http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (11): 2269-2272, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.2269.2272

Prediction of the Subcellular Localization Based on the Fusion Sequence Features

Ronghui Wu and Linchun Wang
School of Information Science and Engineering, Hunan University, Changsha, 410082, China

Abstract: The subcellular localization is quite crucial to the research of the protein functional characteristic, it is quite useful to understand the mechanism of cellular activities, so the prediction of the subcellular localization of apoptosis proteins is an significant research area in the post-genomics era. Although, much effort has been put into the methodology developing for the problem by now, much to our chagrin, these methods which have been put forward more or less have some defects. So, it is very urgent to develop a more effective method to predict the subcellular localization of the protein. This study proposed a new feature extraction method based on the fusion features of the protein sequences, this method not only contains the information of the Amino Acid Composition (AAC), but also have the location information of the amino acid residues in the protein sequence, at the same time, it also considers the biochemical properties, as well as the relationship between the different amino acids in the sequence. Then, the Nearest Neighbor (NN) algorithm is selected as the classification algorithm to predict the subcellular localization of protein in two different datasets. The results of the experiment show that the overall predictive effects are improved by the method proposed compared with some existing methods.

Keywords: Subcellular localization, sequence encoding, amino acid composition, the nearest neighbor algorithm

INTRODUCTION

With the human genome project successfully implemented, a lot of protein sequences and DNA sequences have been produced. In the public databases, the protein sequences have explosively increased. In addition, the protein plays the crucial role for all biological processes in the organism (Zhao, 2013a). So, it is urgent to develop an effective method to predict the functions of the protein. However, the subcellular localization is a vital step for the further study of the PPIs, the protein function and some potential roles in the biological cells.

Although, many biochemical experimental methods proposed by different researchers to determine the subcellular localization, but these methods still have some disadvantages, such as the time-consuming and the high-cost. So, it is urgent to design the high-efficiency computational coding methods for the problem. However, the protein sequences contain many information about the structures and functions and the structure information plays role in other places (Bose *et al*., 2012; Zhao, 2013b), the protein sequence information play an important role in the prediction of subcellular localization, protein structure, the protein function and so on, so the coding methods based on the protein sequence is especially important. At the same time, many sequence-based methods were proposed. Such as the amino acid composition, which is proposed by the Nakashima *et al*. (1986) and Nakashima and Nishikawa (1994), this method is mainly based on the frequencies of occurrence of twenty kinds of amino acids in the protein sequence. Then, there are other methods were proposed to predict the subcellular localization (Guo *et al*., 2013), for example the dipeptide composition coding method (Bhasin and Raghava, 2004) which statistics the frequencies of occurrence of the two consecutive amino acids residues. However, these methods still have some defects, for instance, the AAC not consider the location information in the protein sequence and correlation information. To deal with the problem, the Pseudo Amino Acid Composition (PAAC) method and some hybrid approaches were put forward (Chou, 2001; Lin and Wang, 2011; Xiao *et al*., 2005; Liu *et al*., 2010). To take into the account the sequence profile composition information and the compositions information of the whole sequence Pierleoni *et al*. (2006) proposed a coding method. Recently Mooney *et al*. (2011) propose a new protein sequence representation method, in this method, the whole sequences were mapped into single properties. Nevertheless, most of theses approaches are not reliable and robust in some cases and the containing information is not comprehensive, so the forecast result is not always good.

In order to get better prediction, here proposed a novel sequence-based method, which not only based on the weighed AAC, but also consider the correlation

Corresponding Author: Ronghui Wu, School of Information Science and Engineering, Hunan University, Changsha, 410082, China

information between the amino acid residues in the sequence, at the same time, this coding method also take into account the relevant physicochemical properties of the amino acids. Finally, this study selects the nearest neighbor algorithm as the classification algorithm and selects the CL317 and ZW225 as the dataset, Empirical studies have shown that the novel coding approach have better prediction performance compared with some existing methods.

THE CODING OF PROTEIN SEQUENCE

In this study, a new sequence-based coding method is proposed, this method Named Fusion Feature Coding Method (FFCM), which contains two part, the first one is Weighed Amino Acid Composition (WAAC), the other one is the autocorrelation information of the proteins in the protein sequence, in this part the Auto Covariance (AC) (Guo *et al*., 2008) is selected as the correlation function. The following 20 ordered alphabet (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y), respectively stand for the twenty amino acids. Consider a protein chain of L amino acid residues:

$$p_1\, p_2\, p_3\, p_4\, p_5 \cdots p_L \tag{1}$$

The FFCM of the protein sequence x could be defined as follows:

$$x = (x_1 x_2, \ldots\ldots x_{20}, x_{21}, x_{22} \ldots\ldots x_{20+\lambda})^T, (\lambda < L) \tag{2}$$

$$x_i = \begin{cases} \dfrac{a_i}{\sum_{i=1}^{20} a_i + w\sum_{k=1}^{\lambda} b_k}, (1 \le i \le 20) \\ \dfrac{wb_k}{\sum_{i=1}^{20} a_i + w\sum_{k=1}^{\lambda} b_k}, (20+1 \le i \le 20+\lambda) \end{cases} \tag{3}$$

where, the first 20 element $x_1 x_2, \ldots.x_{20}$ is the weighed amino acid composition, which statistics the occurrence frequency and location of the amino acid in the sequence, as the above assumption about the protein sequence, calculated as follows:

$$a_i = \frac{1}{L}\sum_{j=1}^{L} v_{i,j} \cdot j \tag{4}$$

where, the j mean the location of amino acid residue in the sequence, $j \in [1, L]$ and if the i-th amino acid appear in the j-th position of the sequence, the $v_{i,j} = 1$, otherwise, the $v_{i,j} = 0$. The next λ dimensional of the above vector is λ sequentially related factors, which reflects the sequence order correlation between all the λ most contiguous residues along a protein chain:

$$b_k = \frac{1}{L-k}\sum_{i=1}^{L-k} J_{i,i+k}, 1 \le k \le \lambda \tag{5}$$

$$J_{i,i+k} = \frac{1}{4}\sum_{j=1}^{4}[H_j(p_i) - \frac{1}{L}H_j(p_i)] \times [H_j(p_{i+k}) - \frac{1}{L}H_j(p_i)] \tag{6}$$

where, the $H_1(p_i)$, $H_2(p_i)$, $H_3(p_i)$ and $H_4(p_i)$ are, respectively the normalized value of the hydrophobicity, polarity, polarizability and the volume of side chains. In the study, it is found by preliminary tests that the optimal value for λ is 30, for the w is 5.

DATA SET

In order to access the predictive performance of the new coding method, there are two valid standard datasets are selected as the datasets, one is called CL317, which is provided by Chen and Li (2007a), in the dataset, there are 317 apoptosis proteins which are classified into six subcellular localizations:112 cytoplasmic proteins, 55 membrane proteins, 34 mitochondrial proteins, 17 secreted proteins, 52 nuclear protein and 47 endoplasmic reticulum proteins. The other is called ZW225, which is provided by Zhang *et al*. (2006), this dataset contains 225 proteins datasets and are classified into four subcellular localizations:41 nuclear proteins, 70 cytoplasmic proteins, 25 mitochondrial proteins and 89 membrane proteins.

EVALUATION MEASURE

Here, the Leave-one-out-cross-validation (LOOCV) (Zheng *et al*., 2011) is selected to detect the new coding method, in order to further evaluate the approach, the following evaluation indicators are also selected:accuracy (ACC), sensitivity (SN), precision (PE) and Matthews Correlation Coefficient (MCC) (Matthews, 1975):

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{7}$$

$$SN = \frac{TP}{TP + FN} \tag{8}$$

$$PE = \frac{TP}{TP + FP} \tag{9}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}} \tag{10}$$

where, TP, TN, FP and FN denote true positive, true negative, false positive and false negative, respectively.

RESULTS AND DISCUSSION

Here, the FFCM is analyzed and compared with several existing methods. This method contains comprehensive feature information, such as the composition information, location information and the correlation features and in the experiment, two classic standard dataset are selected as the training and testing data set. So, the FCCM have achieved better predictive performance compared with other existing approaches on the whole.

Table 1 and 2, respectively show the sensitivity of the proposed FCCM and the existing methods on the two dataset. In contrast, the FCCM has the higher sensitivity, although effects on the individual data sets are slight inferior.

Table 3 and 4, respectively demonstrate the overall predictive performance of the FCCM and other methods, the selected evaluation indicators are ACC and MCC, as listed in the table, regardless of the CL317 or the ZW225, the FCCM achieved better accuracy. In addition, the FCCM also have higher MCC in most of the datasets.

Table 1: Predicting results of SN on the dataset CL317 by the LOOCV

	SN (%)					
Predictor	Cy	Me	Nu	En	Mi	Se
Chen and Li (2007a)	81.3	81.8	82.7	83.0	85.3	88.2
Chen and Li (2007b)	91.1	89.1	73.1	87.2	79.4	58.8
Liao *et al.* (2010)	88.4	85.5	78.9	91.5	76.5	58.5
FFCM	92.8	83.8	86.0	97.8	60.0	76.5

Cy: Cytoplasmic, Me: Membrane, Nu: Nuclear, En: Endoplasmic, Mi: Mitochondrid and Se: Secreted, respectively

Table 2: Predicting results of SN on the dataset ZW225 by the LOOCV

	SN (%)			
Predictor	Me	Cy	Nu	Mi
Zhang *et al.* (2006)	93.3	90.0	63.4	60.0
Chen and Li (2007a)	91.0	92.9	73.2	68.0
Zhang *et al.* (2009)	92.1	87.1	73.2	64.0
FFCM	89.8	87.1	78.0	72.0

Me: Membrane, Cy: Cytoplasmic, Nu: Nuclear and Mi: Mitochondrid, respectively

Table 3: Predicting results of MCC and ACC on the datasets CL317 by the LOOCV

	MCC (%)						
Predictor	Cy	Me	Nu	En	Mi	Se	ACC (%)
Chen and Li (2007a)	0.80	0.77	0.73	0.90	0.74	0.68	82.7
Chen and Li (2007b)	0.80	0.83	0.69	0.91	0.77	0.65	84.2
Liao *et al.* (2010)	-	-	-	-	-	-	83.6
FFCM	0.83	0.84	0.89	0.91	0.67	0.77	87.5

Cy: Cytoplasmic, Me: Membrane, Nu: Nuclear, En: Endoplasmic, Mi: Mitochondrid and Se: Secreted, respectively

Table 4: Predicting result of ACC and MCC on the datasets ZW225 by the LOOCV

	MCC (%)				
Predictor	Me	Cy	Nu	Mi	ACC (%)
Zhang *et al.* (2006)	93.3	90.0	63.4	60.0	83.1
Chen and Li (2007a)	91.0	92.9	73.2	68.0	85.8
Zhang *et al.* (2009)	92.1	87.1	73.2	64.0	84.0
FFCM	81.4	73.6	78.0	78.1	84.8

Me: Membrane, Cy: Cytoplasmic, Nu: Nuclear and Mi: Mitochondrid, respectively

CONCLUSION

In this study, a sequence-based approach is proposed to predict the subcellular localization of apoptosis proteins based on the fusion features and the nearest neighbor algorithm. As previously expected, the approach obviously improves the prediction performance compared with several existing methods. However, due to the limited knowledge, so the further study should put into the improvement of the prediction performance.

REFERENCES

Bhasin, M. and G.P.S. Raghava, 2004. ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. Nucleic Acids Res., 32: W414-W419.

Bose, P.K., N. Paitya, S. Bhattacharya, D. De and S. Saha et al., 2012. Influence of light waves on the effective electron mass in quantum wells, wires, inversion layers and superlattices. Quantum Matter, 1: 89-126.

Chen, Y.L. and Q.Z. Li, 2007a. Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo-amino acid composition. J. Theor. Biol., 248: 377-381.

Chen, Y.L. and Q.Z. Li, 2007b. Prediction of the subcellular localization of proteins. J. Theor. Biol., 45: 775-783.

Chou, K.C., 2001. Prediction of protein cellular attributes using pseudo amino acid composition. Proteins Struct. Funct. Genet., 43: 246-255.

Guo, Y., L. Yu, Z. Wen and M. Li, 2008. Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. Nucleic Acids Res., 36: 3025-3030.

Guo, Z., S. Yang, Q. Hu and L. Peng, 2013. A transverse and longitudinal encoding of protein sequence and its application. J. Comput. Theor. Nanosci., 10: 271-275.

Liao, B., B.Y. Liao and Q.G. Zeng, 2010. A novel method for similarity analysis and protein subcellular localization prediction. Bioinformatics, 26: 2678-2683.

Lin, J. and Y. Wang, 2011. Using a novel adaboost algorithm and chous pseudo amino acid composition for predicting protein subcellular localization. Prot. Peptide Lett., 18: 1219-1225.

Liu, T., X. Zheng, C. Wang and J. Wang, 2010. Prediction of subcellular location of apoptosis proteins using pseudo amino acid composition: An approach from auto covariance transformation. Prot. Peptide Lett., 17: 1263-1269.

Matthews, B.W., 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta Protein Struct., 405: 442-451.

Mooney, C., Y.H. Wang and G. Pollastri, 2011. SCLpred: Protein subcellular localization prediction by N-to-1 neural networks. Bioinformatics, 27: 2812-2819.

Nakashima, H. and K. Nishikawa, 1994. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. J. Mol. Biol., 238: 54-61.

Nakashima, H., K. Nishikawa and O. Tatsuo, 1986. The folding type of a protein is relevant to the amino acid composition. J. Biochem., 99: 153-162.

Pierleoni, A., P.L. Martelli, P. Fariselli and R. Casadio, 2006. BaCelLo: A balanced subcellular localization predictor. Bioinformatics, 22: e408-e416

Xiao, X., S. Shao, Y. Ding, Z. Huang, Y. Huang and K.C. Chou, 2005. Using complexity measure factor to predict protein subcellular location. Amino Acids, 28: 57-61.

Zhang, L., B. Liao, D. Li and W. Zhu, 2009. A novel representation for apoptosis protein subcellular localization prediction using support vector machine. J. Theor. Biol., 259: 361-365.

Zhang, Z.H., Z.H. Wang, Z.R. Zhang and Y.X. Wang, 2006. A novel method for apoptosis protein subcellular localization prediction combining encoding based on grouped weight and support vector machine. FEBS Lett., 580: 6169-6174.

Zhao, Q., 2013a. Allodynamic regulation of protein activity. Quantum Matter, 2: 144-152.

Zhao, Q., 2013b. Nature of protein dynamics and thermodynamics. Rev. Theor. Sci., 1: 83-101.

Zheng, L.L., S. Niu, P. Hao, K. Feng, Y. D. Cai and Y. Li, 2011. Prediction of protein modification sites of pyrrolidone carboxylic acid using mRMR feature selection and analysis. PloS One, Vol. 6. 10.1371/journal.pone.0028221