Journal of Bioinformatics and Computational Biology
Vol. 18, No. 4 (2020) 2050022 (20 pages)

DOI: 10.1142/S0219720020500225

Structural profile matrices for predicting structural properties of proteins

Nuh Azginoglu[*,§], Zafer Aydin[†,¶] and Mete Celik[‡,‖]

[*]*Department of Computer Engineering*
Nevsehir Haci Bektas Veli University
Nevsehir 50300, Turkey

[†]*Department of Computer Engineering*
Abdullah Gul University, Kayseri 38080, Turkey

[‡]*Department of Computer Engineering*
Erciyes University, Kayseri 38039, Turkey
[§]*nuh@nevsehir.edu.tr*
[¶]*zafer.aydin@agu.edu.tr*
[‖]*mcelik@erciyes.edu.tr*

Received 3 May 2020
Accepted 13 May 2020
Published 10 July 2020

Predicting structural properties of proteins plays a key role in predicting the 3D structure of proteins. In this study, new structural profile matrices (SPM) are developed for protein secondary structure, solvent accessibility and torsion angle class predictions, which could be used as input to 3D prediction algorithms. The structural templates employed in computing SPMs are detected by eight alignment methods in LOMETS server, gap affine alignment method, ScanProsite, PfamScan, and HHblits. The contribution of each template is weighted by its similarity to target, which is assessed by several sequence alignment scores. For comparison, the SPMs are also computed using Homolpro, which uses BLAST for target template alignments and does not assign weights to templates. Incorporating the SPMs into DSPRED classifier, the prediction accuracy improves significantly as demonstrated by cross-validation experiments on two difficult benchmarks. The most accurate predictions are obtained using the SPMs derived by threading methods in LOMETS server. On the other hand, the computational cost of computing these SPMs was the highest.

Keywords: Protein structure prediction; secondary structure; solvent accessibility; torsion angle; structural profile matrix.

1. Introduction

Proteins are building blocks of living beings consisting of amino acid sequences linked together by peptide bonds. They conform to their three-dimensional (3D) structures

§Corresponding author.

through a process known as folding, which is part of protein synthesis that takes place in cells. Proteins carry out their functions by participating in biochemical reactions. Although the function is typically related to a small set of amino acids that form functional groups, the overall 3D shape of a protein constitutes the conditions for its interaction with other molecules in such a way that a hand fits into a glove. Therefore, understanding the 3D structure provides essential information about the function of a protein.

A four-level hierarchy has been proposed by Linderstrom-Lang[1] to categorize the structure of proteins. These are primary, secondary, tertiary, and quaternary structures. The primary structure refers to the amino acid sequence, secondary structure refers to patterns of hydrogen bonds, tertiary structure refers to the 3D structure of a single amino acid chain, and quaternary structure refers to the 3D structure of multiple chains. Most of the proteins do not have a quaternary structure. Due to technological advancements in gene sequencing, the amino acid sequence of proteins are discovered and stored in databases rapidly. On the other hand, the experimental determination of secondary, tertiary, and quaternary structures takes place at a slower rate. For this reason, computational prediction of structural information has been considered as an alternative to experimental techniques.[2]

Predicting the 3D structure of proteins is a challenging problem. Therefore first, various information about the target protein are collected or predicted such as sequence and structural profiles, secondary structure, solvent accessibility, torsion angles, and contact maps.[3] Typically, sequence and structural profiles are derived by sequence alignment algorithms and structural properties are predicted using machine learning algorithms, all of which can be employed as input to more sophisticated energy minimization algorithms for solving the 3D conformation of proteins.[4]

Computational prediction of structural properties also employs sequence and structural profiles as the input signal. To date, most of the prediction algorithms used sequence profiles derived by alignment methods called PSI-BLAST and HHblits. In recent works of Aydin *et al.*,[5,6] it has been shown that deriving better structural profiles can improve the accuracy of prediction algorithms considerably. The term "better" includes the use of advanced alignment algorithms and template scoring techniques which assign weights to template proteins.

In this paper, we develop novel structural profile matrix (SPM) estimation methods for protein structure prediction tasks including secondary structure, solvent accessibility, and torsion angle class prediction. An SPM includes a set of discrete probability distributions each showing the propensity of a target amino acid to be in one of the structural classes (e.g. in secondary structure prediction the class can be one of the labels from the alphabet {H, E, L}). To find template proteins that will be used to construct structural profiles, we employ eight threading methods in LOMETS server[7,8] (i.e. dPPAS, dPPAS2, Env-PPAS, MUSTER, PPAS, wdPPAS, wMUSTER, and wPPAS), gap affine alignment algorithm,[9] ScanProsite, which is the motif mining method for finding Prosite motifs,[10] PfamScan method,[11] which searches an amino acid sequence against a library of HMM-profiles in Pfam database,

and HHblits, which computes an HMM-profile for the target and aligns it against HMM-profiles of the proteins in Protein Data Bank (PDB). SPMs are also computed using the Homolpro method,[12] a state-of-the-art method that uses BLAST alignments.[13] Structural properties are predicted using DSPRED, a two-stage hybrid classifier that employs dynamic Bayesian networks (DBNs) and a support vector machine (SVM).[14]

The reasons for selecting the aforementioned methods are as follows. LOMETS is selected because it includes multiple alignments and score terms that contain information about 3D structure of the templates. This may help to detect structurally close templates more accurately. PfamScan method searches against a database of protein families that are organized based on domain information and can capture structural and functional context of the proteins. ScanProsite can search for Prosite motifs that include profiles and patterns that identify protein domains, families and functional sites. Gap affine method performs pairwise alignments between target and templates and allows customizing the score function to include information related to secondary structure, solvent accessibility and torsion angle classes, which could be useful to detect distant templates. HH-blits starts with multiple alignments and performs alignments between HMM-profiles using sequence as well as secondary structure similarity information and has the potential to detect structurally distant templates. Homolpro is selected for comparison with other methods as it uses pairwise alignments between amino acid sequences using amino acid similarity information only and is simpler than the other methods. DSPRED is selected as it is developed by Aydin *et al.*[14] earlier, which enables to have full control over the prediction models.

2. Methods

In this section, first, the datasets used for predicting structural properties of proteins are explained. This is followed by sections that explain how the SPMs are derived. Finally, the DSPRED method is described briefly.

2.1. *Datasets*

In this study, three different datasets are used. PDB99 is the template dataset employed for computing the SPMs (excluding LOMETS). The 3D structure information of the proteins in PDB99 is available in PDB. In this dataset, the percentage of sequence identity between any two proteins is less than or equal to 99%. PDB99 contains 35,603 proteins and 8,934,723 amino acids.[6,15] The other two datasets are CB513[16] and EVAset,[17] which are used to assess the accuracy of protein secondary structure, solvent accessibility, and torsion angle class prediction by DSPRED and various SPM methods. The CB513 contains 513 proteins and 84,119 amino acids. The EVAset includes 2876 proteins and 584,595 amino acids after proteins shorter than 30 amino acids are removed. CB513 dataset is available in the website that

contain Jpred's distribution material http://www.compbio.dundee.ac.uk/jpred/legacy/data/. Both CB513 and EVAset are difficult benchmarks. Detailed information for the preparation of the datasets used can be obtained from our previous study.[6] In addition to these datasets, the LOMETS method uses its own PDB template library that contains 72,172 PDB files.

2.2. *Generating structural profile matrix*

An SPM contains structural class estimates for the amino acids of a protein sequence, in which the classes are represented as soft labels in the form of probability scores that sum to 1. The number of rows of an SPM is equal to the number of amino acids in the protein and the number of columns is the number of structural classes. There are three classes for secondary structure prediction ({H, E, L}), two classes for solvent accessibility prediction ({e, b}), and seven classes for torsion angle prediction ({L, A, M, B, E, G, O}). The definition of torsion angle class labels can be found in Aydin *et al.*[18] Therefore, the dimension of an SPM is $N \times 3$ for secondary structure, $N \times 2$ for solvent accessibility, and $N \times 7$ for torsion angle class prediction where N is the number of amino acids in the target protein.[6] Once an SPM is computed, it is sent as input to the DSPRED program to predict the structural class labels of the amino acids in the target. In this study, two different SPMs are used as inputs to the DSPRED program: SPM-1 and SPM-2. SPM-1 is calculated using the HHBlits program. SPM-2 is calculated by Pfamscan, gap affine, LOMETS, or ScanProsite. Only one of these SPM-2s are used in DSPRED at a time. In addition to SPM-1 and SPM-2, there is also the SPM computed by the Homolpro method, which is employed as a post-processing block after DSPRED.[6] This SPM is only used in experiments that incorporate Homolpro to DSPRED in which SPM-1 and SPM-2 are excluded by setting their weights to zero.

2.2.1. *Generating SPM-1 using HHBlits*

The SPM-1 is generated starting from an HHblits[19] alignment, which is detailed in Aydin *et al.*[6] and summarized in Fig. 1.

2.2.2. *Generating SPM-2 using PfamScan*

The first type of SPM-2 is derived using the PfamScan method,[11] which aligns the target sequence against the HMM-profiles using HMMER.[20] Details of SPM-2 computation by PfamScan can be found in Aydin *et al.*[6] and is summarized in Fig. 7.

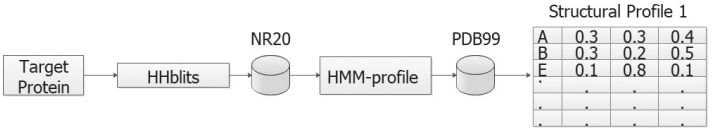


Fig. 1. Computing SPM-1 using HHBlits.

2.2.3. *Generating SPM-2 using gap affine alignment method*

The SPM computed by gap affine alignment method is based on the pairwise alignment between target and templates[9] and is known as alignment with affine gap scores. It uses local dynamic programming matrices denoted as M, I_x, and I_y. The alignment can have gaps both in target (i.e. query) and the template (i.e. hit) proteins. The sizes of the matrices are $K+1$ by $L+1$ where K and L are the lengths of the target and template proteins, respectively. The targets represent proteins in the test datasets (i.e. test sets in cross-validation experiments on CB513 and EVAset), and the templates include proteins in the PDB99 data set (Sec. 2.1).

M, I_x, and I_y are first initialized to contain zeros. Then the matrix elements are updated recursively using Eqs. (1)–(3) starting from the top left corner until reaching the bottom right corner of each matrix

$$M(i,j) = \max\begin{Bmatrix} M(i-1,j-1) + W_S(x_i, y_j) \\ I_x(i-1,j-1) + W_S(x_i, y_j) \\ I_y(i-1,j-1) + W_S(x_i, y_j) \end{Bmatrix}, \tag{1}$$

$$I_x(i,j) = \max\begin{Bmatrix} M(i-1,j) - d \\ I_x(i-1,j) - e \end{Bmatrix}, \tag{2}$$

$$I_y(i,j) = \max\begin{Bmatrix} M(i,j-1) - d \\ I_y(i,j-1) - e \end{Bmatrix}, \tag{3}$$

where x represents the amino acid sequence of the target protein, y denotes the amino acid sequence of the template protein, x_i is the ith amino acid of the target, y_j is the jth amino acid of the template, $M(i,j)$ is the (i,j)th element of M matrix and includes the best score up to (i,j) given that x_i is aligned with y_j, $I_x(i,j)$ is the (i,j)th element of I_x and includes the best score up to (i,j) given that the x_i is aligned to a gap, $I_y(i,j)$ is the (i,j)th element of I_y and includes the best score up to (i,j) given that y_j is aligned to a gap, $W_S(x_i, y_j)$ is the similarity score for aligning x_i with y_j, d is the gap opening penalty and e is the gap extension penalty. In this work, d is set to 8 and e to 1. $W_S(x_i, y_j)$ is computed as

$$\begin{aligned} W_S = {} & w_1 s(x_i, y_j) + w_2 S_{\mathrm{SS}}(x_i, y_j) + w_3 S_{\mathrm{SA}}(x_i, y_j) \\ & + w_4 S_{\mathrm{TA}}(x_i, y_j), \end{aligned} \tag{4}$$

where $s(x_i, y_j)$ is the similarity score for aligning amino acid x_i with y_j, $S_{\mathrm{SS}}(x_i, y_j)$ is the similarity score for aligning the predicted secondary structure label of x_i with the true secondary structure label of y_j, $S_{\mathrm{SA}}(x_i, y_j)$ is the similarity score for aligning the predicted solvent accessibility label of x_i with the true solvent accessibility label of y_j, $S_{\mathrm{TA}}(x_i, y_j)$ is the similarity score for aligning the predicted torsion angle label of x_i with the true torsion angle label of y_j, w_1 up to w_4 denote the weight of each score term, which are selected as 0.25. In this work, prediction scores are obtained using the SVM stage of the DSPRED method. For this purpose, a seven-fold cross-validation experiment is performed on CB513 and a 10-fold cross-validation

experiment on EVAset using DSPRED with SPM-1 only. For $s(x_i, y_j)$, the BLOSUM62 matrix is used[21,22] and the remaining similarity scores are computed according to Eqs. (5) and (7).

$$S_{\mathrm{SS}}(x_i, y_j) = 1 - \frac{\{(\rho_H - t_H)^2 + (\rho_E - t_E)^2 + (\rho_L - t_L)^2\}}{3}, \tag{5}$$

$$S_{\mathrm{SA}}(x_i, y_j) = 1 - \frac{\{(\rho_e - t_e)^2 + (\rho_b - t_b)^2\}}{2}, \tag{6}$$

$$\begin{aligned} S_{\mathrm{TA}}(x_i, y_j) = 1 - \{&(\rho_L^T - t_L^T)^2 + (\rho_A - t_A)^2 + (\rho_M - t_M)^2 \\ &+ (\rho_B - t_B)^2 + (\rho_E^T - t_E^T)^2 + (\rho_G - t_G)^2 + (\rho_O - t_O)^2\}/7, \end{aligned} \tag{7}$$

where ρ_H, ρ_E, and ρ_L represent the predicted helix, strand, and loop probabilities, respectively, for x_i; t_H, t_E, and t_L denote the true helix, strand, and loop probabilities, respectively, for y_j. A one-hot encoding is used to represent the secondary structure label of y_j since the true label is known. For example, if the actual secondary structure label of y_j is helix then $t_H = 1$, $t_E = 0$, and $t_L = 0$. Similarly, ρ_e, t_e, ρ_b, t_b represent scores for solvent accessibility prediction and ρ_L^T up to t_O denote scores for torsion angle class prediction.

The recurrence relations for M, I_x, and I_y can be represented by a finite state automaton that contains a state for each of M, I_x, and I_y, in which a transition corresponds to a score update and an alignment corresponds to a path through these states. While computing M, I_x, and I_y matrices, the states that are chosen by the maximum operator and the previous states in Eqs. (1)–(3) are also stored. Once the matrices M, I_x, and I_y are computed, the element with maximum value is found and the path that corresponds to the maximum scoring alignment is found by backtracking the states.[9] During this procedure, each time one visits the M state, this aligns x_i with y_j. If the I_x state is visited, then x_i of target is aligned to a gap symbol and similarly if I_y state is visited, y_j of the template is aligned to a gap symbol.

In the next step, the alignment between target and template is used to compute SPM-2. For this purpose, Eq. (8) is used to compute the frequency of occurrence weight from the alignment score, which is added to the corresponding element of the SPM. Once all the weights are accumulated, the rows are normalized so that the sum of the elements in each row is 1. Figure 2 summarizes the steps of SPM-2 computation using gap affine alignment method.

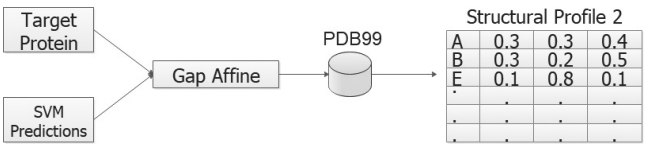


Fig. 2. Generating SPM-2 with gap affine alignment method.

2.2.4. *Generating SPM-2 with LOMETS method*

The third SPM-2 generation method that is developed in the scope of this paper employs the alignment methods in LOMETS server of Zhang lab, in which the target protein is aligned with the known proteins in the NCBI database[23] using PSI-Blast[13] to identify structurally similar templates. LOMETS produces the best 10 alignments from a total of eight threading methods including PPAS, Env-PPAS, wPPAS, dPPAS, dPPAS2, wdPPAS, MUSTER, and wMUSTER.[8,24–26]

The LOMETS program produces 80 alignment files for each target protein corresponding to 80 templates. A sample alignment result is given in Fig. 3 in which the template protein is 1an4A. According to this output, the first part that starts with '>' shows the target protein, and the second part shows the template protein.

In the next step, the PDB files that contain 3D structure coordinate information of the template proteins are downloaded. Then, the secondary structure and solvent accessibility labels of these templates are derived by the DSSP program and torsion angle labels are derived by a program called *phi_psi_linux*. Then the SPM are computed as in Sec. 2.2.3.

The following actions are taken when generating SPMs by LOMETS:

- Because the structures of proteins with the same PDB ID are similar, those templates with the same PDB ID as the target are not used when computing the SPM. For example, if the target protein 1a0iA is aligned with 1a0iB or 1a0iC by LOMETS, these templates are omitted during SPM computation.
- A given template can be detected repeatedly by the eight methods in LOMETS. Such repetitions are not eliminated in order to increase the weight of the templates that repeat.
- For some proteins, the PDB chain is represented by more than one character such as 1dynA1 and 1dynA2 (with chain IDs A1 and A2). Because the structure label information of such proteins is not accessible directly by the DSSP program the amino acid and label sequences are downloaded based on the letter codes in the chain (i.e. 1dynA). Then the downloaded sequence is locally aligned with the amino acid sequences of 1dynA1 and 1dynA2 and the label assignments are made based on these alignments. If these proteins cannot be aligned then they are not used to compute the SPM.
- For some templates, the amino acid sequence provided by LOMETS and the one downloaded by DSSP program have different lengths. In that case, the

```
>P1;target
sequence:target: : : : : : : :
---MKRESHKHAEQARRNRLAVALHELASLIPAEWKQQNVSAAPSKATTVEAACRYIRHLQQNGST*
>P1;1an4A
structure:1an4A: : : : : : : :
MDEKRRAQHNEVERRRRDKINNWIVQLSKIIPDSS-MESTKSGQSKGGILSKASDYIQELRQSNHR*
```

Fig. 3. A sample alignment output of the LOMETS program.

redundancy in the DSSP output is determined (if there is any) and the remaining parts are aligned based on LOMETS.

- In some cases, chain breaks occur in DSSP output. For such cases, gaps are included to the relevant positions of the DSSP output.
- In a given alignment, if there is space in the target, the relevant characters are deleted both from target and template. If there are gaps in the template protein, they are retained as gaps in the label sequences of the template.
- X (unknown amino acid), B, and Z amino acids can occur in the primary structures of some proteins in the data sets (especially in CB513). If the amino acid X is present at the beginning of the protein sequence, then it is replaced with M because of the start codon can commonly correspond to amino acid M. If the protein sequence contains an X that is not at the beginning of the protein, zeros are added to the corresponding positions of the SPM. The amino acid B may be either E or Q, and amino acid Z can be D or N.[27] For this reason, amino acids B are replaced by E and Z's are replaced by D.

Taking these conditions into account, the label sequence of the template protein is aligned with the target based on the alignments between the amino acid sequences in the LOMETS output. Then, zero matrices are formed and the matrix values are updated according to the alignments. Generating SPM-2 with LOMETS is summarized in Fig. 4.

2.2.5. *Generating SPM-2 using ScanProsite*

The fourth SPM-2 generation method that is developed in the scope of this paper employs Prosite's motif mining method to compute SPMs.[10]

The ScanProsite program searches for patterns in the Prosite database. A sample output of the ScanProsite program is shown in Fig. 5. According to this output, the protein contains patterns PS00001, PS00005, PS00006, PS00007, and PS00008. The location of these patterns in the protein sequence is also indicated in the output. For example, the PS00001 pattern includes 'NLTK' between the 40th and 43rd amino acids of the protein sequence.

After the patterns in the protein are found, the PDB IDs of the templates that contain those patterns are obtained from the link (https://prosite.expasy.org/cgi-bin/prosite/get-prosite-entry?PSxxxx) by a python script. The xxxx region in the

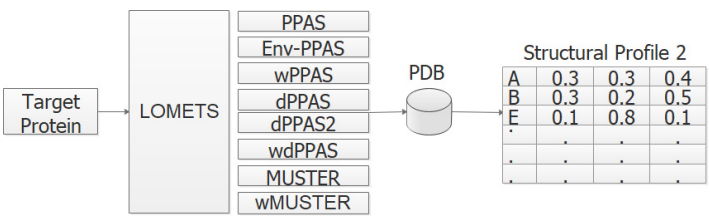


Fig. 4. Generating SPM-2 using LOMETS method.

```
>1delb-2-AUTO.1.all : PS00001 ASN_GLYCOSYLATION N-glycosylation site.
    40 - 43  NLTK
>1delb-2-AUTO.1.all : PS00005 PKC_PHOSPHO_SITE Protein kinase C phosphorylation site.
    24 - 26  TrK
    97 - 99  SvR
>1delb-2-AUTO.1.all : PS00006 CK2_PHOSPHO_SITE Casein kinase II phosphorylation site.
    24 - 27  TrkE
    42 - 45  TklE
    48 - 51  TimE
>1delb-2-AUTO.1.all : PS00007 TYR_PHOSPHO_SITE Tyrosine kinase phosphorylation site.
    25 - 33  RkefEgidY
>1delb-2-AUTO.1.all : PS00008 MYRISTYL N-myristoylation site.
    12 - 17  GVfaAN
```

Fig. 5. Sample output for the ScanProsite program.

link is updated for all the patterns found and the script is run separately for each. Then, the link http://www.rcsb.org/pdb/files/fasta.txt?structuredList=PDB_name is used to download the amino acid sequences of all the chains of the PDB proteins. If a chain is repeated multiple times in a protein only one of them is downloaded to prevent bias, and the others are ignored. In the next step, the ScanProsite program is run for each PDB template, and the location of the consistent pattern is determined. Thus, for each protein in the data set, the ScanProsite patterns, their positions on the target and the PDB templates are determined. Similar to SPM-2 generation by gap affine and LOMETS methods, the 3D coordinate information in PDB is downloaded for the templates and the secondary structure and solvent accessibility labels are derived by the DSSP program and the torsion angle information using a script that processes the PDB file. This is followed by SPM computation as in Sec. 2.2.3. Generating SPM-2 using ScanProsite method is summarized in Fig. 6.

2.2.6. *Eliminating hit proteins using percentage of sequence identity threshold*

The similarities between the template proteins and the target protein are taken into account when computing SPMs. For this purpose, a total of 9% of sequence identity thresholds (PSITs) are defined: 20% up to 100%, with increments of 10%. For each threshold value, if the percentage of sequence identity score between a template and the target is greater than the threshold, then that template is excluded from SPM computation. Defining those thresholds allows to analyze the predictive performance of SPMs at different target template similarities, in which a low PSIT considers distant templates only and a high PSIT considers structurally close as well as distant templates in computing the SPM. As a result of this thresholding, a total of nine different SPMs are computed for each PSIT, for a given target and SPM derivation method. Table 1 contains the number of templates at each PSIT for the methods used to derive SPM-2s for CB513 benchmark (i.e. sum of templates for 513 targets).

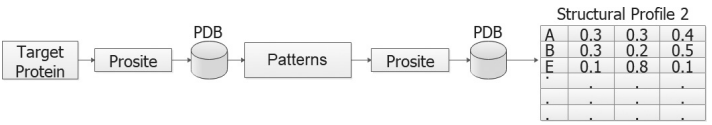


Fig. 6. Generating SPM-2 using ScanProsite.

Table 1. Total number of templates used to compute SPM-2 at each PSIT value for CB513 targets.

	20%	30%	40%	50%	60%	70%	80%	90%	100%
Homolpro	1	174	1521	3850	6270	7670	8126	8559	10,844
HHblits	52,471	71,912	77,249	79,050	80,632	81,376	81,727	82,163	84,932
PfamScan	1250	13,586	29,726	45,913	55,150	71,345	76,511	79,315	105,308
LOMETS	1192	1914	3140	4363	5055	6242	6623	6719	8122
ScanProsite	32,220	46,666	57,472	67,693	73,828	79,384	82,809	85,198	96,788
Gap affine	4,269,196	5,586,341	7,936,430	10,723,640	11,919,684	16,479,761	17,928,897	18,109,037	18,264,339

According to this table LOMETS and Homolpro methods obtain the least number of templates. They are followed by PfamScan, ScanProsite, HHblits and gap affine methods. Due to incorporating score terms for structural properties, the gap affine method selects many templates with short alignments with the target whose alignment score is eventually positive. At PSIT=100% all the templates in PDB99 are used for all the targets. Although this includes noise into the computation of the SPM, it is suppressed by weighting the template proteins as explained in Sec. 2.2.7.

2.2.7. *Weighting template proteins*

An SPM is initialized to contain zeros. Then, based on the alignments between targets and templates, weighted frequency of occurrence counts is added to the matrix elements. Finally, the rows of the SPM are normalized so that the sum of scores in each row (i.e. for each amino acid) is equal to one. As a result of this process, SPM contains probability scores for different class types that could be assigned to amino acids of the target. If there are no templates aligned to a particular amino acid of the target, then the corresponding row of the SPM contains all zeros. Two types of weighting schemes are implemented. The first one is the weighting used for SPM-1 and is detailed in Ref. 6. The second weighting approach is used to compute SPM-2 and is summarized in the following equation:

$$w = \begin{cases} I^a & \text{if } I \neq 0, \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

where w is the frequency of occurrence weight of the template amino acid that is aligned to an amino acid of the target, I is the percentage of sequence identity score between target and template obtained by Blastp alignment,[13] a is the integer power parameter that enables to amplify the contribution of structurally close templates as compared to distant ones. The a parameter prevents the contribution of a few structurally close templates from being suppressed by many structurally distant templates and is selected according to Table 2. In other words, it amplifies the contribution of structurally close templates.

2.3. *DSPRED*

In this paper, protein secondary structure, solvent accessibility, and torsion angle classes are predicted by the DSPRED program, which is a two-stage hybrid classifier that includes DBNs and a SVM.[6,14] The steps of the DSPRED is shown in Fig. 7. The weights w_1, w_2, w_3, and w_4 are selected the same as in our previous study.[6] To be specific, in experiments that use SPM-1 only the weight terms in Fig. 7 are selected as

Table 2. a parameter values with respect to *PSIT*.

PSIT	20	30	40	50	60	70	80	90	100
a	1	2	3	4	5	6	7	8	9

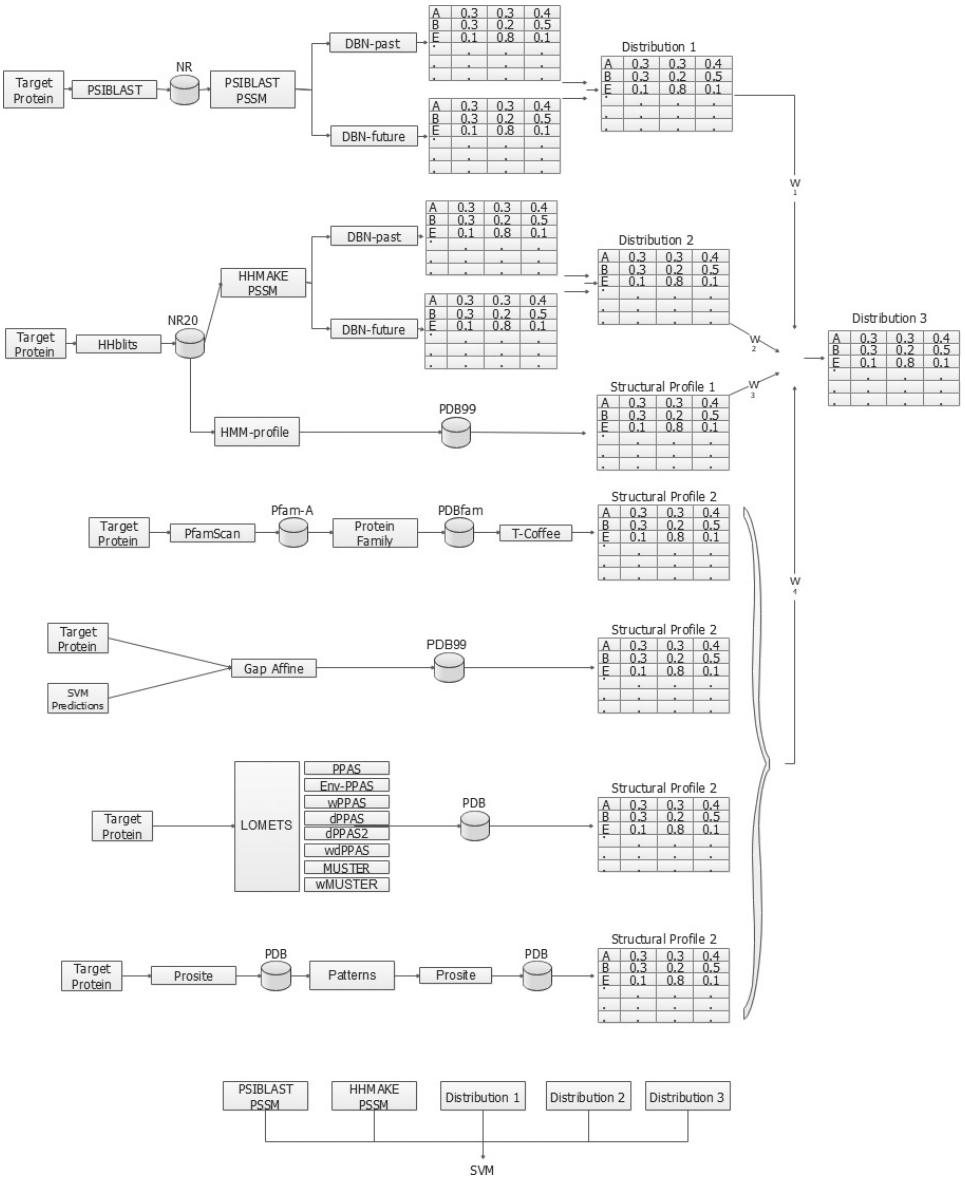


Fig. 7. The steps of the DSPRED method (modified from Fig. S5 in Ref. 6).

$w_1 = w_2 = w_3 = 1/3$ and $w_4 = 0$. In Homolpro experiments, $w_1 = w_2 = 1/2$ and $w_3 = w_4 = 0$. The SPM derived by Homolpro is used as a post-processing block after DSPRED as explained in Ref. 6. In experiments that employ SPM-2 (including the SPMs derived by PfamScan, gap affine, LOMETS, and ScanProsite) $w_1 = w_2 = w_3 = w_4 = 1/4$ (i.e. SPM-1 and SPM-2 are used together).

3. Results and Discussion

In this paper, PfamScan[6] and Homolpro[6,12] which is a state-of-the-art method are compared with LOMETS, Gap Affine, and ScanProsite methods all of which are used to compute SPMs for DSPRED. Our previous study[6] can be reviewed for detailed information on how Homolpro is used. A 10-fold cross-validation experiment is performed on EVAset and a seven-fold cross-validation on CB513 benchmark. Tables 3 and 4 include the overall secondary structure prediction accuracy (i.e. Q3) and segment overlap (SOV) measure, respectively, on EVAset; Tables 5 and 6 summarize the overall secondary structure prediction accuracy (i.e. Q3) and SOV measure, respectively, on CB513; Tables 7 and 8 contain the overall solvent accessibility prediction accuracy (i.e. Q2) and SOV measure, respectively, on EVAset. Based on these experiments, it can be said that the gap affine method gives the best results for low PSIT values (i.e. distant target template pairs), followed by LOMETS and the PfamScan methods. LOMETS method is the most successful in all prediction methods when PSIT is between 40–90%. The gap affine method yields a similar SOV result to the LOMETS method for PSIT = 80% in secondary structure experiments on CB513 and is slightly better than LOMETS for PSIT = 90%. At PSIT = 100%,

Table 3. Secondary structure prediction overall accuracy (Q_3) of DSPRED on EVAset.

	20%	30%	40%	50%	60%	70%	80%	90%	100%
Homolpro (Blastp)	82.89	82.94	83.58	84.46	85.14	85.66	85.94	86.20	88.88
SPM-1 (HHblits)	**83.84**	**84.70**	85.57	86.15	86.68	86.90	86.80	86.69	89.64
SPM-1 (HHblits) & SPM-2 (PfamScan)	83.75	84.51	85.27	85.76	86.23	86.38	86.35	86.47	**90.61**
SPM-1 (HHblits) & SPM-2 (ScanProsite)	83.67	84.30	84.95	85.40	85.80	85.91	85.81	85.74	88.75
SPM-1 (HHblits) & SPM-2 (Gap Affine)	83.81	**84.70**	85.53	86.18	86.75	86.99	86.94	86.87	89.70
SPM-1 (HHblits) & SPM-2 (LOMETS)	83.70	84.61	**85.85**	**86.68**	**87.22**	**87.42**	**87.42**	**87.38**	90.22

Table 4. Secondary structure prediction segment overlap measure (SOV) of DSPRED on EVAset.

	20%	30%	40%	50%	60%	70%	80%	90%	100%
Homolpro (Blastp)	78.24	78.29	78.71	79.48	80.17	80.72	81.04	81.28	84.29
SPM-1 (HHblits)	**79.30**	**80.00**	80.91	81.55	82.13	82.48	82.38	82.23	85.57
SPM-1 (HHblits) & SPM-2 (PfamScan)	**79.13**	79.74	80.56	80.99	81.57	81.69	81.53	81.59	**86.37**
SPM-1 (HHblits) & SPM-2 (ScanProsite)	79.09	79.58	80.23	80.61	81.07	81.25	81.11	80.95	84.29
SPM-1 (HHblits) & SPM-2 (Gap Affine)	79.11	**79.96**	80.82	81.60	82.29	82.64	82.64	82.47	85.53
SPM-1 (HHblits) & SPM-2 (LOMETS)	79.06	79.90	**81.29**	**82.08**	**82.69**	**82.98**	**82.92**	**82.77**	85.75

Table 5. Secondary structure prediction overall accuracy (Q_3) of DSPRED on CB513.

	20%	30%	40%	50%	60%	70%	80%	90%	100%
Homolpro (Blastp)	81.65	81.80	83.29	85.04	86.60	87.43	87.87	88.29	91.87
SPM-1 (HHblits)	82.78	84.21	85.63	86.81	87.70	87.91	87.82	87.58	91.67
SPM-1 (HHblits) & SPM-2 (PfamScan)	82.59	84.06	85.47	86.65	87.66	88.02	88.31	88.54	**93.17**
SPM-1 (HHblits) & SPM-2 (ScanProsite)	82.69	83.79	84.87	86.00	86.84	86.99	87.03	86.86	90.66
SPM-1 (HHblits) & SPM-2 (Gap Affine)	**83.06**	**84.45**	85.85	87.23	88.23	88.60	88.63	88.47	92.00
SPM-1 (HHblits) & SPM-2 (LOMETS)	82.93	84.40	**86.08**	**87.66**	**88.75**	**89.09**	**89.02**	**89.02**	92.75

Table 6. Secondary structure prediction segment overlap measure (SOV) of DSPRED on CB513.

	20%	30%	40%	50%	60%	70%	80%	90%	100%
Homolpro (Blastp)	77.30	77.51	79.24	81.22	83.01	83.94	84.62	84.88	89.05
SPM-1 (HHblits)	78.35	80.33	81.42	83.06	84.20	84.25	84.37	84.36	88.87
SPM-1 (HHblits) & SPM-2 (PfamScan)	78.20	79.74	81.12	82.24	83.69	84.08	84.40	84.63	**90.13**
SPM-1 (HHblits) & SPM-2 (ScanProsite)	78.40	79.57	80.49	81.92	82.99	83.02	83.19	83.16	87.45
SPM-1 (HHblits) & SPM-2 (Gap Affine)	**78.70**	80.27	81.89	83.47	84.79	85.06	**85.23**	**85.22**	89.06
SPM-1 (HHblits) & SPM-2 (LOMETS)	78.66	**80.49**	**81.98**	**83.87**	**84.90**	**85.35**	**85.23**	85.07	89.63

the PfamScan method has performed better than all the other techniques. Only in the solvent accessibility experiment performed on EVAset, the LOMETS method performed better than PfamScan. Although the ScanProsite method has sometimes performed better than certain methods, it has never been the most successful for any PSIT value. Homolpro only passed the ScanProsite at high similarity rates and lagged behind all the other methods.

Table 7. Solvent accessibility prediction overall accuracy (Q_2) of DSPRED on EVAset.

	20%	30%	40%	50%	60%	70%	80%	90%	100%
Homolpro (Blastp)	79.90	79.90	79.90	79.92	80.13	80.63	80.95	81.23	86.82
SPM-1 (HHblits)	80.57	81.45	82.25	83.02	83.55	83.81	83.82	83.71	87.92
SPM-1 (HHblits) & SPM-2 (PfamScan)	80.46	81.26	82.04	82.59	83.07	83.11	83.29	83.59	**88.91**
SPM-1 (HHblits) & SPM-2 (ScanProsite)	80.37	81.14	81.85	82.35	82.80	82.93	82.85	82.86	86.87
SPM-1 (HHblits) & SPM-2 (Gap Affine)	80.47	81.30	82.17	82.84	83.45	83.76	83.81	83.88	87.94
SPM-1 (HHblits) & SPM-2 (LOMETS)	**80.60**	**81.55**	**82.74**	**83.50**	**84.15**	**84.38**	**84.37**	**84.43**	88.41

Table 8. Solvent accessibility prediction segment overlap measure (SOV) of DSPRED on EVAset.

	20%	30%	40%	50%	60%	70%	80%	90%	100%
Homolpro (Blastp)	57.72	57.72	57.72	57.76	58.06	58.92	59.05	59.28	68.12
SPM-1 (HHblits)	59.21	60.57	61.98	62.89	63.73	64.20	64.40	64.92	**75.53**
SPM-1 (HHblits) & SPM-2 (PfamScan)	58.95	60.48	61.95	62.91	63.58	63.63	63.86	64.41	74.49
SPM-1 (HHblits) & SPM-2 (ScanProsite)	58.87	60.32	61.62	62.37	62.80	62.79	62.68	62.69	70.22
SPM-1 (HHblits) & SPM-2 (Gap Affine)	58.95	**60.89**	62.42	63.47	64.16	64.40	64.62	64.59	72.58
SPM-1 (HHblits) & SPM-2 (LOMETS)	**59.39**	60.74	**62.59**	**64.07**	**65.46**	**65.73**	**65.86**	**65.91**	73.63

Table 9. Q_3 and Q_2 Z-Scores (LOMETS - Pfam).

	20%	30%	40%	50%	60%	70%	80%	90%	100%
CB513 Q3 SS3*	1.8460	1.9132	3.5814	6.1907	6.9305	6.8929	4.5931	3.119	−3.367
EVAset Q3 SS3	−0.7323	1.4963	8.9212	14.4302	15.7746	16.6648	17.1372	14.5935	−7.1624
EVAset Q2 SA2**	1.9115	4.0298	9.9356	13.1114	15.7731	18.6099	15.8592	12.3909	−8.5253

Notes: *Secondary structure, **Solvent accessibility.

To assess whether the Q3 results of the LOMETS method is significantly better than the PfamScan method, Z-tests are performed. The Z-score and p-values of these hypothesis tests are provided in Tables 9 and 10. Based on these tests, the performance improvement obtained by LOMETS over PfamScan method is statistically significant for PSIT >40%. For PSIT <40%, the two methods performed comparably.

Table 11 includes the torsion angle class prediction results by DSPRED when SPM-1 only is used as the structural profile method as well as when SPM-1 and SPM-2 are used together. Since the CB513 dataset has no torsion angles labels, this experiment was performed only in the EVAset. In the latter, the best SPM-2 method is used in each PSIT. Based on these results, using SPM-2 together with SPM-1 performs significantly better than employing SPM-1 alone in DSPRED.

Table 10. Q_3 and Q_2 P-Values (LOMETS - Pfam).

	20%	30%	40%	50%	60%	70%	80%	90%	100%
CB513 Q3 SS3*	0.06432	0.05614	0.00034	<0.00001	<0.00001	<0.00001	<0.00001	0.0018	0.00076
EVAset Q3 SS3	0.4654	0.13362	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001
EVAset Q2 SA2**	0.05614	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001

Notes: *Secondary structure, **Solvent accessibility.

Table 11. Torsion angle class prediction accuracy on EVAset.

	20 % (G*)	30% (G)	40% (L**)	50% (L)	60% (L)	70% (L)	80% (L)	90% (L)	100% (P***)
SPM-1 Q_7	**73.83**	**75.23**	76.44	77.43	78.28	78.51	78.27	78.18	83.35
SPM-1 & SPM-2 Q_7	73.77	75.16	**76.95**	**78.25**	**79.18**	**79.50**	**79.51**	**79.53**	**83.89**
SPM-1 SOV	**69.93**	**71.51**	72.88	74.01	74.88	74.98	74.68	74.55	80.25
SPM-1 & SPM-2 SOV	69.85	71.41	**73.53**	**74.80**	**75.95**	**76.28**	**76.23**	**76.21**	**80.91**

Notes: *Gap affine, **LOMETS, ***PfamScan.

As a result, gap affine method should be preferred for secondary structure prediction and the LOMETS method should be preferred for solvent accessibility prediction at low similarity regions. The LOMETS method should be preferred for other similarity regions both in secondary structure and solvent accessibility prediction except for PSIT = 100%, in which PfamScan is the best.

When the results are examined, the gap affine method is effective at low similarity rates where the reliability of amino acid sequence similarity to distinguish structurally similar templates from dissimilar ones is reduced. Gap affine method compensates this by using structural information (including secondary structure, solvent accessibility and torsion angle classes) in addition to amino acid similarity to compute the alignments. When the similarity level of the templates is increased the success rates of the other methods become higher than this method because they use multiple alignments between query and templates and they employ more advanced alignment techniques such as alignments between sequence and sequence profiles. Although LOMETS is a successful method in general, it is seen that it lags behind other methods with very low and very high similarity rates. The reason behind the behavior at low similarity rates could be because LOMETS has high precision since it uses 3D structure information and tries to exclude templates with noisy alignments, which could otherwise be useful for estimating the SPM. For this reason, typically there are few templates only with low similarity with the target in LOMETS alignments. In this range, the predictions of structural properties obtained by DSPRED for the query and employed in gap affine method could be more accurate than the structural information used by the threading methods of LOMETS because gap affine method reaches more templates than LOMETS. The reason behind the behavior at high similarity rates is that alhough LOMETS uses top 10 alignments from eight different threading methods as template proteins, after pruning the templates that have the same PDB ID as the target, generally, less than 80 template proteins remain in total, and the similarity of many of these with the target protein remains below 100%. On the other hand, the PfamScan method can reach more closely related templates from the same domain family in this high similarity region, which increases the success rate of PfamScan. While the PfamScan method is very successful in the high similarity rate, the high elimination rate in the low similarity

rate causes the success rate of this method to decrease. ScanProsite uses Prosite patterns, and this method has a lower success rate than all the other methods. In this method, the site patterns do not seem to give good results. Similar to ScanProsite, the success rate of Homolpro method is also lower than the other methods. This could be because Homolpro uses pairwise alignments of sequences obtained by blastp program by employing sequence similarity information only to detect templates.

In this paper, we provide a comparative analysis of the methods for computing SPM, which are used to predict various structural properties of proteins. We especially compare threading-based alignment methods in LOMETS (that use 3D structure information to detect templates) to other methods that use alignments of sequences and sequence profiles. We find that using more sophisticated alignment techniques (such as HHblits and LOMETS together) improves the accuracy of predicting structural properties by up to 1.5% as compared to using HHblits only. Although the improvements achieved are statistically significant, they may not be at a level to provide a huge impact in terms of predicting the 3D structure of proteins. Nonetheless, the analyses provided will help the researchers to select the appropriate template detection method at different similarity intervals and optimize (i.e. fine tune) their structure prediction tasks better.

4. Conclusion

In this paper, we generated new SPMs using LOMETS alignments, Gap Affine alignments and a motif search method called ScanProsite. We compared the performance of these methods to the recently developed SPM technique that uses PfamScan method. Incorporating the new SPMs into DSPRED classifier, we obtained improvements in protein secondary structure, solvent accessibility and torsion angle class predictions, which shows that the proposed SPM technique is effective for structure prediction tasks. As a future work, we are planning to employ other alignment techniques for deriving SPMs such as those in LOMETS2, which is the new version of LOMETS. Furthermore the new SPM techniques can be employed in prediction methods that use deep learning.

Acknowledgments

The experiments calculations reported in this paper were partially performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA resources). This work was supported by 3501 TUBITAK National Young Researchers Career Award [Grant Number 113E550].

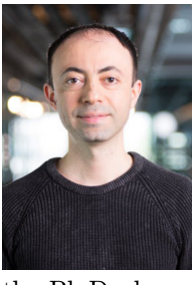
References

1. Linderstrøm-Lang KU, *Lane Medical Lectures: Proteins and Enzymes*, Vol. 6, Stanford University Press, 1952.

2. Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J, Critical assessment of methods of protein structure prediction (casp) round xiii, *Proteins Struct Funct Bioinform* **87**(12):1011–1020, 2019.
3. Fayech S, Essoussi N, Limam M, Data mining techniques to predict protein secondary structures, *2013 5th Int Conf Modeling, Simulation and Applied Optimization (ICMSAO)*, IEEE, 2013, pp. 1–5.
4. Kuhlman B, Bradley P, Advances in protein structure prediction and design, *Nat Rev Mol Cell Biol* **20**:681–697, 2019.
5. Aydin Z, Baker D, Noble WS, Constructing structural profiles for protein torsion angle prediction, *Proc Int Joint Conf Biomedical Engineering Systems and Technologies*, Vol. 3, SCITEPRESS-Science and Technology Publications, 2015, pp. 26–35.
6. Aydin Z, Azginoglu N, Bilgin HI, Celik M, Developing structural profile matrices for protein secondary structure and solvent accessibility prediction, *Bioinformatics* **35**(20):4004–4010, 2019.
7. Zhang Y, I-TASSER server for protein 3D structure prediction, *BMC Bioinform*, 2008.
8. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y, The i-tasser suite: Protein structure and function prediction, *Nat Meth* **12**(1):7, 2015.
9. Durbin R, Eddy SR, Krogh A, Mitchison G, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, 1998.
10. Sigrist CJ, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, Bucher P, Prosite: A documented database using patterns and profiles as motif descriptors, *Briefings Bioinform* **3**(3):265–274, 2002.
11. Finn RD *et al.*, The pfam protein families database: Towards a more sustainable future, *Nucleic Acids Res* **44**(D1):D279–D285, 2016.
12. Magnan CN, Baldi P, SSpro/ACCpro 5: Almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity, *Bioinformatics* **30**(18):2592–2597, 2014.
13. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ, Gapped blast and psi-blast: A new generation of protein database search programs, *Nucleic Acids Res* **25**(17):3389–3402, 1997.
14. Aydin Z, Singh A, Bilmes J, Noble WS, Learning sparse models for a dynamic Bayesian network classifier of protein secondary structure, *BMC Bioinform* **12**(1):154, 2011.
15. Wang G, Dunbrack RL, PISCES: A protein sequence culling server, *Bioinformatics*, 2003.
16. Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ, Jpred: A consensus secondary structure prediction server, *Bioinformatics (Oxford, England)* **14**(10):892–893, 1998.
17. Koh IY *et al.*, Eva: Evaluation of protein structure prediction servers, *Nucleic Acids Res* **31**(13):3311–3315, 2003.
18. Aydin Z, Baker D, Noble WS, Template scoring methods for protein torsion angle prediction, *Int Joint Conf Biomedical Engineering Systems and Technologies*, Springer, 2015, pp. 206–223.
19. Remmert M, Biegert A, Hauser A, Söding J, Hhblits: Lightning-fast iterative protein sequence searching by hmm-hmm alignment, *Nat Meth* **9**(2):173, 2012.
20. Eddy S, HMMER3: A new generation of sequence homology search software, available at: http://hmmer.janelia.Org, 2010.
21. Henikoff S, Henikoff JG, Amino acid substitution matrices from protein blocks, *Proc Natl Acad Sci* **89**(22):10915–10919, 1992.
22. Eddy SR, Where did the blosum62 alignment score matrix come from?, *Nat Biotechnol* **22**(8):1035, 2004.
23. Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I, GenBank, *Nucleic Acids Res* **47**:D94–D99, 2018.

24. Wu S, Zhang Y, Lomets: A local meta-threading-server for protein structure prediction, *Nucleic Acids Res* **35**(10):3375–3382, 2007.
25. Wu S, Zhang Y, Muster: Improving protein sequence profile–profile alignments by using multiple sources of structure information, *Proteins Struct Funct Bioinform* **72**(2):547–556, 2008.
26. Yan R, Xu D, Yang J, Walker S, Zhang Y, A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction, *Sci Rep* **3**:2619, 2013.
27. IUPAC, IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN). Nomenclature and symbolism for amino acids and peptides, Recommendations 1983, *Biochem J* **219**(2):345–373, 1984.

Nuh Azginoglu received his Bachelor of Science (B.Sc.), Master of Science (M.Sc.) and the Ph.D. degrees in computer engineering from Erciyes University, Kayseri, Turkey in 2010, 2013, and 2019, respectively. He is currently a faculty member of the Department of Computer Engineering, Nevsehir Haci Bektas Veli University, Nevsehir, Turkey. His recent research interests focus on bioinformatics, data mining and deep learning.



Zafer Aydin received his Bachelor of Science (B.Sc.) and Master of Science (M.Sc.) degrees with high honor from the Electrical and Electronics Engineering Department of Bilkent University in 1999 and 2001, respectively. He then enrolled in the Ph.D. program of the same department and worked as a teaching assistant for one year. Starting from 2002, he worked as a Graduate Research Assistant in School of Electrical and Computer Engineering at Georgia Institute of Technology, Atlanta GA USA and received the Ph.D. degree in 2008. As a result of maintaining an interest in bioinformatics research, he worked as a post-doctoral fellow for three years in Noble Research Lab, which is part of the Genome Sciences Department at University of Washington, Seattle, WA USA. From September 2011 to February 2014, he worked as an Assistant Professor in Electrical and Electronics Engineering Department of Bahcesehir University, Istanbul, Turkey. Currently he is an Assistant Professor in Computer Engineering Department of Abdullah Gul University, Kayseri, Turkey.

Mete Celik received the B.Sc. degree in control and computer engineering and the M.Sc. degree in electrical engineering from Erciyes University, Kayseri, Turkey, in 1999 and 2001, respectively, and the Ph.D. degree in computer science from the University of Minnesota, Minneapolis, USA, in 2008. He is currently a faculty member of the Department of Computer Engineering, Erciyes University, Turkey. His research interests include data analysis, big data, spatial databases, spatial data mining, spatio-temporal data mining, bioinformatics and location-based services. He is a member of the IEEE and ACM.