Prediction of DNA-binding residues in local segments of protein sequences with Fuzzy Cognitive Maps

Abdollah Amirkhani, Mojtaba Kolahdoozi, Chen Wang, and Lukasz Kurgan

Abstract— While protein-DNA interactions are crucial for a wide range of cellular functions, only a small fraction of these interactions was annotated to date. One solution to close this annotation gap is to employ computational methods that accurately predict protein-DNA interactions from widely available protein sequences. We present and empirically test first-of-its-kind predictor of DNA-binding residues in local segments of protein sequences that relies on the Fuzzy Cognitive Map (FCM) model. The FCM model uses information about putative solvent accessibility, evolutionary conservation and relative propensities of amino acid to interact with DNA to generate putative DNA-binding residues. Empirical tests on a benchmark dataset reveal that the FCM model secures AUC = 0.72 and outperforms recently released hybridNAP predictor and several popular machine learning methods including Support Vector Machines, Naïve Bayes and k-Nearest Neighbor. The improvements in the predictive performance result from an intrinsic feature of FCMs that incorporate relations between the input features, besides the relations between the inputs and output that are modelled by other algorithms. We also empirically demonstrate that use of a short sliding window results in further improvements in the predictive quality. The funDNApred webserver that implements the FCM predictor is available at http://biomine.cs.vcu.edu/servers/funDNApred/.

Index Terms— Proteins, DNA, protein-DNA interactions, DNA-binding residues, Fuzzy Cognitive Maps.

1 Introduction

Proteins carry out many cellular functions by interacting with a wide range of ligands, including DNA [1-5]. Molecular-level analysis of the protein-DNA interactions facilitates their classification, decoding of the underlying physics, and discovery of patterns that define specificity of the protein-DNA recognition [3, 6, 7]. The number of DNA-binding proteins was recently estimated to be on average at 3% of proteins in eukaryotic organisms and 5% in the animals species [8]. Given that we already sequenced 27 million eukaryotic proteins (source: UniProt resource [9, 10] as of March 2, 2018) and assuming conservative estimates we should expect to know 3% of 27 million = 810 thousand DNA-binding eukaryotic proteins. Unfortunately, UniProt annotates only about 45 thousand proteins that interact with DNA, even when we include both experimental and computational, homology-derived results. This reveals that a significant majority of these interactions remains to be discovered. One solution is to use the available data on the protein-DNA interactions to devise computational models that accurately predict DNA interactions from protein structures and sequences [11].

Prediction of the protein-DNA interactions can be done at three levels: whole protein, residue and at the atomic scale [12]. At the coarsest whole protein level we predict whether or not a given protein binds DNA. At the residue level we predict which residues in the protein sequence interact with DNA. At the highest resolution level, we consider interactions between individual atoms of proteins and DNA. The resolution of the prediction is typically determined based on the available data, i.e., whether or not both protein structure and sequence or only the sequence are available. Predictions that rely on the protein structure are limited to a relatively small number of proteins for which the three-dimensional structures are available. Protein Data Bank (PDB) [13, 14], the worldwide database of protein structures, includes 128 thousand structures for 42 thousand distinct proteins (as of March 2, 2018). This is a small fraction of the 109 million currently sequenced proteins that can be obtained from UniProt (as of March 2, 2018). Although a high-quality predicted structure could be used instead of the native structure, this would reduce quality of the predictions and, more importantly, would not solve the problem of the low coverage. Recent works found that the overall structural coverage that includes native and predicted structures ranges between a few and 30%, depending on the considered organisms [15]. The total coverage for the human proteins is at about 28% [16]. On the other hand, sequence only-based approaches can be applied to all available protein sequences. These methods

- *A. Amirkhani and M. Kolahdoozi are with the Department of Electrical Engineering, Iran University of Science and Technology, Tehran, Iran. E-mails: amirkhani@ieee.org, mojtaba_kolahdooz@elec.iust.ac.ir.*
- *C. Wang and L.A. Kurgan are with the Department of Computer Science, Virginia Commonwealth University, Richmond, VA 23284. E-mails: wangc27@mymail.vcu.edu, lkurgan@vcu.edu*
- *Corresponding author: Lukasz Kurgan*

can be used to make predictions at the residue and whole protein levels. We focus on the sequence-based predictors that provide results at the finer, residue level.

The sequence-based predictors are empirically designed and tested using datasets of protein sequences with the annotated DNA-binding residues. The annotations of the DNA-binding residues are primarily extracted from the structures of the protein-DNA complexes. As of March 2, 2018, PDB includes structures for 4,475 protein-DNA complexes, allowing us to derive sufficiently large and diverse datasets to build and test predictive tools. A few reviews have summarized and compared the sequence-based predictors of DNA-binding residues [11, 17-19]. These predictors include (in chronological order) DBS-Pred [20], DBS-PSSM [21], BindN [22], DNABindR [23, 24], DP-Bind [25, 26], DISIS [27], ProteDNA [28], BindN+ [29], NAPS [30], MetaDBSite [31], DisoRDPbind [32, 33], DRNApred [34] and hybridNAP [12].

Virtually all of these methods predict DNA binding residues from the whole protein sequence. A recently released exception is hybridNAP that can be used to predict the interacting residues in local segments of at least three consecutive residues in the input protein chain. Such approach allows prediction of the DNA binding for fragments of protein sequence and parts of the whole sequence. This latter option is particularly useful when inputs that are required for these predictors cannot be produced for the entire sequence. An example scenario where these inputs may not be available for the whole sequence is when multiple sequence alignment does not provide sufficiently deep profile to generate position specific scoring matrix (PSSM) for some of the residues. Recent review reveals that PSSM is commonly used to make predictions. More specifically, 9 out of out of 14 predictors (including the four newest methods) surveyed in a recent review utilize PSSM [12].

The authors of the abovementioned predictors have used a wide range of machine learning algorithms to empirically generate their predictive models. Predictive models for ProteDNA [28], BindN+[29], DP-Bind [25, 26] were derived with the Support Vector Machine (SVM) algorithm. The k-Nearest Neighbor (kNN) algorithm was used to derive models for DISIS [27], DBS-Pred [20], and DBS-PSSM [21]. Another popular algorithm is regression which was utilized for DRNApred [34] DisoRDPbind [32, 33], and hybridNAP [12]. One common characteristic of these predictive models is that they map predictive inputs (typically in the form of numerical features extracted from the sequence) into the annotation of DNA-binding residues. However, they do not exploit the fact that some of these inputs can be mutually related.

To this end, we develop a novel tool that provides accurate prediction of DNA-binding residues for full protein sequences and local sequence fragments by utilizing mutual relations between inputs. We apply fuzzy cognitive map (FCM) model [35] to address this goal. This model was used only once on the past to perform computations of protein sequences at the residue level. This was the context of content of secondary protein structure [36]. FCMs express relations between inputs and output as well as relations between inputs. They are a powerful predictive tool [37, 38] which is extensively used to build predictive models in medicine [39-42] and in numerous other fields [43-45]. To the best of our knowledge, we are the first to apply FCMs to predict protein-ligand interactions.

2 Materials and Methods

2.1 Datasets

We develop and test our predictive tool using datasets of proteins with the annotated DNA-binding residues. We rely on the datasets that were published recently alongside the hybridNAP method [12] and the recent assessment of tools that predict the DNA-binding residues [19]. The training and validation datasets, which we use to empirically design our predictor, are sourced from [12] and were extracted based on the annotated proteins from the BioLiP database [46]. BioLiP is a semi-manually curated database of protein-ligand interactions that are extracted from PDB. This database labels a residue as a DNA-binding if the distance between an atom of this residue and an atom of DNA in the protein-DNA complex < 0.5Å plus the sum of the Van der Waals radii of the two atoms.

We borrowed the DNA_T test dataset from [19] to evaluate and compare predictive performance of our tool. This benchmark dataset was developed specifically to compare predictors of the DNA-binding residues. It contains 47 DNA-binding proteins and 9106 residues including 875 DNA-binding residues and 8231 non-DNA-binding residues, resulting in the 1 to 9.4 ratio of DNA-binding to non-DNA-binding residues. Importantly, we ensure that the protein sequences in the DNA_T dataset are dissimilar to the proteins in the training and validation datasets. We use BLASTCLUST [47] to remove the training and validation proteins that share sequence similarity > 30% with the proteins in TEST_T. This ensures a fair comparison with other methods, such as the recently released hybridNAP.

After removing similarity to DNA_T, the dataset extracted from BioLiP includes 18,995 protein sequences with a total of 32,055 DNA-binding residues. We balanced the number of DNA-binding and non-DNA-binding residues in this dataset to ease computational learning of the predictive model. We include all native DNA-binding residues and we randomly subsample the same number of the non-DNA-binding residues. We randomly split the resulting set of 64,110 residues into two subsets: 70% is used for the training dataset and the remaining 30% for the validation dataset. Only the final model, which we optimize by maximizing its predictive performance on the validation set, is used to perform predictions on the TEST_T dataset. The training and test datasets are available at http://biomine.cs.vcu.edu/servers/funDNApred/.

We emphasize that the training, validation and test datasets benefit from high-quality annotations of DNA-binding residues that were performed in [12, 19]. In contrast to older studies that consider one protein-DNA complex per protein to annotate binding residues, we combine annotations coming from potentially multiple complexes that cover the same protein in order to provide a more complete set of the DNA binding residues. First, we map all protein sequences to UniProt with the help of SIFTS [48]. Next, we

transfer the DNA-binding annotations from the multiple BioLiP/PDB protein chains that are linked to the same (unique) UniProt protein. As we show in [12], this results in 19.7% increase in the number of annotated DNA-binding residues when compared with the best case scenario that represents the approach that prior works took to annotate the binding residues, i.e., when chains with the highest number of the DNA-binding residues are utilized.

2.2 Overview for the Predictive Model

Our method makes predictions for individual residues in an input protein sequence. It can be used to predict all residues in the given chain as well as a selected subset of residues, i.e., a local segment of adjacent residues in the input protein sequence. The prediction process consists of three steps:

- The i^{th} input residue is represented by a small set of numeric features (concepts)
- The values of these features are input into the FCM model
- The FCM model computes the prediction and outputs propensity for DNA-binding $P(i)$, i.e., higher value suggests a higher likelihood that i^{th} residue interacts with DNA

Several studies investigated sequence-derived features that are commonly used to characterize and predict DNA-binding residues [12, 49-52]. A recent article that summarized these studies concludes that the most frequently used features are evolutionary conservation (ECO), relative solvent accessibility (RSA), and relative propensity of specific amino acids (AAs) for the DNA-binding (RAA) [12]. ECO is relevant since residues that interact with DNA are typically conserved across homologous protein sequences [53]. RSA quantifies accessibility of residues to the solvent that surrounds proteins that is normalized to the size of specific AAs. The use of RSA stems from the fact that protein-DNA interaction occurs on the protein surface. Finally, these studies also suggest that the type of the interacting AAs and their immediate neighbours in the protein sequence can be also used to determine relative propensity for the DNA binding. Consequently, we use these three features as inputs for the FCM.

RAA is quantified with relative difference in abundance of a given AA type between the DNA-binding residues and the corresponding non-DNA-binding residues on the protein surface. We consider only the surface to eliminate a confounding factor related to a bias in composition of AAs in the protein core; these residues typically do not bind DNA and can be identified with RSA. RAA is defined as the difference between fractions of a given AA type among the DNA-binding residues and among the surface non-DNA-binding residues divided by the fraction among the non-DNA-binding residues. The positive (negative) RAA values denote enrichment (depletion) among the DNA-binding residues compared to the non-DNA-binding residues on the surface. We compute the relative differences using Composition Profiler program [54]. Motivated by [12], we use a weighted average of the RAA values for the residue that is predicted (with weight = 0.5) and its two neighbors in the sequence (with weights = 0.25) as one of the inputs for the FCM model.

RSA values are derived from the protein structure, typically using the DSSP program [55]. However, since our sole input is the protein sequence we have to substitute the native RSA with putative RS. We generate the putative RSA directly from the sequence with a very fast and accurate ASAquick program [56]. More precisely, we divide the absolute surface area predicted with ASAquick by the maximal value of surface area of a given AA, which we obtain from [57], to compute the putative RSA values.

ECO is computed from the multiple sequence alignment generated with HHBlits [58] against the redundancy reduced UniProt20 database ver. 2015_06 using the default parameters. We use the alignment to produce n×20 matrix of position-specific frequencies p_{AAi} where AA_i represents the 20 AA types and n is the protein sequence length. Next, we used this matrix to calculate evolutionary conservation based on formula from [59]:

$$ECO = \frac{log \sum_{i=1}^{20} p_c^2(i)/p_0(i)}{log \sum_{i=1}^{20} p_c(i)/p_0(i)} \quad (1)$$

where i is position in the sequence and $p_0(i)$ is the BLOSUM62 background distribution for the i^{th} position [60]. Like in [12], we use the hidden Markov model-based position-specific frequencies generated with HHblits rather than the PSSM-based scores since they provide a faster to compute and better measure of evolutionary conservation [61].

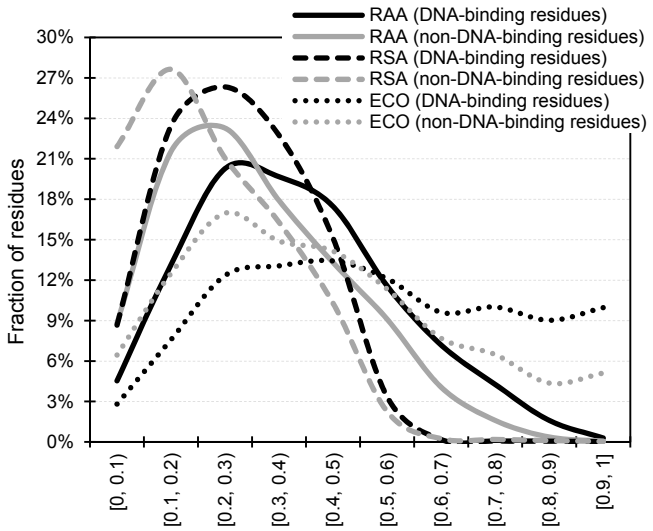


Fig.1. Distributions of the values of the three features (RAA, RSA and ECO) for the DNA-binding residues (black curves) and non-DNA-binding residues (grey curves) in the training dataset.

We empirically analyze whether these three features can differentiate between DNA-binding and non-DNA-binding residues. Fig. 1 compares distributions of the values of these features between the native DNA binding residues (black curves) and native non-DNA binding residues

(grey curves) in the training dataset. As expected, we observe that the RAA values are much higher for the native DNA-binding residues when compared to the non-DNA-binding residues (solid lines in Fig. 1). The Mann-Whitney test reveals that the two distributions are significantly different (p-value < 0.0001). Similarly, the RSA values (dashed lines) and ECO values (dotted lines) for the residues that interact with DNA are substantially larger. Again, the Mann-Whitney test shows that the differences between the corresponding distributions for the DNA-binding and non-DNA-binding residues are statistically significant (p-value < 0.0001). These results agree with the analysis in [12], and they justify the use of these three features in our FCM model.

2.3 Assessment of the Predictive Model

The predictions generated by our model are real-values that quantify propensity for the DNA binding. The predictive quality of these propensities is measured with the area under the ROC curve (AUC). The curve is obtained by plotting TPrate = sensitivity = TP/(TP+FN) versus FPrate = 1 – specificity = 1 – TN/(TN+FP), where TP (TN) is the count of the correctly predicted DNA-binding (non-DNA-binding) residues, FP is the number of the native non-DNA-binding residues that have been incorrectly predicted as DNA-binding residues, and FN is the number of the native DNA-binding residues that have been incorrectly predicted as non-DNA-binding. The TPrate and FPrate values are established by thresholding the propensities using every unique value of the output propensity. For each threshold, we assume the residues with propensities > threshold as DNA binding and the remaining residues as non-DNA binding. We use the AUC value as a criterion that we optimize (maximize) in the process of learning the FCM model from the training dataset.

We also evaluate and compare binary predictions on the TEST_T dataset. We generate these predictions from the real-value propensities. Residues that are predicted with propensities > a given threshold are predicted as DNA binding while the remaining residues as set as non-DNA-binding. We set the threshold value such that the resulting binary predictions have FPrate = 10%. This corresponds to the prediction where the FPrate is similar to the rate of native DNA binding residues in the TEST_T dataset. Ensuring that the methods evaluated on TEST_T are set to the same FPrate allows for a robust side-by-side comparison of their binary measures of predictive performance. We assess the binary predictions with four measures: sensitivity, specificity, accuracy and Matthews correlation coefficient (MCC):

$$sensitivity = TPrate = \frac{TP}{TP+FN} \quad (2)$$

$$specificity = 1 - FPrate = \frac{TN}{TN+FP} \quad (3)$$

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

$$MCC = \frac{TN*TP-FN*FP}{\sqrt{(TP+FP)*(TP+FN)*(TN+FN)*(TN+FP)}} \quad (5)$$

Sensitivity determines the predictive quality for the native DNA-binding residues, specificity for the native non-DNA-binding residues, while accuracy measures the overall predictive quality. MCC is suitable to evaluate imbalanced datasets, such as TEST_T set where 9.4% of residues are DNA-binding. Values of MCC range between -1 and 1 and should be interpreted like other correlation coefficients. The same measures were used in the past to assess predictors of the DNA-binding residues [11, 12, 18, 19].

2.4 Fuzzy Cognitive Maps

FCMs were first proposed by Kosko [35]. The FCM model is a graph composed of nodes and edges. The nodes are used to model features (concepts) relevant to a given application area. In our case these are the three predictive features (ECO, RSA and RAA) and the output feature that denotes propensity for DNA-binding. The edges express causal relationships between features. For our project, they quantify relations between the predictive features and output as well as relations between the three predictive features. The causal relations between features are determined either by specialists or by means of learning from data [62-65].

The values of features in the FCM are determined by a vector $F = [F_1, F_2, \ldots, F_N]$ where $F_i \in [0\ 1]$, i=1…N, and N is the number of features. N = 4 for our FCM model. The causal relationships between nodes are defined with an $N \times N$ dimensional matrix W:

$$W = \begin{bmatrix} 0 & w_{12} & \ldots & w_{1N} \\ w_{21} & 0 & \ldots & w_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ w_{N1} & w_{N2} & \ldots & 0 \end{bmatrix} \quad (6)$$

where $w_{ij} \in [-1\ 1]$ is a weight that quantifies strength and direction of a relation for an edge from i^{th} to j^{th} feature. These causal relations are defined as follows:

- If $w_{ij} > 0$ then an increase in value of feature F_i leads to an increase in value of feature F_j that is proportional to $|w_{ij}|$
- If $w_{ij} < 0$ then an increase in value of feature F_i leads to reduction in value of feature F_j that is proportional to $|w_{ij}|$
- If $w_{ij} = 0$ then there is no causal relationship between feature F_i and F_j

The entries on the diagonal of W are zero since inclusion of relations of a feature with itself can lead to instability. Fig. 2 shows a sample FCM with 3 features and 4 edges together with its weight matrix W.

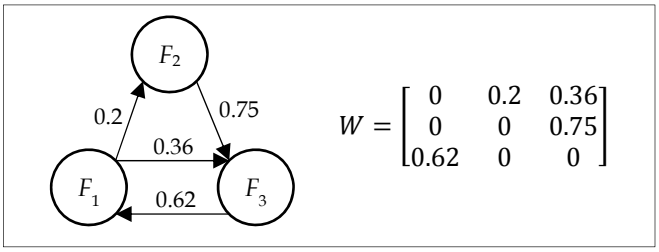


Fig.2. A sample FCM with 3 features and the corresponding weight matrix W.

FCMs are executed iteratively to update values of the features. The value of each feature in the $(t+1)^{th}$ iteration is

determined based on the weight matrix W and the values in the t^{th} iteration for the features connected to that feature as follows:

$$F_i(t+1) = \varphi(Fi(t) + \sum_{j=1, j\neq i}^{N} F_j(t) w_{ji}) \quad (7)$$

In Eq.7, $F_i(t)$ is the value of the i^{th} feature (node in the FCM graph) in the t^{th} iteration and $\varphi()$ is a transfer function for which the values are limited to [0, 1] interval. While several different transfer functions were proposed, the sigmoid transfer function are preferred [66]. Thus, we also apply the sigmoid function:

$$\varphi(x) = \frac{1}{1+e^{-\alpha x}} \quad (8)$$

with a typically used value of parameter $\alpha = 5$ [67].

2.5. Learning Fuzzy Cognitive Maps

The task of training an FCM determines the weight matrix W from the training dataset. The coefficients in the matrix are typically learned using an algorithm [68]. While earlier methods relied on Hebbian [69, 70] and genetic algorithms [62, 63, 71, 72], newer methods apply other strategies including ant colony optimization [73] and particle swarm optimization (PSO) [74] algorithms. We apply the PSO algorithm motivated by its successful applications in several recent studies [75-77].

PSO is an evolutionary algorithm based on social interactions between particles that possess swarm intelligence, which was proposed by Kennedy and Eberhart [78]. In PSO, every particle is represented by a vector in d dimensional problem space. The initial population of particles in this space is initialized by a random position vector $X_i = (x_{i1}, x_{i2}, \ldots, x_{id})$ and velocity vector $V_i = (v_{i1}, v_{i2}, \ldots, v_{id})$. The algorithm uses a fitness function to find out whether the particles are close to an optimal solution. We define the optimal solution as such that produces maximal AUC value on the training dataset based on the values of the output feature. Each particle is associated with two vectors: $pbest_i$ (the best position of the i^{th} particle in the course of its displacements) and gbest (the best vector for all particles). The Eqs. 9 and 10 are used to modify the two vectors in the course of the optimization performed by the PSO algorithm:

$$v_i(t+1) = \omega v_i(t) + c_1 r_1 \big(pbest_i(t) - x_i(t)\big) + c_2 r_2 \big(gbest_i(t) - x_i(t)\big) \quad (9)$$

$$x_i(t+1) = x_i(t) + v_i(t+1) \quad (10)$$

where ω is the learning coefficient, c_1, $c_2 \in [1.5, 2]$ (with uniform distribution) are used to guide a trade-off between the positions of the best particle solution and the best global solution, and r_1 and r_2 are random numbers in the [0, 1] interval [79]. Use of a large learning coefficient value results in a more global search while a smaller value makes the search more local. Based on [80], we gradually reduce the learning coefficient value in subsequent iterations using Eq. 11:

$$\omega = \omega_{max} - t \frac{\omega_{max} - \omega_{min}}{T} \quad (11)$$

where ω_{min} and ω_{max} are the minimal and maximal value of the learning coefficient, respectively, and T is the total number of iterations. We set the values of the parameters as follows: ω_{min} *=0.2*, ω_{max}*=-0.3*, and T *=2000.*

To sum up, we use PSO to optimize values in matrix W. In each iteration of this optimization the FCM model and Eq. 7 are used to produce the value of the output feature (propensity for the DNA binding) from the values of the three input features (RSA, RAA and ECO) for each residue in the training dataset. The resulting predictions (values of the output feature generated by FCM) are utilized to compute AUC. Next, the AUC value is used to update the PSO search using Eqs. 9 and 10 and the process iteratively repeats until convergence. We define convergence as either of the following two conditions: 1) AUC value on the validation dataset decreases over 30 consecutive iterations, which suggests overfitting into the training dataset; and 2) AUC value on the training dataset does not increase over 30 consecutive iterations. The solution is the matrix W that results in the maximal AUC value.

We also study impact of neighboring residues on the predictive performance of the FCM-based prediction. To accomplish that we use a sliding window, where the predictions for the residues that are adjacent to the currently predicted residue are used together with the predictions for this residue to compute the final propensities for the DNA binding. The FCM predictions for the residues inside the window are combined using a weighted sum where the values of these weights $c_k \in [0, 1]$ are included into the PSO optimization (along with the matrix W):

$$DNA\ binding\ propensity = c_k * FCM_{out\_k} \quad (12)$$

where k = 1, 2.., K is the index of a residue in the window, K={3, 5, 7, …} is the window size, and FCM_{out_k} is the value of the output feature in FCM for the k^{th} residue in the window. The window with K = 3 includes the predicted residue in the center + one immediate neighbor on each side in the sequence, with K = 5 includes residue in the center + two neighbors on each side, etc. The process of prediction with the FCM model for K = 3 is depicted in Fig. 3.

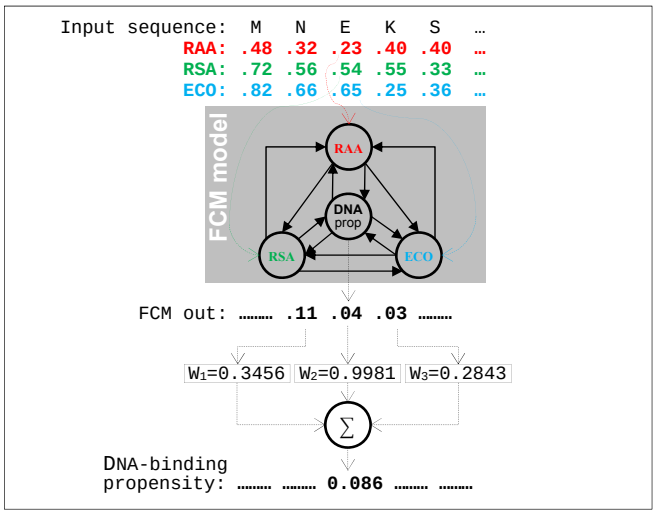


Fig. 3. Flow of prediction with the FCM-based predictive model.

The PSO optimization generated the following sets of weights: [0.3456, 0.9981, 0.2843] for K = 3; [0.2095, 0.1790, 0.9997, 0.05, 0.3150] for K = 5; and [0.3637, 0.0505, 0.2750, 0.9965, 0.0766, 0.1476, 0.3648] for K = 7. As expected, the

weight for the residue in the center of the window is the largest. This is the residue for which the DNA-binding propensity is ultimately predicted. The flanking residues are associated with lower weight value. The cumulative values of these weights are similar on the left and right side of the window. This again is an anticipated result since protein sequence has no particular direction.

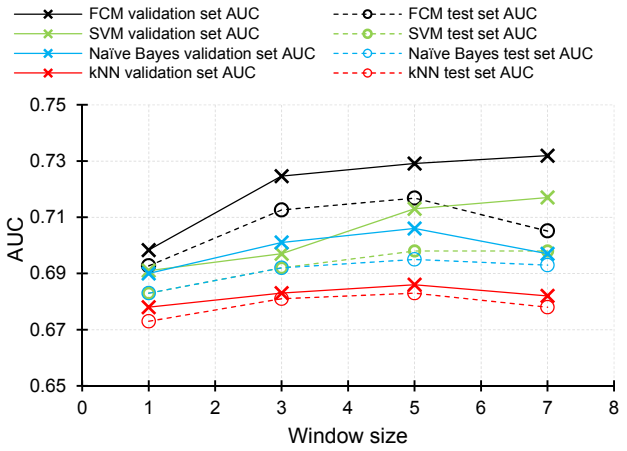


Fig. 4. Comparison of predictive performance measured with AUC for FCM, SVM, Naïve Bayes and kNN models that make predictions for a single residue and using short residue windows.

3 Results and Discussion

3.1 Selection of Window Size

We empirically compare the FCM that makes predictions for a single residue with the FCM models that use windows. Black lines and markers in Fig. 4 summarize the results on the validation dataset and compares them with the corresponding results on the test dataset. The FCM that does not utilize the window secures AUC =0.70 on the validation dataset and 0.69 on the TEST_T dataset. Use of the window to process the FCM's outputs increases the predictive performance to 0.72 for window size 3 and to 0.73 for window sizes 5 and 7, when tested on the validation dataset. We do not consider longer windows since the FCM model already did not register improvements between sizes 5 and 7. The same FCMs on the TEST_T dataset obtains AUC values equal 0.71 (K=3), 0.72 (K=5) and 0.71 (K=7). Altogether, the results reveal that window size K = 5 is the best choice.

3.2 Comparison of Protein-fragment based Predictors of DNA-binding Residues

We focus comparative assessment on the methods that predict DNA-binding residues in protein fragments. We compare the predictive quality of the FCM-based predictor with a selection of machine learning algorithms and the recent predictor of DNA-binding residues in local sequence segments, hybridNAP [12]. The selection of the machine learning algorithms is motivated by their use to implement the whole-sequence predictors of the DNA-binding residues. We include SVM that was used to implement three whole-sequence predictors: ProteDNA [28], BindN+[29], DP-Bind [25, 26]; kNN that was utilized by another three whole-sequence predictors: DISIS [27], DBS-Pred [20], and DBS-PSSM [21]; and Naïve Bayes that was used to develop DNABindR [23, 24]. These three algorithms are used with the same inputs that are available to the FCM-based predictor: ECO, RAA and RSA. For each algorithm, we compute the predictive model on the training dataset and make predictions on the TEST_T dataset.

Similar to FCM, we selected the best window size for each of the three other algorithms. Fig. 4 shows that trends in the AUC values for SVM, kNN and Naïve Bayes are similar to the trend for FCM. Namely, the AUCs improve when increasing the window size from 1 to 3, and from 3 to 5. However, the results on the validation dataset for the window size 7 (solid lines in Fig. 4) are either lower (for Naïve Bayes and kNN) or similar (for SVM) to the results for the window size 5. Moreover, the results on the TEST_T dataset (dashed lined in Fig. 4) follow the same pattern, i.e., the AUCs for window size 7 are either the same or worse when compared to the results for window size 5. We conclude that the considered predictors provide the best results for the window size equal 5. Consequently, Table 1 compares results on the TEST_T dataset for the window size = 5 and compares them to the results based on a single residue and window size = 3. Moreover, we compare with the results of the regression-based hybridNAP. Inclusion of this model covers two regression-based whole-sequence tools: DRNApred [34] and DisoRDPbind [32, 33]. Finally, we compare these methods to a baseline predictor which generates random numbers. The binary assessment (accuracy, sensitivity and MCC) is performed at a fixed FPrate = 10% to facilitate side-by-side comparison of these values. We note that specificity of all methods = 90% given the fixed value of the FPrate. This FPrate was selected to mimic the rate of the native DNA-binding residues in the TEST_T dataset.

Table 1 reveals that the FCM-based solution outperforms all other considered machine learning algorithms. When predicting using a single residue, FCM model secures AUC = 0.693 and sensitivity = 30.5%. To compare, the other machine learning algorithms obtain AUC ≤ 0.683 and sensitivity ≤ 25.5%. The differences in accuracy are relatively small due to the imbalanced nature of the dataset. The substantial increase by 5% in sensitivity produced by our solution translates to 0.5% increase in accuracy since positive instances (DNA-binding residues) constitute about 10% of the TEST_T dataset. We also observe a visible increase in the MCC, which equals 0.19 for the FCM compared to 0.14 for the second best machine learning algorithm. When compared to hybridNAP, the FCM-based predictor features higher AUC (0.693 vs. 0.685), sensitivity (30.5% vs. 28.8%) and MCC (0.19 vs. 0.17). The improvements of FCM over the other methods stem from the intrinsic to the FCM model use of relations between the predictive inputs. The other algorithms rely solely on the relations between the inputs and the output.

TABLE 1. COMPARISON OF THE PREDICTIVE PERFORMANCE OF THE FCM PREDICTOR WITH MACHINE LEARNING MODELS AND THE RECENT RELEVANT PREDICTOR, HYBRIDNAP, ON THE TEST_T DATASET. THE LAST LINE SHOWS BASELINE RESULTS PRODUCED WITH A RANDOM PREDICTOR. SPECIFICITY OF ALL METHODS IS SET TO 90% (FPRATE = 10%).

Inputs	Algorithm	AUC	Accuracy [%] at FPrate=10%	Sensitivity [%] at FPrate=10%	MCC at FPrate=10%
Single residue (no window)	**FCM**	**0.693**	**84.5**	**30.5**	**0.189**
	SVM	0.683	83.8	25.5	0.143
	Naïve Bayes	0.683	83.7	24.6	0.135
	kNN	0.673	83.9	25.3	0.144
Window size = 3	**FCM**	**0.713**	**84.3**	**31.0**	**0.190**
	SVM	0.692	84.1	28.5	0.169
	Naïve Bayes	0.692	83.8	25.7	0.145
	kNN	0.681	83.8	25.4	0.142
Window size = 5	**FCM**	**0.717**	**84.5**	**32.6**	**0.203**
	SVM	0.698	84.0	28.0	0.165
	Naïve Bayes	0.695	83.9	26.5	0.152
	kNN	0.683	83.7	25.3	0.140
hybridNAP (no window)		0.685	84.2	28.8	0.170
Baseline (random predictor)		0.494	82.3	9.8	-0.001

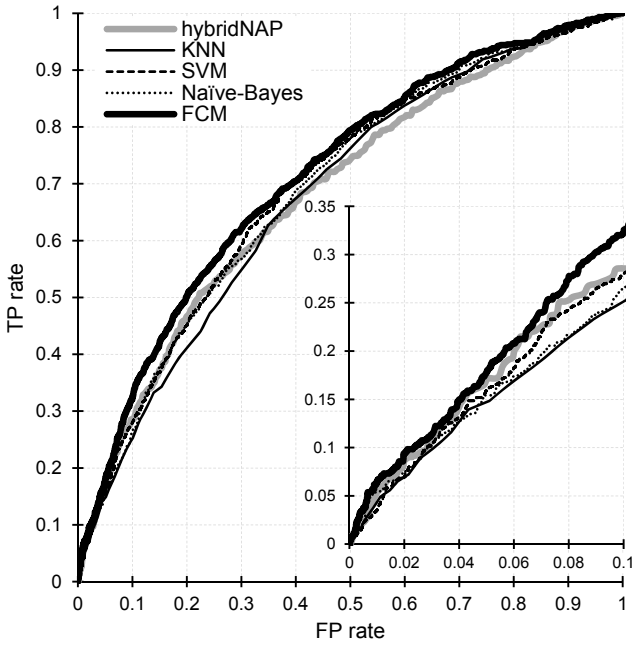


Fig. 5. ROC curves for the hybridNAP and the FCM, SVM, Naïve Bayes and kNN predictors that use window size = 5 on the TEST_T dataset. An inset in the bottom right corner shows an enlarged version of the ROC curve for the FPrate range between 0 and 0.1.

Table 1 also quantifies improvements due to the use of the window-based prediction. The predictive quality of the FCM method improves with the use of the window. In general, the results for the window size = 3 are better than when the window is not used. Similarly, results for the window size = 5 are better than for the window size = 3. We note that use of larger windows is not expected to lead to further improvements (see Fig. 4). The FCM's AUC grows from 0.693 (no window) to 0.717 (window size = 5). Similarly, its sensitivity and MCC increase from 30.5% to 32.6% and from 0.19 to 0.20, respectively. The other methods also register increases. The AUC and sensitivity of the best of the three considered machine learning algorithm, SVM, improve from 0.683 to 0.698 and from 25.5% to 28%, respectively. This reveals that the use of the relevant information for the adjacent residues is helpful. Moreover, the lowest predictive performance of kNN could be explained by the low similarity between the proteins in the training and TEST_T datasets.

Fig. 5 compares the ROC curves for hybridNAP and the FCM, SVM, Naïve Bayes and kNN predictors that use window size = 5. The curve for the FCM method is above the other curves for the entire FPrate range. The inset in the lower right corner, which focuses on the low FPrate values, reveals that FCM and hybridNAP provide similar results for FPrate < 0.06. However, FCM provides a visible advantage for the FPrates between 0.06 and 0.8. Interestingly, Naïve Bayes and kNN-based predictors maintain similarly low predictive quality when FPrate<0.1.

A side-by-side comparison of the best FCM model with window size = 5 and hybridNAP shows a substantial advantage for the former model. Table 1 shows that the FCM model secures AUC = 0.171 vs. 0.693 for hybridNAP. We also compare sensitivity values for different levels of FPrates. Based on Fig. 5, we observe that FCM registers 3.8% improvement in sensitivity when FPrate = 10% (sensitivity = 32.6% for FCM vs. 28.8% for hybridNAP), 3.9% increase when FPrate = 20% (sensitivity = 50.6% vs. 46.7%), and 4.6% improvement when FPrate = 30% (sensitivity = 62.2% vs. 57.6%). This confirms that the improvements in sensitivity over hybridNAP are consistent over a wide range of the FPrates.

We investigate statistical significance of differences in the predictive performance between the FCM method and the other considered predictors. We generate results on a diverse collection of protein sets to evaluate whether the improvements offered by the FCM model are robust, i.e., whether they consistent over the considered collection of protein sets. More specifically, we select half of the proteins in the TEST_T dataset at random and without replacement, and we evaluate AUC on these proteins. We repeat this 100 times and report the corresponding averages

and standard errors in Fig. 6. We also run paired *t*-test to assess whether the differences between the 100 pairs of results between the FCM predictor and each of the other four algorithms are statistically significant. The corresponding *p*-values are shown at the top of the Fig. 6. The results demonstrate that the increases in the AUC provided by the FCM are statistically significant when compared with hybridNAP and the methods that rely on the SVM, Naïve Bayes and kNN algorithms. This conclusion holds for every configuration, including approaches with and without the window.

We also assess statistical significance of differences between different versions of the FCM models. The AUC of the FCM that uses window size = 5 is significantly better than the AUC of the version with shorter window (*p*-value = $4.41*10^{-29}$) and without the window (*p*-value = $9.14*10^{-41}$). Similarly, the version that applies window size = 3 obtains significantly higher AUC than the window-less version (*p*-value = $1.54*10^{-35}$). We note that although the magnitude of these improvements is modest (the averages over the 100 experiments are 0.718 vs. 0.713 vs. 0.695), the significance analysis reveals that the corresponding differences are consistent over many diverse protein sets, which on average share only 50% of data.

TABLE 2. COMPARISON OF THE PREDICTIVE PERFORMANCE OF THE SEGMENT-BASED PREDICTORS (FCM MODEL AND HYBRIDNAP) WITH THE WHOLE-SEQUENCE PREDICTORS ON THE TEST_T DATASET. THE LAST LINE SHOWS BASELINE RESULTS PRODUCED WITH A RANDOM PREDICTOR. SPECIFICITY OF ALL METHODS IS SET TO 90% (FPRATE = 10%).

Type of algorithms	Algorithm	AUC	Accuracy [%] at FPrate=10%	Sensitivity [%] at FPrate=10%	MCC at FPrate=10%
Segment and whole-sequence algorithms	FCM	0.717	84.5	32.6	0.203
	hybridNAP	0.685	84.2	28.8	0.170
Whole-sequence algorithms	BindN+	0.797	85.5	43.7	0.293
	DBS-PSSM	0.796	86.0	48.3	0.329
	DP-Bind	0.797	85.6	43.9	0.295
Baseline (random predictor)		0.494	82.3	9.8	-0.001

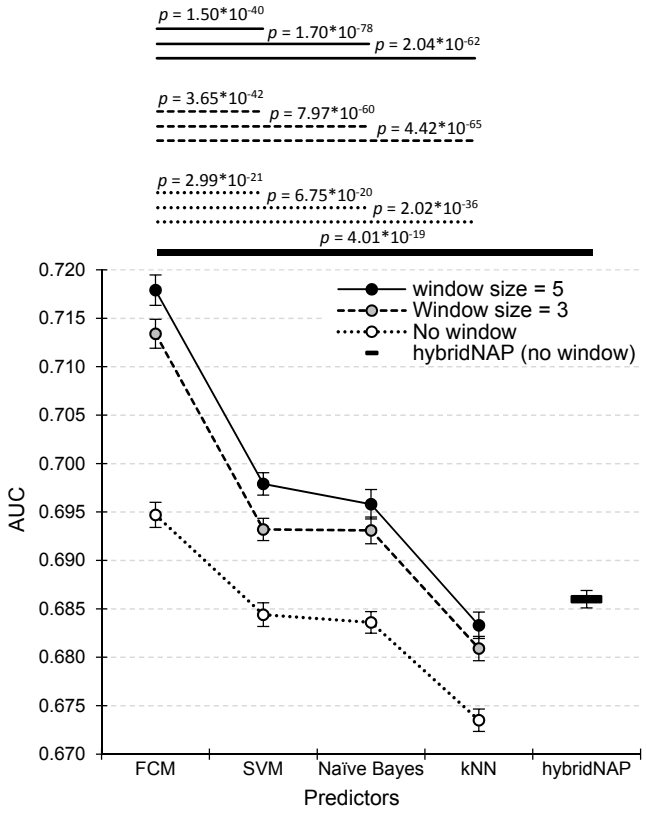


Fig. 6. Analysis of statistical significance of differences in AUCs between FCM and the other predictors including hybridNAP, SVM, Naïve Bayes and kNN on the TEST_T dataset. Solid, dashed and dotted lines represent results secured with window sizes 5, 3 and 1 (no window), respectively.

3.3 Comparison of whole-sequence based Predictors of DNA-binding Residues

We also perform comparative assessment for the whole-sequence methods that predict DNA-binding residues in complete protein sequences. We compare the FCM method and hybridNAP that predict at the sequence fragment and the whole-sequence levels with three popular whole-sequence predictors of the DNA-binding residues that were recently evaluated on the same TEST_T dataset in [12, 19]. These three methods include BindN+ [29], DBS-PSSM [21], and DP-Bind [25, 26]. The results are summarized in Table 2. The AUC for the FCM model is 0.717 while the other fragment-based method, hybridNAP, has a lower AUC = 0.685. Moreover, AUCs for BindN+, DBS-PSSM, and DP-Bind are 0.797, 0.796 and 0.797, respectively. The sensitivity of the whole-sequence based methods, which equals 48% for DBS-PSSM and 44% for BindN+ and DP-Bind, is also higher than the sensitivity for the fragment-based methods that equals 33% for the FCM model. We conclude that, as expected, the whole-sequence based methods provide more accurate results but at a cost of forcing the predictions over the entire protein sequence. The main reason for the higher accuracy is the fact that the whole-sequence based predictors use more information (whole sequence vs. a short fragment) and a much higher number of features. Specifically, BindN+, DP-Bind, and DBS-PSSM use 286, 140 and 100 and features, respectively. To compare, the FCM-based predictor uses an order of magnitude fewer features, i.e., 15 features when window size is set to 5. Consequently, the whole-sequence based methods require substantially longer runtime. They calculate hundreds of features and they also use computationally demanding PSI-BLAST algorithm [81] to derive PSSM, which

in turn is used to compute some of these features. As a result, the whole-sequence based methods take several minutes to generate prediction for an average size protein sequence. To compare, running the FCM model requires only several seconds. Moreover, the FCM model can be used to provide predictions for small segments of the protein chain (say, segments of five consecutive residues when using the version with window size = 5). This is useful when some of the inputs (e.g., evolutionary conservation that requires well-defined position-specific frequencies or PSSM scores, which in turns require sufficiently deep multiple sequence alignment) are not available for some of the residues in the input protein sequence. Importantly, Table 2 demonstrates that the recently released hybridNAP method, which like FCM can be used to predict small segments [12], is outperformed by the FCM-based solution.

4 Summary and Conclusions

We present and empirically test a new FCM-based method for the prediction of DNA-binding residues in local segments of protein sequences. This is the first application of the FCM to model protein-ligand interactions in protein sequences. The FCM model takes three sequence-derived features (RAA, putative RSA, and ECO) in a short sliding window to derive real-valued propensities for the DNA binding. The model was parametrized using the PSO algorithm.

The empirical tests on a recently published benchmark dataset reveal that FCM outperforms several other fragment based approaches that include popular machine learning algorithms (SVM, Naïve Bayes and kNN) and the recently released predictor of DNA binding residues, hybridNAP. These improvements stem from an intrinsic feature of the FCM model which considers not only the relations between inputs and the output (like the other models do) but also relations between the input features. We also demonstrate that the best results are achieved for the window size of 5 and that the improvements offered by our solution are robust.

Although this study focuses on the prediction of DNA-binding residues, the novel FCM-based architecture can be extended to predict other types of interactions, such as protein-protein and protein-RNA interactions. These extensions will be the subject of future work.

Lastly, we make the proposed here predictor available online as a convenient and free webserver named funDNApred (Fuzzy Cognitive Map approach to DNA residues prediction). The computations are performed on the server side and the end user only needs to enter the input protein chain to acquire the predictions. The server accepts up to 10 sequences at the time and it delivers the results via email and in the web browser window. The funDNApred webserver is available at http://biomine.cs.vcu.edu/servers/funDNApred/.

Acknowledgment

The authors thank Dr. Jian Zhang from Xinyang Normal University for the access to datasets. This work was supported in part by the Robert J. Mattauch Endowed Chair funding to L.K.

References

[1] T. Siggers, and R. Gordan, "Protein-DNA binding: complexities and multi-protein codes," *Nucleic Acids Res,* vol. 42, no. 4, pp. 2099-111, Feb, 2014.

[2] R. Sathyapriya, M. S. Vijayabaskar, and S. Vishveshwara, "Insights into Protein-DNA Interactions through Structure Network Analysis," *Plos Comp Biology,* vol. 4, no. 9, Sep, 2008.

[3] P. Prabakaran, J. G. Siebers, S. Ahmad, M. M. Gromiha, M. G. Singarayan, and A. Sarai, "Classification of protein-DNA complexes based on structural descriptors," *Structure,* vol. 14, no. 9, pp. 1355-67, Sep, 2006.

[4] N. R. Steffen, S. D. Murphy, L. Tolleri, G. W. Hatfield, and R. H. Lathrop, "DNA sequence and structure: direct and indirect recognition in protein-DNA binding," *Bioinformatics,* vol. 18 Suppl 1, pp. S22-30, 2002.

[5] B. F. Pugh, and D. S. Gilmour, "Genome-wide analysis of protein-DNA interactions in living cells," *Genome Biol,* vol. 2, no. 4, pp. REVIEWS1013, 2001.

[6] S. Jones, P. van Heyningen, H. M. Berman, and J. M. Thornton, "Protein-DNA interactions: A structural analysis," *J Mol Biol,* vol. 287, no. 5, pp. 877-96, Apr 16, 1999.

[7] D. Lejeune, N. Delsaux, B. Charloteaux, A. Thomas, and R. Brasseur, "Protein-nucleic acid recognition: statistical analysis of atomic interactions and influence of DNA structure," *Proteins,* vol. 61, no. 2, pp. 258-71, Nov 1, 2005.

[8] V. Charoensawan, D. Wilson, and S. A. Teichmann, "Genomic repertoires of DNA-binding transcription factors across the tree of life," *Nucleic acids research,* vol. 38, no. 21, pp. 7364-7377, 2010.

[9] R. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L. S. Yeh, "UniProt: the Universal Protein knowledgebase," *Nucleic Acids Res,* vol. 32, no. Database issue, pp. D115-9, Jan 01, 2004.

[10] C. The UniProt, "UniProt: the universal protein knowledgebase," *Nucleic Acids Res,* vol. 45, no. D1, pp. D158-D169, Jan 4, 2017.

[11] J. Si, R. Zhao, and R. Wu, "An overview of the prediction of protein DNA-binding sites," *Int J Mol Sci,* vol. 16, no. 3, pp. 5194-215, 2015.

[12] J. Zhang, Z. Ma, and L. Kurgan, "Comprehensive review and empirical analysis of hallmarks of DNA-, RNA- and protein-binding residues in protein chains," *Brief Bioinform*, Dec 15, 2017.

[13] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Res,* vol. 28, no. 1, pp. 235-42, Jan 1, 2000.

[14] H. M. Berman, G. J. Kleywegt, H. Nakamura, and J. L. Markley, "The Protein Data Bank at 40: reflecting on the past to prepare for the future," *Structure,* vol. 20, no. 3, pp. 391-6, Mar 7, 2012.

[15] M. J. Mizianty, X. Fan, J. Yan, E. Chalmers, C. Woloschuk, A. Joachimiak, and L. Kurgan, "Covering complete proteomes with X-ray structures: a current snapshot," *Acta Crystallogr D Biol Crystallogr,* vol. 70, no. Pt 11, pp. 2781-93, Nov, 2014.

[16] Q. C. Zhang, D. Petrey, L. Deng, L. Qiang, Y. Shi, C. A. Thu, B. Bisikirska, C. Lefebvre, D. Accili, T. Hunter, T. Maniatis, A. Califano, and B. Honig, "Structure-based prediction of protein-protein interactions on a genome-wide scale," *Nature,* vol. 490, no. 7421, pp. 556-60, Oct 25, 2012.

[17] C. Kauffman, and G. Karypis, "Computational tools for protein-DNA interactions," *Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery,* vol. 2, no. 1, pp. 14-28, Jan-Feb, 2012.

[18] R. Nagarajan, S. Ahmad, and M. M. Gromiha, "Novel approach for selecting the best predictor for identifying the binding sites in DNA binding proteins," *Nucleic Acids Res,* vol. 41, no. 16, pp. 7606-14, Sep, 2013.

[19] J. Yan, S. Friedrich, and L. Kurgan, "A comprehensive comparative review of sequence-based predictors of DNA- and RNA-binding residues," *Brief Bioinform,* vol. 17, no. 1, pp. 88-105, May 1, 2016.

[20] S. Ahmad, M. M. Gromiha, and A. Sarai, "Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information," *Bioinformatics,* vol. 20, no. 4, pp. 477-486, 2004.

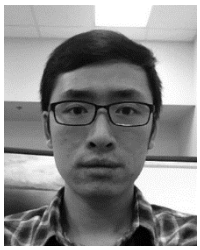
[21] S. Ahmad, and A. Sarai, "PSSM-based prediction of DNA binding sites in proteins," *BMC bioinformatics,* vol. 6, no. 1, pp. 33, 2005.

[22] L. Wang, and S. J. Brown, "BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences," *Nucleic acids research,* vol. 34, no. suppl 2, pp. W243-W248, 2006.
[23] C. Yan, M. Terribilini, F. Wu, R. L. Jernigan, D. Dobbs, and V. Honavar, "Predicting DNA-binding sites of proteins from amino acid sequence," *BMC bioinformatics,* vol. 7, no. 1, pp. 1, 2006.
[24] J.-h. Lee, M. Hamilton, C. Gleeson, C. Caragea, P. Zaback, J. D. Sander, X. Li, F. Wu, M. Terribilini, and V. Honavar, "Striking similarities in diverse telomerase proteins revealed by combining structure prediction and machine learning approaches." p. 501.
[25] S. Hwang, Z. Gou, and I. B. Kuznetsov, "DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins," *Bioinformatics,* vol. 23, no. 5, pp. 634-636, 2007.
[26] I. B. Kuznetsov, Z. Gou, R. Li, and S. Hwang, "Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins," *PROTEINS: Structure, Function, and Bioinformatics,* vol. 64, no. 1, pp. 19-27, 2006.
[27] Y. Ofran, V. Mysore, and B. Rost, "Prediction of DNA-binding residues from sequence," *Bioinformatics,* vol. 23, no. 13, pp. i347-i353, 2007.
[28] W. Y. Chu, Y. F. Huang, C. C. Huang, Y. S. Cheng, C. K. Huang, and Y. J. Oyang, "ProteDNA: a sequence-based predictor of sequence-specific DNA-binding residues in transcription factors," *Nucleic Acids Research,* vol. 37, pp. W396-W401, Jul 1, 2009.
[29] L. Wang, C. Huang, M. Q. Yang, and J. Y. Yang, "BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features," *BMC Systems Biology,* vol. 4, no. 1, pp. 1, 2010.
[30] M. B. Carson, R. Langlois, and H. Lu, "NAPS: a residue-level nucleic acid-binding prediction server," *Nucleic acids research,* vol. 38, no. suppl 2, pp. W431-W435, 2010.
[31] J. Si, Z. Zhang, B. Lin, M. Schroeder, and B. Huang, "MetaDBSite: a meta approach to improve protein DNA-binding sites prediction," *BMC systems biology,* vol. 5, no. Suppl 1, pp. S7, 2011.
[32] Z. Peng, C. Wang, V. N. Uversky, and L. Kurgan, "Prediction of Disordered RNA, DNA, and Protein Binding Regions Using DisoRDPbind," *Methods Mol Biol,* vol. 1484, pp. 187-203, 2017.
[33] Z. Peng, and L. Kurgan, "High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder," *Nucleic Acids Res,* vol. 43, no. 18, pp. e121, Oct 15, 2015.
[34] J. Yan, and L. Kurgan, "DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues," *Nucleic Acids Res,* vol. 45, no. 10, pp. e84, Jun 02, 2017.
[35] B. Kosko, "Fuzzy cognitive maps," *International Journal of Man-Machine Studies,* vol. 24, no. 1, pp. 65-75, 1986/01/01/, 1986.
[36] L. A. Kurgan, W. Stach, and J. Ruan, "Novel scales based on hydrophobicity indices for secondary protein structure," *J Theor Biol,* vol. 248, no. 2, pp. 354-66, Sep 21, 2007.
[37] H. J. J. Song, C. Y. Y. Miao, R. Wuyts, Z. Q. Q. Shen, M. D'Hondt, and F. Catthoor, "An Extension to Fuzzy Cognitive Maps for Classification and Prediction," *IEEE Transactions on Fuzzy Systems,* vol. 19, no. 1, pp. 116-135, Feb, 2011.
[38] W. Stach, L. A. Kurgan, and W. Pedrycz, "Numerical and linguistic prediction of time series with the use of fuzzy cognitive maps," *IEEE Transactions on Fuzzy Systems,* vol. 16, no. 1, pp. 61-72, Feb, 2008.
[39] A. Amirkhani, E. I. Papageorgiou, A. Mohseni, and M. R. Mosavi, "A review of fuzzy cognitive maps in medicine: Taxonomy, methods, and applications," *Computer Methods and Programs in Biomedicine,* vol. 142, pp. 129-145, Apr, 2017.
[40] J. Subramanian, A. Karmegam, E. Papageorgiou, N. Papandrianos, and A. Vasukie, "An integrated breast cancer risk assessment and management model based on fuzzy cognitive maps," *Computer Methods and Programs in Biomedicine,* vol. 118, no. 3, pp. 280-297, 2015/03/01/, 2015.
[41] A. Amirkhani, M. R. Mosavi, K. Mohammadi, and E. I. Papageorgiou, "A novel hybrid method based on fuzzy cognitive maps and fuzzy clustering algorithms for grading celiac disease," *Neural Computing and Applications*, pp. DOI: 10.1007/s00521-016-2765-y, 2016.
[42] E. I. Papageorgiou, A. S. Billis, C. Frantzidis, E. I. Konstantinidis, and P. D. Bamidis, "A preliminary fuzzy cognitive map - based desicion support tool for geriatric depression assessment," in 2013 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2013, pp. 1-8.
[43] E. I. Papageorgiou, and K. Poczeta, "Application of Fuzzy Cognitive Maps to Electricity Consumption Prediction," in 2015 Annual Meeting of the North American Fuzzy Information Processing Society, 2015, pp. 1-6.
[44] E. I. Papageorgiou, K. Poczeta, and C. Laspidou, "Application of Fuzzy Cognitive Maps to water demand prediction," *2015 IEEE International Conference on Fuzzy Systems (Fuzz-IEEE 2015)*, 2015.
[45] R. Sarala, G. Zayaraz, and V. Vijayalakshmi, "Fuzzy cognitive map-based reasoning for prediction of multi-stage attacks in risk assessment," *International Journal of Intelligent Engineering Informatics,* vol. 4, no. 2, pp. 151-167, 2016.
[46] J. Yang, A. Roy, and Y. Zhang, "BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions," *Nucleic Acids Research,* vol. 41, no. D1, pp. D1096-D1103, 2013.
[47] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J Mol Biol,* vol. 215, no. 3, pp. 403-10, Oct 5, 1990.
[48] S. Velankar, J. M. Dana, J. Jacobsen, G. van Ginkel, P. J. Gane, J. Luo, T. J. Oldfield, C. O'Donovan, M. J. Martin, and G. J. Kleywegt, "SIFTS: Structure Integration with Function, Taxonomy and Sequences resource," *Nucleic Acids Res,* vol. 41, no. Database issue, pp. D483-9, Jan, 2013.
[49] R. Sathyapriya, M. S. Vijayabaskar, and S. Vishveshwara, "Insights into protein-DNA interactions through structure network analysis," *PLoS Comput Biol,* vol. 4, no. 9, pp. e1000170, Sep 05, 2008.
[50] S. Dey, A. Pal, M. Guharoy, S. Sonavane, and P. Chakrabarti, "Characterization and prediction of the binding site in DNA-binding proteins: improvement of accuracy by combining residue composition, evolutionary conservation and structural parameters," *Nucleic Acids Res,* vol. 40, no. 15, pp. 7150-7161, Aug, 2012.
[51] S. Ahmad, O. Keskin, A. Sarai, and R. Nussinov, "Protein-DNA interactions: structural, thermodynamic and clustering patterns of conserved residues in DNA-binding proteins," *Nucleic Acids Res,* vol. 36, no. 18, pp. 5922-5932, Oct, 2008.
[52] W. Wang, J. Liu, Y. Xiong, L. Zhu, and X. Zhou, "Analysis and classification of DNA-binding sites in single-stranded and double-stranded DNA-binding proteins using protein information," *IET Syst Biol,* vol. 8, no. 4, pp. 176-83, Aug, 2014.
[53] N. M. Luscombe, and J. M. Thornton, "Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity," *J Mol Biol,* vol. 320, no. 5, pp. 991-1009, Jul 26, 2002.
[54] V. Vacic, V. N. Uversky, A. K. Dunker, and S. Lonardi, "Composition Profiler: a tool for discovery and visualization of amino acid composition differences," *BMC Bioinformatics,* vol. 8, pp. 211, 2007.
[55] W. Kabsch, and C. Sander, "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers,* vol. 22, no. 12, pp. 2577-2637, 1983.
[56] E. Faraggi, Y. Q. Zhou, and A. Kloczkowski, "Accurate single-sequence prediction of solvent accessible surface area using local and global features," *Proteins,* vol. 82, no. 11, pp. 3170-3176, Nov, 2014.
[57] M. Z. Tien, A. G. Meyer, D. K. Sydykova, S. J. Spielman, and C. O. Wilke, "Maximum Allowed Solvent Accessibilites of Residues in Proteins," *PLOS ONE,* vol. 8, no. 11, pp. e80635, 2013.
[58] M. Remmert, A. Biegert, A. Hauser, and J. Soding, "HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment," *Nat Methods,* vol. 9, no. 2, pp. 173-5, Feb, 2012.
[59] J. Fischer, C. Mayer, and J. Soding, "Prediction of protein functional residues from sequence by probability density estimation," *Bioinformatics,* vol. 24, no. 5, pp. 613-20, Mar 1, 2008.
[60] Y. Dou, X. Zheng, J. Yang, and J. Wang, "Prediction of catalytic residues based on an overlapping amino acid classification," *Amino Acids,* vol. 39, no. 5, pp. 1353-61, Nov, 2010.
[61] S. L. Li, K. Yamashita, K. M. Amada, and D. M. Standley, "Quantifying sequence and structural features of protein-RNA interactions," *Nucleic Acids Res,* vol. 42, no. 15, pp. 10086-10098, 2014.
[62] W. Stach, L. Kurgan, W. Pedrycz, and M. Reformat, "Learning fuzzy cognitive maps with required precision using genetic algorithm approach," *Electronics Letters,* vol. 40, no. 24, pp. 1519-1520, Nov 25, 2004.
[63] W. Stach, L. Kurgan, W. Pedrycz, and M. Reformat, "Genetic learning of fuzzy cognitive maps," *Fuzzy Sets and Systems,* vol. 153, no. 3, pp. 371-401, 2005/08/01/, 2005.
[64] E. I. Papageorgiou, "Learning Algorithms for Fuzzy Cognitive Maps-A Review Study," *IEEE Transactions on Systems Man and Cybernetics Part C-Applications and Reviews,* vol. 42, no. 2, pp. 150-163, Mar, 2012.

[65] G. Napoles, E. Papageorgiou, R. Bello, and K. Vanhoof, "Learning and Convergence of Fuzzy Cognitive Maps Used in Pattern Recognition," *Neural Processing Letters,* vol. 45, no. 2, pp. 431-444, Apr, 2017.
[66] S. Bueno, and J. L. Salmeron, "Benchmarking main activation functions in fuzzy cognitive maps," *Expert Systems with Applications,* vol. 36, no. 3, pp. 5221-5229, 2009.
[67] D. Grant, and K. M. Osei-Bryson, "Using Fuzzy Cognitive Maps to Assess MIS Organizational Change Impact," in 38th Annual Hawaii International Conference on System Sciences, 2005, pp. 263c-263c.
[68] W. Stach, L. Kurgan, and W. Pedrycz, "Expert-Based and Computational Methods for Developing Fuzzy Cognitive Maps," *Fuzzy Cognitive Maps: Advances in Theory, Methodolologies, Tools and Applications,* vol. 247, pp. 23-41, 2010.
[69] E. Papageorgiou, C. Stylios, and P. Groumpos, "Fuzzy Cognitive Map Learning Based on Nonlinear Hebbian Rule," *AI 2003: Advances in Artificial Intelligence: 16th Australian Conference on AI, Perth, Australia, December 3-5, 2003. Proceedings*, T. D. Gedeon and L. C. C. Fung, eds., pp. 256-268, Berlin, Heidelberg: Springer Berlin Heidelberg, 2003.
[70] W. Stach, L. Kurgan, and W. Pedrycz, "Data-Driven Nonlinear Hebbian Learning Method for Fuzzy Cognitive Maps," *2008 IEEE International Conference on Fuzzy Systems, Vols 1-5*, pp. 1977-1983, 2008.
[71] W. Stach, W. Pedrycz, and L. A. Kurgan, "Learning of fuzzy cognitive maps using density estimate," *IEEE Trans Syst Man Cybern B Cybern,* vol. 42, no. 3, pp. 900-12, Jun, 2012.
[72] W. Stach, L. Kurgan, and W. Pedrycz, "A divide and conquer method for learning large Fuzzy Cognitive Maps," *Fuzzy Sets and Systems,* vol. 161, no. 19, pp. 2515-2532, Oct 1, 2010.
[73] Y. Chen, L. Mazlack, and L. Lu, "Learning fuzzy cognitive maps from data by ant colony optimization," in Proceedings of the 14th annual conference on Genetic and evolutionary computation, Philadelphia, Pennsylvania, USA, 2012, pp. 9-16.
[74] Y. G. Petalas, K. E. Parsopoulos, and M. N. Vrahatis, "Improving fuzzy cognitive maps learning through memetic particle swarm optimization," *Soft Computing,* vol. 13, no. 1, pp. 77-94, Jan, 2009.
[75] P. Oikonomou, and E. I. Papageorgiou, "Particle Swarm Optimization Approach for Fuzzy Cognitive Maps Applied to Autism Classification," *Artificial Intelligence Applications and Innovations, Aiai 2013,* vol. 412, pp. 516-526, 2013.
[76] H. Nasiriyan-Rad, A. Amirkhani, A. Naimi, and K. Mohammadi, "Learning Fuzzy Cognitive Map with PSO Algorithm for Grading Celiac Disease," *2016 23rd Iranian Conference on Biomedical Engineering and 2016 1st International Iranian Conference on Biomedical Engineering (Icbme)*, pp. 336-341, 2016.
[77] J. L. Salmeron, S. A. Rahimi, A. M. Navali, and A. Sadeghpour, "Medical diagnosis of Rheumatoid Arthritis using data driven PSO-FCM with scarce datasets," *Neurocomputing,* vol. 232, pp. 104-112, Apr 5, 2017.
[78] R. Eberhart, and J. Kennedy, "A new optimizer using particle swarm theory," in Micro Machine and Human Science Symposium, 1995, pp. 39-43.
[79] M. R. Sierra, and C. A. C. Coello, "Improving PSO-Based multi-objective optimization using crowding, mutation and epsilon-dominance," *Evolutionary Multi-Criterion Optimization,* vol. 3410, pp. 505-519, 2005.
[80] M. R. Bonyadi, and Z. Michalewicz, "Particle Swarm Optimization for Single Objective Continuous Space Problems: A Review," *Evol Comput,* vol. 25, no. 1, pp. 1-54, Spring, 2017.
[81] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Res,* vol. 25, no. 17, pp. 3389-402, Sep 1, 1997.

Abdollah Amirkhani received the MSc and PhD degrees (with honors) in electrical engineering from Iran University of Science and Technology (IUST), Tehran, in 2012 and 2017, respectively. He earned the Outstanding Student Award (2015) among all PhD students of electrical engineering in Iran. He is currently a lecturer in the school of computer engineering at IUST. His research interests are in soft computing, fuzzy cognitive maps, data mining and machine learning.

Mojtaba Kolahdoozi received the BSc degree in electronics engineering from the University of Tabriz in 2014 and the MSc degree in digital electronics from Iran University of Science and Technology (IUST) in 2017. He is currently working as a research assistant in the Fuzzy Logic laboratory at IUST. His research interests include fuzzy logic, evolutionary computation, image processing, data mining and bioinformatics.



Chen Wang received the PhD degree in Computer Science from the Virginia Commonwealth University and the MSc degree in Information and Communication Engineering from the Harbin Institute of Technology, China, in 2018 and 2011, respectively. He is currently a postdoctoral fellow at the Columbia University. His research interests include machine learning and statistics in bioinformatics, in particular characterization and prediction of protein structures, functions, and protein-ligand interactions. He has served as a reviewer for several journals including IEEE Transactions on Geoscience and Remote Sensing, IEEE Access, and EURASIP Journal on Advances in Signal Processing. He is a student member of SPIE.

Lukasz A. Kurgan received the MSc degree in robotics from AGH University of Science and Technology in 1999, and the PhD degree in computer science from the University of Colorado at Boulder in 2003. He is the Robert J. Mattauch Endowed Professor and vice-Chair in the Computer Science Department at the Virginia Commonwealth University. He is Associate Editor-in-Chief of *Biomolecules*, structural bioinformatics area Editor for *BMC Bioinformatics* and serves on the editorial boards of several other journals including *International Journal of Molecular Sciences* and *Current Protein and Peptide Science*. He has published over 130 journal articles in venues such as *Chemical Reviews*, *Nucleic Acids Research*, *Bioinformatics*, *Science Signaling*, *Cellular and Molecular Life Sciences*, and *Briefings in Bioinformatics*. His research interests are in structural bioinformatics of proteins and RNAs, intrinsic disorder, structural genomics, and protein-ligand/drug interactions. His research was funded by NSF, CIHR, NSERC, and Alberta Cancer Foundation. He is a fellow of AIMBE, fellow of the Kosciuszko Foundation Collegium of Eminent Scientists, and a senior member of ACM.