ISSN 2278-3091
Volume 8, No.1.3, 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse3981.32019.pdf
https://doi.org/10.30534/ijatcse/2019/3981.32019

Local Protein Structures to Bridge Sequence-Structure Knowledge

[1]Rohayanti Hassan, [1]Asraful Syifaa' Ahmad, [2]Mahmud Imrona, [3]Shahreen Kasim
[1]School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia, rohayanti@utm.my
[2]School of Computing, Telkom University, 40257 Bandung, West Java, Indonesia
[3]Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor, Malaysia

ABSTRACT

Protein sequences can be classified based on their structure similarity and/or common evolutionary origin called structural class. Information on structural class is readily available, easing the protein structure and protein function probing. SCOP and CATH are two prominent classification schemes used to assign the structural class of proteins. Both schemes determine the structural class manually base on known protein tertiary structures. However, the quantity of known protein sequences is growing exponentially with respect to the quantity of known tertiary proteins structures. Although SCOP and CATH are examples of well-established databases that contain more reliable information of structural class, yet the lack of known structural class of protein due to the laborious wet-lab experimental routine limits the high-throughput structural class assignment. The fact that this is a tedious and time-consuming manually-determined method has further limited the structural class assignment. As a consequence, the assignment of structural class by computational method suffers from the arbitrated statistical infer-ence. Thus, this study aims to provide a structural class prediction method that can acquire the knowledge of local protein structures, derived from known excessive primary sequences, in order to produce high-throughput sequence-structure class assignment instead of the laborious experimental based method. This structural class prediction method is termed as SVM-LpsSCPred..

Key words : Protein structural class, local protein structure, support vector machine.

1. INTRODUCTION

Due to the laborious manually-determined schemes, several computational methods has been explored in order to produce high-throughput sequence-structure class assignment. These methods utilized the knowledge of known secondary structure contents and arrangements which is available in a larger quantity and is well-known compared to tertiary structures. The investigation begin with the threshold-based classification method has been used to assign the structural class for corresponding protein sequence [1]-[2]. However, no unified quantitative measurement was used to set those threshold values which in turn lead to arbitrated statistical inference.

Currently, the structural class is predicted using more sophisticated method which basically integrates two mechanisms: firstly, the amino acids of protein sequences are represented by features vector and secondly the features vector is then served into classification method to predict the corresponding structural class. However, it is a challenging task to predict the structural class for protein sequences that is characterized by low-identity to each other. Most related studies are primarily focused on complex features vector. Advanced representations such as merging the amino acid composition with its evolutionary and neighborhood information, pseudo-amino acids that considered the effects of sequence order [3]-[4] and multi composite features [5] resulted in a more accurate prediction.

Meanwhile, structural class based on known domains are listed 110,800 times in the recent SCOP [22] database version 1.75 as stated in June, 2009 and 128,688 counts in version 3.3 release of the CATH [22] database as stated in July, 2009. This shows a huge gap between known sequence and known structural class in which only 1-2% of the sequences can be assigned to the corresponding structural class. However, the knowledge of known structural class from SCOP and CATH are frequently used as a standard of truth for classification method even though both schemes show inconsistent structural class assignments for some protein sequences. This study conducted a preliminary experiment onto RS126 dataset. As depicted in Table 1, SCOP and CATH produced different structural classes' assignment for 13 protein sequences of RS126 dataset. These differing assignments could lead to wrong classes as well as overestimate error.

Thus, local protein structure was introduced to incorporate with SVM and termed as SVM-LpsSCPred in order to bridge the sequence-structure knowledge. By using only a simple features vector, this method can still precisely predict the structural class. This paper is organized as follows. In section 2, the materials and methods used are explained. The experimental results of comparative evaluation proposed in this paper is presented in section 3. Finally, our work of this paper is summarized in the last section.

Table 1: Inconsistent structural class assignment between SCOP and CATH for 13 sequences from RS126

PDB ID	SCOP	CATH	PDB ID	SCOP	CATH
1cdt	mixed	all-β	3hmg	all-β	mixed
1eca	all-α	mixed	4rxn	small	all-β
1il58	mixed	all-α	5lyz	mixed	all-α
1il8a	mixed	all-β	6hir	small	all-β
1sh1	small	all-β	8adh	all-β	mixed
1pyp	all-β	mixed	9wga	mixed	all-β
3ebx	mixed	all-β			

2. MATERIALS AND METHODS

Based on the aforementioned deficiencies in structural class prediction method, this study proposes an improved method designated as SVM-LpsSCPred that aims to overcome the insufficient sequence-structure knowledge due to the lack of protein sequence identity also by only using a simpler feature vector. Figure 1 demonstrates the analogy of structural class prediction by the proposed method. Initially, SCOP and CATH predicted the Azurin Electron Transport protein (PDB ID: 2ccw) as all-β class throughout the tedious manual experimental routine. However, in using the restricted Chou's threshold-based classification method [2] that bases on the whole protein sequences, the class was still unknown.

This study has introduced the use of local protein structures which were derived from the fragmentations of secondary structures. These local protein structures were then represented by a simpler features vector known as dihedral angles. In order to avoid the inconsistent standard of truth, the classes of local protein structures were determined using Chou's threshold-based classification method [2]. Subsequently, Support Vector Machine (SVM) was implemented to predict the structural class. To evaluate the performances of the proposed method, three measurement metrics were used: accuracy (acc), fraction of true positive (tpr) and fraction of false positive (fpr). The success rate (similarity rate) was then evaluated against SCOP and CATH as well as other state-of-the-art methods. In the following section, the detail framework of SVM-LpsSCPred is presented

2.1 Dataset Preparation

Protein sequences used in this study were taken from RS126 dataset [6]. RS126 is a popular benchmark dataset widely used in various structural protein predictions and is still being continuously use. Furthermore, RS126 comprises of low-identity sequences with no sequences sharing more than 24% identity. The low-identity sequences are indeed becoming an interest to many researchers due to their low prediction accuracy.

This study primarily focused on introducing the local protein structure which comprised of local secondary structures fragments, to predict the structural class. The experiments began with the investigation of the effects of different secondary structure assignments to structural class prediction. This secondary structure can be assigned from a given sequence using either secondary structure assignment method (SSAM) or secondary structure prediction method (SSPM)5. SSAMs use atomic coordinate patterns to annotate the secondary structure which is limited to three states: Helix, Strand and Coil. The knowledge of atomic coordinate is however unnecessary for SSPMs. The methods used under SSAM were DSSP [7] and STRIDE [8], both of which uses knowledge-based algorithm11 and are publicly accessible, while NNSSP [10] and PHD [9] were methods used under SSPM. In this case, NNSSP uses Nearest Neighbor [12] algorithm while PHD uses multi-layer Neural Network [13] algorithm.

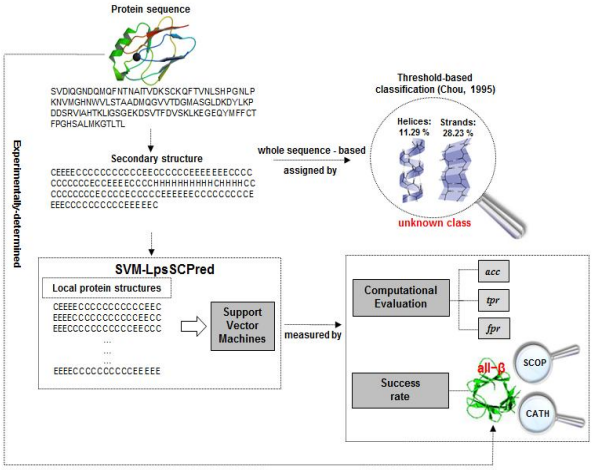

Figure 1: The structural class assignment/prediction for Azurin Electron Transport protein (PDB ID: 2ccw) using SCOP, CATH, Chou's threshold-based classification method [2] and proposed method termed as SVM-LpsSCPred.

The amino acids or residues in every sequences were represented by dihedral angles score, denoted as da_{aa} [14]. This score embodies the protein dihedral angles attribute that could be retrieved from PDB. The latent patterns within dihedral angles had the potential to reveal the hindered structural class owing to the strong inheritance between dihedral angles and secondary structure contents [15]. Secondary structure was folded based on the conformation of protein atom coordinates and backbone, while dihedral angle was used to define the protein backbone. Furthermore, dihedral angles score was chosen as it is not influenced by the lack of protein sequence identity [14].

This study aims to predict the structural class based on the knowledge of 1-dimensional secondary structure contents which is known to be available in a large quantity. The structural class is limited to three states: all-α, all-β and mixed class. At the initial stage, Chou's threshold-based classification method [2] was used to assign the structural class for the local protein structures. However, this method classified the structural class into four groups: all-α, all-β, α+β and α/β. The structural class for α+β and α/β was then grouped into mixed class as no directionality knowledge could be used

to discern both classes was available in the 1-dimensional secondary structure. In determining the success rate, the predicted structural class was compared to structural class that is accessed from SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/) and CATH (http://www.cathdb.info/) online databases.

2.2 Structural Class Prediction Using Local Protein Structures

Each sequence was fragmentized, using sliding window method [16], into local protein structures, *lps*, each of which consisted of dihedral angles score, secondary structure and structural class information. An exhaustive scan for different local protein structures length, *fl*, was made in order to find the values that gave the best results to which the *fl* was then fixed to 19 continuous residues as proposed by [16]. In order to be entered into SVM, each local protein structure was transformed into features vector and class. Features vector was represented by the average of dihedral angles score using the following formula:

$$F_{k,fl} = \sum_{k=1}^{Nlen-fl} \frac{\sum_{l=k}^{k+(fl-1)} lps.da_l}{fl} \quad (1)$$

where *lps.da* denoted the local protein structure of dihedral angles score while *k* and *l* denoted the indices. *Nlen* denoted the number of amino acids in each sequence. As multiclass SVM [17] was implemented, the feature class, *Labk,fl* was denoted as 1 if structural class was all-α, 2 if structural class was all-β or 3 if structural class was mixed. For comparison with other classification methods, similar features vector and class were served into Neural Network (NN) [18], Naïve Bayesian (NB) [19] and K-Nearest Neighbor (KNN) [14]. The structural class of the protein sequence was then determined based on the dominant predicted class of its local protein structures respectively.

3. RESULT AND DISCUSSION

3.1. Improvement on the Success Rate of Structural Class Prediction

By using RS126 sequence dataset, the effectiveness of the proposed method was tabulated (Table 2). Results indicated that the success rate improved from 18.2% to 42.7% compared to the earlier threshold-based classification method2. Compared to SCOP, the best success rate was achieved by DSSP with 80.0% followed by STRIDE with 76.4%, PHD 63.4% and NNSSP 53.6%. Similarly using CATH, DSSP also showed the best success rate with 92.7%., followed by STRIDE, PHD and NNSSP with 90.9%, 71.8% and 55.5% respectively.

Table 2 show that the highest success rate based on the threshold-based classification method [2] for both SCOP and CATH was only 50%. A low success rate of threshold-based classification method [2] was caused by the restrictive minimal or maximal margin in predicting the structural class. To make matters worse, the interspersed and segregated nature of secondary structure lead to the unsteady proportions of Helices and Strands which was the most essential prediction criteria in threshold-based classification [2]. The proposed method succeeded in improving the threshold-based classification method [2] by bridging the sequence-structure knowledge using local protein structures which later SVM exploited to reveal the hindered structural class. On the other hand, compared to SSPMs, SSAMs demonstrated a better success rate compared to SCOP and CATH due to their similar manually-determined method in discerning their particular structural targets [20].

Table 2: An increment of success rate (%) presented by SVM-LpsSCPred compared to threshold-based classification method [2]

Method	Category	Secondary structure	Success rate against SCOP	Success rate against CATH
SVM-LpsSCPred	SSAM	DSSP	80.0	92.7
		STRIDE	76.4	90.9
	SSPM	PHD	63.4	71.8
		NNSSP	53.6	55.5
Threshold-based classification method [2]	SSAM	DSSP	50.0	50.0
		STRIDE	49.1	48.2
	SSPM	PHD	45.5	44.5
		NNSSP	34.5	33.6

3.2. Analysis of Secondary Structure Assignments

Table 3 show the results of the extended analysis on the effects of different secondary structure assignment methods on structural class prediction. The results were yielded from SVM-LpsSCPred which was evaluated using acc, tpr and fpr. DSSP in SSAM category proved to be the best performer in all metrics with acc of 87.1%, tpr of 88.4% and fpr of 1.1%. This is followed by STRIDE, also in SSAM category, with acc of 85.6%, tpr of 85.6% and fpr of 2.1%. While SSPM has been proven to excel in secondary structure prediction, it also showed a competent acc, tpr and fpr for structural class prediction [21,22]. NNSSP under SSPM achieved 80.6% in acc, 80.2% in tpr and 4.4% in fpr, while PHD performed with a slight decrement of 0.4% in acc and 0.1% in tpr. SSAMs also demonstrated more superior performance compared to SSPMs.

3.3. Prediction Prone Towards CATH

A large portion of SCOP structural classes disagrees with CATH's [23]. This argument is supported by our findings as in Table 1. Prior to that, the proposed method demonstrated better success rate towards CATH compared to SCOP for all cases. This is due to the architecture of CATH named Topology that uses secondary structures information as well as their topological connections in classifying the structural class [25]. A manual classification scheme posed by SCOP that uses only tertiary structures limited the sequence-structure assignment. Besides, CATH [24] had integrated the manual classification with a semi-automatic

hierarchical classification which in the latest version 3.3 has successfully discerned up to 128,688 structural domains, outperforming the SCOP version 1.79 by 14%.

Further findings also indicated that the proposed method could predict the structural class of 39 sequences as shown in Table 4, which initially were categorized as unknown by threshold-based classification method [2]. Table 4 also shows that the proposed method match aligned with CATH rather than SCOP. In addition to a higher similarity rate to CATH, the proposed method might facilitate as an automatic structural class prediction method specifically for low-identity sequences.

Table 3: Performance of SVM-LpsSCPred in different secondary structure assignment methods

Category	Secondary structure	*acc* (%)	*tpr* (%)	*fpr* (%)
SSAM	DSSP	87.1	88.4	1.1
	STRIDE	85.6	85.6	2.1
SSPM	PHD	80.2	80.1	4.4
	NNSSP	80.6	80.2	4.4

3.4. Comparison to Other Classification Methods

The experiments were further focused to test the effect of different classifiers in discriminating the latent patterns of local protein structures to predict the structural class. As shown in Table 5, SVM, using similar features vector and label of local protein structures, four classifiers were evaluated using acc and the results are compared with NN, NB and KNN.

The results of these four classifiers were also cross-validated using similar 10 folded datasets. Results indicated that SVM outperformed the rest of the classifiers with 87.1% acc. This is followed by KNN 80.2%, NN 76.7% and NB 70.3%. The superior acc by SVM is in line with previous studies 4,21 which achieved over 80% acc.

The superiority of SVM was centered by its ability to: (i) map the input features vector into high dimensional features space and (ii) seek an optimized linear division where the n-separated hyperplane were constructed, n denoted the labels of structural class. In this SVM, a model was created using Radial Basis kernel function which was defined as follows:

$$K(\vec{y}_i, \vec{y}_j) = \exp(\frac{-r \| \vec{y}_i - \vec{y}_j \|^2}{2\sigma^2}) \quad (2)$$

Where $\vec{y}_i$ was labels and $\vec{y}_j$ was input vector. The input vector will be the center of the RBF and σ will determine the area of influence this input vector has over the data space. A larger value of σ will give a smoother decision surface and a more regular decision boundary since the RBF with large σ will allow an input vector to have a strong influence over a larger area.

Table 4: SVM-LpsSCPred succeeds to predict the unknown structural class (formerly derived from threshold-based classification method2) for 39 sequences of RS126

PDB ID	SCOP	CATH	SVM-LpsSCPred	PDB ID	SCOP	CATH	SVM-LpsSCPred
1azu	all-β	all-β	all-β	2sodb	all-β	all-β	all-β
1bbpa	all-β	all-β	all-β	3blm	mixed	mixed	mixed
1bmv1	all-β	all-β	all-β	3cd4	all-β	all-β	all-β
1bmv2	all-β	all-β	all-β	3cln	all-α	all-α	all-α
1cbh	mixed	mixed	mixed	3hmga	all-β	mixed	mixed
1fdl	all-β	all-β	all-β	3pgm	mixed	mixed	mixed
1fkf	mixed	mixed	mixed	4bp2	all-α	all-α	all-α
1fxia	mixed	mixed	mixed	4cms	all-β	all-β	all-β
1l58	mixed	all-α	all-α	4rhv1	all-β	all-β	all-β
1pyp	all-β	mixed	mixed	4rhv3	all-β	all-β	all-β
1r092	all-β	all-β	all-β	4sgbi	small	mixed	mixed
1rbp	all-β	all-β	all-β	4ts1a	mixed	mixed	mixed
1rhd	mixed	mixed	mixed	4xiaa	mixed	mixed	mixed
2alp	all-β	all-β	all-β	5er2e	all-β	all-β	all-β
2cyp	all-α	all-α	all-α	5hvpa	all-β	all-β	all-β
2gn5	all-β	all-β	all-β	5ldh	mixed	mixed	mixed
2ltnb	all-β	all-β	all-β	5lyz	mixed	all-α	all-α
2paba	all-β	all-β	all-β	6cpp	all-α	all-α	all-α
2pcy	all-β	all-β	all-β	6hir	small	all-β	all-β
2rspa	all-β	all-β	all-β				

Table 5: Comparison amongst classification methods

Classification method	*acc (%)*
SVM	87.1
K-Nearest Neighbor	80.2
Neural Network	76.7
Naïve Bayesian	70.3

4. CONCLUSION

SVM is a powerful classifier, while the local protein structures input bases are able to enrich the knowledge between known protein sequences and known structural classes. In this study, the advantages of both elements have been integrated to precisely predict the structural class. The integration is known as SVM-LpsSCPred that has been developed to solve the problems of insufficient known structural knowledge as well as low success rate which are posed by the former threshold-based classification method. Based on a higher similarity rate to CATH, the proposed method might facilitate as an automatic structural class prediction method specifically for low-identity sequences. It is anticipated that more influenced features vector can be adopted in the future works.

ACKNOWLEDGEMENT

This work is supported by MyMaster Scholarship of the Ministry of Education Malaysia, RMC UTM, G-Heart scheme under the Gates Scholars Foundation and GUP grant, with Vot No: 16H73. We would like to thank Universiti Tun Hussein Onn Malaysia for supporting this research under the Contract

Grant-Endowment (Vot number: A066), also, thanks to Gates IT Solution Sdn Bhd for the whole support

REFERENCES

1. Raghuraj R and Lakshminarayanan S, **"Variable predictive model based classification algorithm for effective separation of protein structural classes,"** *Comput. Biol. Chem.* vol. 32, p. 302–306, 2008. https://doi.org/10.1016/j.compbiolchem.2008.03.009
2. Chou KC, **"Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes"**, *Bioinformatics* vol. 21, p. 10–19, 2005. https://doi.org/10.1093/bioinformatics/bth466
3. Chen C, Zhou X, Tian Y, Zou X and Cai P, **"Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network"**, *Anal. Biochem.* vol. 357, p. 116–121, 2006. https://doi.org/10.1016/j.ab.2006.07.022
4. Zhou X, Bin Chen C, Li ZC and Zou XY**, "Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes"**, *J. Theor. Biol.* vol. 248, p. 546–551, 2007. https://doi.org/10.1016/j.jtbi.2007.06.001
5. Kurgan LA, Zhang T, Zhang H, Shen S and Ruan J. **"Secondary structure-based assignment of the protein structural classes"**, *Amino Acids* vol. 35, p. 551–564, 2008 https://doi.org/10.1007/s00726-008-0080-3
6. Altun G, Hu H-J, Gremalschi S, Harrison RW and Pan YA, **"Feature Selection Algorithm Based on Graph Theory and Random Forests for Protein Secondary Structure Prediction",** in *Bioinformatics Research and Applications, Third International Symposium, ISBRA 2007, Atlanta, GA, USA, May 7-10, 2007. Proceedings* 4463/2007, p. 590–600 (Springer Berlin Heidelberg, 2007). https://doi.org/10.1007/978-3-540-72031-7_54
7. Carter P, Andersen CAF and Rost B, **"DSSPcont: Continuous secondary structure assignments for proteins"**, *Nucleic Acids Res.* vol. 31, p. 3293–3295, 2003. https://doi.org/10.1093/nar/gkg626
8. Heinig M, and Frishman D, "STRIDE: **A web server for secondary structure assignment from known atomic coordinates of proteins",** *Nucleic Acids Res.* vol. 32, p. W500–W502, 2004. https://doi.org/10.1093/nar/gkh429
9. Burkhard R, **"How to Use Protein 1-D Structure Predicted by PROFphd**", *Proteomics Protoc. Handbook.* p. 875–901, 2005. https://doi.org/10.1385/1-59259-890-0:875
10. King RD *et al.*, **"Is it better to combine predictions**?", *Protein Eng.* vol. 13, p. 15–9, 2000. https://doi.org/10.1093/protein/13.1.15
11. Gu F, Chen H and Ni J, **"Protein structural class prediction based on an improved statistical strategy**", *BMC Bioinformatics* vol. 9, p. 1–9, 2008. https://doi.org/10.1186/1471-2105-9-S6-S5
12. Salamov AA and Solovyev VV, **"Protein secondary structure prediction using local alignments**", *J. Mol. Biol.* vol. 268, p. 31–36, 1997. https://doi.org/10.1006/jmbi.1997.0958
13. Deng J, Li K, Irwin GW and Fei M, **"Two-stage RBF network construction based on particle swarm optimization"**, *Trans. Inst. Meas. Control*, vol. 35, p. 25–33, 2013. https://doi.org/10.1177/0142331211403795
14. Colubri A *et al.*, **"Minimalist Representations and the Importance of Nearest Neighbor Effects in Protein Folding Simulations**", *J. Mol. Biol.* vol. 363, p. 835–857, 2006. https://doi.org/10.1016/j.jmb.2006.08.035
15. Wang J and Zheng X, "**Comparison of protein secondary structures based on backbone dihedral angles"**, *J. Theor. Biol.* vol. 250, p. 382–387, 2008. https://doi.org/10.1016/j.jtbi.2007.10.013
16. Chen K, Kurgan L, and Ruan J, **"Optimization of the Sliding Window Size for Protein Structure Prediction",** *2006 IEEE Symp. Comput. Intell. Bioinforma. Comput. Biol.* p. 1–7, 2006. https://doi.org/10.1109/CIBCB.2006.330959
17. Liu T, Zheng X and Wang J. **"Prediction of protein structural class for low-similarity sequences using support vector machine and PSI-BLAST profile",** *Biochimie* vol. 92, p. 1330–1334, 2010. https://doi.org/10.1016/j.biochi.2010.06.013
18. Pugalenthi G, Tang K, Suganthan PN and Chakrabarti S, **"Identification of structurally conserved residues of proteins in absence of structural homologs using neural network ensemble**", *Bioinformatics* vol. 25, p. 204–210, 2009. https://doi.org/10.1093/bioinformatics/btn618
19. Chinnasamy A, Sung W-K and Mittal A. **"Protein structure and fold prediction using Tree-Augmented naïve Bayesian classifier"**, *J. Bioinform. Comput. Biol.* vol. 3, p. 803–19, 2005. https://doi.org/10.1142/S0219720005001302
20. Csaba G, Birzele F and Zimmer R, **"Systematic comparison of SCOP and CATH: a new gold standard for protein structure analysis**", *BMC Struct. Biol.* vol. 9, p. 23, 2009. https://doi.org/10.1186/1472-6807-9-23
21. Kurgan L, Cios K and Chen K, **"SCPRED: Accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences**", *BMC Bioinformatics* vol. 9, p. 226, 2008. https://doi.org/10.1186/1471-2105-9-226
22. Greene LH *et al.* "**The CATH domain structure database: New protocols and classification levels give a more comprehensive resource for exploring evolution"**, *Nucleic Acids Res.* vol. 35, p. D291–D297, 2007. https://doi.org/10.1093/nar/gkl959
23. Marsden RL and Orengo CA, "**The classification of protein domains",** *Methods Mol Biol* vol. 453, p. 123–146, 2008. https://doi.org/10.1007/978-1-60327-429-6_5

24. Cuff AI *et al.* **"The CATH classification revisited-architectures reviewed and new ways to characterize structural divergence in superfamilies**", *Nucleic Acids Research* vol. 37, p. D310–D314, 2009. https://doi.org/10.1093/nar/gkn877
25. Shi JY and Zhang YN**, "Fast SCOP classification of structural class and fold using secondary structure mining in distance matrix",** in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 5780 LNBI, p. 344–353 (Springer, Berlin, Heidelberg, 2009). https://doi.org/10.1007/978-3-642-04031-3_30