Enhanced Prediction of Intrinsically Disordered Regions with the Optimized Data

Irem Ersöz Kaya, Ayça Çakmak Pehlivanlı, and Turgay Ibrikci

***Abstract*—A protein that lacks a three-dimensional (3-D) structure in its intrinsic state has been called natively unfolded or intrinsically disordered. The observation that many intrinsically disordered protein regions play a key role in many essential functions has promoted increased interest in studies on the structural identification of intrinsically disordered proteins in the field of bioinformatics. Since amino acid sequence have been widely used for the determination of protein structure, it has been theorized that the sequence could also determine disorder. To improve the quality of prediction, recent studies have focused on finding more useful features and developing more robust predictors. Machine learning techniques are ideally used for extracting the complex relationships and correlations hidden in large data sets. In the study, several features of the chosen proteins were combined together in different ways to obtain an optimized dataset and prediction was accomplished by using the most common method, SVM, resulting in significant increase in success rate with the modeled data. Besides, the feature selection method, ERGS, was used to explore the optimum features that have the adequate information on finding disorder. In the research, 37 attributes were found to be the most influential features in predicting disordered regions.**

***Index Terms*—Effective range based gene selection, intrinsically disordered proteins, support vector machine, structural prediction.**

I. Introduction

The rapid pace of genomic sequencing and the abundance of the sequence data have motivated the studies on structure prediction of the protein from its amino acid sequence. These predictions are then used to deduce function. This scheme assumes that tertiary structures are a prerequisite for function [1]. Generally, the loss of protein function is associated with the lack of specific three-dimensional (3D) structure. However, this view ignores numerous proteins that utilize unfolded or incompletely folded regions for function [2]. Proteins that lacks 3D structures in its intrinsic state has been called natively unfolded or intrinsically disordered [3]. Many proteins that have disordered regions of amino acid sequences are also involved in a variety of biological functions [4].

Intrinsically disordered protein regions have been found to play a key role in many vital functions including DNA binding, cell signaling, protein modification and also in some diseases such as Alzheimer and Parkinson diseases [5]. These discoveries have induced interest in identification of intrinsically disordered proteins. Afterwards, they are thought to be used in drug design, protein expression and functional recognition.

Since amino acid sequence attributes have been widely used for determining protein structure, it has been theorized that the sequence would also determine disorder. Thus, as alternatives to experimental methods, several computational methods have been suggested for disorder prediction based on the amino acid sequence. Most preferred machine learning techniques can be cited as neural networks [6]-[8] and support vector machines [9], [10].

In most studies, the input patterns have been mostly derived from a variety of sequence properties that characterize disorder namely, flexibility, amino acid frequency, complexity, charge, and secondary structure [11], [12]. To improve the quality of prediction, recent studies have sought to find more useful features and to develop more robust predictor [7], [10], [13] For instance, Su *et al.* have used a condensed position specific scoring matrix profile by merging associated columns of the matrix concerned with several physicochemical properties of amino acids [14]. They achieved rather successful prediction via training a Radial Basis Function (RBF) neural network with their proposed PSSMP patterns. Position specific scoring matrix profiles obtained from Psi-Blast execution involve the evolutionary information with the level of position conservation. As a result specific disorder prediction tools were developed such as PONDRs [5], [15], DisEMBL [16] GlobPlot [17] DISOPRED2 [2], FoldIndex [18], RONN [7], DisPRO [8], PreLink [6] and DisPSSMP [14].

Machine learning techniques are ideal for use in extraction of complex relationships and correlations hidden in large data sets, since, the methods are well suited for the typically multidimensional, noisy, and complex data of computational biology. In this study, our aim is to determine an efficient computational way that can provide information about the structural class of ordered and disordered proteins using different biochemical and physical features of amino acids. To achieve an accurate prediction of disordered data, several features of the chosen proteins were combined together in different ways to obtain an optimized dataset and Support Vector Machine (SVM) that is one of the successful methods in computational prediction was used to explore the success rate with the optimized data. Subsequently, the dimension of the data was reduced by using the most informative features that were selected by the Effective Based Range Gene Selection method [18].

Manuscript received July 2, 2013; revised September 24, 2013.

Irem Ersöz Kaya is with the Department of Software Engineering, Mersin University, Mersin, Turkey (e-mail: iremer@ mersin.edu.tr).

Ayça Çakmak Pehlivanlı is with the Department of Statistics, Mimar Sinan Fine Arts University, Istanbul, Turkey (e-mail: ayca.pehlivanli@msgsu.edu.tr).

Turgay Ibrikci is with the Department of Electrical-Electronics Engineering, Cukurova University, 01330, Adana, Turkey (e-mail: ibrikci@cu.edu.tr).

II. Datasets and Methods

A. Datasets

In this study, two data sets were used. The first data set, containing 80 completely ordered proteins, was obtained from PONDR® website by Yang *et al.* For the second data set, 79 out of 91 proteins from a study conducted by Uversky *et al.* were used [19]. They reported 91 completely disordered proteins defined through spectroscopic methods in the literature. The compositions of the two protein sets named as CO80 and CD79, respectively, are shown in Table I.

TABLE I: The Compositions of Datasets

	CO80	**CD79**
Number of Chains	80	79
Number of Ordered Regions	80	0
Number of Disordered Regions	0	79
Number of Ordered Residues	16568	0
Number of Disordered Residues	0	14462
Total Residues	16568	14462

A dataset constituted from the two chosen protein sets, was developed by balancing equal amounts of completely ordered (CO) and completely disordered (CD) protein sets, named as COD159. This dataset was used for training and then subsequently used to test the methods to measure the performances.

In the study, an input pattern was derived by using the several physicochemical properties, the evolutionary knowledge and the compositions of amino acids in the window, as given in the study of Ersöz Kaya *et al.* [20]. The 120 attributes of each pattern were chosen based on the following criteria: forty-nine (49) property scales of amino acids from the literatures and AAIndex; the measure of sequence complexity called K2 entropy; the first order statistics of 20 known amino acids; the compositions for 30 different property groups of amino acids; and the 20 columns of the position-specific scoring matrix representing probabilities of conservation against mutations to 20 different amino acids. An attribute value for an amino acid in a given position is calculated by averaging the values regarding the related information of all the amino acids in the window.

The final pattern comprise of 50 composition-based attributes, 50 property-based attributes and 20 evolution-based attributes. Whole list that contain the contents and the references of all properties are given in the thesis of Ersöz Kaya [21]. For each attribute, the values of the input patterns were rescaled between 0-1 by min-max normalization technique.

In order to achieve successful models in predicting the protein structure, it is preferable to consider neighboring amino acids in the knowledge representation [12]. Thus, information about an amino acid in a protein was obtained by using all the amino acids surrounding it within a predetermined residue length window. In addition, the real value representations of amino acids and the average information of all the amino acids within the window were used to represent each feature with only one attribute in a pattern. Thus, dependency on the window size and the nuisance of dimensionality were prevented.

B. Performance Measure

Sensitivity, Specificity, Accuracy, and Matthews' Correlation Coefficient are widely used indices to quantify the prediction performance [14]. Unfortunately, all given measures are fairly affected by the relative frequency of the target, and they are not sufficient in an isolated evaluation.

Therefore, probability excess has been proposed as an unbiased measure for evaluating the performance of prediction [7]. Using probability excess provides an independent measure opportunity of the relative class frequencies by means of the evaluation of sensitivity and specificity values cooperatively, sensitivity + specificity − 1. Probability excess measure can be graphed by a plot of sensitivity versus specificity. It is defined by the following equation;

$$\text{Probability Excess } (probEx) = \frac{TPxTN - FPxFN}{(TP + FN)x(TN + FP)} \quad (1)$$

According to the probability excess criteria, the values greater than 0.5 reveal is an acceptable prediction performance.

C. Support Vector Machines

A support vector machine is a statistical learning technique based on the Structural Risk Minimization (SRM) principle. The algorithm which was introduced by Vapnik is originally designed to deal with linear binary classification problems [22]. The basic idea of the SVM is to find an optimum hyperplane that yields the largest margin between the two classes. Afterwards, SVM was extended to solve nonlinear cases by applying kernel techniques [23].

Let $\{(x_i, y_i) | i = 1, \ldots, n\}$ where $x_i \in R^n$ denotes the input vectors and $y_i \in \{+1, -1\}$ specifies the class labels belonging to the samples. Solving the following quadratic optimization problem provides the optimum separating hyperplane;

$$\text{minimize} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \quad (2)$$

$$\text{subject to} \quad y_i\left(\langle w \cdot x_i \rangle - b\right) \geq 1 - \xi_i \quad \text{if } y_i = 1$$

$$y_i\left(\langle w \cdot x_i \rangle - b\right) \leq 1 - \xi_i \quad \text{if } y_i = -1$$

$$\xi_i \geq 0$$

where ξ_i is called a slack variable used for measuring the occurred error at point (x_i, y_i). C is the penalty parameter that trades off the margin size for the number of misclassified data points.

The solution of the quadratic optimization problem is given by the saddle point of the Lagrange function [24].

$$L(w,b,\alpha)=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}\xi_i - \sum_{i,j=1}^{n}\alpha_i\left(y_i\left(\langle w\cdot x_i\rangle-b\right)-1+\xi_i\right)-\sum_{i=1}^{n}\mu_i\xi_i \quad (3)$$

α_i denotes Lagrange multipliers in (3) where $\alpha_i \geq 0$. The problem is simplified into the Lagrangian dual problem by using Karush-Kuhn-Tucker (KKT) conditions [25].

To find the optimal solution, a dual Lagrangian must be maximized with respect to nonnegative multipliers α_i, as given in (4)

$$\text{maximize} \quad L_D=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j\langle x_i\cdot x_j\rangle \quad (4)$$

$$\text{subject to} \quad \sum_{i=1}^{n}(\alpha_i y_i)=0$$

$$0\leq\alpha_i\leq C$$

The solution occurs when $w^*=\sum y_i\alpha_i^* x_i$ [26]. This leads to expressing the optimal decision function as

$$f(x)=\operatorname{sgn}\left(\sum_{i=1}^{n}y_i\alpha_i^*\langle x_i\cdot x\rangle+b^*\right) \quad (5)$$

To succeed in nonlinear classification tasks, the input space is mapped into a higher-dimensional (Hilbert) feature space via a nonlinear mapping function, $\phi: R^n\rightarrow H$ [25]. The computation of the dot product $\langle x_i\cdot x\rangle=\langle\phi(x_i)\cdot\phi(x)\rangle$ can be reduced by using a kernel function that returns the inner product of the feature mapping, stated as $K(x_i,x)=\langle\phi(x_i)\cdot\phi(x)\rangle$. Thus, the decision function can be expressed as

$$f(x)=\operatorname{sgn}\left(\sum_{i=1}^{n}y_i\alpha_i^*K(x_i,x)+b^*\right) \quad (6)$$

Each nonzero α_i indicates that the corresponding x_i is a support vector.

D. Effective Range Based Gene Selection

In this study, an input pattern is defined by using 120 features as given detailed above. Besides the main purpose of the study, it is also aimed to select the features that are potentially useful to distinguish disorder and order proteins. Selection subset of features also effects the computational time and complexity of this high dimensional dataset.

In the selection process of more informative features subset, a novel and efficient feature selection and ranking approach called Effective Range Based Gene Selection (ERGS) which is introduced by Chandra and Gupta in 2011 is applied [18]. They described ERGS algorithm based on effective range which is defined using statistical inference theory [27].

According to statistical inference theory, there is a negative correlation between magnitude of the confidence interval and reliability of the estimate. Since presence of outliers and high class variance causes wider range of confidence interval, ERGS was designed to overcome these problems. ERGS algorithm also uses Chebyshev's inequality which gives a distribution independent bound.

The effective range takes into account of statistical inference theory, Chebyshev's inequality and class prior probabilities. The ERGS algorithm basically tries to find the feature weights by using this effective range of each class. The discrimination is determined according to these weights. If the decision boundaries between classes are far away from each other, the weights that are calculated for features will be higher. Effective ranges of these features do not overlap or have smaller overlapping area.

There are two major differences between the ERGS and most of the existing feature selection algorithms; there is no search requirements and iterative process for feature subset selection in the nature of ERGS algorithm.

III. Results

One of the goals of this study was to measure the effectiveness of the SVM method on locating the disordered regions of proteins that exist within a modeled dataset. For this purpose, the sequence of chosen proteins was first modeled by way of a scheme in which several features of amino acids are combined together by locally coding at each position in order to obtain an optimized protein dataset. The method was then implemented on the dataset labeled COD159. The results obtained were compared to the eleven methods presented in the literature as publicly available.

In the study, the dataset was partitioned into six different subsets, each of which had approximately equal sequence lengths and a balanced number of ordered and disordered residues. One of the subsets was designated as the validation set to find the optimum parameters. The remaining subsets were used to derive the prediction performance scores via a 5 fold cross validation. The subsets are called CV1, CV2, CV3, CV4 and CV5. While performing the cross validation, the success of the training performance for four combined subsets was validated against the remaining subset, with the process being repeated for each subsequent subset. The success rates from the five subsets were then averaged to obtain a measure of the model's performance.

Before the performance trials of the methods, the series of runs on the validation dataset were executed to determine the optimal values for the parameter C and σ of the SVM. For this purpose it is performed grid search on penalty parameter C and kernel parameter σ using cross validation. Initially, the penalty parameter C was tried by increasing the values between 1 and 100 in steps of 10. Because of more favorable results in low levels of C values, the increment reduced to 1 between 1 and 30.

During the trials, trainings were performed for all values of C accompanying each value of σ, which was initially

assigned values between 1 and 30 in increments of 5. The interval was later diminished to 0.5, regarding to the highest values obtained for the success rate. Finally, C and σ was set at 10 and 3, respectively.

In this study, the SVM was trained with the exponential kernel function for mapping the decision surface. After the learning phase, which was performed with the selected parameters, the testing phase was executed on the test data of each subset.

The results of the SVM with the modeled data (SVM_COD159) were given in terms of the following measures: Sensitivity (*Sens*), Specificity (*Spec*), Accuracy (*Acc*), Matthew's Correlation Coefficient (*Mcc*) and Probability Excess (*probEx*). However, they were arranged according to the order of the *probEx* criteria in the tables.

TABLE II: THE PERFORMANCE OF SVM_COD159

	CV1	CV2	CV3	CV4	CV5	Mean
Acc	0.913	0.712	0.834	0.813	0.825	0.819
probEx	0.825	0.423	0.667	0.626	0.650	0.638
Mcc	0.826	0.427	0.669	0.626	0.656	0.641
Spec	0.918	0.781	0.866	0.810	0.894	0.854
Sens	0.907	0.643	0.801	0.816	0.756	0.786
AUC	0.971	0.781	0.906	0.904	0.920	0.897

The performances attained after testing the modeled data are presented in Table II. The average of the accuracy rates of the SVM_COD159 reaches a level of 81.9% yielding the *probEx* values over 0.5 for all subsets except subset 2.

TABLE III: THE PERFORMANCE COMPARISON OF SVM_COD159 VS. THE ELEVEN PREDICTION TOOLS

Methods	*Sens*	*Spec*	*Acc*	*Mcc*	*probEx*
SVM_COD159	0.786	0.854	0.819	0.641	0.638
DisPSSMP	0.825	0.765	0.795	0.589	0.590
RONN	0.675	0.888	0.782	0.580	0.563
FoldIndex	0.722	0.815	0.769	0.540	0.536
DISOPRED2	0.469	0.981	0.725	0.543	0.449
PONDR	0.632	0.782	0.707	0.420	0.414
DisPro	0.383	0.982	0.683	0.467	0.365
DisEMBL(465)	0.348	0.978	0.663	0.430	0.327
PreLink	0.319	0.991	0.655	0.430	0.310
DisEMBL(hot)	0.502	0.749	0.626	0.260	0.251
DisEMBL(coil)	0.719	0.446	0.583	0.170	0.165
GlobPlot	0.308	0.821	0.565	0.151	0.129

The performance of the SVM with the modeled data was compared with several known structural prediction tools. Performance evaluations were conducted on the results obtained with the aid of eleven disorder prediction tools referenced in the literature - PONDRs, DisEMBL, GlobPlot, DISOPRED2, FoldIndex, RONN, DisPRO, PreLink and DisPSSMP. In Table III, the order of success for each of the methods was given with regard to the measure, *probEx*.

The four methods - SVM, DisPSSMP, RONN, and FoldIndex - performed significantly better than the other methods. These are the only methods that have a probability excess value over 0.5. According to the results, it can be seen that the SVM_COD159, DisPSSMP and FoldIndex accomplished a more balanced prediction than the others. This indicates the tendency to predict order i.e. under-prediction of disorder. For example, PreLink yields the best score of specificity in predicting disorder with 99%, but it attains this at the expense of missing estimated data, with a significant number of disordered residues. On the other hand, the DisEMBL (coil) has a good performance of sensitivity but reveals a significant over-prediction.

Consequently, a comparison of the performances of the various methods demonstrates that the SVM_COD159 provides a considerable increase in classification performance with respect to the other common methods given in the literature. As is evident from these results, it can be concluded that the SVM with the modeled data is an effective way to achieve accurate predictions of disorder in proteins with the modeled data

In this research, feature selection was carried out as an alternative study. The purpose here was to find the more effective features in determining a disordered structure. In this way, a more productive and successful study could also be achieved by removing the features that do not carry adequate information. For this reason, the new and successful ERGS method of feature selection was chosen.

In order to select these features, the ERGS method was implemented on the validation set. At the end of the implementation, it was determined that prediction success was increased when the 37 attributes with average factor values larger than 0.4 were used. As a consequence, the data set was rearranged in such a way as to accommodate only these 37 features, and titled ERCOD159. The success values of the SVM with the new reduced data set are given in Table IV.

TABLE IV: THE PERFORMANCE COMPARISON OF SVM_COD159 VS. SVM_ERCOD159

	COD159	ERCOD159
Acc	0,819	0,831
probEx	0,638	0,662
Mcc	0,641	0,664
Spec	0,854	0,862
Sens	0,786	0,800
AUC	0,897	0,910

Upon examination of the values given in Table IV, it was seen that an increase in the success rate by approximately 1% was achieved when compared to the results of the execution on the data set containing all 120 characteristics. Even though this figure appears to be small, an increase in the *probEx* value demonstrates that more balanced results were obtained for each of the two classes. When comparing the results in Table IV, it can be seen that the value for AUC increased from 0.897 to 0.91. It is obvious that the elimination of the unnecessary features enhances the learning capability of the method. In addition to an increased rate of success and more balanced scores, factor reduction also results in lessening computational time and memory

consumption. The results are also confirmed by the ROC curve (Fig. 1). Fig. 1 illustrates that ERGS yields better classification performance than the cases in which the feature selection was not performed.

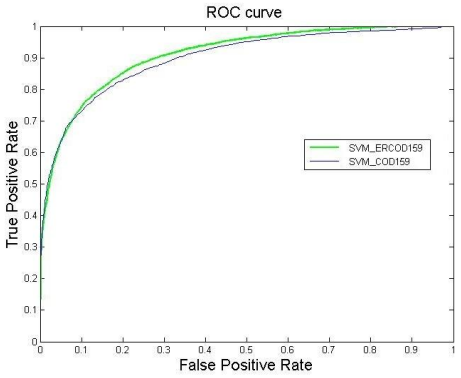

Fig. 1. The ROC curves for SVM_COD159 and SVM_ERCOD159.

According to the findings obtained in the study, it can be said that the 37 features selected carry sufficient information necessary to identify the disordered regions. The designated traits identified by the ERGS results are listed in the supplementary file (Appendix).

IV. Conclusion

Studies on structural genomics indicate that numerous protein segments fail to fold into a fixed 3D structure under physiological conditions. Contrary to the structure–function paradigm, these proteins comprise the disordered regions, yet exhibit function. Over the years, it has been confirmed that these proteins play key roles in vital functions and also in some diseases, such as Alzheimer's disease, Parkinson disease, and certain types of cancer. Thus, structure prediction and functional characterization of the proteins has become a challenging area of research.

Since amino acid sequence attributes have been widely used for determining protein structure, it has been theorized that the sequence would also determine disorder. Hence, as alternatives to experimental methods, several computational methods have been suggested for disorder prediction based on the amino acid sequence. Recent studies have focused on finding more useful features and developing more robust predictors to improve the quality of prediction.

A dataset was constructed organizing the amino acid sequence according to several physicochemical properties, evolutionary knowledge, and compositions, and a method was developed for predicting disorder regions within the modeled dataset in the study. Eleven specific tools, including PONDRs, DisEMBL, GlobPlot, DISOPRED2, FoldIndex, RONN, DisPRO, PreLink, and DisPSSMP, were used for comparison to test the classification performance of the method.

When compared with these most common disorder predictors, the proposed SVM_COD159 method achieves the best performance on prediction of disordered regions in proteins which were represented by specific features. This method outperforms the other ten methods significantly, without either under-predicting or over-predicting the disordered regions. Furthermore, the method accomplishes significantly more balanced classifications.

The results of this study demonstrated that SVM with optimal parameters can be used successfully in disorder prediction. The model provides a significant increase in success among the other eleven tools. As a result, the demanding problem of accurate prediction was significantly improved.

Furthermore, feature selection was carried out as an alternative research in the study. The most informative features for separating the disordered/ordered regions in proteins were determined by using the ERGS method. The obtained results corroborate much of the previous findings in literature. The new data set that was created with the selected features increased the prediction accuracy of the proposed method. This indicates that the selection procedure resulted in eliminating correlation and in discovering the necessary-sufficient properties that can be used to predict intrinsically disordered structure in proteins.

Appendix

The features found with ERGS was given by the following list.

X(2)	Amino acid composition
X(3)	Amino acid distribution
X(5)	Localized electrical effect
X(10)	Polarity [Grantham]
X(11)	Polarity [Zimmerman]
X(24)	Beta-sheet propensity derived from designed sequences
X(27)	Turn propensity
X(30)	Normalized frequency of alpha-helix
X(32)	Normalized frequency of beta-turn
X(33)	Hydrophilicity scale
X(34)	Average flexibility indices
X(35)	Normalized flexibility parameters (B-values), average
X(36)	Location parameters of the fit of the B factors
X(37)	Hydrophobicity index [Engelman et al.]
X(38)	Hydrophobicity index [Prabhakaran]
X(43)	Hydrophobicity
X(54)	First order statistics of E
X(72)	Disorder Promoting (R+K+E+P+S)
X(74)	Flexibility (G+T+R+S+N+Q+D+P+E+K)
X(76)	Beta Strand Related (E+B)
X(78)	Best Helix Formers (E+M+A+L)
X(81)	Best Sheet Breakers (S+G+K+P+D+E)
X(83)	External (R+N+D+Q+E+H+K)
X(85)	Acidic (D+E)
X(95)	Most Hydrophilic (D+E+N+Q+R+K)
X(98)	Helix Propensity (A+E+Q+H+K+M+L+R)
X(101)	Position-specific score for A
X(102)	Position-specific score for C
X(103)	Position-specific score for D
X(104)	Position-specific score for E
X(106)	Position-specific score for G
X(107)	Position-specific score for H
X(108)	Position-specific score for I
X(112)	Position-specific score for N
X(115)	Position-specific score for R
X(116)	Position-specific score for S
X(117)	Position-specific score for T

REFERENCES

[1] C. B. Anfinsen, "Principles that Govern the Folding of Protein Chains," *Science*, vol. 181, pp. 223–230, 1973.

[2] J. J. Ward, J. S. Sodhi, L. J. Mcguffin, B. F. Buxton, and D. T. Jones, "Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life," *J. Mol. Biol.*, vol. 337, pp. 635–645, 2004.

[3] A. K. Dunker, C. J. Oldfield, J. Meng, P. Romero, J. Y. Yang, J. W. Chen, V. Vacic, Z. Obradovic, and V. N Uversky, "The Unfoldomics Decade: An Update on Intrinsically Disordered Proteins," *BMC Genomics*, vol. 9, no. 2, 2008.

[4] R. M. Williams, Z. Obradovic, V. Mathura, W. Braun, E. C. Garner, J. Young, S. Takayama, C. J. Brown, and A. K. Dunker, "The Protein Non-Folding Problem: Amino Acid Determinants of Intrinsic Order and Disorder," in *Proc. Pacific Symposium on Biocomputing*, 2001, pp. 89-100.

[5] J. Bellay, S. Han, M. Michaut, T. Kim, M. Costanzo, B. J. Andrews, C. Boone, G. D. Bader, C. L. Myers, and P. M. Kim, "Bringing Order to Protein Disorder Through Comparative Genomics and Genetic Interactions," *Genome Biology*, vol. 12, R14, 2011.

[6] K. Coeytaux and A. Poupon, "Prediction of Unfolded Segments in a Protein Sequence Based on Amino Acid Composition," *Bioinformatics*, vol. 21, no. 9, pp. 1891–1900, 2005.

[7] R. Z. Yang, R. Thomson, P. Mcneil, and R. M. Esnouf, "RONN: The Bio-basis Function Neural Network Technique Applied to the Detection of Natively Disordered Regions in Proteins," *Bioinformatics*, vol. 21, pp. 3369-3376, 2005.

[8] J. Hecker, J. Y. Yang, and J. Cheng, "Protein Disorder Prediction at Multiple Levels of Sensitivity and Specificity," *BMC Genomics*, vol. 9, no. 1, S9, 2008.

[9] M. J. Mizianty, W. Stach, K. Chen, K. D. Kedarisetti, F. M. Disfani, and L. Kurgan, "Improved Sequence-Based Prediction of Disordered Regions with Multilayer Fusion of Multiple Information Sources," *Bioinformatics*, vol. 26, pp. 489-496, 2010.

[10] I. Nishikawa, Y. Nakajima, M. Ito, S. Fukuchi, K. Homma, K. Nishikawa, "Computational Prediction of O-linked Glycosylation Sites That Preferentially Map on Intrinsically Disordered Regions of Extracellular Proteins," *Int. J. Mol. Sci.*, vol. 11, pp. 4991-5008, 2010.

[11] A. K. Dunker *et al.*, "Intrinsically Disordered Protein," *Journal of Molecular Graphics and Modeling*, vol. 19, pp. 26-59, 2001.

[12] P. Radivojac, Z. Obradovic, D. K. Smith, G. Zhu, S. Vucetic, C. J. Brown J. D. Lawson, and A. K. Dunker, "Protein Flexibility and Intrinsic Disorder," *Protein Science*, vol. 13, no. 1, pp. 71-80, 2004.

[13] L. Wang and U. H. Sauer, "OnD-CRF: Predicting Order and Disorder in Proteins Using Conditional Random Fields," *Bioinformatics*, vol. 24, pp. 1401–1402, 2008.

[14] C. Su, C. Chen, and Y. Ou, "Protein Disorder Prediction by Condensed PSSM Considering Propensity for Order or Disorder," *BMC Bioinformatics*, vol. 7, pp. 319, 2006.

[15] K. Peng, S. Vucetic, P. Radivojac, C. J. Brown, A. K. Dunker, and Z. Obradovic, "Optimizing Long Intrinsic Disorder Predictors with Protein Evolutionary Information," *Journal of Bioinformatics and Computational Biology*, vol. 3, no. 1, pp. 35-60, 2005.

[16] R. Linding, L. J. Jensen, F. Diella, P. Bork, T. J. Gibson, and R. B. Russell, "Protein Disorder Prediction: Implications for Structural Proteomics," *Structure*, vol. 11, no. 11, pp. 1316-1317, 2003.

[17] R. Linding, R. B. Russell, V. Neduva, and T. J. Gibson, "Globplot: Exploring Protein Sequences for Globularity and Disorder," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3701–3708, 2003.

[18] B. Chandra and M. Gupta, "An Efficient Statistical Feature Selection Approach for Classification of Gene Expression Data," *J Biomed Inform*, vol. 44, no. 4, pp. 529-35, 2011.

[19] V. N. Uversky, J. R. Gillespie, and A. L. Fink, "Why Are "Natively Unfolded" Proteins Unstructured Under Physiologic Conditions?" *Proteins*, vol. 41, pp. 415-427, 2000.

[20] I. E. Kaya, T. Ibrikci, and O. K. Ersoy, "Prediction of Disorder with New Computational Tool: BVDEA," *Expert Systems with Applications*, vol. 38, no. 2, pp. 14451-14459, 2011.

[21] I. E. Kaya, "Computational Prediction of Disordered Regions in Proteins," Ph.D. dissertation, Dept. Electrical and Electronics Eng, Cukurova Univ., Adana, TURKEY, 2008.

[22] V. Vapnik, and A. Lerner, "Pattern Recognition Using Generalized Portrait Method," *Autom. Remote Control*, vol. 24, pp. 774–780, 1963.

[23] C. Cortes and V. N. Vapnik, "Support Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273-297, 1995.

[24] P. K. Dash, S. R. Samantary, and P. Ganapati, "Fault Classification and Section Identification of An Advanced Series-Compensated Transmission Line Using Support Vector Machine," *IEEE Trans. Power Delivery*, vol. 22, pp. 67–73, 2007.

[25] B. Demir and S. Ertürk, "Improving SVM Classification Accuracy Using A Hierarchical Approach For Hyperspectral Images," *IEEE International Conference on Image Processing*, Cairo, Egypt, Nov, 2009.

[26] Z. G. Jiang, H. G. Fu, and L. J. Li, "Support Vector Machine for Mechanical Faults Classification," *J. Zhejiang Univ. Sci.*, vol. 6, no. 5, pp. 433–439, 2005.

[27] C. R. Rao, *Linear Statistical Inference and Its Application*, John Wiley and Sons, 1965.

Irem Ersöz Kaya received the B.Sc. degree in physics engineering from Hacettepe University, Turkey in 1997. Her M.Sc. and Ph.D. degree in computer science from Cukurova University (Turkey) in 2003 and 2008, respectively. She is currently an assistant professor at the Department of Software Engineering in Tarsus Technology Faculty and the director of Tarsus Vocational School at Mersin University.

She teaches programming languages, artificial intelligence, artificial neural networks, software design, and her research interest is in the area of bioinformatics, machine learning and artificial neural networks.

Ayça Çakmak Pehlivanlı received BSc degree in statistics from Hacettepe University, Turkey (1997), MSc degree in computer science from Syracuse University, USA (2001) and Phd degree in computer science from Çukurova University, Turkey (2008).
She is currently an assistant professor at the Department of Statistics, Mimar Sinan Fine Arts University and her main research interest is the application of machine and statistical learning learning theory to medical data and bioinformatics.

Turgay Ibrikci received his BS degree in physics (Cukurova University, Adana, Turkey), MSc in computer science (Nova Southeastern University, Fort Laudardale, Florida, USA), and PhD in Electrical and Electronics Engineering Department (Cukurova University). Currently, he is an assistant professor at Electrical-Electronics Engineering Department, Cukurova University.

He had internatioanl experience as a visiting researcher at Computational Neuro Engineering Lab (CNEL), University of Florida (1999), at the Neurosignal Analysis Lab (NAL), University of Texas, Health Science Center (2001 and 2004) and at the Institute of Bioinformatics, University of Georgia (2011). His research interests include machine learning, bioinformatics, protein structures, and medical image processing.