J Comput Aided Mol Des (2009) 23:163–169
DOI 10.1007/s10822-008-9248-x

ProSeg: a database of local structures of protein segments

Yoshito Sawada · Shinya Honda

Received: 8 July 2008 / Accepted: 26 September 2008 / Published online: 16 October 2008
© The Author(s) 2008. This article is published with open access at Springerlink.com

Abstract Integration of knowledge on the sequence-structure correlation of proteins provides a basis for the structural design of artificial novel proteins. As one of strategies, it is effective to consider a short segment, whose size is in between an amino acid and a domain, as a correlation unit for exploring the structure-to-sequence relationship. Here we report the development of a database called **ProSeg**, which consists of two sub-databases, Segment DB and Cluster DB. Segment DB contains tens of thousands of segments that were prepared by dividing the primary sequences of 370 proteins using a sliding *L*-residue window ($L = 5, 9, 11, 15$). These segments were classified into several thousands of clusters according to their three-dimensional structural resemblance. Cluster DB contains much cluster-related information, which includes image, rank, frequency, secondary structure assignment, sequence profile, etc. Users can search for a suitable cluster by inputting an appropriate parameter (i.e., PDB ID, dihedral angles, or DSSP symbols), which identifies the backbone structure of a query segment. Analogous to a language, **ProSeg** could be regarded as a 'structure-sequence dictionary' that contains over 10,000 'protein words'. **ProSeg** is freely accessible through the Internet (http://riodb.ibase.aist.go.jp/proseg/).

Keywords Sequence-structure correlation · Database · Short peptide segment · Classification · Clustering · PDB · Protein word · Protein design

Y. Sawada · S. Honda (✉)
National Institute of Advanced Industrial Science and Technology (AIST), Central 6, Tsukuba 305-8566, Japan
e-mail: s.honda@aist.go.jp

Introduction

Integration of knowledge of on the sequence-structure correlation of proteins not only offers help on improving methods for predicting structure of natural proteins but also provides a basis for the structural design of artificial novel proteins. Many studies have been carried out so far to address this issue, and they can be classified into two major categories (Table 1). One is an amino acid-based approach, which examines the correspondence of a single (or a few) amino acid(s) with a partial structure [1, 2]. In this approach, the number of 'correspondence rules' necessary to describe the correlation does not become large due to the limited number of available amino acids (ca 20 types). However, each of these correspondence rules is polysemous. Hence, the sequence-structure correlation is expressed as probabilistic, and not deterministic. This increases ambiguity in the correlation as the chain-length increases. The other category is a domain-based approach, which examines the correspondence between a domain sequence and its structure [3–6]. Since the correspondence between a domain sequence and its structure is generally tight, this approach mostly allows to 'translate' a sequence into a structure in a monosemous manner. However, the number of possible sequences for a typical domain size is astronomical. Therefore, in the case of a domain-based approach, it is absolutely impossible to prepare comprehensive number of correspondence rules for all units. To circumvent the respective disadvantages of the above two approaches, we think it would be worthwhile to consider a medium size that is in between an amino acid (small) and a domain (large) as a correlation unit, and explore the structure-to-sequence relationship using a short segment-based approach.

In the previous study [7], we classified local structures of protein segments by means of exhaustive clustering

Table 1 Conceptual summary of conventional approaches to a sequence-structure correlation in proteins

Correlation unit	Amino acid	Domain
Size of unit	Small, fixed ($L = 1$)	Large, variable ($L > 50$)
Number of unit type	A few (20)	Astronomic ($>100^{65}$)
Correspondence to conformation	Polysemous, probabilistic, context-dependent	Monosemous, deterministic, context-independent
Performance	Applicable to any type of sequence, but the correlation is ambiguous	Correlation is tight and clear, but impossible to prepare comprehensive 'rules' for all units

analyses and investigated the structural diversity of proteins. Our results showed that the structures of protein segments occupy only tiny regions of the protein universe, where they are distributed in a dense-and-sparse manner. In addition, their diversity follows a power-law distribution. These results suggested that the organization of proteins is based on certain mathematical guidelines using a limited number of local structures. Moreover, analysis of the clusters of classified segments revealed that the limitation of the number of local structures is not attributed only to the conformational preference of single residues. These features are attractive outcomes because they are quite similar to the features normally found in the structure of natural languages.

Besides the general nature of the structural diversity of proteins, the clustering analyses also provided us with numerous distinct structural motifs, including known canonical ones. However, the number was simply too huge to depict most of them in the limited space of the previous paper. Therefore, here we have developed a new database called **ProSeg**, which contains the entire results of the clustering analyses along with various characteristics of the classified clusters. In addition, a web-based interface has been implemented to facilitate easy user access. Because of the underlying exhaustive clustering analyses, **ProSeg** will be able to provide the essential physicochemical properties of almost all backbone structures that a short segment is able to form. Hence, **ProSeg** would be useful for many applications in protein science.

Methods

Segments to be classified were prepared by dividing the primary sequences of 370 proteins using a sliding L-residue window ($L = 5$, 9, 11 or 15). For example, for a 100-residue polypeptide, 92 segments, each 9-residue long, were produced. These 370 proteins were selected as a set of non-redundant representative proteins from the Culled Protein Data Bank (PDB) (version: Dec. 13, 2001; resolution <1.6 Å; R-factor < 0.2; sequence identity <25%) [8]. The divided segments were classified into a number of clusters by a single-pass clustering (SP) method [9] or a 3D mesh gridding (3DMesh) method. Structural dissimilarity (D or D_{issim}) between the segments is defined on the basis of backbone dihedral angles. This parameter was used as a distance scale in the SP method. The threshold value (D_{th}) is set to 30° or 40°, which is a parameter responsible for making a new cluster in the SP method. After clustering, various structural and sequence properties of the clusters were analyzed in order to characterize each one of them. The root mean square (RMS) deviations in Euclid distance were calculated by a least-squares superimposition of the backbone atoms (C, CA, and N) of the classified segments. Detailed procedures of clustering and mathematical definitions of cluster properties were described previously [7].

Clustering and related calculations were performed on Pentium-based Linux PCs using original programs compiled by Intel FORTRAN Compiler for Linux. In some cases, MUSASHI (http://musashi.sourceforge.jp/), DSSP [10], and Rasmol [11] programs were used. The database functionality was implemented using the Oracle 10 g software, operating on AIX 5.3 L Unix and IBM p570 server system (POWER5^{+} × 4, 16 GB RAM). The web interface was developed using HTML, Perl, and FORTRAN.

Results

Data contents

ProSeg consists of two sub-databases, Segment DB and Cluster DB, which are connected to each other through an internal index in the Cluster DB (Fig. 1). Currently, 78,622 5-residue long segments, 76,694 9-residue long segments, 75,744 11-residue long segments, and 73,876 15-residue long segments are archived in the Segment DB. These segments were classified into thousands of clusters. The number of resultant clusters depended on the difference in the segment length, clustering methods and clustering parameters used for the analysis. In the Cluster DB, 2,217 clusters were formed using the conditions: $L = 5$, SP, $D_{\mathrm{th}} = 30°$; 10,494 clusters were formed using the conditions: $L = 9$, SP, $D_{\mathrm{th}} = 30°$; 4,179 clusters were formed using the conditions: $L = 9$, SP, $D_{\mathrm{th}} = 40°$; 1,449 clusters were formed using the conditions: $L = 9$, 3DMesh; 17,096

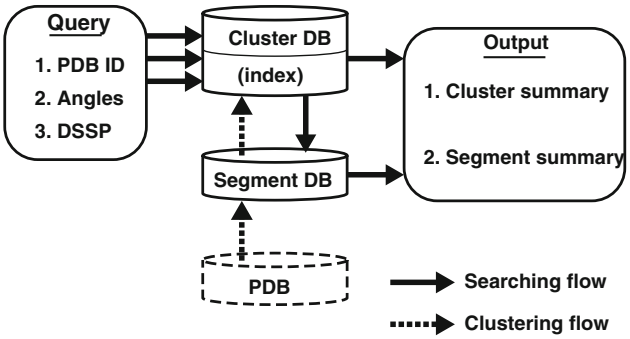

Fig. 1 Schematic overview of **ProSeg**

clusters were formed using the conditions: $L = 11$, SP, $D_{th} = 30°$; and 30,187 clusters were formed using the conditions: $L = 15$, SP, $D_{th} = 30°$. The following segment-specific properties are also available: the segment length, PDB ID, chain ID, the number of the central residue, segment sequence (1-letter code), secondary structure assignment by DSSP, radius of gyration (R_G), total number of hydrogen bonds (HB) between backbones, number of intra-segment HB, number of inter-segment HB, backbone dihedral angles (ϕ, ψ, ω), and Cartesian coordinates in PDB format.

In the Cluster DB, the following properties are summarized for every cluster: clustering conditions (clustering method, threshold value, segment length, etc.), number of assigned segments (M_r), cluster ranking, normalized frequency and a list of classified segments. Each cluster is accompanied by a cluster center (CC), which is defined as a center of mass (i.e., centroid) of a cluster in the multi-dimensional space of the protein universe. The CC corresponds to a segment having a fictitious conformation whose dihedral angles are the averages of the dihedral angles of a set of segments that were assigned to the cluster. The Cluster DB also contains the properties of CC, such as the averaged dihedral angles (ϕ, ψ, ω) and an image of the CC. For every cluster, the following structural properties, corresponding to the averaged parameters of a set of assigned segments, are available: averaged number of HB, averaged R_G, and dominant secondary structure. The RMS deviations derived from the superimposing calculations are also recorded. Sequence properties, such as the amino acid frequency for each position of a segment, are also available. Additionally, the sequence profile, i.e., position specific scoring matrix (PSSM), and averaged Kullback–Leibler relative entropy (KL) are available for every cluster.

For every segment, the internal index in the Cluster DB, which links the two sub-databases, contains a set of identification codes for the segment and its cluster, structural dissimilarity of the segment to its CC, and RMS deviation between the segment and its CC.

Search capabilities

In order to gain information of interests, users can search a cluster (or clusters) by inputting an appropriate parameter that identifies the backbone structure of a query segment. Currently, there are three ways to identify the backbone structure of the segment (Fig. 1). In the case where the structural data of a protein that contains the query segment are deposited in the PDB, the users are recommended to specify the PDB accession code (PDB ID) of the protein, identification code (Chain ID) of the polypeptide chain to which the segment belongs, and number of the central residue of the segment. In **ProSeg**, the number of a residue is reassigned to an integer by simply counting from the N-terminal end of a polypeptide chain having a defined structure. After inputting these parameters, the program searches for the corresponding record from the stored data in **ProSeg**, temporarily cached PDB data, or original files in the PDB via the Internet. This means that **ProSeg** can accept any PDB ID of the latest version of the PDB as a query. In the case where the structural data is not available in the PDB, the users are allowed to input a set of backbone dihedral angles of the query segment. This function enables the users to inquire for any type of conformation regardless of the presence or absence of the PDB data. If the users feel troublesome to input the dihedral angles of the query segment, they can specify the type of the secondary structure of the query segment by inputting the DSSP symbols. When the users have an unpublished structural data of a protein, they may upload the data for the protein in PDB format into **ProSeg** in order to calculate and obtain a set of dihedral angles of the protein. When a query is executed using the PDB ID or dihedral angles, a list of clusters with CCs close to the backbone structure of the query segment is displayed. When a query is executed using the DSSP symbols, a list of clusters with CCs showing secondary structure identical to that of the query segment is displayed. These lists will facilitate in accessing a desired cluster and gaining further information.

Web interface

User interface consists of several HTML pages (Fig. 2). In the 'Main' page, the users can put in a query to **ProSeg** in the manner explained above. In the 'Search Options' page, the users can change the default settings for "Clustering conditions", "Output options", or "Criteria for searching targets". Currently, the default setting of the clustering condition, which specifies a target set of clusters for a query, is as follows: $L = 9$, SP, $D_{th} = 30°$. In the 'Cluster List' page, clusters close to the users' query, as judged by the D_{issim} value, are listed with an image, rank, M_r, RMS deviation, R_G, averaged number of HB, averaged KL, and

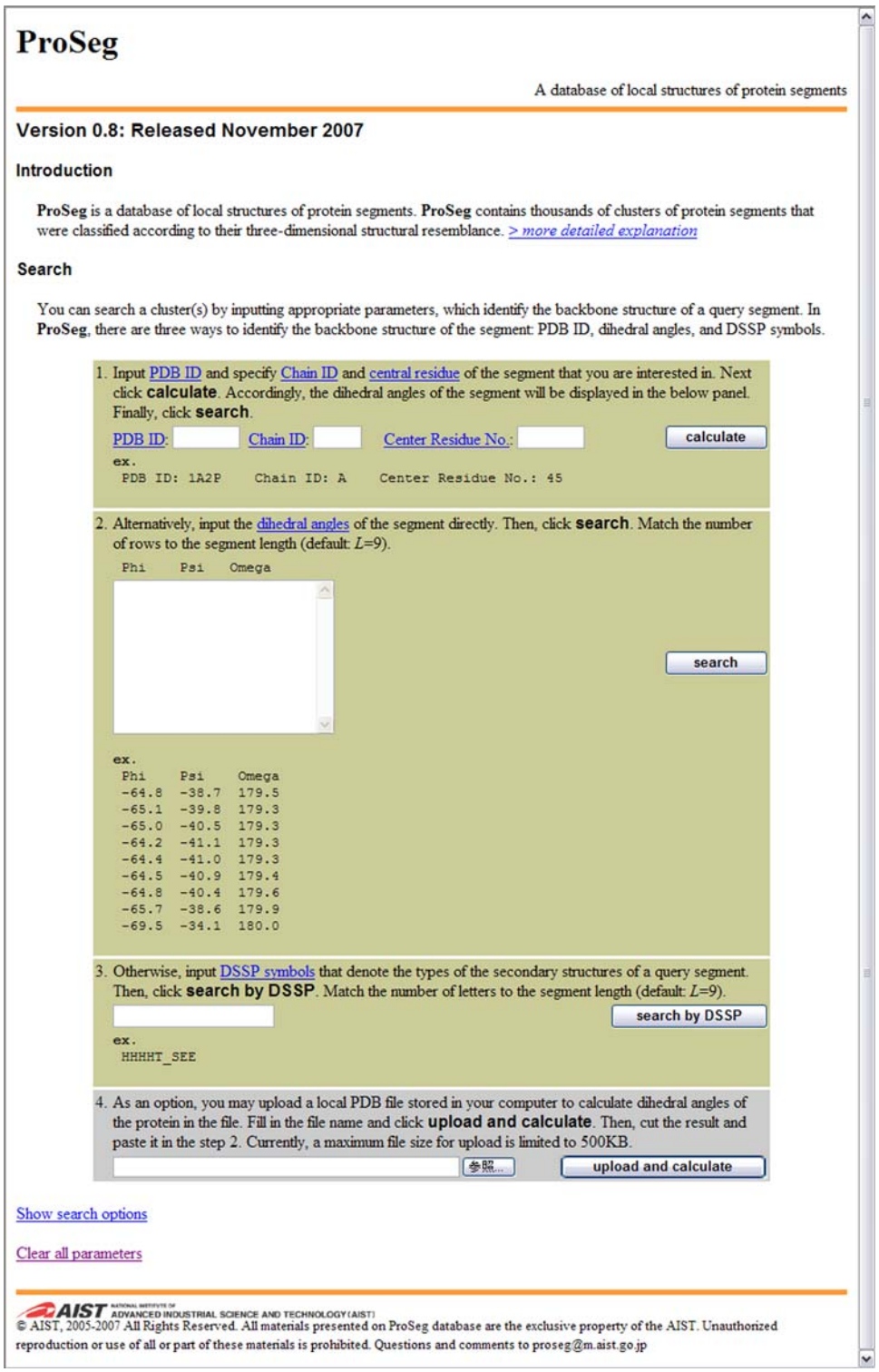

Fig. 2 Web entrance of **ProSeg** (http://riodb.ibase.aist.go.jp/proseg/)

dominant secondary structure. In the 'Cluster Details' page, the users can brows an image of the CC, an image of the superimposed segments and a full set of characteristics of each cluster including a set of dihedral angles of the CC, sequence profile, and frequency counts of amino acids. In the 'Segment List' page, segments that are classified into the same cluster are listed with several characteristics including the PDB ID of the parent protein, number of the central residue, amino acid sequence, and structural dissimilarity (i.e., dihedral angle distance) to the CC. The list is downloadable in the plain text format. Most of the columns of these lists in **ProSeg** are furnished with a sortable

function. In the 'Segment Details' page, a full set of characteristics for each segment is summarized. The Cartesian coordinate file in the PDB format can be downloaded from this page. In the Glossary page, several technical terms and abbreviations are explicated. Tutorials and FAQs for using the **ProSeg** database are available in the Overview page.

Discussion

Features

Currently, not a few databases associated with protein structures are available. Compared to these databases, however, **ProSeg** possesses several distinctive features. Most of the taxonomic databases—such as SCOP [12], CATH [13] and FSSP [14]—accumulate data of domains, and not of short segments. In contrast to the domain-based databases, which are informative for understanding the divergent evolution of proteins, **ProSeg**, by focusing on short segments, may be helpful in shedding lights on the convergent evolution of proteins. To our knowledge, only a limited number of short segment-based databases are available in the public domain. For example, MSDmotif summarizes small 3D structure motifs (http://www.ebi.ac.uk/msd-srv/msdmotif/), in which characteristics of 35 motifs including their dihedral angles, sequence statistics and ligand binding can be reviewed. ArchDB is a compilation of structural classifications of loops extracted from known protein structures [15]. I-sites library contains a set of sequence patterns that strongly correlate with protein structure at the local level [16]. LSBSP1 and LSBSP2 contain large sets of sequence profiles for short segments, and these databases have been implemented in the integrated computational system PrISM.1 for predicting local structures [17, 18]. Presently, the last two databases cannot be accessed directly and freely. DPFS stores 1.1 million clusters of local conformations of 8-residue segments and these clusters are further classified into structural clusters and functional clusters [19]. As described in the previous section, **ProSeg** is different from these short segment-based databases in the underlying fundamental concept, method of classification, number of classified categories (i.e., clusters), amount of stored information, and flexibility of searching. Especially, searching a database by inputting the structure of a query segment of that is of interest to the user is accomplished only with **ProSeg**.

One of the significant features of **ProSeg** is that the program does not directly search segments in the database but search for clusters whose CC is close to the users' query. These clusters have been predetermined by exhaustive clustering analyses, so that when the users submit the query, time-consuming calculations to compute the dissimilarities of tens of thousands of segments to the query segment become unnecessary. Although one may feel that the strategy lacks discretion, the strategy is actually guaranteed by the fact that the structures of protein segments occupy only tiny regions of the multi-dimensional space of the protein universe, and they are distributed in a dense-and-sparse manner, as described in our previous study [7]. Therefore, this feature in **ProSeg** improves the performance of the database without reducing the coverage of the structural space of the protein universe.

ProSeg contains several sets of clusters, which were obtained by changing a method and/or parameters in the clustering calculations. Therefore, the users can select an appropriate set of clusters in the 'Search Options' page according to their need. For example, a shorter segment length ($L = 5$ or 9) may be adequate for analyzing a single structural motif because this length corresponds to the typical size of secondary structure elements. In contrast, a longer segment length ($L = 11$ or 15) may be adequate for analyzing a long-range correlation between two and more structural motifs. Thus, it appears that **ProSeg** can effectively serve to satisfy broad purposes as information on multiple length peptide segments are archived in this database. Another example is the threshold value (D_{th}). A loose threshold value, such as 40°, would be suitable for analyzing rare structures because it increases the number of segments in the cluster and enhances the statistical significance of the cluster properties for rare structures. In contrast, a strict threshold value, such as 30°, would be suitable for analyzing common structures. Because the number of segments in the cluster for common structures is very large, reducing this number by omitting segments having relatively dissimilar structure will rather improve the reliability of the cluster properties. Selecting an appropriate threshold value is, therefore, recommended because the difference in frequency between the common structures and rare structures is occasionally more than 10^5 times [7].

Examples of structural motifs

Thousands of clusters are archived in **ProSeg**. Some of these clusters appear to consist of known structural motifs. In alternative case, these clusters themselves would be regarded as new structural motifs of proteins. Figure 3 shows several examples of clusters, which were obtained using the condition: $L = 15$, SP, $D_{\mathrm{th}} = 30°$. First example is a helix-strand motif. The helix-portion in the cluster CS01.03.02.000015 ends with Gly, which is followed by a strand-portion. The sequence profile of the cluster shows that the amino acid residues between the helix and the strand are biased. The Gly is known as a residue at the C′-

position of a Schellmann motif, one of C-terminal capping motifs [20]. The residue after the Gly tends to be either Ile or Val with medium preference. Next two amino acids strongly inclined to be Pro and Val, respectively. The last Val is incorporated in the strand-portion, which is reasonable because of the high preference of a branched side-chain amino acid in β-sheet structure. Consequently, in the cluster, we can recognize a motif in which three residues, -Gly[Ile,Val]Pro-, connect a helix with a strand. In addition to the above, another helix-strand motif is identified in **ProSeg**. The cluster CS01.03.02.000034 is less common than the first example but still appears frequently in the protein universe, as indicated by its rank (the 34th among 30,187 clusters). In the motif of the second example, different consensus sequence (i.e., -GlyAlaAspXaa-) connects a helix with a strand. In the case of a strand-helix motif, a small residue (Ala, Cys or Thr) and Pro tend to occur in the region connecting the strand and the helix, as shown in the cluster CS01.03.02.000036. The residue after the Pro strongly inclined to be either Asp or Ser, which is common and known as an Ncap residue of an N-terminal capping motif [20]. At least two types of helix–helix motifs are found in **ProSeg**. The first type of helix–helix motif consists of canonical C-terminal and N-terminal capping motifs, and as shown in the cluster CS01.03.02.000041, an aliphatic or a hydrophobic residue (Ile, Val, Met or Leu) connects the two helix-portions. The second type of helix–helix motif lacks a canonical C-terminal capping motif in the first helix, and as shown in the cluster CS01.03.02.000025, an aromatic residue, such as Trp, is likely to reside at the last spiral of the first helix and appears to interact with the hydrophobic surface of the second helix. The angle formed between the axes of the two helices is not acute as compared with those of the former examples. One example of strand–strand motifs is a bended strand. The cluster CS01.03.02.000093 indicates that, His is likely to occur at the middle of the bended strand (or between two strands). The last example is a β-hairpin structure. Two anti-parallel strands in the β-hairpin are connected by a four-residue loop having $\alpha_R\alpha_R\gamma_R\alpha_L$ conformation, as shown in the cluster CS01.03.02.000138, which can be classified as a type 4:4 hairpin according to Thornton's nomenclature [21]. The sequence deviation of this cluster is significant as indicated by the value of KL. This suggests that the consensus sequence found in a four-residue loop is valuable information for protein engineering and protein design (discussed later). In conclusion, although they were produced using only backbone parameters, the clusters archived in **ProSeg** illustrate 3D structural features involving side-chains and provide worthwhile information about a structural motif of interest to the user.

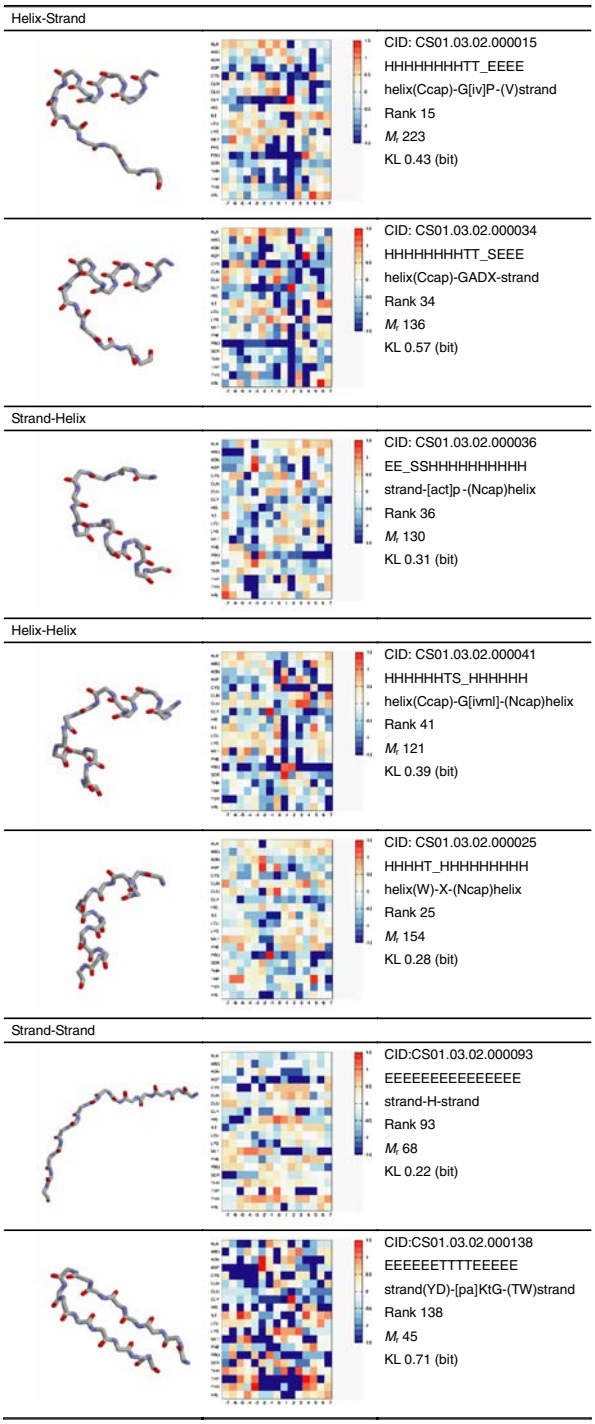

Fig. 3 Examples of structural motifs available in **ProSeg**. Image in the left column corresponds to the backbone structure of the segment that is closest to the respective cluster center. Matrix figure in the middle column represents the sequence profile of the cluster. Properties in the right column includes the ID number of a cluster, secondary structure assignment by DSSP symbols, simplified notation of motif, cluster ranking, number of segments in a cluster, and Kullback-Leibler relative entropy. In the simplified motif notation, the uppercase and lowercase letters denote amino acids that tend to occur at the corresponding position with strong and medium preferences, respectively

Applications

Since an understanding of the structure-to-sequence relationship of a protein is one of the principal themes in protein science, the extensive information stored in **ProSeg** will be useful for many applications in the protein research such as in molecular evolution, folding mechanism, structural prediction and molecular design. Especially, **ProSeg** is informative for various structure-based protein engineering due to its capability to respond to a structural query. In fact, we have recently succeeded in designing a novel small, linear peptide consisting of only 10 amino acids by using a sequence profile introduced from the early prototype of **ProSeg**. The profile used there is essentially the same as that of the β-hairpin structure in the cluster CS01.03.02.000138 in Fig. 3, which is strongly inclined to a particular consensus sequence. This novel peptide having this consensus sequence folds into a unique 3D structure in water, and exhibits a reversible and cooperative structural transition upon thermal denaturation [22]. These protein-like features of small peptides will be fundamental to the future development of new types of small, stable, and specific ligands for therapeutic use.

Conventionally, the structure of a protein molecule used to be analogized with the grammar of a language [23]. Moreover, we previously found that the distribution of local structures of protein segments showed a behavior that was formulated well by Zipf's law [7], suggesting that this resemblance is not just a metaphor, and that the structure of a protein and the structure of a language probably share common structural rules and have a quantitative correlation. If one supposes that a letter, a word and a sentence correspond, respectively, to an amino acid, a short segment and a domain, **ProSeg** would be regarded as a 'structure-sequence dictionary' that contains over 10,000 'protein words'. We anticipate that **ProSeg** will aid us to learn a 'foreign language' encoded in proteins.

Acknowledgments The authors are grateful to Mr. Keiichi Tsukamoto (Mitsubishi Space Software Co., Ltd., MSS) and Mr. Hiroaki Ishikawa (MSS) for collaboration in developing the database system, and to Mr. Hiroki Matsumoto (Nagaoka University of Technology) for assistance in programming an earlier version of the clustering programs. The authors are also grateful to Dr. Miyuki Ishimura (AIST), Dr. Hisayuki Morii (AIST), and Dr. Kentaro Tomii (AIST) for helpful comments on the specification of the search capabilities and web interface. We thank the Tsukuba Advanced Computing Center in AIST for the use of its computer facilities. This work was funded in part by grants from the New Energy and Industrial Technology Development Organization (NEDO) of Japan and the Japan Science and Technology Agency (JST).

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

1. Chou PY, Fasman GD (1974) Biochemistry 13:211. doi:10.1021/bi00699a001
2. Garnier J, Osguthorpe DJ, Robson B (1978) J Mol Biol 120:97. doi:10.1016/0022-2836(78)90297-8
3. Blundell TL, Sibanda BL, Sternberg MJ et al (1987) Nature 326:347. doi:10.1038/326347a0
4. Jones DT, Taylor WR, Thornton JM (1992) Nature 358:86. doi:10.1038/358086a0
5. Sonnhammer EL, Eddy SR, Durbin R (1997) Proteins 28:405. http://dx.doi.org/10.1002/(SICI)1097-0134(199707)28:3<405::AID-PROT10>3.0.CO;2-L
6. Schaffer AA, Wolf YI, Ponting CP et al (1999) Bioinformatics 15:1000. doi:10.1093/bioinformatics/15.12.1000
7. Sawada Y, Honda S (2006) Biophys J 91:1213. doi:10.1529/biophysj.105.076661
8. Wang G, Dunbrack RL Jr (2003) Bioinformatics 19:1589. doi:10.1093/bioinformatics/btg224
9. Richards JA, Jia X (1999) Remote sensing digital image analysis. Springer-Verlag, New York
10. Kabsch W, Sander C (1983) Biopolymers 22:2577. doi:10.1002/bip.360221211
11. Sayle RA, Milner-White EJ (1995) Trends Biochem Sci 20:374. doi:10.1016/S0968-0004(00)89080-5
12. Murzin AG, Brenner SE, Hubbard T et al (1995) J Mol Biol 247:536
13. Orengo CA, Michie AD, Jones S et al (1997) Structure 5:1093. doi:10.1016/S0969-2126(97)00260-8
14. Holm L, Ouzounis C, Sander C et al (1992) Protein Sci 1:1691
15. Espadaler J, Fernandez-Fuentes N, Hermoso A et al (2004) Nucleic Acids Res 32:D185. doi:10.1093/nar/gkh002
16. Bystroff C, Baker D (1998) J Mol Biol 281:565. doi:10.1006/jmbi.1998.1943
17. Yang AS, Wang LY (2002) Bioinformatics 18:1650. doi:10.1093/bioinformatics/18.12.1650
18. Yang AS, Wang LY (2003) Bioinformatics 19:1267. doi:10.1093/bioinformatics/btg151
19. Tendulkar AV, Joshi AA, Sohoni MA et al (2004) J Mol Biol 338:611. doi:10.1016/j.jmb.2004.02.047
20. Aurora R, Rose GD (1998) Protein Sci 7:21
21. Sibanda BL, Thornton JM (1991) Methods Enzymol 202:59. doi:10.1016/0076-6879(91)02007-V
22. Honda S, Yamasaki K, Sawada Y et al (2004) Structure 12:1507. doi:10.1016/j.str.2004.05.022
23. Anonym (2002) Nat Struct Biol 9:713