Journal of Biomolecular Structure & Dynamics, ISSN 0739-1102
Volume 28, Issue Number 3, (2010)

SMpred: A Support Vector Machine Approach to Identify Structural Motifs in Protein Structure Without Using Evolutionary Information

http://www.jbsdonline.com

Ganesan Pugalenthi[1]
Krishna Kumar Kandaswamy[2, 3]
P. N. Suganthan[4,*]
R. Sowdhamini[5]
Thomas Martinetz[2]
Prasanna Kolatkar[1,*]

[1]Laboratory of Structural Biochemistry, Genome Institute of Singapore, 60 Biopolis Street, Singapore 138672
[2]Institute for Neuro- and Bioinformatics, University of Lübeck, 23538 Lübeck, Germany
[3]Graduate School for Computing in Medicine and Life Sciences, University of Lübeck, 23538 Lübeck, Germany
[4]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, 639798
[5]National Centre for Biological Sciences, UAS-GKVK campus, Bellary Road, Bangalore 560 065, India

Abstract

Knowledge of three dimensional structure is essential to understand the function of a protein. Although the overall fold is made from the whole details of its sequence, a small group of residues, often called as structural motifs, play a crucial role in determining the protein fold and its stability. Identification of such structural motifs requires sufficient number of sequence and structural homologs to define conservation and evolutionary information. Unfortunately, there are many structures in the protein structure databases have no homologous structures or sequences. In this work, we report an SVM method, SMpred, to identify structural motifs from single protein structure without using sequence and structural homologs. SMpred method was trained and tested using 132 proteins domains containing 581 motifs. SMpred method achieved 78.79% accuracy with 79.06% sensitivity and 78.53% specificity. The performance of SMpred was evaluated with MegaMotifBase using 188 proteins containing 1161 motifs. Out of 1161 motifs, SMpred correctly identified 1503 structural motifs reported in MegaMotifBase. Further, we showed that SMpred is useful approach for the length deviant superfamilies and single member superfamilies. This result suggests the usefulness of our approach for facilitating the identification of structural motifs in protein structure in the absence of sequence and structural homologs. The dataset and executable for the SMpred algorithm is available at http://www3.ntu.edu.sg/home/EPNSugan/index_files/SMpred.htm.

Key words: Protein folding; Structural motifs; Support vector machine; Fingerprint; Protein function.

Introduction

The overall fold provides suitable scaffold for a protein to perform its biological function. According to the Anfinsen hypothesis, the information necessary to achieve 3d structure of a protein in a given environment is contained in its amino acid sequence (1). Currently with the extraordinary upsurge in computational hardware and tools, determinations of protein structure and function from sequence and evolutionary data by modeling have become very routines (2-13). Previous studies revealed that many protein domains adopt the same fold structures even if they have statistically insignificant sequence similarity (14, 15). This indicates the existence of a small group of residues that play crucial roles in determining the fold and its stability, although the overall fold is made from the whole details of its sequence (16, 17). These residues, often termed as structural motifs, are conserved during evolution in both sequence and structure and may form the common structural core by maintaining a particular spatial pattern (16, 18, 19). Identification of such structural motifs is potentially useful in protein structure prediction, protein engineering, modelling experiments, mutation studies, distant homology detection, *etc* (20-23).

[*]E-mail: EPNSugan@ntu.edu.sg

Sequentially conserved residues, often termed as sequence motifs, are useful in understanding the conservational variation and have been successfully linked to functionally important sites indicating higher selection pressure on them (24-26). A common approach to identify such sequentially conserved motifs is to measure the amino acid sequence conservation from multiple sequence alignments of evolutionary related protein sequences, based on the assumption that they are relatively conserved during evolution (24, 26, 27). Several algorithms and databases have been developed for discovering the sequentially conserved motifs and scanning the sequence database using motifs (22, 24, 28-30).

Unlike sequentially conserved motifs, identification of structural motif is not an easy task as it has two major issues. The first one is the requirement of sufficient number of structural homologs to define the conservation of structural features. Many protein structures reported in the protein structure database do not have sufficient number of homologous sequences and structures (18). The second one is the quality of alignment (31, 32). The accuracy of structural motif identification depends upon the quality of the alignments which drops when more structures are aligned.

There have been several methods reported in the recent past for the identification of structural motifs from protein structures. PROMOTIF (33), SPASM (34), Spratt (35), DAVROS (36) are some of the methods that identify and analyze structural motifs for protein structures related at the family level. SMotif server identifies set of structural motifs from structurally aligned protein structures by examining the conservation of amino acid preference and other important structural features like secondary structural content, hydrogen bonding pattern and residue packing (37). Recently a neural network method has been reported to identify structurally conserved residues from a single protein structure using homologous sequences and high quality multiple sequence alignment (38). The importance of structural motifs is further underscored by MegaMotifBase database which provides a compilation of structural motifs related at the family and superfamily level (39).

Although there have been efficient methods to detect structural motifs, none of these methods are specifically dedicated to the identification of structural motifs from a single protein structure without using an evolutionary information derived from homologous sequences and/or structures. In this work, we present an SVM approach, SMpred, to identify structural motifs from a single protein structure in the absence of homologous sequences and/or structures.

Materials and Methods

Datasets

The dataset for this work was obtained from MegaMotifBase database (39). MegaMotifBase provides a comprehensive collection of structural motifs for 1194 superfamilies. Structural motifs for each superfamily in this database was identified using SMotif algorithm from the structurally aligned superfamily members by measuring the conservation of sequence and structural features such as secondary structural content, hydrogen bonding pattern and residue packing. In addition, this database provides structural motifs for the individual structure by consulting the structural alignments.

In this work, we considered 132 protein domains from 132 superfamilies that contain at least five structural members (Please see supplementary material). Each protein domain has less than 20% sequence identity with other protein domains. Structural motifs for each structure were obtained from the MegaMotifBase database. Positive dataset was constructed using 581 structural motifs containing 1880 residues. The minimum and maximum length of the motifs is 3 residues and 22 residues respectively. Remaining 10683 residues that reside in the non-motif regions were considered for the negative dataset.

Training dataset: 1307 residues were randomly selected from 1880 residues for the positive dataset. Equal number of non-motif residues randomly selected from the negative dataset.

Test dataset: Test dataset consists of remaining 573 residues from the positive dataset and 9376 non-motif residues from the negative set.

Evaluation dataset: We created another dataset of 188 proteins containing 1161 structural motifs that are not present in training and testing dataset. Structural motifs for each protein were obtained from MegaMotifBase database (39). The minimum and maximum length of the motifs is 3 residues and 19 residues respectively. The performance of SMpred was evaluated with MegaMotifBase using this dataset.

Features

Each residue in the dataset is represented by 100 features (Please see supplementary material). For each residue, spatial neighbors were identified from the protein structure. Spatial neighbors were defined as residues that are present within 5A distance from a given residue in the 3d structure (40). The details of the each feature used in this study are mentioned below.

Amino Acid Type and Structural Features: For each residue, amino acid type is represented in the form of binary variables (0 or 1). Structural features such as solvent accessibility, secondary structures, hydrogen bonds and residue compactness were computed from the individual protein structure using the JOY package (41).

Frequency of Amino Acids and Functional Group in Spatial Neighbors: For each residue, amino acid composition was computed from its spatial neighbours. In addition, we categorized 20 amino acids into 10 functional groups based on the presence of side chain chemical group such as phenyl (F/W/Y), carboxyl (D/E), imidazole (H), primary amine (K), guanidino (R), thiol (C), sulfur (M), amido (Q/N), hydroxyl (S/T) and nonpolar (A/G/I/L/V/P) (40). Frequencies of 10 functional groups in the spatial neighbours were calculated for each residue.

Structural Features in Spatial Neighbors: The content of structural features such as secondary structure, hydrogen bond, residue compactness and solvent accessibility were computed from spatial neighbors of each residue in the dataset.

Physicochemical Properties: 12 physico chemical properties were obtained from AAINDEX database (42). The selected physico-chemical properties include molecular weight, hydrophobicity, hydrophilicity, hydration potential, refractivity, average accessible surface area, free energy transfer, flexibility, residue volume, mutability, melting point, optical activity, side chain volume, polarity, and isoelectric points. For each residue, physico-chemical property value was calculated as the sum of physico-chemical property value for all spatial neighbors of a given residue, divided by the number of spatial neighbors.

SVM Binary Classification

Support Vector Machine (SVM) has been successfully used to solve various problems in Bioinformatics. For example, SVM has been used in predicting protein subcellular location (43), membrane protein type (44, 45), protein structural class (46), specificity of GalNAc-transferase (47), HIV protease cleavage sites in protein (48), beta turn types (49), protein signal sequences and their cleavage sites (50), alpha turn types (51), B-cell epitope (52), protein structural classes (53) catalytic triads of serine hydrolases (54). SVM is a supervised machine learning method which is based on the statistical learning theory (55). When used as a binary classifier, SVM constructs a hyperplane in a kernel feature space that acts as the decision surface between the two classes. This is achieved by maximizing the margin of separation between the hyperplane and those points nearest to it. The details of

the formulation and solution methodology of SVM for binary classification task can be found elsewhere (55, 56). Only relevant details are provided here.

Let $x_i \in R^n$, $i = 1, 2,\ldots,n$ be input training instances and $y_i \in \{+1, -1\}$ be their corresponding target class. Let N be the total number of instances.

Decision on class affiliation can be made depending upon the sign of the function $f(x)$:

$$f(x) = \sum_{i=1}^{m} y_i \alpha_i K(x_i, x_j) + b \quad [1]$$

where m is the number of input instances having non-zero positive values of the Lagrange multipliers (α_i)(usually a subset of n known as the support vectors) obtained by solving a quadratic optimization problem and b is the bias term.

K (x_i, x_j) denotes the kernel function. In present study, simulations were performed using the RBF function, defined by

$$K(x_i, x_j) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right) \quad [2]$$

where γ is the RBF kernel parameter

All the computations were performed using LIBSVM – 2.81 standard package (57). Various user-defined parameters *i.e*., kernel parameter, γ and regularization parameter, C were optimized employing a grid search.

In order to identify the prominent features that separate the positive and negative classes, we used Info Gain algorithm with the ranker method. This method was implemented using Weka 3.5 (58). We calculate the information gain for each feature, and rank them according to this measure, which indicates the gain of information.

Performance Evaluation of SVM

A 10-fold cross-validation experiment was adopted to evaluate the performance of SVM models (52). The dataset was randomly divided into 10 subsets. The training and testing were carried out 10 times for each model using one distinct set for testing and the remaining nine for training. The performance of the model was reported as the average performance over 10 sets.

For the purpose of assessing the generalization capabilities, we calculated the accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and Matthew's Correlation Coefficients (MCC).

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)} \quad [3]$$

$$\text{Sensitivity} = \frac{TP}{(TP + FN)} \quad [4]$$

$$\text{Specificity} = \frac{TN}{(TN + FP)} \quad [5]$$

$$\text{PPV} = \frac{TP}{TP + FP} \quad [6]$$

$$\text{NPV} = \frac{TN}{TN + FN} \quad [7]$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TN + FN)(TP + FN)(TN + FP)}} \quad [8]$$

where TP-True Positive, FP-False Positive, TN-True Negative, FN-False Negative, PPV-Positive Predictive Value, NPV-Negative Predictive Value.

The Matthew's correlation coefficient ranges from $-1 \leq MCC \leq 1$. A value of MCC = 1 indicates the best possible prediction while MCC = -1 indicates the worst possible prediction (or anti-correlation). Finally, MCC = 0 would be expected for a random prediction scheme.

Results and Discussion

Classification Result for SMpred

The SMpred classifier was trained using the dataset containing 1307 residues that reside in motif region (positive samples) and 1307 residues residing in non-motif regions (negative samples) while the performance of the classifier was tested on the dataset containing remaining 573 residues residing in motif region (positive samples) and 9376 residues residing in non-motif regions.

Our method achieved 78.79% accuracy with 79.06% sensitivity and 78.53% specificity in the test dataset using all features. We applied a feature reduction protocol to eliminate the redundant features. As seen in Table I, feature selection (reduction) generally does not deteriorate the classification performance much until the number of features decreases to 10. It can be seen in Table I that the usage of smaller number of features only leads to a very small decrease in the specificity rate but overall accuracy and sensitivity rate is significantly improved. With 10 features, we obtained 80.80% accuracy with 84.47% sensitivity and 77.14% specificity.

Identification of structural motifs by SMpred involves three steps. (i) SMpred accepts protein structure in the PDB format as an input. (ii) It computes physicochemical properties, sequence and structural features for each residue and its spatial neighbors. (iii) It identifies structural motifs and displays the motifs in a convenient tabular format. The steps involved in the structural motif identification is shown in Figure 1.

Evaluation of SMpred with MegaMotifBase

We evaluated the performance of SMpred with the MegaMotifBase database using 188 protein structures. 1161 structural motifs from 188 structures were obtained from MegaMotifBase. Out of 1161 motifs, SMpred correctly identified 1503 motifs.

Table I
Classification results achieved on the test dataset using different feature subsets.

Feature subset	Sensitivity (%)	Specificity (%)	MCC	PPV	NPV	Accuracy (%)
10	84.47	77.14	0.617	78.69	83.20	80.80
20	83.07	80.98	0.640	81.36	82.67	82.02
30	82.37	81.50	0.638	81.66	82.18	81.93
40	81.33	79.58	0.609	79.93	80.96	80.45
50	77.31	79.76	0.570	79.24	77.81	78.53
60	80.80	80.63	0.614	80.66	80.73	80.71
70	78.36	79.06	0.574	78.91	78.47	78.70
80	77.84	78.36	0.562	78.24	77.91	78.09
90	78.01	78.18	0.562	78.14	78.01	78.09
All features	79.06	78.53	0.575	78.64	78.91	78.79

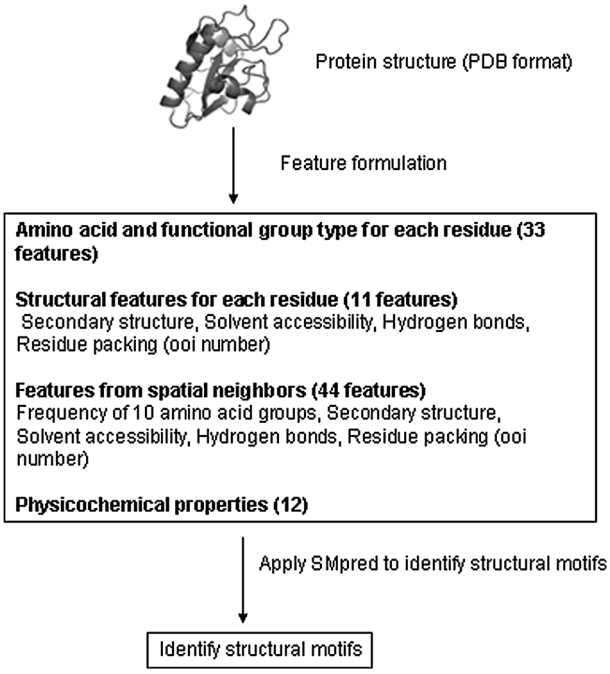

Figure 1: Steps involved in the identification of structural motifs in protein structure.

In 93 proteins, SMpred correctly identified all the motifs listed in the MegaMotifBase. More than 70% of the motifs were identified in the remaining 95 proteins. Table II shows the performance of MegaMotifBase and SMpred in 10 superfamilies. The complete list of 188 proteins is provided in supplementary material.

Motif Identification for Length-deviant Superfamilies

Length variation in proteins has been the object of several analysis and many groups have performed independent studies on the domain length variation (59). It has been shown that large length difference exists between the members of a protein structural domain superfamily and such expansion correlates with the presence of introns and accretion of functional and structural motifs (59). It has also been suggested that additional lengths may have functional or a structural role such as thermal stability, subunit interaction, substrate specificity, *etc*., (59, 60).

Most algorithms that identify motifs from the alignments consider only those regions that are common to all members in the alignment (*i.e*., gapless column in the alignment) and ignore the additional length. The best example for the length-deviant superfamily is Ferritin like superfamily (SCOP superfamily code: 47240) which has an average size of 250 residues (32). This superfamily includes a small domain such as ruberythrin (Domain code: 1dvba1) which has 147 residues and larger domain such as methane monooxyganenase hydroxylase/MMO (PDB code: 1mty; Chain D) which has 512 residues. Structure guided alignment generated for this superfamily shows that although individual domains vary in size between 147 and 512 residues, only 127 residue sites (residues 101 to 255 in case of 1mtyd) are common to all members of this superfamily due to many insertions and deletions.

In MegaMotifBase, 8 structural motifs were reported for 1mtyd (61). We observed that these 8 motifs are located between residues 101 and 255. The extra length in this protein due to insertion is completely ignored in MegaMotifBase. SMpred when applied to this protein identified 15 motifs (Table III). Out of 15 motifs, 7

Table II

Evaluation of SMpred with MegaMotifBase. This table shows number of motifs identified by MegaMotifBase and SMpred in 10 multimember superfamilies. Common motif is a number of motifs that are recognized by both methods.

Domain code	Superfamily	No of Motifs in MegaMotifBase	No of Motifs identified by SMpred	Common Motifs
1lyqa-	E set domains	2	5	2
1jv2a2	Integrin domains	3	4	3
1qfja1	Riboflavin synthase domain-like	5	6	5
1pmla-	Kringle-like	2	3	2
1ayj--	Scorpion toxin-like	3	4	3
1e4ea2	Glutathione synthetase ATP-binding domain-like	8	9	8
1b87a-	Acyl-CoA N-acyltransferases	6	9	6
1i52a-	Nucleotide-diphospho-sugar transferases	8	11	8
1dxxa1	Calponin-homology domain, CH-domain	4	5	4
1hzpa2	Thiolase-like	6	8	6

Table III

Structural motifs in methane monooxyganenase hydroxylase/MMO (PDB code: 1mty; Chain D). "Yes" indicates the presence of motif in the additional length regions and "No" indicates the presence of motif in the regions which are conserved across the superfamily members.

No	Motifs identified by SMpred	Motifs reported in MegaMotifBase	Occurrence in additional length
1	81-84	-	Yes
2	98-125	110-122	No
3	137-142	133-151	No
4	178-181	-	No
5	186-188	-	No
6	197-222	199-202; 204-206;220-222	No
7	231-241	229-232;235-239	No
8	249-254	252-254	No
9	270-291	-	Yes
10	304-309	-	Yes
11	346-348	-	Yes
12	354-356	-	Yes
13	437-439	-	Yes
14	446-448	-	Yes
15	466-468	-	Yes

motifs are present in the region between 101 and 255. Remaining 8 motifs occur in the extra length region (Figure 2).

We analyzed the identified motifs to understand their possible roles in the extra length regions. The crystal structure of 1mty is composed of six subunits namely B, C, D, E, G and H. Subunit D interacts with subunits B and G (61). Our analysis shows that these additional motifs may have significant role in the subunit interaction. For example, motifs 270-291 and 466-468 play a role in the subunit interaction by forming hydrogen bonds at the interface. The observed hydrogen bonds include: hydrogen bond between Asn272 (motif 270-291) and Tyr148 (Chain G), hydrogen bond between Cys466 (motif 466-468) and Asp71 (Chain B) and hydrogen bond between Gln467 (motif 466-468) and Asp50 (Chain G). In addition, a hydrophobic interaction was observed between Val438 in chain D (motif 437-439) and Val164 in chain G. Previous study suggested that residues 348-363 serve as building block fragments which are critical for achieving the native fold (40). As seen in Table III, SMpred identified two motifs (motifs 346-348 and 344-356) that fall between residues 348 and 363. This result shows that the motifs identified by SMpred in extra length region could play an important role in the subunit interaction and fold stability.

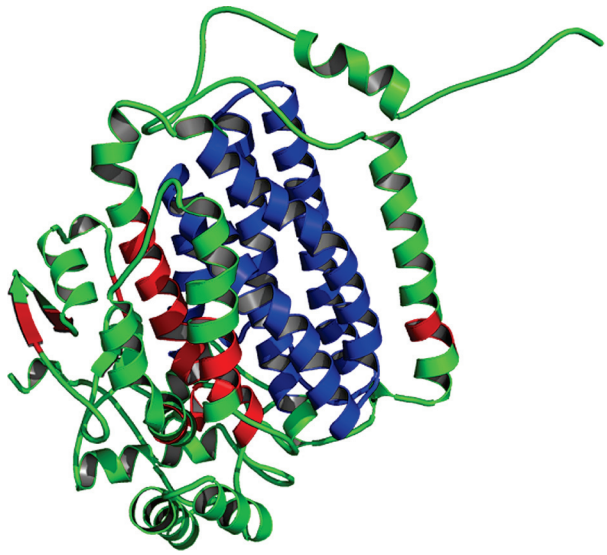

Figure 2: Motifs identified by SMpred in methane monooxyganenase hydroxylase/MMO (PDB code: 1mty; Chain D). Regions which are conserved across all superfamily members are shown in blue. Additional length is shown in green. Motifs identified by SMpred in the additional length are marked in red.

Motif Identification for Single Member Superfamilies by SMpred

A majority of the entries in the protein structural database has no structural homologs. For example, out of 1194 superfamilies reported in PASS2 database (32), 544 superfamilies are single member superfamilies for which it is hard to identify motif due to the lack of structural homologs. MegaMotifBase provides structural motifs for each single member superfamily in the form of sequence-structural templates. These motifs were identified from the sequentially conserved segments that have high content of structural features such as secondary structure, hydrogen bond, solvent inaccessible residues and residue packing. It should be noted that the structural features were derived from single protein structure and there is no guarantee that these features are conserved. Therefore, motif derivation is error prone.

Since SMpred is capable of identifying structural motifs without using sequence or structural homologs, this method can be used to identify motifs for single member

Table IV
Performance of SMpred in single member superfamily (Superfamily: N-acetylmuramoyl-L-alanine amidase-like; SCOP superfamily code: 55846; Domain code; 1LBA--)

No	SMpred	SMotif	MegaMotifBase
1	13-19	14-18	-
2	31-33	-	-
3	46-50	47-50	-
4	-	-	66-88
4	77-82	76-81	76-84
5	102-107	103-106	-
6	119-121	-	-

superfamilies. To assess the performance of SMpred, we selected N-acetylmuramoyl-L-alanine amidase-like superfamily (SCOP superfamily code: 55846) containing bacteriophage T7 lysozyme domain (domain code: 1LBA--) as a structural member. This superfamily has been reported as a single member superfamily in MegaMotifBase database which corresponds to SCOP 1.63 release (17). As shown in Table IV, MegaMotifBase listed two structural motifs (residue numbers in PDB: 66-88 and 76-84) for this protein whereas SMpred identified 6 structural motifs. Out of 6 motifs, only one motif overlaps with MegaMotifBase definition.

Fortunately, this superfamily has expanded with 8 additional structural members in the recent SCOP database (SCOP 1.75) (17). This gives us an opportunity to verify whether the motifs identified by SMpred are significant or not. We aligned T7 lysozyme domain with 8 additional structural members using STAMP program (63) which superposes the protein structures and subsequently generates multiple structural alignment. SMotif algorithm when applied on the structural alignment identified 4 structural motifs. As seen in Table IV, SMpred correctly identified all the four motifs from single protein structure. This shows that SMpred is able to capture conserved structural features from single protein structure without the knowledge of homologous sequences and structures.

We performed brief analysis to understand the possible roles of the identified motifs. The bacteriophage T7 lysozyme is a bifunctional protein that cuts amide bonds in the bacterial cell wall and binds to and inhibits transcription by T7 RNA polymerase (64). As shown in Table IV, motifs 13-19 and 46-50 were identified by SMotif and SMpred. It has been shown that His17 and Tyr46 are required for amidase activity. In addition, Tyr46 may also have significant structural role (64). Further, SMpred identified two motifs (motifs 31-33 and 119-121) which are not recognized by SMotif and MegaMotifBase. Previous study has suggested that Arg30, Glu31 and Arg33 play roles in the inhibition of T7 RNA polymerase (64). This result shows the capability of SMpred in recognizing the significant structural motifs from single protein structure in the absence of sequence and structural homologs.

An Example: Sedolisin

Although the structural motifs are generally associated with structural role, some of them might have functional role. For example, structural motifs play both structural and functional roles in sedolisin. Sedolisin (pdbcode 1ga6) belongs to a family of carboxyl serine peptidases with a unique catalytic triad consisting of Glu80, Asp84 and Ser287 (65). The structure of sedolisin comprised a single domain consisting of a 7 stranded parallel beta sheet flanked by a number of helices.

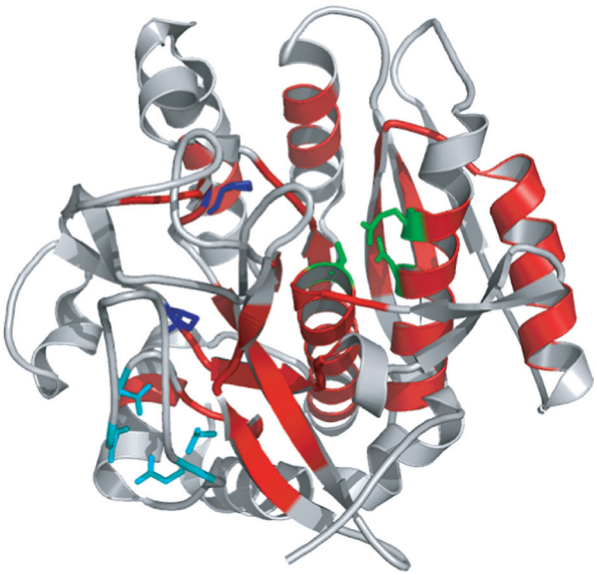

Figure 3: Structural motifs identified by SMpred in Sedolisin (pdbcode 1ga6). Structural motifs are shown in cyan. The catalytic triad (Glu80, Asp84 and Ser287) is shown green with sticks representation. Two prolines (blue) and metal binding residues (cyan) are shown in sticks.

SMpred identified 14 structural motifs in Sedolisin (30-49; 79-92; 99-102; 114-119; 129-135; 151-154; 163-167; 188-194; 199-206; 216-222; 260-267; 284-300; 314-316 and 353-356) (Figure 3). It has been reported that sedolisin structure has two proline residues, Pro192 and Pro260, in areas crucial to the preservation of the fold (61). Both prolines (motifs 188-194 and 284-300) were detected by SMpred. Two of the identified structural motifs (motifs 79-92 and 284-300) contain catalytic residues Glu80, Asp84 and Ser287. Glu80 forms hydrogen bond with Ser287 and also interact, through its side chain, with Asp84. Previous study reported that mutation in Asp84 leads to a 10^4 fold decrease in catalytic activity whereas mutation in Ser287 leads to complete loss of catalytic activity (65). This suggests that these two structural motifs play catalytic role in sedolisin. In some cases, structural motifs provide optimal environment for the protein to perform its function. For example, sedolisin has five metal binding residues (328,329,344,346

and 348) which are located in flexible loop regions. It is a more reliable assumption that motif 314-316 and motif 353-356 could play a role in orientating these residues suitable for metal binding.

Conclusion

Structural motifs play crucial roles in protein structure and function. We presented an SVM method, SMpred, to identify structural motifs from single protein structure without using sequence or structural homologs and their alignments. The performance of SMpred was compared with MegaMotifBase database. Our analysis showed that SMpred is a suitable method to identify structural motifs in length deviant superfamily. Successful recognition of structural motifs in single member superfamily showed that SMpred is a useful approach to identify structural motifs for proteins that have no sequence and structural homologs. The supplementary material and SMpred codes are available at http://www3.ntu.edu.sg/home/EPNSugan/index_files/SMpred.htm.

Acknowledgments

GP and PK acknowledge the financial support and infrastructure offered Genome Institute of Singapore. KKK acknowledges the support by the Graduate School for Computing in Medicine and Life Sciences funded by Germany's Excellence Initiative [DFG GSC 235/1].

References

1. C. B. Anfinsen. *Science 181*, 223–230 (1973).
2. M. Parthiban, M. B. Rajasekaran, S. Ramakumar, and P. Shanmughavel. *J Biomol Struct Dyn 26,* 535-547 (2009).
3. K. Sujatha, A. Mahalakshmi, D. K. Y. Solaiman, and R. Shenbagarathai. *J Biomol Struct Dyn 26*, 771- 779 (2009).
4. R. Chattopadhyaya and A. Pal. *J Biomol Struct Dyn 25*, 357-371 (2008).
5. D. Josa, E. F. F. da Cunha, T. C. Ramalho, T. C. S. Souza, and M. S. Caetano. *J Biomol Struct Dyn 25*, 373-376 (2008).
6. J. Dasgupta and J. K. Dattagupta. *J Biomol Struct Dyn 25*, 495-503 (2008).
7. A. Bagchi and T. C. Ghosh. *J Biomol Struct Dyn 25*, 517-523 (2008).
8. S. Subramaniam, A. Mohmmed, and D. Gupta. *J Biomol Struct Dyn 26*, 473-479 (2009).
9. S. Suma Mohan, J. J. P. Perry, N. Poulose, B. G. Nair, and G. Anilkumar. *J Biomol Struct Dyn 26*, 455-464 (2009).
10. R. Vinekar and I. Ghosh. *J Biomol Struct Dyn 26*, 741-754 (2009).
11. S. Mishra. *J Biomol Struct Dyn 27*, 283-291 (2009).
12. U. B. Sonavane, S. K. Ramadugu, and R. R. Joshi. *J Biomol Struct Dyn 26*, 203-214 (2008).
13. S. K. Singh, S. R. Choudhury, S. Roy, and D. N. Sengupta. *J Biomol Struct Dyn 26*, 235-245 (2008).
14. C. Chothia. *Nature 357*, 543-544 (1992).
15. C. A. Orengo, D. T. Jones, and J. M. Thornton. *Nature 372*, 631-634 (1994).
16. S. Chakrabarti, K. Venkatramanan, and R. Sowdhamini. *Protein Eng 16(11),* 791-3 (2003).
17. T. J. Hubbard, A. G. Murzin, S. E. Brenner, and C. Chothia. *Nucleic Acids Res 25(1)*, 236-239 (1997).
18. S. Chakrabarti, G. Manohari, G. Pugalenthi, and R. Sowdhamini. *In Silico Biol 6(4)*, 311-9 (2006).
19. G. Pugalenthi, P. N. Suganthan, R. Sowdhamini, and S. Chakrabarti. *Bioinformatics 23(5)*, 637-638 (2007).
20. S. Chakrabarti and R. Sowdhamini. *FEBS Lett 569*, 31-36 (2004).
21. S. Chakrabarti, J. John, and R. Sowdhamini. *J Mol Model 10(1)*, 69-75 (2004).
22. S. Chakrabarti, A. P. Anand, N. Bhardwaj, G. Pugalenthi, and R. Sowdhamini. *Nucleic Acids Res 33*, W274-6 (2005).
23. R. Unger and J. L. Sussman. *Journal of Computer-Aided Molecular Design 7(4)*, 457-472 (1993).
24. T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble. *Nucleic Acids Res 37*, W202-8 (2009).
25. M. A. Saqi and M. J. Sternberg. *Protein Eng 7*, 165-171 (1994).
26. W. R. Taylor. *J Theor 119(2)*, 205-18 (1986).
27. T. Marschall and S. Rahmann. *Bioinformatics 25(12)*, i356-64 (2009).

28. M. C. Frith, N. F. Saunders, B. Kobe, and T. L Bailey. *PLoS Comput Biol 4(4)*, e1000071 (2008).
29. N. Hulo, A. Bairoch, V. Bulliard, L. Cerutti, E. De Castro, P. S. Langendijk-Genevaux, M. Pagni, and C. J. Sigrist. *Nucleic Acids Res 34*, D227-D230 (2006).
30. J. G. Henikoff, E. A. Greene, S. Pietrokovski, and S. Henikoff. *Nucleic Acids Res 28*, 228-230 (2000).
31. G. Pugalenthi, K. Tang, P. N. Suganthan, and S. Chakrabarti. *Bioinformatics 25(2)*, 204-210 (2009).
32. A. Bhaduri, G. Pugalenthi, and R. Sowdhamini. *BMC Bioinformatics 2*, 5-35 (2004).
33. E. G. Hutchinson and J. M. Thornton. *Protein Sci 5*, 212-220 (1996).
34. G. J. Kleywegt. *J Mol Biol 285*, 1887-1897 (1999).
35. I. Jonassen, I. Eidhammer, D. Conklin, and W. R Taylor. *Bioinformatics 18*, 362-367 (2002).
36. K. B. Murray, W. R. Taylor, and J. M. Thornton. *Proteins 57*, 365-380 (2004).
37. G. Pugalenthi, P. N. Suganthan, R. Sowdhamini, and S. Chakrabarti. *Bioinformatics 23(5)*, 637-638 (2007).
38. G. Pugalenthi, K. Tang, P. N. Suganthan, and S. Chakrabarti. *Bioinformatics 25(2)*, 204-210 (2009).
39. G. Pugalenthi, P. N. Suganthan, R. Sowdhamini, and S. Chakrabarti. *Nucleic Acids Res 36*, D218-21 (2008).
40. G. Pugalenthi, K. K. Kumar, P. N. Suganthan, and R. Gangal. *Biochem Biophys Res Commun 367(3)*, 630-4 (2008).
41. K. Mizuguchi, C. M. Deane, T. L. Blundell, M. S. Johnson, and J. P. Overington. *Bioinformatics 14*, 617-623 (1998).
42. S. Kawashima, H. Ogata, and M. Kanehisa. *Nucleic Acids Res 27*, 368-369 (1999).
43. K. C. Chou and Y. D. Cai. *J Biol Chem 277*, 45765-45769 (2002).
44. Y. D. Cai, G. P. Zhou, and K. C.Chou. *Biophysical Journal 84*, 3257-3263 (2003).
45. Y. D. Cai, R. Pong-Wong, K. Feng, J. C. H. Jen, and K. C. Chou. *Theoretical Biology 226*, 373-376 (2004).
46. Y. D Cai, X. J. Liu, X. B. Xu, and K. C. Chou. *Computers & Chemistry 26*, 293-296 (2002).
47. Y. D. Cai, X. J. Liu, X. B. Xu, and K. C. Chou. *Peptides 23*, 205-208 (2002).
48. Y. D. Cai, X. J. Liu, X. B. Xu, and K. C. Chou. *J Comput Chem 23*, 267-274 (2002).
49. Y. D. Cai, X. J. Liu, X. B. Xu, and K. C. Chou. *J Pept Sci 8*, 297-301 (2002).
50. Y. D. Cai, S. Lin, and K. C.Chou. *Peptides 24*, 159-161(2003).
51. Y. D. Cai, K. Feng, Y. X. Li, and K. C. Chou. *Peptides 24*, 629-630 (2003).
52. J. Chen, H. Liu, J. Yang, and K. C. Chou. *Amino Acids 33(3)*, 423-8 (2007).
53. X. D. Sun and R. B. Huang. *Amino Acids 30*, 469–475 (2006).
54. Y. D Cai, G. P. Zhou, C. H. Jen, S. L. Lin, and K. C. Chou. *Journal of Theoretical Biology 228*, 551-557 (2004).
55. C. Cortes and V. Vapnik. *Machine Learning 20*, 273-297 (1995).
56. V. Vapnik. New York, Wiley (1998).
57. C. C. Chang and C. J. Lin. (2001) www.csie.ntu.edu.tw/~cjlin/libsvm.
58. I. H. Witten, and E. Frank. Morgan Kaufmann, San Francisco, CA (2000).
59. S. Sandhya, S. S. Rani, B. Pankaj, M. K. Govind, B. Offmann, N. Srinivasan, and R. Sowdhamini. *PLoS One 4(3)*, e4981 (2009).
60. S. Sandhya, B. Pankaj, M. K. Govind, B. Offmann, N. Srinivasan, and R. Sowdhamini. *BMC Struct Biol 31*, 8-28 (2008).
61. A. C. Rosenzweig, H. Brandstetter, D. A. Whittington, P. Nordlund, S. J. Lippard, and C. A. Frederick. *Proteins 29(2)*, 141-52 (1997).
62. A. Barzilai, S. Kumar, H. Wolfson, and R.Nussinov. *Proteins 56(4)*, 635-49 (2004).
63. R. B. Russell and G. J. Barton. *J Mol Biol 244*, 332–350 (1994).
64. X. Cheng, X. Zhang, J. W. Pflugrath, and F. W. Studier. *Proc Natl Acad Sci U S A 91(9)*, 4034-8 (1994).
65. A. Wlodawer, M. Li, Z. Dauter, A. Gustchina, K. Uchida, H. Oyama, B. M. Dunn, and K. Oda. *Nat Struct Biol 8(5)*, 442-6 (2001).

Date Received: March 31, 2010

Communicated by the Editor Ramaswamy H. Sarma