Hindawi
Mathematical Problems in Engineering
Volume 2018, Article ID 8087391, 7 pages
https://doi.org/10.1155/2018/8087391

Research Article

A Low Computational Complexity Scheme for the Prediction of Intrinsically Disordered Protein Regions

Hao He and Jiaxiang Zhao

College of Electronic Information and Optical Engineering, Nankai University, Tianjin 300350, China

Correspondence should be addressed to Jiaxiang Zhao; zhaojx@nankai.edu.cn

Received 27 October 2017; Revised 23 January 2018; Accepted 6 March 2018; Published 5 April 2018

Academic Editor: Aimé Lay-Ekuakille

We employ the Rayleigh entropy maximization to develop a novel IDP scheme which requires computing only five features for each residue of a protein sequence, that is, the Shannon entropy, topological entropy, and the weighted average values of three propensities. Furthermore, our scheme is a linear classification method and hence requires computing simpler decision curves which are more robust as well as using fewer learning samples to compute. The simulation results of our scheme as well as some existing schemes demonstrate its effectiveness.

1. Introduction

Accurately identifying intrinsically disordered proteins (IDPs) which have at least a region lacking a unique 3D structure with a dynamic conformational ensemble [1, 2] is vital to obtain more effective drug designs, better protein expressions, and functional annotations. This is because it is confirmed that some of these intrinsically disordered proteins are involved in some of the most important regulatory functions in the cell [3], which have a great impact on diseases such as Alzheimer's disease, Parkinson's disease, and certain types of cancer [4]. It is essential to investigate the IDPs through the computation of the amino acid sequence of a protein [4]. This is because it is often difficult to purify and crystallize the disordered protein regions [5], which creates great problems for the disordered protein regions identification with the experimental approaches. Furthermore, experimental approaches for the disordered protein regions identification are usually both expensive and time-consuming [4].

Many IDP schemes have been proposed in the past decades, which can be roughly classified into two categories. (1) The first category is to exploit the amino acid propensity scales of the protein sequences for IDPs, such as FoldIndex [6], GlobPlot [7], IUPred [8], and FoldUnfold [9]. These schemes utilize the amino acid propensity scales to compute parameters such as the ratio of mean net charges with the mean hydropathy, the relative propensity of an amino acid residue, and the interresidue contacts for IDPs. These IDP schemes are simple but not accurate enough in general [10]. (2) The second category is to employ machine learning techniques for the IDPs. The examples of these include PONDR®s [11], RONN [12], DISOPRED2 [13], BVDEA [4], and DisPSSMP [14]. Many of these schemes are based on the artificial neural networks as well as support vector machine (SVM) which in general require computing a lot of features of a given protein sequence for IDPs. The computation of these features of a protein sequence could be expensive and time-consuming. More recently, MetaPrDOS [15] and Meta-Disorder predictor [16] which use several different predictors and their trade-off to yield an optimal decision for IDPs are also reported.

In this paper, we employ the Rayleigh entropy maximization to develop a novel IDP scheme which requires computing only five features for each residue of a protein sequence, that is, the Shannon entropy, topological entropy, and the weighted average values of three propensities. In contrast with most existing IDP schemes which need to compute no less than 30 features for each residue of a protein sequence, our scheme with a similar performance greatly reduces the computational complexity. Furthermore, our scheme based on the linear classification method has simpler decision

curves which are more robust and require fewer learning samples to compute. Our scheme is trained and tested by the dataset DIS803 with 10-fold cross-validation, firstly. The dataset DIS803 is comprised of 803 protein sequences with 1254 disordered regions and 1343 ordered regions, which include 92423 disordered and 315503 ordered residues. As a comparison, we run our scheme together with some existing schemes, such as PONDR [11], FoldIndex [6], DISOPRED2 [13], RONN [12], and DISPRO [17] on the datasets PU159 and R80 which are comprised of 239 protein sequences with 183 disordered regions and 231 ordered regions. They are comprised of 18111 disordered and 46477 ordered residues, respectively. The simulation results suggest that only our scheme, BVDEA [4], and DisPSSMP [14] have PE (probability excess) values exceeding 0.5 for both datasets PU159 and R80. Our scheme is at least as accurate as BVDEA [4] and DisPSSMP [14] and requires computing only 5 features for each residue of a protein sequence, while the other two need to compute 188 and 120 features for each residue, respectively. In addition, both BVDEA [4] and DisPSSMP [14] are based on nonlinear classification methods which require computing the complex decision curves that are less robust in general.

2. A Brief Review of Some Notations

In a protein sequence, the complexity denotes how a sequence can be rearranged in many different ways [18]. It has been demonstrated that the low complexity regions are more likely to be disordered than ordered [12]. Shannon entropy and topological entropy are two parameters to measure the complexity of a sequence. To begin with, let us first recall some notations.

Given a protein sequence $\{w(j),\ 1 \le j \le N\}$ of length N, the *Shannon entropy* is

$$H_S(w) = -\sum_{k=1}^{20} f_k \log_2 f_k, \tag{1}$$

where f_k for $1 \le k \le 20$ is defined as

$$f_k = \frac{\sum_{j=1}^{N} K(j)}{N}, \quad K(j) = \begin{cases} 1, & w(j) = \Phi(k) \\ 0, & w(j) \ne \Phi(k) \end{cases} \tag{2}$$

with $\Phi = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$ being an ordered set of 20 amino acid symbols.

The complexity function $p_w(n)$ representing the total number of different n-length subwords of w $(1 \le n \le N)$ is defined as [19]

$$p_w(n) = |\{u : |u| = n\}|, \tag{3}$$

where a subword u of a length n is one of any n consecutive symbols of w. $|u|$ denotes the length of u. For example, for a given sequence $w = MSTEAS$, the subwords of length 2 are

$$\{MS, ST, TE, EA, AS\}, \tag{4}$$

which yields

$$p_w(n) = 5. \tag{5}$$

Given a finite protein sequence w of length N, let n be the unique integer satisfying $20^n + n - 1 \le |w| < 20^{n+1} + (n+1) - 1$ and $w_1^{20^n+n-1}$ denote the first $20^n + n - 1$ consecutive symbols of w; that is, $w_1^{20^n+n-1} = w(1) \cdots w(20^n + n - 1)$.

The *topological entropy* of w is

$$H_{\text{top}}(w) = \frac{\log_{20} p_{w_1^{20^n+n-1}}(n)}{n}, \tag{6}$$

where $p_{w_1^{20^n+n-1}}(\cdot)$ is defined in (3). Thus, we have $H_{\text{top}}(w) = 1$ when the subwords of $w_1^{20^n+n-1}$ run over all the possible subwords of length n. On the other hand, $w_1^{20^n+n-1}$ is a repetition sequence comprising a single letter which suggests $H_{\text{top}}(w) = 0$. Similar to [19], we also compute the average of the topological entropy of w as

$$H_{\text{top}}(w) = \frac{1}{N - (20^n + n - 1) + 1} \cdot \sum_{l=1}^{N-(20^n+n-1)+1} \frac{\log_{20} p_{w_l^{20^n+n-1+l-1}}(n)}{n}. \tag{7}$$

The *Rayleigh entropy maximization* [20] of $\mathbf{X} = [\mathbf{x}_1\ \mathbf{x}_2\ \cdots\ \mathbf{x}_{N_s}]$, where N_s represents the total number of all the samples and $\mathbf{x}_j$ $(1 \le j \le N_s)$ represents the features of the jth sample, is to compute the projection direction $\mathbf{W}$ which optimizes the cost function

$$J(\mathbf{W}) = \frac{\mathbf{W}^T \mathbf{S}_B \mathbf{W}}{\mathbf{W}^T \mathbf{S}_W \mathbf{W}}. \tag{8}$$

$\mathbf{S}_W$ and $\mathbf{S}_B$ in (8) are, respectively, defined as

$$\mathbf{S}_W = \sum_{i=1}^{2} \sum_{j=1, \mathbf{x}_j \in \mathbf{X}_i}^{N_i} (\mathbf{x}_j - \mathbf{m}_i)(\mathbf{x}_j - \mathbf{m}_i)^T \tag{9}$$

$$\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T \tag{10}$$

$$\mathbf{m}_i = \frac{1}{N_i} \sum_{j=1, \mathbf{x}_j \in \mathbf{X}_i}^{N_i} \mathbf{x}_j, \tag{11}$$

where N_i is the number of samples in the ith class and $\mathbf{X}_i$ is the set of samples in the ith class.

Using the Lagrange method, the optimal $\mathbf{W}$ and the corresponding optimal projection $\mathbf{Y}$ of $\mathbf{X}$ on the direction of $\mathbf{W}$ are given as

$$\mathbf{W} = \mathbf{S}_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2) \tag{12}$$

$$\mathbf{Y} = \mathbf{W}^T \mathbf{X}. \tag{13}$$

3. The Computation of the Optimal Projection Direction

In this section, we compute the Shannon entropy and the topological entropy of the dataset DIS803 from DisProt [21] (http://www.disprot.org/). Then, choosing Remark 465,

Table 1: Bulky hydrophobic and aromatic amino acid.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
hyd_aro	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	1	1	1

Deleage/Roux, and Bfactor(2STD) propensities provided by the GlobPlot NAR paper [7] (http://globplot.embl.de/html/propensities.html), we compute the weighted average values of these propensities of the dataset DIS803. Finally, utilizing the computed Shannon entropy, topological entropy, and the weighted average values of three propensities of the dataset DIS803, we derive the optimum projection direction $\mathbf{W}$ defined in (8). The procedure proceeds as follows:

(1) Let w be a protein sequence. We choose a window of length N to extract N consecutive residues from w. Therefore, we assume the length of w to be N. Using (1), we can compute the Shannon entropy of w. To compute the topological entropy of w, we first map w to the propensities as follows. We map bulky hydrophobic (I, L, V) as well as aromatic (F, W, Y) amino acid residues defined in [10] to 1 and the rest of residues to 0. We use $\overline{w}$ to represent the mapped sequence of w. Table 1 lists all the amino acid residues and their corresponding mapping values.

Then, utilizing (7), we compute the average topological entropy of w as

$$H_{\text{top}}(w) = \frac{1}{N-(2^n+n-1)+1} \cdot \sum_{l=1}^{N-(2^n+n-1)+1} \frac{\log_2 p_{\overline{w}_l^{2^n+n-1+l-1}}(n)}{n}, \tag{14}$$

where the parameter n here satisfies $2^n+n-1 \le |\overline{w}| < 2^{n+1}+(n+1)-1$. $\overline{w}_l^{2^n+n-1+l-1}$ denotes the lth 2^n+n-1 consecutive symbols of $\overline{w}$; that is, $\overline{w}_l^{2^n+n-1+l-1} = \overline{w}(l)\cdots\overline{w}(2^n+n-1+l-1)$.

For example, Table 1 suggests that the protein sequence

$$w = MSTEASVSYAALILADAEQEITSEKLLAITKAAGA \tag{15}$$

is mapped to

$$\overline{w} = 0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,1\,1\,1\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,1\,1\,0\,1\,0\,0\,0\,0\,0\,0. \tag{16}$$

Therefore, we have $|\overline{w}| = 35$ which yields $n = 4$ satisfying $2^n+n-1 \le |\overline{w}| < 2^{n+1}+(n+1)-1$. Substituting $n = 4$ into 2^n+n-1, we obtain $\overline{w}_1^{2^n+n-1} = \overline{w}_1^{19}$. Thus, from (14), the topological entropy of w is 0.8508.

(2) For this protein sequence w of length N, we also compute the weighted average values of Remark 465, Deleage/Roux, and Bfactor(2STD) propensities defined in the GlobPlot NAR paper [7]:

$$M_p(w) = \frac{1}{N}\sum_{l=1}^{N} \widetilde{w}^p(l) \cdot \ln(l+1), \quad p = 1,2,3, \tag{17}$$

where $\widetilde{w}^p(l)$ with $1 \le l \le N$ represents the values of the pth propensity of w. We use the pth propensity of w with $p = 1,2,3$ to denote Remark 465, Deleage/Roux, and Bfactor(2STD) propensities, respectively. The weight $\ln(j+1)$ in (18) is identical to the sum function of the GlobPlot NAR paper [7].

For example, Remark 465 propensity of the sequence in (15) is

$$\widetilde{w}^1 = -0.1113\ 0.2627\ -0.1297\ 0.5214\ 0.1739\cdots. \tag{18}$$

From (18), it follows that $M_1(w)$ of Remark 465 propensity is 0.1551. Similarly, $M_2(w)$ and $M_3(w)$, respectively, corresponding to the Deleage/Roux and Bfactor (2STD) propensities are −0.4255 and −0.1368.

(3) For a general protein sequence w of length L, we use a sliding window of length N ($N < L$) to extract N consecutive residues $\mathbf{w}_j = w(j)\cdots w(j+N-1)$, $1 \le j \le L-N+1$. For this sliced $\mathbf{w}_j$, we compute the Shannon entropy $H_S(\mathbf{w}_j)$, the topological entropy $H_{\text{top}}(\mathbf{w}_j)$, and $M_p(\mathbf{w}_j)$ for $p = 1,2,3$ defined in (18). Define a 5×1 vector $\mathbf{v}_j$ to be

$$\mathbf{v}_j = \left[H_S(\mathbf{w}_j)\ H_{\text{top}}(\mathbf{w}_j)\ M_1(\mathbf{w}_j)\ M_2(\mathbf{w}_j)\ M_3(\mathbf{w}_j)\right]^{\mathrm{T}}. \tag{19}$$

Thus, we can compute the feature matrix of the protein sequence w of length L as

$$\mathbf{F} = \left[\mathbf{x}_1\ \mathbf{x}_2\ \cdots\ \mathbf{x}_L\right], \tag{20}$$

where vector $\mathbf{x}_l (1 \le l \le L)$ is

$$\mathbf{x}_l = \begin{cases} \dfrac{1}{l}\displaystyle\sum_{j=1}^{l} \mathbf{v}_j, & 1 \le l \le N \\ \dfrac{1}{N}\displaystyle\sum_{j=l-N+1}^{l} \mathbf{v}_j, & N < l \le L-N+1 \\ \dfrac{1}{L-l+1}\displaystyle\sum_{j=l-N+1}^{L-N+1} \mathbf{v}_j, & l > L-N+1 \end{cases} \tag{21}$$

with $\mathbf{v}_j$ for $1 \le j \le L-N+1$ being defined in (19).

For the protein sequence w of (15), we choose the size of window $N = 20$ and compute the 10th and 30th residues of w. $\mathbf{x}_{10}$ and $\mathbf{x}_{30}$ are

$$\mathbf{x}_{10} = \frac{1}{10}\sum_{j=1}^{10} \mathbf{v}_j, \qquad \mathbf{x}_{30} = \frac{1}{6}\sum_{j=11}^{16} \mathbf{v}_j, \tag{22}$$

where $\mathbf{v}_j$ is defined in (19).

(4) Utilizing 10-fold cross-validation [22], we randomly divide the dataset DIS803 into ten subsets of approximately equal size. The protocol uses nine subsets as the training dataset to build a model and the remaining 10th subset for testing. Using the training dataset of 10-fold cross-validation [22], we can compute the feature matrix

$$\mathbf{X} = \left[\mathbf{F}_1 \ \mathbf{F}_2 \ \cdots \ \mathbf{F}_{N_s}\right] \tag{23}$$

$$\mathbf{F}_i = \left[\mathbf{x}_{i_1} \ \mathbf{x}_{i_2} \ \cdots \ \mathbf{x}_{i_{L_i}}\right], \tag{24}$$

where N_s is the total number of the protein sequences of the training dataset. $\mathbf{F}_i$ defined in (20) with $1 \le i \le N_s$ is the feature matrix of the ith protein sequence whose length is L_i. Of all the residues of the training dataset obtained from DIS803 through 10-fold cross-validation described above, we divide it into two disjoint subsets: one comprised of all the disordered residues and the other of all the ordered residues of the training dataset. Let N_{dis} and N_{ord}, respectively, denote the number of all the disordered and all the ordered residues of the training dataset. $\mathbf{X}_{\text{dis}}$ and $\mathbf{X}_{\text{ord}}$, respectively, represent the feature matrices defined in (23) corresponding to all the disordered and all the ordered residues of the training dataset. From (11), it follows that

$$\mathbf{m}_{\text{dis}} = \frac{1}{N_{\text{dis}}}\sum_{j=1}^{N_{\text{dis}}} \mathbf{X}_{\text{dis}}^j, \qquad \mathbf{m}_{\text{ord}} = \frac{1}{N_{\text{ord}}}\sum_{j=1}^{N_{\text{ord}}} \mathbf{X}_{\text{ord}}^j, \tag{25}$$

where $\mathbf{X}_{\text{dis}}^j$ and $\mathbf{X}_{\text{ord}}^j$ represent the jth column vector in $\mathbf{X}_{\text{dis}}$ and $\mathbf{X}_{\text{ord}}$, respectively. Using $\mathbf{m}_{\text{dis}}$ and $\mathbf{m}_{\text{ord}}$, $\mathbf{S}_W$ in (9) can be calculated as

$$\mathbf{S}_W = \sum_{j=1}^{N_{\text{dis}}} \left(\mathbf{X}_{\text{dis}}^j - \mathbf{m}_{\text{dis}}\right)\left(\mathbf{X}_{\text{dis}}^j - \mathbf{m}_{\text{dis}}\right)^T + \sum_{j=1}^{N_{\text{ord}}} \left(\mathbf{X}_{\text{ord}}^j - \mathbf{m}_{\text{ord}}\right)\left(\mathbf{X}_{\text{ord}}^j - \mathbf{m}_{\text{ord}}\right)^T. \tag{26}$$

From (12), the projection direction is

$$\mathbf{W} = \mathbf{S}_W^{-1}\left(\mathbf{m}_{\text{dis}} - \mathbf{m}_{\text{ord}}\right). \tag{27}$$

The projection $\mathbf{Y}$ can be computed by (13). Finally, using linear searching in $\mathbf{Y}$, we can obtain the threshold of classification.

4. The Simulation Results

We employ the Rayleigh entropy maximization shown in the previous sections to develop an IDP scheme which requires computing only five features for each residue of a protein sequence, that is, the Shannon entropy, topological entropy, and the weighted average values of three propensities. In contrast, computing no less than 30 features is demanded by most existing schemes, such as PONDR [11], DISOPRED2 [13], RONN [12], DISPRO [17], BVDEA [4], and DisPSSMP [14], for the IDP identification. Furthermore, our scheme is based on the linear classification method which requires fewer learning samples to compute the simple decision curves that are more robust.

In order to train and test our scheme, the sequences in the dataset DIS803 are randomly split into ten subsets of approximately equal size to conduct a 10-fold cross-validation. The dataset DIS803 is comprised of 803 protein sequences. The results of our scheme with different window sizes are shown in Table 2. We use Sens., Spec., PE, and MCC to abbreviate sensitivity, specificity, probability excess, and Matthews' correlation coefficient, respectively. In addition, the values on probability excess and Matthews' correlation coefficient with different window sizes are shown in Figure 1. When the window size is larger than 35, the values tend to be smooth. Thus, we present our results with the window size of 35 in subsequent simulations.

As a comparison, we run our scheme together with some of the best known schemes, such as PONDR [11], FoldIndex [6], DISOPRED2 [13], RONN [12], and DISPRO [17], on the datasets PU159 and R80 which are comprised of 239 protein sequences with 183 disordered regions and 231 ordered regions. Dataset PU159 consists of P80 and U79 [23] where P80 and U79 with 80 completely ordered and 79 completely disordered proteins, respectively, are from PONDR® web site [23, 24]. Dataset R80 is from RONN [12] and contains 80 proteins with 183 disordered regions and 151 ordered regions.

Considering the classification method used, we use DISREM as the abbreviation of our scheme. The simulation

Table 2: Performance on dataset DIS803 with different window sizes.

Window sizes	11	15	19	23	27	31	35	39	43	47	51	55	59	63
Sens.	0.6892	0.7188	0.7309	0.7622	0.7691	0.7765	0.7747	0.7566	0.7754	0.7490	0.7716	0.7802	0.8052	0.7961
Spec.	0.7371	0.7546	0.7651	0.7543	0.7560	0.7581	0.7647	0.7835	0.7658	0.7826	0.7669	0.7609	0.7426	0.7451
PE	0.4263	0.4734	0.4960	0.5166	0.5251	0.5346	0.5394	0.5400	0.5411	0.5315	0.5386	0.5412	0.5478	0.5412
MCC	0.3679	0.4105	0.4317	0.4451	0.4524	0.4606	0.4663	0.4730	0.4681	0.4658	0.4664	0.4667	0.4671	0.4625

Table 3: Performance comparison with existing schemes on dataset PU159.

Schemes	Sens.	Spec.	PE	MCC
DISREM	0.821	0.757	0.577	0.576
DisPSSMP	0.825	0.765	0.590	0.589
BVDEA	0.796	0.785	0.581	0.586
RONN	0.675	0.888	0.563	0.580
FoldIndex	0.722	0.815	0.536	0.540
DISOPRED2	0.469	0.981	0.449	0.543
PONDR	0.632	0.782	0.414	0.420
DISPRO	0.383	0.982	0.365	0.467

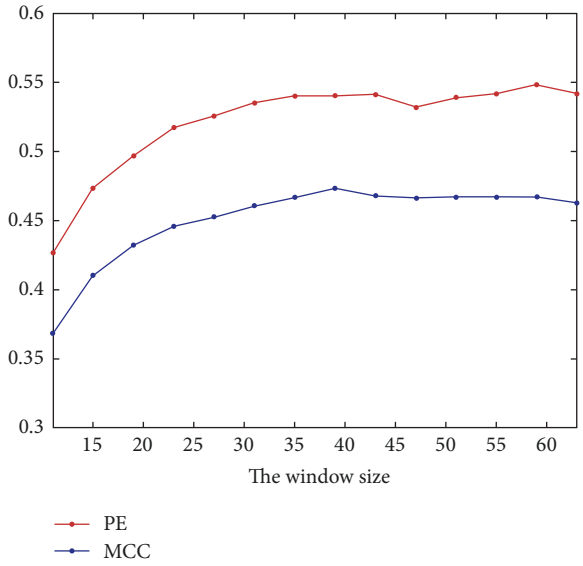

Figure 1: The performance with different window sizes on PE and MCC.

results listed in Tables 3 and 4 show that the IDP identification accuracy of our scheme is approximately accurate as those of BVDEA [4] and DisPSSMP [14] whose performance exceeds the rest of the schemes mentioned above on the datasets PU159 and R80. From Tables 3 and 4, it is suggested that only our scheme, BVDEA [4], and DisPSSMP [14] have PE (probability excess) values exceeding 0.5 for both datasets PU159 and R80. To achieve these PE values, our scheme requires computing only 5 features of each residue, while computing 188 and 120 features for each residue of a protein sequence is demanded by DisPSSMP [14] and BVDEA [4], respectively.

Furthermore, unlike nonlinear classification of DisPSSMP [14] and BVDEA [4] which require computing the complex decision curves, our scheme is based on the Rayleigh entropy maximization which is the linear classification method. Therefore, our scheme has simpler decision curves to compute and hence decision curves are more robust and require fewer learning samples than those of DisPSSMP [14] and BVDEA [4].

Table 4: Performance comparison with existing schemes on dataset R80.

Schemes	Sens.	Spec.	PE	MCC
DISREM	0.736	0.888	0.625	0.507
DisPSSMP	0.767	0.848	0.615	0.463
BVDEA	0.817	0.728	0.545	0.451
RONN	0.603	0.878	0.481	0.395
DISPRO	0.418	0.993	0.411	0.578
DISOPRED2	0.405	0.972	0.377	0.470
PONDR	0.557	0.816	0.373	0.278
FoldIndex	0.488	0.811	0.299	0.224

5. Conclusions

In this paper, we compute the Shannon entropy, the topological entropy, and the weighted average values of three propensities to develop a criterion based on Rayleigh entropy maximization for predicting the intrinsically disordered regions of a protein. Compared with several existing schemes, the identification accuracy of our scheme is at least as accurate as those schemes whose performance exceeds the rest of the compared schemes. Particularly, in contrast with those schemes that require computing no less than 30 features, our scheme only relies on computing five features.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] J. Yan, M. J. Mizianty, P. L. Filipow, V. N. Uversky, and L. Kurgan, "RAPID: Fast and accurate sequence-based prediction of intrinsic disorder content on proteomic scale," *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, vol. 1834, no. 8, pp. 1671–1680, 2013.

[2] V. N. Uversky, "The mysterious unfoldome: Structureless, underappreciated, yet vital part of any given proteome," *Journal of Biomedicine and Biotechnology*, vol. 2010, Article ID 568068, 14 pages, 2010.

[3] P. E. Wright and H. J. Dyson, "Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm," *Journal of Molecular Biology*, vol. 293, no. 2, pp. 321–331, 1999.

[4] I. Ersöz Kaya, T. Ibrikci, and O. K. Ersoy, "Prediction of disorder with new computational tool: BVDEA," *Expert Systems with Applications*, vol. 38, no. 12, pp. 14451–14459, 2011.

[5] C. J. Oldfield, E. L. Ulrich, Y. Cheng, A. K. Dunker, and J. L. Markley, "Addressing the intrinsic disorder bottleneck in structural proteomics," *Proteins: Structure, Function, and Genetics*, vol. 59, no. 3, pp. 444–453, 2005.

[6] J. Prilusky, C. E. Felder, and T. Zeev-Ben-Mordehai, "FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded," *Bioinformatics*, vol. 21, no. 16, pp. 3435–3438, 2005.

[7] 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.

[8] Z. Dosztányi, V. Csizmok, P. Tompa, and I. Simon, "IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content," *Bioinformatics*, vol. 21, no. 16, pp. 3433-3434, 2005.

[9] O. V. Galzitskaya, S. O. Garbuzynskiy, and M. Y. Lobanov, "FoldUnfold: Web server for the prediction of disordered regions in protein chain," *Bioinformatics*, vol. 22, no. 23, pp. 2948-2949, 2006.

[10] F. Orosz and J. Ovádi, "Proteins without 3D structure: Definition, detection and beyond," *Bioinformatics*, vol. 27, no. 11, pp. 1449–1454, 2011.

[11] 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.

[12] Z. R. 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, no. 16, pp. 3369–3376, 2005.

[13] 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," *Journal of Molecular Biology*, vol. 337, no. 3, pp. 635–645, 2004.

[14] C.-T. Su, C.-Y. Chen, and Y.-Y. Ou, "Protein disorder prediction by condensed PSSM considering propensity for order or disorder," *BMC Bioinformatics*, vol. 7, article 319, 2006.

[15] T. Ishida and K. Kinoshita, "Prediction of disordered regions in proteins based on the meta approach," *Bioinformatics*, vol. 24, no. 11, pp. 1344–1348, 2008.

[16] A. Schlessinger, M. Punta, G. Yachdav, L. Kajan, and B. Rost, "Improved disorder prediction by combination of orthogonal approaches," *PLoS ONE*, vol. 4, no. 2, Article ID e4433, 2009.

[17] J. Cheng, M. J. Sweredoski, and P. Baldi, "Accurate prediction of protein disordered regions by mining protein structure data," *Data Mining and Knowledge Discovery*, vol. 11, no. 3, pp. 213–222, 2005.

[18] E. A. Weathers, M. E. Paulaitis, T. B. Woolf, and J. H. Hoh, "Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein," *FEBS Letters*, vol. 576, no. 3, pp. 348–352, 2004.

[19] D. Koslicki, "Topological entropy of DNA sequences," *Bioinformatics*, vol. 27, no. 8, pp. 1061–1067, 2011.

[20] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.-R. Muller, "Fisher discriminant analysis with kernels," in *Proceedings of the 9th IEEE Signal Processing Society Workshop on Neural Networks for Signal Processing (NNSP '99)*, pp. 41–48, Madison, Wis, USA, August 1999.

[21] M. Sickmeier, J. A. Hamilton, T. LeGall et al., "DisProt: The database of disordered proteins," *Nucleic Acids Research*, vol. 35, no. 1, pp. D786–D793, 2007.

[22] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, vol. 2, pp. 1137–1143, San Mateo, Calif, 1995.

[23] V. N. Uversky, J. R. Gillespie, and A. L. Fink, "Why are "natively unfolded" proteins unstructured under physiologic conditions?" *Proteins: Structure, Function, and Genetics*, vol. 41, no. 3, pp. 415–427, 2000.

[24] "PONDR® Predictors of Natural Disordered Regions," http://www.pondr.com/.