Advance Publication (e190002)
doi: 10.2142/biophysico.bppb-v19.0002

Biophysics and Physicobiology
https://www.jstage.jst.go.jp/browse/biophysico/

Regular Article

Information quantity for secondary structure propensities of protein subsequences in the Protein Data Bank

Ryohei Kondo[1], Kota Kasahara[2], Takuya Takahashi[2]

[1] *Graduate School of Life Sciences, Ritsumeikan University, Kusatsu, Shiga 525-8577, Japan*
[2] *College of Life Sciences, Ritsumeikan University, Kusatsu, Shiga 525-8577, Japan*

Received December 10, 2021; Accepted February 2, 2022;
Released online in J-STAGE as advance publication February 8, 2022
Edited by Haruki Nakamura

Elucidating the principles of sequence–structure relationships of proteins is a long-standing issue in biology. The nature of a short segment of a protein is determined by both the subsequence of the segment itself and its environment. For example, a type of subsequence, the so-called chameleon sequences, can form different secondary structures depending on its environments. Chameleon sequences are considered to have a weak tendency to form a specific structure. Although many chameleon sequences have been identified, they are only a small part of all possible subsequences in the proteome. The strength of the tendency to take a specific structure for each subsequence has not been fully quantified. In this study, we comprehensively analyzed subsequences consisting of four to nine amino acid residues, or *N*-gram ($4 \le N \le 9$), observed in non-redundant sequences in the Protein Data Bank (PDB). Tendencies to form a specific structure in terms of the secondary structure and accessible surface area are quantified as information quantities for each *N-gram*. Although the majority of observed subsequences have low information quantity due to lack of samples in the current PDB, thousands of *N*-grams with strong tendencies, including known structural motifs, were found. In addition, machine learning partially predicted the tendency of unknown *N*-grams, and thus, this technique helps to extract knowledge from the limited number of samples in the PDB.

Key words: protein structure, structural bioinformatics, *N*-gram, protein folding

Significance

Although recent great successes of the protein structure predictions, the molecular principles of sequence–structure relationships are not fully understood. In this study, we comprehensively measured the tendency to form a specific structure for subsequences, or *N*-grams, observed in the Protein Data Bank. Because the current dataset is too sparse to cover the 20^N variation of subsequences, we applied information quantity as a measure of the tendency. Higher information quantities indicate the stronger tendency observed in more diverse protein families. As a result, we discovered several thousand subsequences that almost always form a specific secondary structure regardless of its surrounding.

Introduction

How a protein sequence determines its structure and function is a key question in biology [1-3]. To elucidate the principles of sequence–structure relationships, biologists have tackled discovering rules from the enormous amount of

Corresponding author: Kota Kasahara, College of Life Sciences, Ritsumeikan University, 1-1-1 Noji-higashi, Kusatsu, Shiga 525-8577, Japan. ORCID iD: https://orcid.org/0000-0003-0207-6271, e-mail: ktkshr@fc.ritsumei.ac.jp

structural data of proteins deposited in the Protein Data Bank (PDB) [4-7]. Statistical studies on the PDB have achieved success for a wide range of prediction tasks from protein sequences, such as predictions of secondary structures [8], solvent-accessible residues [9,10] and intrinsic disorder [11,12]. In the tasks to predict structural features for a short segment of protein sequence or subsequence, three levels of information hidden in the sequence are used. The first is the propensity of amino acids composing the target segments [13]. The second is context-dependent information, which means that information from the regions other than the target site [14,15]. The interactions with the other regions separated along the sequence from the target segment, including non-local interactions, create challenges for predictions [16-18]. The third is evolutionary information, such as the sequence profile generated from a multiple sequence alignment [19,20]. State-of-the-art prediction methods based on the sequence profile drastically improve the predictions [8]. Toward understanding the principles of sequence–structure relationships, the distinction of the contributions of the information to determine the structures is of importance. At the standpoint of Anfinsen's dogma, the first two levels of information are enough to predict the structures. Trials to maximize the prediction performances without evolutionary information provide indispensable insights [21]. In addition, differences in roles of the first and second types of the information should be dissected. Although the first type of information, i.e., a short segment of the amino acid sequence, has been studied for a long time, the principles have not been fully clarified. When we consider 20 types of standard amino acids, subsequences consisting of N residues have 20^N variations of the sequence. Different sequences should exhibit different physical properties. Characterizing the universe of N-residue subsequences, or N-gram, is the key to understanding the principles of sequence–structure relationships [22].

A prominent example of well-characterized classes of N-gram sequences is chameleon sequences, which are subsequences with the capability to form distinct secondary structures depending on the environment [23,24]. Chameleon sequences have a weak tendency to form a specific structure, and their structures are determined by their surroundings rather than by their sequence. On the contrary, non-chameleon sequences have a strong tendency to form specific structures. Previous studies have successfully identified many chameleon sequences by taking advantage of the wealth of data in the PDB based on the binary classification of subsequences into chameleon. However, the strength of the tendency to form a specific structure in each subsequence can be considered as a spectrum in principle. Classifying the subsequences into chameleon may overlook the quantitative features of the sequence–structure relationships.

In this study, we quantified the strength of the tendency to form a specific structure of each subsequence in terms of the information quantity calculated from the statistics on the PDB. The information quantity is the negative logarithm of the expected probability for a probability distribution. When segments with the same subsequence more frequently form the same secondary structure than expected, this subsequence is considered to have a high information quantity. Alternatively, when segments with the same subsequences take various secondary structures by chance, this subsequence has no information about its secondary structure.

In addition to the secondary structures, we also assessed the relative accessible surface area (rASA) of the residues. We aimed to characterize the N-grams observed in the PDB in terms of information quantity for their structures encoded in each N-gram sequence. We collected N-grams ($4 \leq N \leq 9$) from the PDB and evaluated the propensity for secondary structure (helical, beta, and coil structures) and rASA. We found thousands of N-grams with strong tendencies for these structural features. They are potential candidates for structural motifs of proteins. We also found that a machine learning technique can be partially applied to estimate the tendency of unknown N-grams.

Materials and Methods

Dataset Construction

The method overview is shown in Figure 1. We analyzed the same dataset as in our previous study [25]. This dataset is a subset of the PDB [26] snapshot in 2017 filtered by the following criteria: i) X-ray structures with a resolution better than or equal to 3.0 Å, ii) the number of atoms is less than one million, and iii) the redundancy in the dataset is eliminated by single-linkage clustering based on the sequence identity ≤ 40 % using CD-HIT software [27]. The data from the PDB were processed using in-house scripts via PDBML [28]. The entry consisting of the highest number of N-gram segments in each cluster was chosen as a representative. The dataset is shown in Supplementary Data S1.

N-gram Analyses

In this article, we define the two terms, the N-gram segment and N-gram sequence, as follows. The former indicates a region consisting of N-successive residues in a polypeptide chain contained in a PDB entry. The latter indicates the amino acid sequence of N-successive residues. An N-gram sequence is a property of each N-gram segment. N-gram segments with the same N-gram sequence were collected from the dataset to characterize an N-gram sequence, and their statistics were assessed.

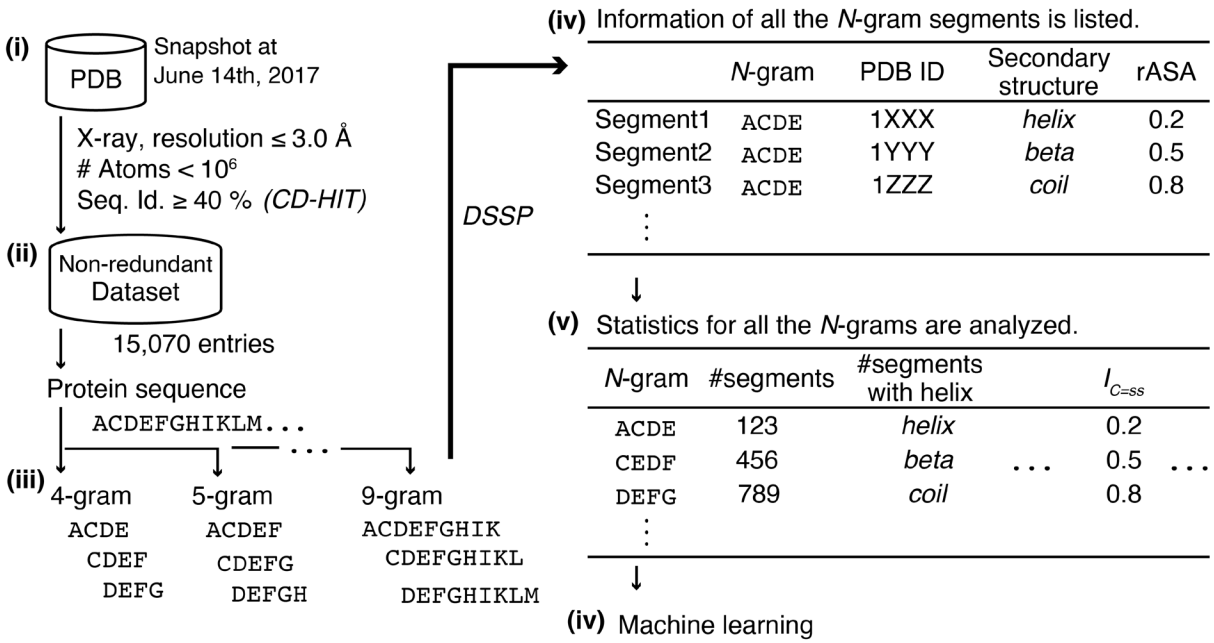

Figure 1 Method overview. The dataset was constructed by extracting entries meeting the criteria from the Protein Data Bank (PDB) with the elimination of redundancy. For each N-gram segment with $4 \leq N \leq 9$ included in the dataset, five properties were assessed including the secondary structure and rASA. Information quantity to predict these properties from the sequence was evaluated in terms of the negative log value of the expected probability to observe the samples in the dataset. In addition, the predictability of the properties was also evaluated using a machine learning technique.

For all the polypeptide chains in the dataset, they were decomposed into N-gram segments by sliding the window of N-successive residues along the sequence. The N value ranged from four to nine. The secondary structure and rASA of each N-gram segment were assigned based on the properties of the residue at the middle of the segment; when N is an even number, the (N/2)-th residue was chosen as a representative. We assessed the following properties of each segment: (i) the secondary structure, (ii) whether highly exposed to solvent or not based on the threshold of rASA ≥ 0.8, and (iii) whether exposed to solvent or not based on the threshold of rASA ≥ 0.2. For property (i), the secondary structure was defined by the DSSP software [29]. The eight classes of secondary structures produced by DSSP were reduced into three classes: *helix* ('H', 'G', and 'I' in the DSSP output), *beta* ('E' and 'S'), and *coil* (other symbols). For properties (ii) and (iii), the rASA was defined as the ASA value of the residue divided by the reference ASA value, which is defined by the ASA of the center residue in the Gly-X-Gly tripeptide.

In this article, the symbols signifying these properties of N-grams are introduced. The three properties are signified as the symbol p_C,
where

$$C \in \{SS, rASA08, rASA02\},$$

and

$$p_{SS} \in \{helix, beta, coil\}$$

$$p_{rASA08} \in \{highly\_exposed, not\_highly\_exposed\}$$

$$p_{rASA02} \in \{exposed,\ not\_exposed\}.$$

Information Quantity of an *N*-gram Sequence

If many segments with the same N-gram sequence always exhibit the same properties, this N-gram sequence is considered to have a strong tendency to that property. Conversely, if the same N-gram sequence can yield diverse properties, such as the chameleon sequences, this N-gram sequence is considered to have a weak tendency. We quantified the strength of the tendency of each N-gram sequence in terms of the expected probability of obtaining an observed

distribution in samples. The expected probability for the *N*-gram sequence *s* for the property *C* is defined as

$$P_C(s) = \frac{n(s)!}{\prod_{k \in p_C} n_k(s)!} \prod_{k \in p_C} f_k(N)^{n_k(s)}, \quad (1)$$

where $n(s)$ signifies the number of *N*-gram segments with the sequence *s* observed in the dataset. $n_k(s)$ is the number of *N*-gram segments with the *N*-gram sequence *s* and the property *k*. $f_k(N)$ denotes the relative frequency of the *N*-gram segments with the property *k* for all sequences with length *N*:

$$f_k(N) = \frac{\sum_{s \in seq(N)} n_k(s)}{\sum_{s \in seq(N)} n(s)}, \quad (2)$$

where $seq(N)$ is a set of all the subsequences with length *N* in the dataset n_{seq}. The quantity $P_C(s)$ indicates the expected probability for the event that $n(s)$ observations of the subsequence *s* in the dataset have the distribution $n_k(s)$ for the property C. The expectation assumed here is as follows: the property k of each observation is determined by a random sampling obeying the distribution $f_k(N)$. When we consider the three-class property, i.e., *C* = *SS*, the distribution $n_k(s)$ have $(3 + n(s) - 1)!/(n(s)!\,(3-1)!$ possibilities as a result of combination with repetition of $n(s)$ samples from three classes. An example for the case of $n(s) = 3$ is shown in Table S1. The expected probability for one of the possibilities of distribution, $P_C(s)$, can be calculated by multiplication of the expected probability of each sample and the number of permutation. A lower expected probability $P_C(s)$ indicates that the *N*-gram sequence *s* is strongly biased toward a specific property in terms of the property *C*. In other words, the *N*-gram sequence *s* with a low $P_C(s)$ value has a high information quantity for its structure. The information quantity is defined as:

$$I_C(s) = -\log P_C(s). \quad (3)$$

Preference to take a specific property *k* by a sequence *s* is assessed by the ratio of relative frequency,

$$R_k(s) = \frac{f_k(s)}{f_k(N)}, \quad (4)$$

Where

$$f_k(s) = \frac{n_k(s)}{n(s)}. \quad (5)$$

The amino acid propensities for a subset of *N*-grams were assessed in terms of the log odds ratio,

$$p_A(S) = \log\left(\frac{f_A^{res}(S)\left(1-f_A^{res}\right)}{f_A^{res}\left(1-f_A^{res}(S)\right)}\right), \quad (6)$$

where f_A^{res} denotes the relative frequency of amino acid *A* (one of the 20 standard amino acids) in the entire dataset, and $f_A^{res}(S)$ is that in the subset *S*.

Cluster Analysis

We filtered enriched *N*-grams for each property based on the criterion $I_C(s) \geq 50$ and $R_k(s) \geq 2.0$. These enriched *N*-grams were analyzed with a hierarchical clustering method based on Ward's method [30]. The distance between *N*-grams was calculated as the Levenshtein distance. We focused on clusters obtained by cutting the dendrograms at arbitrarily determined levels and sequence patterns for each cluster was analyzed using WebLogo [31].

Machine Learning

We examined the predictability of *N*-gram features from their sequences alone by using a machine learning technique, that is, an artificial neural network. The input layer was defined as a 20 × *N*-dimensional binary vector encoding the sequence of the *N*-gram. Each bit corresponds to one of the 20 standard amino acids at the *i*-th residue, where $i \in \{1,2,\ldots,N\}$. Three predictors were constructed for each *N* in the range of $4 \leq N \leq 9$ as follows: i) a three-class classifier for the secondary structure (helix, beta, or coil), ii) a binary classifier for *rASA08* (*highly_exposed* or *not_highly_exposed*), and iii) a binary classifier for *rASA02* (*exposed* or *not_exposed*). The dataset was divided into six subsets by picking PDB entries at random (Supplementary Table S1). The number of samples in each class was balanced. One of the subsets was

further divided into training and test sets for tuning the hyperparameters, including the number of hidden layers, number of nodes in each layer, loss function, and optimizer. The tuning was performed by using the Bayes optimization powered by the *hyperas* library. After that, five-fold cross-validation was performed using the remaining five subsets. The machine learning tasks were performed using *TensorFlow* 1.12.0 with *Keras* 2.2.4 libraries.

Results

Statistics of *N*-gram Segments in the Dataset

We analyzed -gram sequences in the dataset in the range of $4 \leq N \leq 9$. Our dataset consisted of 15,070 non-redundant PDB entries, including 50,231 chains and 11,994,671 residues for 17,367 proteins. For tetra-grams, 156,603 sequences of 20^4 = 160,000 possible sequences were observed in the dataset (Table 1). This means that the current PDB covers nearly 97.9 % of the tetra-gram space. With an increase in *N*, the volume of *N*-gram space was exponentially widened (20^N), and the coverage of the *N*-gram space by observation in the PDB decreased. The diversity of observed *N*-gram sequences was saturated at $N \geq 6$ (Table 1).

Table 1 Statistics of the *N*-gram dataset

N	4	5	6	7	8	9
n_{seq}	156,603	1,615,093	3,518,164	3,840,949	3,836,531	3,803,384
n	11,605,625	11,481,522	11,359,564	11,239,836	11,122,244	11,006,754
f_{helix}	0.4933	0.4950	0.4954	0.4965	0.4966	0.49741704
f_{beta}	0.2305	0.2310	0.2310	0.2309	0.2307	0.230135606
f_{coil}	0.2762	0.2740	0.2737	0.2726	0.2727	0.272447354
$f_{highly\_exposed}$	0.03965	0.03920	0.03896	0.03880	0.03874	0.038693606
$f_{exposed}$	0.4651	0.4631	0.4620	0.4610	0.4603	0.459648867

In the dataset, the majority of sequences have only a few samples of segments. The histogram shows a logarithmic decay of the number of *N*-gram sequences along with an increase in $n(s)$ the number of segments for each sequence (Supplementary Figure S1). The *N*-gram sequences most frequently observed in the dataset for each *N* are summarized in Table 2. The *N*-gram groups are marked as *, †, ‡, §, and ¶ in the table. All the top-10 octa- and nona-grams and some hepta-grams originated from ubiquitin (group ¶). All the other frequently observed octa- and nona-grams originated from a kind of redundancy not removed in our protocol, i.e., redundancy in each PDB entry. For example, subsequences in alpha-hemolysin and ferritin are repeatedly appeared in the dataset (Supplementary Data S2). The group † with poly-H was from the His-tag sequence (PDB ID: 2W5A; Supplementary Figure S2A). Group * indicates Ala-based subsequences. Many of them were from unidentified segments that were artificially assigned as poly-Ala (e.g., PDB ID: 3RFR, 3HDI, 4UYZ, 1H54, and 1GKU; Supplementary Figure S2B). An exception is the molybdenum storage protein subunit beta (PDB ID: 2OGX; Figure 2A). The subsequences in the group ‡ including "*ENLYFQG*" were a part of the recognition sequence for TEV protease. Subsequences in the group § including "*DVLVNNA*" were related to oxidation/reduction enzymes (e.g., PDB ID: 4EGF; Figure 2B). The subsequence "*EELKK*" was the motif named s-helix reported by Stefan et al. [32] (PDB ID: 2OSO; Figure 2C).

Information Quantity of *N*-grams

We analyzed the information quantity of each *N*-gram sequence in terms of $I_C(s)$ defined in Eq. (3). The distributions of $I_C(s)$ for each property *C* and length *N* are shown in Figure 3. In general, the histograms showed a decrease in the frequency along with an increase in, $I_C(s)$ suggesting that only a small part of the *N*-gram sequences had high information quantity. In addition, shorter subsequences had richer information, especially tetra-grams that clearly had higher $I_C(s)$ values than longer ones. Alternatively, distributions with $N \geq 6$ showed similar distributions. In general, because shorter subsequences have a higher number of samples (segments with the same subsequence) in the dataset, they can yield higher values of $I_C(s)$ than longer subsequences. Conversely, longer sequences can encode richer information about structural properties than shorter sequences. The distributions shown in Figure 3 reflect these two effects. Enrichment of information about the secondary structure and rASA in tetra-grams implies that these properties can be encoded in four successive amino acids in many cases, and the abundance of samples is more beneficial than considering longer segments of sequences.

Table 2 The most frequently observed *N*-grams in the dataset

$N=4$		$N=5$		$N=6$		$N=7$		$N=8$		$N=9$	
s	$n(s)$	s	$n(s)$	s	$n(s)$	s	$n(s)$	s	$n(s)$	s	$n(s)$
AALA*	1,604	HHHHH†	636	HHHHHH†	263	ENLYFQG‡	133	AGKQLEDG¶	71	AGKLEDGR¶	71
AAAA*	1,545	AAAAA*	365	ENLYFQ‡	253	LEHHHHH†	105	AKIQDKEG¶	71	AKIQDKEGI¶	71
LAAA*	1,416	NLYFQ‡	364	NLYFQG‡	200	AAAAAAA*	81	DGRTLSDY¶	71	DGRTLSDYN¶	71
ALAA*	1,413	ENLYF‡	272	LVNNAG§	153	VKTLTGK¶	80	DKEGIPPD¶	71	DKEGIPPDQ¶	71
LAAL*	1,257	AALAA*	259	LEHHHH†	152	ENLYFQS‡	76	DQQRLIFA¶	71	DQQRLIFAG¶	71
AAAL*	1,209	LYFQG‡	259	AAAAAA*	123	DVLVNNA§	72	DTIENVKA¶	71	DTIENVKAK¶	71
HHHH†	1,154	AAVAA*	253	NLYFQS	113	AGKQLED¶	71	DYNIQKES¶	71	DYNIQKEST¶	71
LAEA	1,118	AAALA*	228	EHHHHH†	107	AKIQDKE¶	71	EDGRTLSD¶	71	EDGRTLSDY¶	71
LEAL	1,116	LEHHH†	227	TAMIAG	94	DGRTLSD¶	71	EGIPPDQQ¶	71	EGIPPDQQR¶	71
AVAA	1,088	EELKK	210	VLVNNA§	86	DKEGIPP¶	71	ENVKAKIQ¶	71	ENVKAKIQD¶	71

The symbols *, †, ‡, and § denote groups of similar *N*-gram sequences. The *N*-grams with the symbol ¶ are subsequences from the ubiquitin.

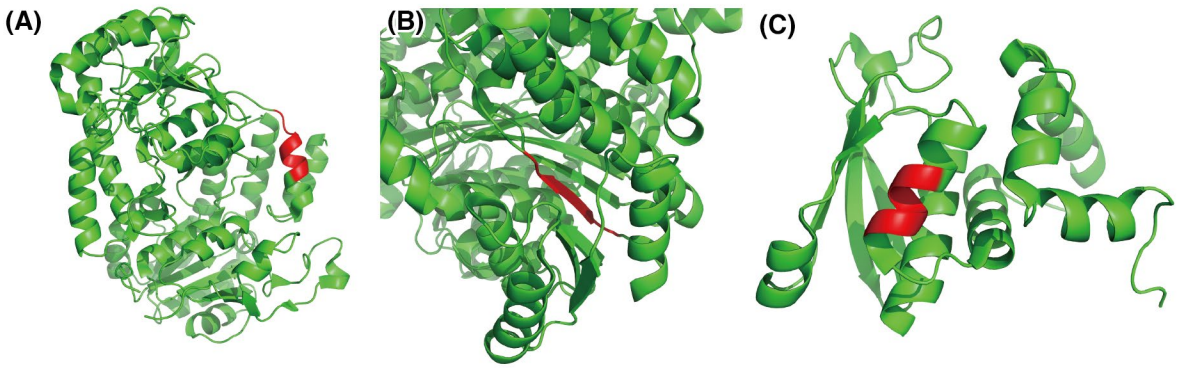

Figure 2 Examples of frequently observed *N*-grams. The focused segments are marked in red. (A) The poly Ala segment in the molybdenum storage protein (PDB ID: 2OGX). (B) The "*DVLVNNA*" segment in an L-xylulose reductase (PDB ID: 4EGF). (C) The "*EELKK*" segment in a member of the vinyl-r-reductase family of proteins (PDB ID: 2OSO).

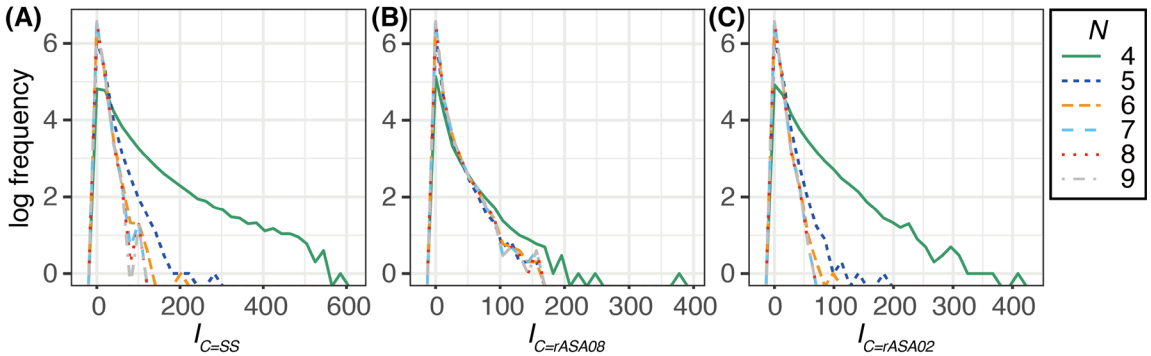

Figure 3 Distributions of $I_C(s)$ for the properties (A) the secondary structure, (B) *rASA08*, and (C) *rASA02*. Green, blue, orange, cyan, red, and gray lines indicate $N = 4$, 5, 6, 7, 8, and 9, respectively.

The property *rASA08,* indicating whether the rASA of the segment ≥ 0.8 or not, showed different behavior from the *rASA02*, which means rASA ≥ 0.2. Whereas $I_C(s)$ of the tetra-grams were still higher than the longer *N*-grams for *rASA08*, the difference between them was smaller than that for *rASA02*. This implies that the property *rASA08* is encoded to longer segments compared to *rASA02*.

Structural Features of the Subsequences

Many subsequences with a strong tendency to form a specific secondary structure were found in the dataset. To filter them, we applied the criteria $I_C(s) \geq 50$, and the probability of forming the secondary structure by the subsequence was two-fold higher than the random ($R_k(s) \geq 2.0$, see Eq. [4]). We found 2,205, 3,919, and 4,505 tetra-grams enriched to form helical, beta, and coil structures, respectively（Supplementary Data S2). The top-5 tetra-grams for each properties are shown in Table 3. For example, the subsequence "*ELAK*" with $I_{C=SS}(s = "ELAK") = 586.3$ had 880 observed segments in the dataset, and 94.5 % of them formed a helix (Supplementary Figures S3A, B, and C). The subsequence "*ELAR*" also yielded a strong tendency to form a helix (Supplementary Figures S3D, E, and F). A cluster analysis shows that motifs constituting an amphiphilic helix, "*EXXK*" and "*EXXR*", were enriched (Supplementary Figure S4). The negatively charged residues often occur at the N-terminal end of the helix due to the macro-dipole effect of the helix [33]. Overall, the tetra-grams with a strong helical tendency favored Ala, Glu, and Leu (Figure 4A). At near the threshold, 93.0 % of 71 observed segments of "*FAQR*" formed a helix and $I_{C=SS}(s = "FAQR") = 50.0$. For the beta structures, Val, Leu, and Ile residues were highly enriched for the tetra-grams with a strong tendency for the beta structures (Figure 4C and Supplementary Figure S5). The tetra-gram with the highest information quantity was $I_{C=SS}(s = "VLVV") = 538.4$ (Supplementary Figures S3G, H, and I). In the case of the coil structures, enrichment of Pro and Gly residues were confirmed (Figure 4E and Supplementary Figure S6). The subsequences with high information quantities, for example $I_{C=SS}(s = "AGAD") = 425.8$ and $I_{C=SS}(s = "LPEG") = 393.9$, appeared at short loops linking two adjacent secondary structural elements (Figures 5A and B). Although these subsequences showed a strong tendency, they did not guarantee the formation of the coil structure. As an example, whereas the representative residue in the segment in PDB ID: 2PF6 met the criteria for the helix (Figure 5C), this short helix appeared distorted, and it was in a loop region. Another example (PDB ID: 3PWQ; Figure 5D) had the "*LPEG*" segment at the kink point in a long helix. If the subsequence with coil tendency appeared within a regular secondary structural element, it might have a distorted conformation.

On the other hand, the majority of *N*-grams had low information quantities; 145,974 tetra-grams (94.5% of all observed tetra-grams) did not meet the criteria ($I_C(s) \geq 50$ and $R_k(s) \geq 2.0$) for secondary structures. There are two factors for a decrease in the information quantity, i.e., lack of samples and weakness of tendency to yield a specific property. To directly assess the latter factor for each *N*-gram sequence, an error of relative frequencies of the secondary structures from the expected values was assessed for each *N*-gram sequence:

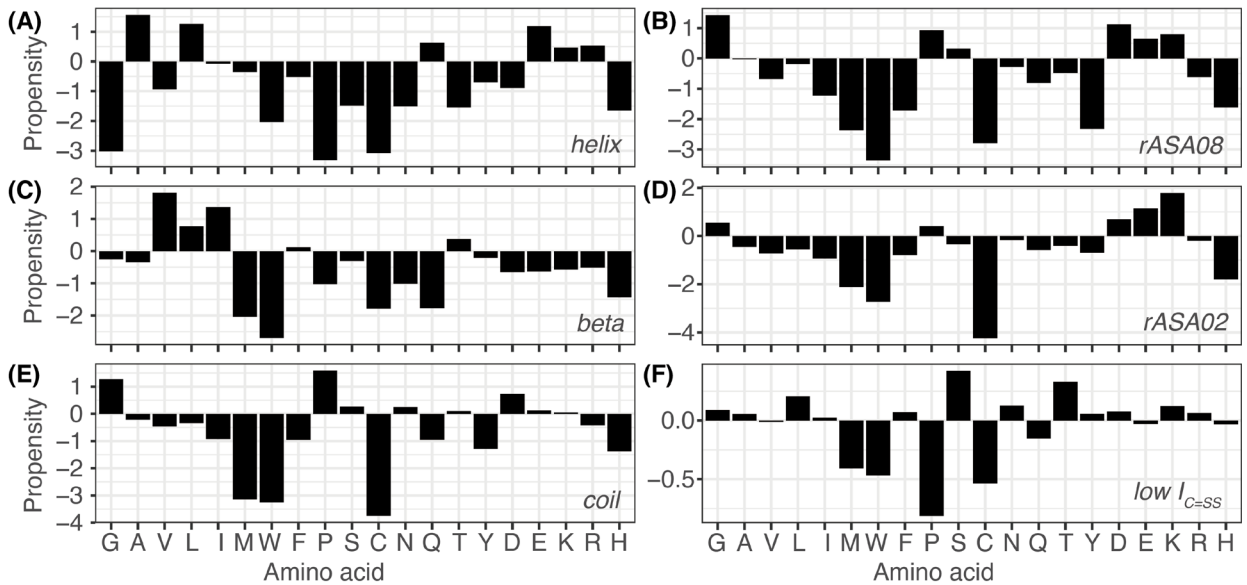

Figure 4 The amino acid propensities for the subsets of tetra-grams with strong tendencies for (A) *helix*, (B) *beta*, (C) *coil*, (D) *highly_exposed*, and (E) *exposed* segments. Panel (F) shows the propensity for the subset of tetra-grams without a strong tendency to form a specific secondary structure.

Table 3 Examples of tetra-grams with a strong secondary structure tendency

s	n	f_{helix}	$I_{C=ss}$	s	n	$RMSE$	$I_{C=ss}$
ELAK	880	0.946	586.3	*ALGG*	605	0.0193	9.499
ELAR	858	0.923	548.1	*GAVA*	549	0.0166	8.943
EALE	867	0.917	544.1	*GLSA*	53	0.0185	9.427
EEAL	899	0.923	539.7	*LSAG*	468	0.0192	9.307
EALR	903	0.917	528.5	*GLLD*	459	0.00678	8.734
s	n	f_{beta}	$I_{C=ss}$	s	n	$f_{highly\_exposed}$	$I_{C=rASA08}$
VLVV	450	0.944	538.4	*KDGK*	530	0.423	377.7
VVVV	384	0.958	481.7	*VDGK*	348	0.420	245.9
VVVG	475	0.884	465.7	*ADDP*	204	0.544	224.3
LVVD	483	0.872	454.3	*PEGY*	226	0.469	193.7
VVVD	529	0.843	452.3	*DGRT*	245	0.385	192.0
s	n	f_{coil}	$I_{C=ss}$	s	n	$f_{exposed}$	$I_{C=rASA02}$
AGAD	723	0.898	425.8	*EKLG*	657	0.945	361.0
LPEG	548	0.842	393.9	*KDGK*	530	0.958	313.5
AGLP	573	0.920	373.8	*AKKL*	544	0.949	304.7
LTPE	490	0.939	365.7	*PEEL*	566	0.935	294.1
LPPG	365	1.00	350.5	*LKEG*	545	0.941	293.6

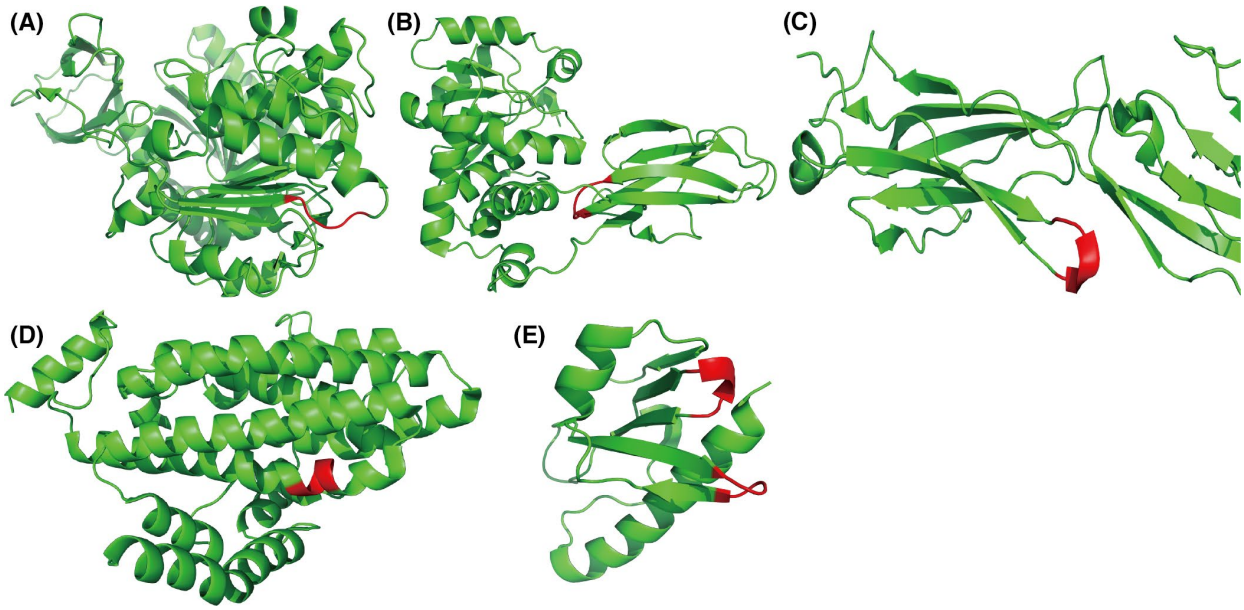

Figure 5 Structures of segments with strong tendencies. The focused segments are marked in red. (A) The "*AGAD*" segment in Lys-gingipain W83 (PDB ID: 4RBM). (B) The "*LPEG*" segment in SEX4 glucan phosphatase (PDB ID: 3NME). (C) The "*LPEG*" segment in the Lutheran blood group glycoprotein (PDB ID: 2PF6). (D) The "*LPEG*" segment in the phenylacetic acid degradation protein paaA (PDB ID: 3PWQ). (E) The two "*KDGK*" segments in an uncharacterized protein (PDB ID: 4EBG).

$$RMSE_{C=SS}(s) = \frac{1}{3}\sqrt{\Sigma_k^{helix,beta,coil}(f_k(s) - f_k(N))^2} \quad , \tag{7}$$

where $f_k(N)$ and $f_k(s)$ denote the relative frequency of N-gram segments with the property k for all the N-grams with the length of N, and that for the segments with the N-gram sequence s, respectively. For tetra-grams in this dataset, the relative frequencies of *helix*, *beta*, and *coil* structures were 0.387, 0.231, and 0.383, respectively. We found 496 tetra-grams yielding a propensity similar to this average distribution with $RMSE_{C=SS}(s) < 0.02$. For example, 604 segments with the subsequence "*ALGG*" were observed in the dataset, and $RMSE_{C=SS}("ALGG") = 0.0193$, $f_{helix}("ALGG") = 0.433$, $f_{beta}("ALGG") = 0.204$, $f_{coil}("ALGG") = 0.363$, and $I_{C=SS}("ALGG") = 9.50$. Other examples are listed in Table 3. For these "extreme" chameleon sequences, small side chains, Leu, Asn, Ser, and Thr, were enriched (Figure 4F).

For the solvent accessibility of segments, we applied two thresholds of rASA: exposed (rASA ≥ 0.2) and highly exposed (rASA ≥ 0.8). The ratios of exposed and highly exposed N-gram segments were 46.0–46.3 % and 3.87–3.96 % for $4 \leq N \leq 9$, respectively (Table 1). We found 703 and 586 tetra-grams with strong tendencies ($I_C(s) \geq 50$ and $R_k(s) \geq 2.0$) to be exposed and highly exposed, respectively. For example, 95.8 % and 42.3 % of 530 observed N-gram segments with "*KDGK*" were exposed and highly exposed, respectively (Table 3). Members of the DUF4467 family of proteins have two "*KDGK*" segments (PDB ID: 4EBG; Figure 5E). Amino acid propensities were similar to these different thresholds. Amino acids preferred to form the coil structure (Gly and Pro), and charged residues (Asp, Glu, and Lys) were enriched for the tetra-grams with strong tendencies to be exposed and highly exposed (Figures 4B and D). However, subsequences enriched for exposed segments and those enriched for highly exposed ones showed distinct features. Only 81 tetra-grams were shared between them. Although the amino acid preferences for the tetra-grams favored in exposed and highly exposed segments were similar, the cluster analysis revealed different sequence patterns (Supplementary Figures S7 and S8). Additionally, for longer subsequences ($N \geq 6$), Gly and Pro were not preferred for *rASA02*, and Ile, Asn, Gln, and Thr were preferred for *rASA02* (Supplementary Figure S9). The two properties, exposed and highly exposed, had different characteristics for longer subsequences. Note that subsequences enriched to high exposure are not necessarily enriched to exposure because enrichment is assessed in terms of dissociation from the background, i.e., $f_k(N)$. For example, the N-gram "*CEMT*" with 48 highly exposed and 50 exposed segments out of 52 observations yielded $I_{highy\_exposed}$("*CEMT*") = 142.6 and $I_{exposed}$("*CEMT*") = 32.34.

Predictions of *N*-grams with Strong Tendencies

We tested the accuracy of predicting whether a subsequence has a strong tendency by using a machine learning technique. In this test, only sequence information encoded as a $20 \times N$-dimensional binary vector and the label about the tendency were given for predictors; no context information of the sequence and no evolutionary information were used. The three-class (for the secondary structure) and binary classifications (for the rASA properties) were performed using a multi-layer neural network. For the binary classification, a positive set consisted of N-grams where more than 80 % of the segments had a particular property, and the negative set consisted of other N-grams that were randomly selected to have the same number of N-grams in the positive set (Supplementary Table S1).

The accuracies for the secondary structure, *rASA08*, and *rASA02* were 62–65 %, 74–75 %, and 71–73 %, respectively (Figure 6). Although the prediction accuracies look worse than the state-of-the-art sequence-based prediction methods [8], the purpose of this study differs from them. This study aimed to predict the tendency of each N-gram sequence rather than to predict the property of a query protein. Our predictions were performed based on the N characters encoding amino acid residues without any context or evolutionary information of the sequences. The fact that the predictions yield better performance compared with the random suggests that the tendency of unobserved N-gram sequences can be predicted based on the data of known N-grams.

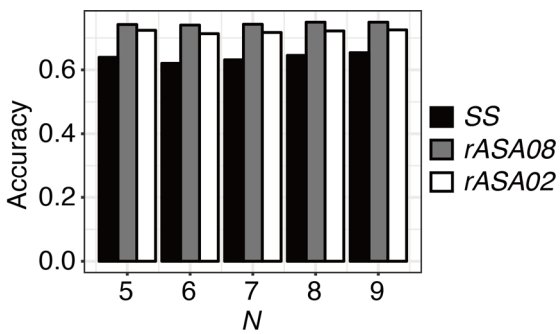

Figure 6 Accuracy of the machine learning prediction tasks. (A) The three-class classification of the secondary structures. (B) The binary classification of highly exposed segments (*rASA08*). (C) The binary classification of exposed segments (*rASA02*).

Discussion

In this study, we comprehensively analyzed protein subsequences consisting of four to nine residues observed in the PDB. Although the *N*-grams in the non-redundant dataset based on the PDB included a majority of all the possible tetra-grams of amino acids, the cases with $N \geq 5$ covered only a tiny subspace of the sequence spaces. In particular, the saturation of the diversity of observed *N*-grams (Table 1) suggests that many *N*-grams with $N \geq 6$ appear in a limited context in the PDB.

To find *N*-grams with a strong tendency to determine their structural properties in the dataset, we applied the information quantities $I_C(s)$ for each *N*-gram sequence *s*. The higher information quantity of an *N*-gram suggests a stronger tendency for the *N*-gram. Tetra-grams clearly have higher information quantities than subsequences with higher *N* values (Figure 3). This is partly because tetra-grams had larger samples (observed segments for each *N*-gram sequence) than longer subsequences did. However, the distribution for the classification, whether highly exposed or not (*rASA08*), showed smaller differences in information quantity between shorter and longer subsequences compared with the other properties. This implies that information of high exposure of segments is encoded in longer subsequences. On the other hand, four residues were sufficient to encode the strong tendencies for the secondary structure and that for exposure (*rASA02*) of segments.

Based on the information quantity, we can filter the *N*-gram sequences with strong tendencies. We applied the arbitrary criteria, $I_C(s) \geq 50$ and $R_k(s) \geq 2.0$, and found many *N*-gram sequences with a strong tendency. Strong tendencies to form helical, beta, and coil structures were found in 2,205, 3,919, and 4,505 tetra-grams, respectively. On average, 80.2 % of segments for these tetra-grams formed their secondary structure. We observed that some of the exceptional segments, which had a sequence with a strong tendency but formed a non-enriched structure, had distorted structures (Figures 5C and D). On the other hand, we also found 496 tetra-grams that evenly formed all three secondary structures. These subsequences had a weak tendency to form a specific secondary structure, that is, a strong tendency to be chameleon sequences. Although many previous studies on chameleon sequences detected the chameleon sequences by finding at least one pair of segments with different secondary structures and the same sequence, our analyses quantified the strength of the tendency to be chameleon sequences. This provides a new viewpoint for the study of chameleon sequences, and the roles of these "extreme" chameleon sequences will be studied in the future. In terms of rASA, we found 703 and 586 tetra-grams with a strong tendency to be exposed (rASA $\geq$ 0.2) and highly exposed (rASA $\geq$ 0.8), respectively. The numbers of informative *N*-grams for exposed and highly exposed segments were smaller than those for the secondary structures.

Although we found many *N*-grams with high information quantity, they are only a small part of all the *N*-grams observed in the PDB. A majority of *N*-grams have a small number of segments, and thus, it is difficult to obtain information about their tendency. However, machine learning techniques provide information about the *N*-grams without a sufficient number of samples based on information about similar subsequences.

Conclusion

A survey of the protein subsequences in the PDB revealed thousands of subsequences with a strong tendency to form a specific secondary structure and to be exposed to the solvent. They included both known and unknown candidates for structural motifs. Although they are only a small part of the entire space of possible subsequences, they provide insight into the protein structure based only on the short array of characters without any context or evolutionary information. Because a major reason for overlooking the tendency for many subsequences is the lack of samples, further growth of the PDB may provide more sequence motifs in the future. In addition, machine learning can partially compensate for the problem due to the lack of known samples. In this study, we focused only on statistics of subsequence and some categorical properties of segments. Further analyses including detailed three-dimensional conformation of each segment and interactions with surrounding segments would provide insight into underlying biophysical mechanisms for enriched subsequences.

Although state-of-the-art sequence-based predictions using evolutionary information have achieved high performances, especially for secondary structures, the principle of how the sequence determines its structures is still unclear. A comprehensive understanding of the strength of the tendency to determine the nature of a segment by the subsequence itself provides indispensable insights for illuminating the principles. In addition, mapping these informative subsequences onto each protein sequence may provide insights into novel proteins.

Conflict of Interest

All the authors declare that they have no conflicts of interest.

Author Contributions

KK, and TT designed this study. RK and KK performed the research. All the authors contributed to the writing of the manuscript.

Acknowledgements

KK is supported by JSPS KAKENHI Grant Numbers JP20K12069 and JP21K06052. The computational resources were provided by the HPCI System Research Project (Project IDs: hp190017, hp190018, hp200063, hp200090, hp210005, and hp210008), the NIG supercomputer at ROIS National Institute of Genetics, and Human Genome Center (the University of Tokyo).

References

[1] Anfinsen, C. B. Principles that govern the folding of protein chains. Science 181, 223–230 (1973). https://doi.org/10.1126/science.181.4096.223
[2] Ash, C. So much more to mucus. Science 370, 420 (2020). https://doi.org/10.1126/SCIENCE.370.6515.418-U
[3] Dill, K. A., MacCallum, J. L. The protein-folding problem, 50 years on. Science 338, 1042–1046 (2012). https://doi.org/10.1126/science.1219021
[4] Kinoshita, K., Nakamura, H. Protein informatics towards function identification. Curr. Opin. Struct. Biol. 13, 396–400 (2003). https://doi.org/10.1016/S0959-440X(03)00074-5
[5] Lobb, B., Doxey, A. C. Novel function discovery through sequence and structural data mining. Curr. Opin. Struct. Biol. 38, 53–61 (2016). https://doi.org/10.1016/j.sbi.2016.05.017
[6] Sudha, G., Nussinov, R., Srinivasan, N. An overview of recent advances in structural bioinformatics of protein-protein interactions and a guide to their principles. Prog. Biophys. Mol. Biol. 116, 141–150 (2014). https://doi.org/10.1016/j.pbiomolbio.2014.07.004
[7] Thornton, J. M., Todd, A. E., Milburn, D., Borkakoti, N., Orengo, C. A. From structure to function: Approaches and limitations. Nat. Struct. Biol. 7 Suppl, 991–994 (2000). https://doi.org/10.1038/80784
[8] Yang, Y., Gao, J., Wang, J., Heffernan, R., Hanson, J., Paliwal, K., et al. Sixty-five years of the long march in protein secondary structure prediction: The final stretch? Brief. Bioinform. 19, 482–494 (2018). https://doi.org/10.1093/bib/bbw129
[9] Ahmad, S., Gromiha, M. M., Sarai, A. Real value prediction of solvent accessibility from amino acid sequence. Proteins 50, 629–635 (2003). https://doi.org/10.1002/prot.10328
[10] Heffernan, R., Dehzangi, A., Lyons, J., Paliwal, K., Sharma, A., Wang, J., et al. Highly accurate sequence-based prediction of half-sphere exposures of amino acid residues in proteins. Bioinformatics 32, 843–849 (2016). https://doi.org/10.1093/bioinformatics/btv665
[11] Fukuchi, S., Hosoda, K., Homma, K., Gojobori, T., Nishikawa, K. Binary classification of protein molecules into intrinsically disordered and ordered segments. BMC Struct. Biol. 11, 29 (2011). https://doi.org/10.1186/1472-6807-11-29
[12] Jones, D. T., Cozzetto, D. DISOPRED3: Precise disordered region predictions with annotated protein-binding activity. Bioinformatics 31, 857–863 (2015). https://doi.org/10.1093/bioinformatics/btu744
[13] Chou, P. Y., Fasman, G. D. Prediction of protein conformation. Biochemistry 13, 222–245 (1974). https://doi.org/10.1021/bi00699a002
[14] Frishman, D., Argos, P. Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. Protein Eng. 9, 133–142 (1996). https://doi.org/10.1093/protein/9.2.133
[15] Ito, M., Matsuo, Y., Nishikawa, K. Prediction of protein secondary structure using the 3d-1d compatibility algorithm. Comput. Appl. Biosci. 13, 415–424 (1997). https://doi.org/10.1093/bioinformatics/13.4.415
[16] Kihara, D. The effect of long-range interactions on the secondary structure formation of proteins. Protein Sci. 14, 1955–1963 (2005). https://doi.org/10.1110/ps.051479505
[17] Kinjo, A. R., Nishikawa, K. Predicting secondary structures, contact numbers, and residue-wise contact orders of native protein structures from amino acid sequences using critical random networks. Biophysics 1, 67–74 (2005). https://doi.org/10.2142/biophysics.1.67
[18] Kurt, N., Mounce, B. C., Ellison, P. A., Cavagnero, S. Residue-specific contact order and contact breadth in single-domain proteins: Implications for folding as a function of chain elongation. Biotechnol. Prog. 24, 570–575 (2008). https://doi.org/10.1021/bp070475v
[19] Rost, B., Sander, C. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. Proc. Natl. Acad. Sci. U.S.A. 90, 7558–7562 (1993). https://doi.org/10.1073/pnas.90.16.7558

[20] Zvelebil, M. J., Barton, G. J., Taylor, W. R., Sternberg, M. J. E. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. J. Mol. Biol. 195, 957–961 (1987). https://doi.org/10.1016/0022-2836(87)90501-8
[21] Heffernan, R., Paliwal, K., Lyons, J., Singh, J., Yang, Y., Zhou, Y. Single-sequence-based prediction of protein secondary structures and solvent accessibility by deep whole-sequence learning. J. Comput. Chem. 39, 2210–2216 (2018). https://doi.org/10.1002/jcc.25534
[22] Vries, J. K., Liu, X., Bahar, I. The relationship between N-gram patterns and protein secondary structure. Proteins 68, 830–838 (2007). https://doi.org/10.1002/prot.21480
[23] Guo, J. T., Jaromczyk, J. W., Xu, Y. Analysis of chameleon sequences and their implications in biological processes. Proteins 67, 548–558 (2007). https://doi.org/10.1002/prot.21285
[24] Li, W., Kinch, L. N., Karplus, P. A., Grishin, N. V. ChSeq: A database of chameleon sequences. Protein Sci 24, 1075–1086 (2015). https://doi.org/10.1002/pro.2689
[25] Kasahara, K., Minami, S., Aizawa, Y. Characteristics of interactions at protein segments without non-local intramolecular contacts in the Protein Data Bank. PLoS One 13, e0205052 (2018). https://doi.org/10.1371/journal.pone.0205052
[26] Berman, H., Henrick, K., Nakamura, H., Markley, J. L. The worldwide Protein Data Bank (wwPDB): Ensuring a single, uniform archive of PDB data. Nucleic Acids Res. 35(SUPPL. 1), D301–D303 (2007). https://doi.org/10.1093/nar/gkl971
[27] Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150–3152 (2012). https://doi.org/10.1093/bioinformatics/bts565
[28] Westbrook, J., Ito, N., Nakamura, H., Henrick, K., Berman, H. M. PDBML: The representation of archival macromolecular structure data in XML. Bioinformatics 21, 988–992 (2005). https://doi.org/10.1093/bioinformatics/bti082
[29] Kabsch, W., Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577–2637 (1983). https://doi.org/10.1002/bip.360221211
[30] Ward, J. H. Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. 58, 236-244 (1963). https://doi.org/10.1080/01621459.1963.10500845
[31] Crooks, G. E., Hon, G., Chandonia, J. M., Brenner, S. E. WebLogo: A sequence logo generator. Genome Res. 14, 1188–1190 (2004). https://doi.org/10.1101/gr.849004
[32] Simm, S., Einloft, J., Mirus, O., Schleiff, E. 50 years of amino acid hydrophobicity scales: Revisiting the capacity for peptide classification. Biol. Res. 49, 31 (2016). https://doi.org/10.1186/s40659-016-0092-5
[33] Sengupta, D., Behera, R. N., Smith, J. C., Ullmann, G. M. The α helix dipole: Screened out? Structure 13, 849–855 (2005). https://doi.org/10.1016/j.str.2005.03.010