MATCH
Communications in Mathematical
and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **75** (2016) 417-430

ISSN 0340 - 6253

Classification of Ordered/Disordered Regions of Intrinsically Disordered Proteins Based on Comprehensive Sequence Analysis and Chou's Pseudo Amino Acid Composition Method

Jia-Feng Yu[1,2,*], En-Si Wu[1,3], Chun-Ling Wang[4], Hong-Mei Wang[4], Ji-Hua Wang[1,4,*]

[1] *Shandong Provincial Key Laboratory of Biophysics, Institute of Biophysics, Dezhou University, Dezhou 253023, China*

[2] *State Key Laboratory of Bioelectronics, Southeast University, Nanjing 210096, China*

[3] *College of Life Science, Shandong Normal University, Jinan 250014, China*

[4] *College of Physics and Electronic Information, Dezhou University, Dezhou 253023, China*

(Received May 26, 2015)

Abstract

Intrinsically disordered proteins (IDPs) are a kind of protein that plays important roles but lack well defined three-dimensional structure. In this paper, comprehensive sequence analysis is performed based on a larger dataset derived from the latest version of Disprot database. The results indicate that there are significant differences between the disordered regions and the ordered regions of IDPs. Further analysis shows that the disordered regions prefer hydrophilic amino acids such as D, E, K, Q, S, T and the ordered regions prefer hydrophobic amino acids such as F, I, L, M, V, W, Y. Then, a classification algorithm for disordered regions and ordered regions is proposed by incorporating the information of sequence composition, sequence order and long range correlation based on Chou's pseudo amino acid composition (PseAAC) method. The results show that the efficiency of the hybrid features can be improved in accordance with the diverse evaluation indices of *ACC*, *MCC* and AUC in comparison to the traditional components composition based numerical features.

[*] Corresponding authors. E-mail: jfyu1979@126.com, jhw25336@126.com

1 Introduction

Intrinsically disordered proteins (IDPs) lack well defined three dimensional structure under physiological conditions, but carries important biological functions [1-2]. The 15 years studies show that IDPs are not rare exceptions, but a new rule [3]. It is estimated that about 10 ~ 35% of prokaryotic and about 15 ~ 45% of eukaryotic proteins contain significant disorder at least 30 residues in length [4]. IDPs play crucial roles in regulation, recognition, signaling, and control of protein-protein interaction networks and they are usually to carry out the functions through binding with other partners [5]. The highly flexibility and random coil-like conformation of IDPs make them a formidable challenge to identify the intrinsically disordered regions (IDRs) and to determine their dynamics dynamic ensembles so called "protein clouds" by experimental methods [6]. Therefore, computational tools for IDPs prediction and analysis have become the main means for IDPs studies. Some IDPs predictors were developed mainly based on simple statistics of amino acid propensity or the physical/chemical properties of amino acids [7-9]. These predictors relied on the sequence features between the ordered regions and disordered regions of IDPs. However, earlier statistics of amino acid propensity only based on the smaller dataset using the limited bioinformatics methods [10], which cannot provide enough input information for IDRs prediction. In the past several years, the number of experimentally verified IDPs has been improved, then it is interesting to mine further the intrinsic features of ordered/disordered regions of IDPs based on the bigger database, which may provide more solid basis for justified protein disorder prediction. On the other hand, recent works have demonstrated that the arranging order of the amino acids and the long range correlation of amino acids play important roles in many protein analysis related problems [11, 12]. Therefore, the influence of long range correlation of amino acids on disordered/ordered regions classifications is studied by incorporating sequence complexity and the PseAAC method in this work. Firstly, a larger dataset with experimentally verified IDPs is constructed, based on which we try to reveal the novel characters of IDRs for the creation of corresponding computational tools. Then comprehensive sequence analysis by incorporating sequence complexity and the PseAAC method is first used for finding novel sequence features for IDRs. The results indicate that the intrinsically ordered region exhibit explicitly different features from the intrinsically disordered regions. Finally, an efficient algorithm is proposed for classifying the ordered regions and the disordered regions of IDPs, which will provide new input information for developing of later IDPs predictors.

2 Materials and methods

2.1 Dataset

The IDPs sequences are downloaded from Disprot 6.01 [13]. A total of 683 IDPs sequences are released in this version. The CD-Hit program [14] is used to exclude the redundant sequences with the threshold of 30%. Those sequences that contain special characters such as B, X and Z are also excluded. Then, a dataset composed of 548 IDPs is obtained. Among the 548 sequences, there are 911 disordered regions and 978 ordered regions, respectively. Traditionally, the regions with the sequence length > 30 amino acids are regarded as IDPs. Therefore, 387 disordered regions and 749 ordered regions are finally obtained.

2.2 Methods for sequence analysis

2.2.1 Sequence components composition

The compositions of the 20 kinds of amino acids and the 400 kinds of dipeptide are used to depict sequence features of IDPs.

2.2.2 Sequence complexity

The sequence complexity is defined as [15]

$$K = -\sum_{i=1}^{N} f_i \log_2 f_i \qquad (1)$$

where f_i represent the usage frequency of the ith kind of amino acids, N = 1, 2, 3, …, 20. Obviously, when K reaches its maximum 4.32, this denotes that the 20 kinds of amino acids are averagely used.

2.2.3 Chou's pseudo amino acid composition

The PseAAC proposed by Chou is a versatile index for displaying the intrinsic information of amino acids order and long range correlations in protein sequences, which has been widely applied to protein classification algorithms [16-20]. However, there is no any use for IDPs in our knowledge. For a protein sequence $R_1R_2R_3\ldots R_L$ with the length of L, the PseAAC is denoted

$$x_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + \omega \sum_{j=1}^{\lambda} \theta_j}, (1 \le u \le 20) \\ \dfrac{\omega\theta_{u-20}}{\sum_{i=1}^{20} f_i + \omega \sum_{j=1}^{\lambda} \theta_j}, (20+1 \le u \le 20+\lambda) \end{cases} \quad (2)$$

where θ_j is called the jth-tier correlation factor that reflects the sequence order correlation, which is written as

$$\begin{cases} \theta_1 = \dfrac{1}{L-1}\sum_{i=1}^{L-1}\Theta(\mathrm{R}_i, \mathrm{R}_{i+1}) \\ \theta_2 = \dfrac{1}{L-2}\sum_{i=1}^{L-2}\Theta(\mathrm{R}_i, \mathrm{R}_{i+2}) \\ \theta_3 = \dfrac{1}{L-3}\sum_{i=1}^{L-3}\Theta(\mathrm{R}_i, \mathrm{R}_{i+3}) \\ \quad\cdots \\ \theta_\lambda = \dfrac{1}{L-\lambda}\sum_{i=1}^{L-\lambda}\Theta(\mathrm{R}_i, \mathrm{R}_{i+\lambda}) \end{cases} \quad (3)$$

The correlation function Θ is given by

$$\Theta(\mathrm{R}_\mathrm{i}, \mathrm{R}_\mathrm{j}) = \frac{1}{3}\left\{ \left[H_1(\mathrm{R}_j) - H_1(\mathrm{R}_i)\right]^2 + \left[H_2(\mathrm{R}_j) - H_2(\mathrm{R}_i)\right]^2 + \left[H_3(\mathrm{R}_j) - H_3(\mathrm{R}_i)\right]^2 \right\} \quad (4)$$

Where H(R) represents the physiochemical parameters of residue R, which should be subjected standard conversion as follows

$$H(i) = \frac{H^0(i) - \sum_{i=1}^{20} \dfrac{H^0(i)}{20}}{\sqrt{\dfrac{\sum_{i=1}^{20}\left[H^0(i) - \sum_{i=1}^{20}\dfrac{H^0(i)}{20}\right]^2}{20}}} \quad (5)$$

where H^0 is the original values of each physiochemical parameter. In this way, each protein sequence transformed into a (20+λ)-D vector, the first 20 components of which reflect the information of the amino composition, and the components from 20+1 to 20+λ reflect the information of sequence order [21].

2.3 Support vector machine

Support vector machine (SVM) has been widely used in prediction related problems of bioinformatics. In this work, the libSVM 3.17 [22] package with RBF kernel function [23] is adopted to accomplish the classification algorithm for ordered and disordered regions.

2.4 Evaluation indices

To evaluate the efficiency of the proposed classification algorithm in this paper, the sensitivity (S_n), specificity (S_p), accuracy (*ACC*), the Matthew's correlation coefficient (*MCC*) are employed respectively. The definition of each evaluation index is described as

$$S_n = \frac{TP}{TP+FN} \tag{6}$$

$$S_p = \frac{TN}{TN+FP} \tag{7}$$

$$\mathrm{ACC} = \frac{TP+TN}{TP+TN+FP+FN} \tag{8}$$

$$MCC = \frac{TP \times TN - FN \times FP}{\sqrt{(TP+FN)\times(TP+FP)\times(TN+FN)\times(TN+FP)}} \tag{9}$$

Where, *TP* and *TN* denote the positive and negative samples that have been correctly predicted respectively, *FP* and *FN* denote the positive and negative samples that have been falsely predicted respectively.

Considering the distribution imbalance of the samples in the training set, the receiver operating characteristic (ROC) curve is also adopted to evaluate the classification results besides the indices mentioned above.

3 Results and discussion

3.1 Sequence complexity analysis

Sequence complexity reflects the degree of each kind of amino acid usage bias in protein sequences. In figure 1, the *K* values of the 387 disordered regions and the 749 ordered regions are calculated respectively. From this figure, it is found that the disordered regions exhibit different patterns with that of the ordered regions. Close observation shows that the K values of the disordered regions are universally lower than that of the ordered regions. As can be seen, most sequences in the ordered regions have the *K* values larger than 4.0, while the *K* values of most disordered regions are less than 4.0. Further analysis shows that the mean value of *K* in disordered regions and ordered regions are 3.70 and 3.94 respectively. In addition, most disordered regions distribute from 3.0 to 4.3 of complexity, more than 90% disordered regions range from 3.1 to 4.2, more than 50% disordered regions is less than 3.9, and some is less 3.0. In contrast, most disordered regions distribute from 3.4 to 4.3 of complexity, more than 90% ordered regions range from 3.5 to 4.3 of complexity, only less than 30% ordered regions is less than 3.9 of complexity, and almost there is no ordered regions when the complexity is less than 3.4. Therefore, figure 1 shows that disordered regions prefer low sequence complexity.

On the other hand, it is noted that the lengths range from 31 to 2369 in ordered regions and from 31 to 1861 in disordered regions. Then it is interesting to investigate whether the differences of sequence complexity of each kind of regions in figure 1 are caused by the diverse lengths. In figure 2, the *K* value as well as the length of each sequence is calculated in ordered and disordered regions respectively. Seen from figure 2, *K* values fluctuate largely when the length is less than 750 in ordered and disordered regions, and the *K* values in disordered regions fluctuate much bigger than in ordered ones, then the *K* values tend to flat and independent of the sequence lengths in the two regions. Figure 2 also shows that the *K* values in disordered regions are always lower than that in the ordered regions, which is consistent with the results in figure 1.

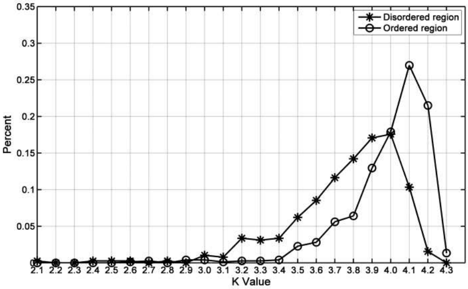

Figure 1. Sequence complexity analysis in ordered/disordered regions of IDPs

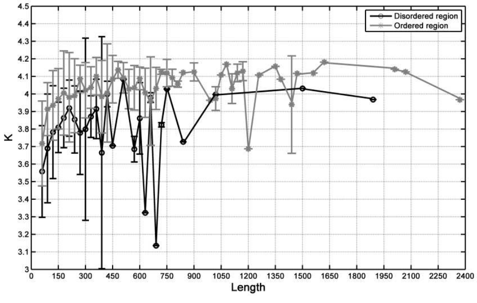

Figure 2. Sequence complexity vs. sequence lengths of ordered/disordered regions

As have been elaborated in equation 1, the *K* value reaches its maximum of 4.32 when the 20 kinds of amino acids are averagely used in protein sequence. On the contrary, the *K* value decreases till zero with the increasing of the amino acids usage bias. Therefore, the results both in figures 1 and 2 indicate that amino acids usage in disordered regions is different from ordered regions.

3.2 Amino acids usage bias analysis in IDPs

Amino acids composition is one of the most adopted numerical features for protein sequence. In the early work by Weathers [24], the usage frequencies of the 20 kinds of amino acids have been deemed as one of the most efficient numerical features for IDPs classifications. In figure 3, the usage frequency of each kind of amino acid is provided. As can be seen from this figure, some amino acids in the disordered regions such as A, D, E, G, K, L, P and S have the frequencies that are greater than 6%, respectively, while the frequencies of C, H, M and W are less than 2%. In the ordered regions, the frequencies of A, E, G, K, L, S and V are greater than 6%, while that of C and W are less than 2%.

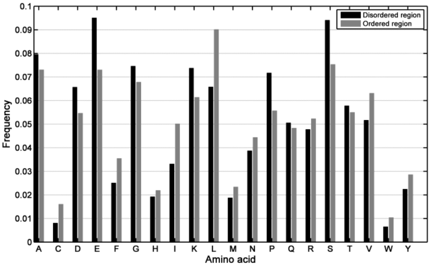

Figure 3. Composition of the 20 kinds of amino acids in ordered/disordered regions

We have shown that the sequence complexities exhibit different tendencies between the disordered regions and ordered regions in figure 1. Observing the curve of the disordered region, it is found that the sequence complexity can be divided into four intervals, i.e. $0 < K \leq 3.1$, $3.1 < K \leq 3.4$, $3.4 < K \leq 4.0$, $4.0 < K \leq 4.3$, which contain 3.10%, 9.82%, 75.19% and 11.89% of the 387 disordered regions respectively. Similarly, the *K* values of the ordered regions can be also divided into four intervals, i.e. $0 < K \leq 3.4$, $3.4 < K \leq 3.8$, $3.8 < K \leq 4.1$, $4.1 < K \leq 4.3$, which contain 2.27%, 17.90%, 57.81% and 22.83% of the 749 ordered regions respectively. Then, the amino acids compositions in different intervals are analyzed both in disordered region and ordered region. The results in table 1 show that in disordered regions, the percentages of A, D, E, G, K, L, P, S are bigger than 6% in most K value intervals, while the percentages of C, H, M W are less than 2%; for interval of $0 < K \leq 3.1$, the amino acids usage bias is extremely displayed that the percentages of almost 10 kinds of amino acids C, E, F, H, I, K, M, N, V, W are less than 2%. In this way, the amino acids compositions in ordered regions can be also analyzed according to table 1. Generally, in ordered regions, the percentages of A, E, G, K, L, S, V are bigger than 6%, while the percentages of C, W are less than 2% in most intervals. Then, table 1 shows that the amino acids compositions usages are consistent with the results of figure 3 in spite of the vibration of the K values.

The correlations of the sequences lengths and the complexity are analyzed in figure 2, according to which the sequences lengths are divided into five intervals both for the disordered regions and the ordered regions, i.e. $30 < L \leq 150$, $150 < L \leq 300$, $300 < L \leq 450$, $450 < L \leq 600$, $600 < L \leq 2400$. Similarly with table 1, the amino acids compositions in each sequence length interval are calculated in table 2. As can be seen from table 2, the amino acids compositions are very consistent with the results of figure 3. Therefore, both tables 1 and 2 indicate that the results of figures 1, 2 and 3 can exhibit the intrinsic features of amino acids compositions.

Table 1. Amino acids composition analysis in different complexity intervals

Regions	Intervals	Amino acids with frequency > 6%	Amino acids with frequency< 2%
Disordered	$0 < K \leq 4.3$ (100%)	A, D, E, G, K, L, P, S	C, H, M, W
	$0 < K \leq 3.1$ (3.10%)	G, P, Q, S, T, Y	C, E, F, H, I, K, M, N, V, W
	$3.1 < K \leq 3.4$ (9.82%)	A, D, E, G, K, P, S	C, F, H, I, M, W, Y
	$3.4 < K \leq 4.0$ (75.19%)	A, D, E, G, K, L, P, S	C, H, M, W, Y
	$4.0 < K \leq 4.3$ (11.89%)	A, D, E, K, L, S	C, W
Ordered	$0 < K \leq 4.3$ (100%)	A, E, G, K, L, S, V	C, W
	$0 < K \leq 3.4$ (2.27%)	E, G, L, P, Q, S	C, F, H, I, M, W
	$3.4 < K \leq 3.8$ (17.09%)	A, E, G, K, L, P, Q, S	C, H, W, Y
	$3.8 < K \leq 4.1$ (57.81%)	A, E, G, K, L, S, V	C, W
	$4.1 < K \leq 4.3$ (22.83%)	A, E, G, K, L, S, V	W

Table 2. Amino acids composition analysis in different sequence length intervals

Regions	Intervals	Amino acids with frequency> 6%	Amino acids with frequency < 2%
Disordered	$30 < L \leq 2400$ (100%)	A, D, E, G, K, L, P, S	C, H, M, W
	$30< L \leq 150$ (72.87%)	A, D, E, G, K, L, P, S	C, H, W
	$150<L\leq 300$ (17.83%)	A, D, E, G, K, L, P, Q, S	C, W
	$300< L \leq 450$ (4.91%)	A, D, E, G, K, L, P, S, T	C, H, M, W
	$450< L \leq 600$ (1.55%)	A, E, G, P, S, T, V	C, H, M, W, Y
	$600<L\leq 2400$ (2.84%)	A, D, E, G, K, P, S	C, H, M, W, Y

Ordered	$30 < L \leq 2369$ (100%)	A, E, G, K, L, S, V	C, W
	$30 < L \leq 150$ (55.54%)	A, E, G, K, L, S, V	C, W
	$150 < L \leq 300$ (20.29%)	A, E, G, L, S, V	C, W
	$300 < L \leq 450$ (11.35%)	A, E, G, K, L, S, V	C, W
	$450 < L \leq 600$ (4.94%)	A, D, E, G, K, L, S, V	C, W
	$600 < L \leq 2400$ (7.88%)	A, E, G, K, L, S, V	C, W

In general, usage frequency is 1/20 = 5% when the amino acid is randomly used. To given more explicit information of amino acids usages bias, we propose a simple index called amino acids preference (*AAP*) to display the differences between the disordered regions and ordered regions, which is given as

$$AAP^{i}{}_{R} = \frac{P^{i}{}_{R}}{P^{i}} - 1. \qquad (10)$$

Where, *i* represents the 20 kinds of amino acids, the right subscript *R* denotes disordered regions or ordered regions, P^i is the frequency of the *i*th kind of amino acid in IDPs (both ordered and disordered regions). Then, *AAP* = 0 indicates that corresponding amino acid is used randomly, and *AAP* > 0 indicates that corresponding amino acid is preferred. In figure 4, we provide the amino acids usage bias analysis results based on *AAP*. Obviously, the amino acids of A, D, E, G, K, P, Q, S and T are more preferred in disordered regions, while C, F, H, I, L, M, N, R, V, W and Y are more preferred in ordered regions. Further analysis shows that most of the amino acids preferred in disordered regions are hydrophilic (D, E, K, Q, S, T) and most of the amino acids preferred in ordered regions are hydrophobic (F, I, L, M, V, W, Y), which may be related to the fact that some IDPs gain its stable 3D structure by exposed its amino acids to bind other partners. Therefore figures 3 and 4 imply that the amino acids usage bias is different between the disordered regions and the ordered regions.

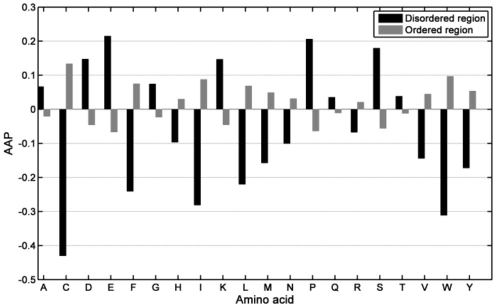

Figure 4. Amino acids usage analysis in ordered/disordered regions

3.3 Classification of disordered/ordered regions based on hybrid sequence features

Above analysis has demonstrated the potentially intrinsic differences between the disordered regions and ordered regions. In this section, the *K* value, usage frequency of the 20 kinds of amino acids (AA), usage frequency of the 400 kinds of dipetides (DAA) and PseAAC are adopted as sequence features in the classification algorithm respectively. It is noted that the hydrophobicity, mass and pK2 of each amino acid are used for PseAAC calculation. The 5-fold cross validation is performed by classifying the training set into five groups randomly, then one is used as testing set and the other four groups are used as training set. In table 3, the classification efficiency of each numerical feature is presented. It is found that the PseAAC achieve the highest classification efficiency compare with AA, DAA and *K*. The *ACC*, *MCC* and AUC for solely using PseAAC are 79.22%, 0.5211 and 0.8467, respectively. In comparison, the same indices for sole AA, DAA and *K* value are (78.34%, 0.4987, 0.8309), (76.76%, 0.4547, 0.8202) and (70.40%, 0.2759, 0.7669), respectively. From the comparison of the ROC curves in figure 5, the same results can be inferred. On the other hand, combination of *K*+AA+DAA+PseAAC achieve the highest *ACC* (79.40%) and *MCC* (0.5262). In summary, the classification efficiency is significantly improved by integrating the PseAAC, which shows that sequence order and long range correlation are important factors for describing the disordered/ordered regions.

Table 3. Classification of disordered/ordered regions based hybrid features

Numerical features	*ACC*(%)	S_n(%)	S_p(%)	*MCC*	AUC
PseAAC	79.22	89.31	59.70	0.5211	0.8467
AA	78.34	89.70	56.33	0.4987	0.8309
DAA	76.76	93.04	45.24	0.4547	0.8202
K	70.40	91.29	29.97	0.2759	0.7669
K+PseAAC	79.05	88.77	60.23	0.5180	0.8449
AA+PseAAC	78.52	88.63	58.95	0.5057	0.8444
DAA+PseAAC	79.05	89.04	59.72	0.5175	0.8443
K+AA+PseAAC	79.13	88.77	60.48	0.5201	0.8465
K+DAA+PseAAC	79.22	89.04	60.22	0.5217	0.8462
AA+DAA+PseAAC	79.05	89.04	59.72	0.5175	0.8444
K+AA+DAA+PseAAC	79.40	89.04	60.74	0.5262	0.8476
AA+DAA	78.69	90.90	55.06	0.5057	0.8342
K+AA	78.96	88.77	59.96	0.5166	0.8398
K+DAA	76.31	88.77	52.20	0.4481	0.8283
K+AA+DAA	79.23	89.31	59.73	0.5212	0.8360

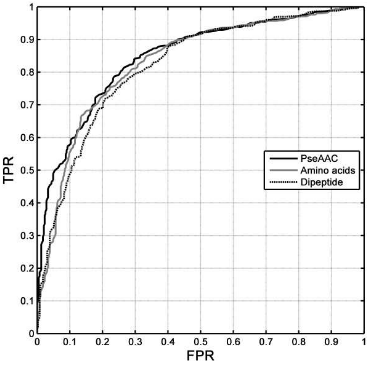

Figure 5. Comparison of the ROC curves of different sequence features

4 Conclusions

The discovery of IDPs challenges the traditional ‘sequence-structure-function’ paradigm [25]. Because lack of stable 3D structure, sequence based prediction has been one of the prerequisites for further understanding the biological significances of IDPs. Up to now, a great number of IDPs predictors have been put forward since the first one proposed in 1997 [26]. However the biological mechanisms of IDPs are not very clear, the sequence characteristics employed in these predictors mainly focus on the sequence components compositions and physiochemical properties of amino acids, it is necessary to provide more comprehensive sequence features for accurate prediction of IDPs. How to propose effective computational methods for demonstrating the features for protein sequences has been an important topic all the time [27, 28]. In this paper, we study the intrinsic sequence features between the disordered regions and ordered regions based on a larger dataset of IDPs. Then, a novel classification algorithm for disordered regions and ordered regions is proposed by incorporating the PseAAC method for the first time. The results show that the sequence order and long range correlation provides complementary information to other components composition based methods.

Acknowledgements: This work was supported by National Natural Science Foundation of China (Projects No. 61302186 and No. 61271378), Shandong Natural Science Foundation (Project No. ZR2010CQ041), the funding from the State Key Laboratory of Bioelectronics of Southeast University and the Taishan Scholars Program of Shandong province.

References

[1] P. E. Wright, H. J. Dyson, Intrinsically unstructured proteins: re–assessing the protein structure–function paradigm, *J. Mol. Biol.* **293** (1999) 321–331.

[2] A. K. Dunker, J. D. Lawson, C. J. Brown, R. M. Williams, P. Romero, J. S. Oh, C. J. Oldfield, A. M. Campen, C. M. Ratliff, K. W. Hipps, J. Ausio, M. S. Nissen, R. Reeves, C. Kang, C. R. Kissinger, R. W. Bailey, M. D. Griswold, W. Chiu, E. C. Garner, Z. Obradovic, Intrinsically disordered protein, *J. Mol. Graph. Model.* **19** (2001) 26–59.

[3] J. Habchi, P. Tompa, S. Longhi, V. N. Uversky, Introducing protein intrinsic disorder, *Chem. Rev.* **114** (2014) 6561–6588.

[4] P. Tompa, Intrsically disordered proteins: a 10-year recap, *Trends Biochem. Sci.* **37** (2012) 509–516.

[5] E. Schad, L. Kalmar, P. Tompa, Exon–phase symmetry and intrinsic structural disorder promote modular evolution in the human genome, *Nucleic Acids Res*. **41** (2013) 4409–4422.

[6] A. K. Dunker, V. N. Uversky, Drugs for ‘protein clouds’: targeting intrinsically disordered transcription factors, *Curr. Opin. Pharm.* **10** (2010) 782–788.

[7] R. Linding, L. J. Jensen, F. Diella, P. Bork, T. J. Gibson, R. B. Russell, Protein disorder prediction: implications for structural proteomics, *Structure* **11** (2003) 1453-1459.

[8] P. Radivojac, Z. Obradovic, C. J. Brown, A. K. Dunker, Prediction of boundaries between intrinsically ordered and disordered protein regions, *Pacific Symposium on Biocomputing* , 2003, pp. 216–227.

[9] J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton, D. T. Jones, Prediction and functional analysis of native disorder in proteins from the three kingdoms of life, *J. Mol. Biol*. **337** (2004) 635–645.

[10] P. Romero, Z. Obradovic, X. Li, E. C. Garner, C. J. Brown, A. K. Dunker, Sequence complexity of disordered protein, *Proteins* **42** (2001) 38–48.

[11] C. Jia, X. Lin, Z. Wang, Prediction of protein S-nitrosylation sites based on adapted normal distribution bi–profile Bayes and Chou’s pseudo amino acid composition, *Int. J. Mol. Sci.* **15** (2014) 10410–10423.

[12] J. F. Yu, X. H. Dou, H. B. Wang, X. Sun, H. Y. Zhao, J. H. Wang, A novel cylindrical representation for characterizing intrinsic properties of protein sequences, *J. Chem. Inf. Model.* **55** (2015) 1261–1270.

[13] M. Sickmeier, J. A. Hamilton, T. LeGall, V. Vacic, M. S. Cortese, A. Tantos, B. Szabo, P. Tompa, J. Chen, V. N. Unersky, Z. Obradovic, A. K. Dunker, DisProt: the database of disordered proteins, *Nucleic Acids Res*. **35** (2007) 786–793.

[14] W. Li, A. Godzik, Cd–hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, *Bioinformatics* **22** (2006) 1658–1659.

[15] C. E. Shannon, A mathematical theory of communication, *Bell Sys. Tech. J.* **27** (1948) 379–423.

[16] H. Lin, H. Wang, H. Ding, Y. L. Chen, Q. Z. Li, Prediction of subcellular localization of apoptosis protein using Chou's pseudo amino acid composition, *Acta Biotheor*. **57** (2009) 321–330.

[17] Z. C. Li, X. B. Zhou, Z. Dai, X. Y. Zou, Prediction of protein structural classes by Chou's pseudo amino acid composition: approached using continuous wavelet transform and principal component analysis, *Amino Acids* **37** (2009) 415–425.

[18] B. Liu, J. Xu, X. Lan, R. Xu, J. Zhou, X. Wang, K. C. Chou, iDNA–Prot|dis: identifying DNA–binding proteins by incorporating amino acid distance–pairs and reduced alphabet profile into the general pseudo amino acid composition, *Plos One* **9** (2014) e106691 (12 pages).

[19] J. Zhang, P. Sun, X. Zhao, Z. Ma, PECM: Prediction of extracellular matrix proteins using the concept of Chou's pseudo amino acid composition, *J. Theor. Biol.* **363** (2014) 412–418.

[20] L. Nanni, S. Brahnam, A. Lumini, Prediction of protein structure classes by incorporating different protein descriptors into general Chou's pseudo amino acid composition, *J. Theor. Biol.* **360** (2014) 109–116.

[21] K. C. Chou, Prediction of protein cellular attributes using pseudo amino acid composition, *Proteins* **43** (2001) 246–255.

[22] C. C. Chang, C. J. Lin, LIBSVM: a library for support vector machines, *ACM T. Intel. Syst. Tec*. **2** (2011) 1–27.

[23] K. M. Chung, W. C. Kao, C. L. Sun, L. L. Wang, C. J. Lin, Radius margin bounds for support vector machines with the RBF kernel, *Neural Comput*. **15** (2003) 2643–2681.

[24] E. A. Weathers, M. E. Paulaitis, T. B. Woolf, J. H. Hoh, Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein, *Febs Lett*. **576** (2004) 348–352.

[25] B. He, K. Wang, Y. Liu, B. Xue, V. N. Uversky, A. K. Dunker, Predicting intrinsic disorder in proteins: an overview, *Cell Res*. **19** (2009) 929–949.

[26] P. Romero, Z. Obradovic, C. Kissinger, J. E. Villafranca, A. K. Dunker, Identifying disordered regions in proteins from amino acid sequence. *Proc. IEEE Int. Conf. Neural Networks* **1** (1997) 90–95.

[27] S. L. Zhang, Y. Y. Liang, Z. G. Bai, A novel reduced triplet composition based method to predict apoptosis protein subcellular localization, *MATCH Commun. Math. Comput. Chem.* **73** (2015) 559–571.

[28] H. L. Hu, F-curve, a graphical representation of protein sequences for similarity analysis based on physicochemical properties of amino acids, *MATCH Commun. Math. Comput. Chem.* **73** (2015) 749-764.