure 4A. The helix 1(1A and 1B) has a bulge and a proline-induced kink leading to an unusual coiled-coil structure. The characteristic heptads repeat was maintained throughout helix 1A from residues 7 to 23. However, Val 24 introduce the one-amino-acid frameshift (= "Skip(+)"). Except Paircoil2, the helix 2 was predicted by all programs. The significance of the present system were manifested by the correct prediction of the helix 1A and 1B. Another example of prediction of coiled coil is shown in Figure 5. N-terminus 364-residues of the variant surface glycoprotein [22] (PDB:1VSG) showed antiparallel α -helical coiled coil as detected by SOCKET. This structure has a kink at the skip of Thr86, but helix 2 and 3 could be predicted by COILS, Marcoil and our system. However, helix 1 was predicted only by only SOSUIcoil.

4. Discussion

The coiled coil is one of the simplest structure units in proteins, and the sequence motif of the heptad repeat of hydrophobic residues is clear. Therefore, several prediction systems have been developed so far on the basis of the heptad repeats. However, there are the inconsistencies of the heptad repeats, called heptad breaks, and the performance of the prediction of the coiled coil reduced by the existence of the heptad breaks. Therefore, we calculated the evaluation function which has various patterns of heptad breaks including all reported heptad breaks (skip(+), stammer, stutter), and also have skip(-) type. Because of this wider coverage of patterns it could also predict some new types of coiled coil. For example, it is reported that some coiled coil composed of regular noncanonical repeats like Hendecad repeats (11-7-11 repeats; β -giardin, 7-11-7 repeats; HPSR2). SOSUIcoil also covers these types of coiled coils. This feature is one of the reasons why the performance of the present method is better than the other methods.

We could identify 43 heptad breaks in 52 coiled coil proteins. When the structure around the heptad breaks was inspected, more than half of the heptad breaks were structurally silent in the meaning that the coiled coil structure was unchanged by the insertion of the heptad break. Furthermore, even if a helix has the kink of the structure, the coiled coil may continue, as seen from the example of Figure 4. The distribution of four types of heptad breaks in the coiled coil has not systematically studied in the previous methods, and the present work is the first extensive analysis of the relationship between the position of the heptad breaks in amino acid sequences and their three dimensional structures.

In the discrimination of coiled coil, we divided the coiled coil structure into three regions: core, medium and outfield. Then, we calculated the average values of several physicochemical properties. The characteristic profiles of physicochemical properties and the success of the prediction indicated that the coiled coil structure is stabilized by the hydrophobic residues in the core regions and the amphiphilic residues in the outfield regions. This fact is probably the reason why the segments around a heptad break can be correctly predicted as the coiled coil. The amphiphilic side chains sticking out to water stabilize the helical structure as discussed previously [23]. The contribution of the amphiphilicity index and the density of glutamic residues to the discrimination score were large, confirming the concept of the stabilization of helices by the amphiphilic residues.

The high performance of the present predictor SOSUIcoil shows that the approach of the 'physical finger print' method was powerful for discriminating the coiled coil from other types of

structures.

It is considered such region play an important role in cellular organisms. For example, it has been pointed that there are structural change in the tail region of myosin molecules depending on the change in the environment. It is well known that particular positions of coiled coil in the tail of myosin are the hot spots of the break coupled with several steps of the motor functions. Therefore, the prediction of the heptad breaks has to be studied in view of the fragile points of the coiled coils. The analyses heptad breaks of large scale data from genome information and GenBank may provide us the useful information which will be reported elsewhere.

The software system SOSUIcoil will be soon available at our homepage:

<http://bp.nuap.nagoya-u.ac.jp/sosui/>

This work was supported by a Grant-in-Aid for the 21st Century COE "Frontiers of Computational Science" from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

- [1] A. Lupas, Coiled coils: new structures and new functions., *Trends Biochem. Sci.*, **21**, 375-382(1996).
- [2] J. Walshaw, Socket: a program for identifying and analysing coiled-coil motifs within protein structures., *J. Mol. Biol.*, **307**, 1427-1450(2001).
- [3] D. A. Parry, Coiled-coils in alpha-helix-containing proteins: analysis of the residue types within the heptad repeat and the use of these data in the prediction of coiled-coils in other proteins., *Biosci. Rep.*, **2**, 1017-1024(1982).
- [4] A. Lupas, Predicting coiled coils from protein sequences., *Science*, **252**, 1162-1164(1991).
- [5] B. Berger, Predicting coiled coils by use of pairwise residue correlations., *Proc. Natl. Acad. Sci. U S A*, **92**, 8259-8263(1995).
- [6] E. Wolf, MultiCoil: a program for predicting two- and three-stranded coiled coils., *Protein Sci.*, **6**, 1179-1189(1997).
- [7] M. Singh, LearnCoil-VMF: computational evidence for coiled-coil-like motifs in many viral membrane-fusion proteins., *J. Mol. Biol.*, **290**, 1031-1041(1999).
- [8] A. V. McDonnell, Paircoil2: improved prediction of coiled coils from sequence., *Bioinformatics*, **22**, 356-358(2006).
- [9] M. Delorenzi, An HMM model for coiled-coil domains and a comparison with PSSM-based predictions., *Bioinformatics*, **18**, 617-625(2002).
- [10] M. Gruber, Historical review: another 50th anniversary--new periodicities in coiled coils., *Trends. Biochem. Sci.*, **28**, 679-685(2003).
- [11] D. A. Parby, Fibrinogen: a preliminary analysis of the amino acid sequences of the portion of the alpha, beta and gamma-chains postulated to form the interdomainal link between globular regions of the molecule., *J. Mol. Biol.*, **120**, 545-551(1978).
- [12] A. D. McLachlan, Periodic features in the amino acid sequence of nematode myosin rod., *J. Mol. Biol.*, **164**, 605-626(1983).
- [13] J. H. Brown, Heptad breaks in alpha-helical coiled coils: stutters and stammers., *Proteins*, **26**,

134-145(1996).

- [14] T. Hirokawa, SOSUI: classification and secondary structure prediction system for membrane proteins., *Bioinformatics*, **14**, 378-379(1998).
- [15] T. Noguchi, PDB-REPRDB: a database of representative protein chains in PDB (Protein Data Bank)., *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **5**, 214-217(1997).
- [16] A. G. Murzin, SCOP: a structural classification of proteins database for the investigation of sequences and structures., *J.Mol. Biol.*, **247**, 536-540(1995).
- [17] J. Kyte, A simple method for displaying the hydropathic character of a protein., *J.Mol. Biol.*, **157**, 105-132(1982).
- [18] S. Mitaku, Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane-water interfaces., *Bioinformatics*, **18**, 608-616(2002).
- [19] K. Imai, Mechanisms of secondary structure breakers in soluble proteins, *Bioinformatics*, **1**, 55-65(2005).
- [20] P. J. Kraulis, MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures, *Journal of Applied Crystallography*, **24**, 946-950(1991).
- [21] C. E. Stebbins, Crystal structure of the GreA transcript cleavage factor from Escherichia coli., *Nature*, **373**, 636-640(1995).
- [22] D. Freymann, 2.9 Å resolution structure of the N-terminal domain of a variant surface glycoprotein from Trypanosoma brucei., *J. Mol. Biol.*, **216**, 141-160(1990).
- [23] N. Uchikoga, Electric charge balance mechanism of extended soluble proteins., *Protein Sci.*, **14**, 74-80(2005).