GMR
Genetics and Molecular Research
ISSN 16765680

Prediction of protein subcellular multisite localization using a new feature extraction method

L.Y. Wang, D. Wang and Y.H. Chen

School of Information Science and Engineering, University of Jinan, Jinan, China

Corresponding authors: D. Wang / Y.H. Chen
E-mail: ise_wangd@ujn.edu.cn / yhchen@ujn.edu.cn

Genet. Mol. Res. 15 (3): gmr.15039013
Received July 21, 2016
Accepted August 1, 2016
Published September 23, 2016
DOI http://dx.doi.org/10.4238/gmr.15039013

ABSTRACT. A basic problem of proteomics is identifying the subcellular locations of a protein. One factor making the problem more complicated is that some proteins may simultaneously exist in two or more than two subcellular locations. To improve multisite prediction quality, it is necessary to use effective feature extraction methods. Here, we developed a new feature extraction method based on the *pK* value and frequencies of amino acids to represent a protein as a real values vector. Using this novel feature extraction method, the multi-label k-nearest neighbors (ML-KNN) algorithm and setting different weights into different attributes' ML-KNN, known as wML-KNN, were employed to predict multiplex protein subcellular locations. The best overall accuracy rate on dataset S1 from the predictor of Virus-mPLoc was 59.92 and 86.04% on dataset S2 from Gpos-mPLoc, respectively.

Key words: Multisite protein; Subcellular localization; Extraction; Multi-label k-nearest neighbors algorithm

INTRODUCTION

Basic biology has shown that a cell is a highly ordered structure whose interior is subdivided into many organelles such as the nucleus, cell wall, ribosomes, mitochondria, and Golgi apparatus, among others, which are collectively referred to as the subcellular location. These subcellular structures in the cell form a large dynamic system. Proteins are among the most important materials in the cell. Proteins and other macromolecules are synthesized, transferred, and activated to function inside this system (Du and Xu, 2013). Protein subcellular locations are also closely related to metabolic pathways, signal transduction, and biological processes, and they play a crucial role in therapeutic target discovery, drug design, and biological research. Knowledge of the subcellular locations of proteins can provide key hints and useful insight into revealing their functions, improving the understanding of the intricate pathways regulating biological processes at the cellular level (Lin et al., 2013). Proteins in cells play very important roles, but can only work normally in specific subcellular locations. Protein subcellular localization is closely linked with function. Proper localization of a protein is a precondition for cell function (Qu et al., 2013).

However, experimental approaches are typically time-consuming, tedious, and costly, and they may lack reproducibility. Particularly, these conventional experiment-based techniques cannot keep pace with the increasing number of novel protein sequences generated using high-throughput next-generation sequencing. Thus, the development of computational methods for the timely and effective identification of newly discovered protein subcellular locations is critical. Over the past two decades, numerous theoretical and computational methods have been developed to predict protein subcellular locations (Zhang et al., 2013).

In this study, we developed a new feature extraction method based on the pK value and frequencies of amino acids to represent a protein as a real values vector. Based on this novel feature extraction method, the multi-label k-nearest neighbors (ML-KNN) algorithm and setting different weights into different attributes' ML-KNN, known as wML-KNN, were employed to predict multiplex protein subcellular locations. Next, several different performance measures for multilabel classifications were used to evaluate this algorithm.

MATERIAL AND METHODS

Dataset

We employed the benchmark dataset S1 for Virus-mPLoc (Chou and Cai, 2005) in this study. This dataset contains both single-site and multisite proteins and was established for viral proteins. It includes 252 locative protein sequences, classified into 6 subcellular locations. Among the 207 different proteins, 165 belong to one subcellular location, 39 to two locations, 3 to three locations, and none to four or more locations. None of the proteins showed ≥25% sequence identity to any other protein in the same subset except viral capsid. The number of proteins in different subcellular locations for dataset S1 is listed in Table 1, which can be downloaded from the website http://www.csbio.sjtu.edu.cn/bioinf/virus-multi/Data.htm.

We used the benchmark dataset S2 for Gpos-mPloc (Su et al., 2005) in this study. This dataset contains both single-site and multisite proteins and is dedicated to Gram-positive bacteria. It includes 523 Gram-positive bacterial protein sequences, classified into 4 subcellular locations. Among the 519 different proteins, 515 belong to one location and 4 to two locations.

None of the proteins show ≥25% sequence identity to any other protein in the same subset (subcellular location). The number of proteins in different subcellular locations for dataset S2 is listed in Table 2; this dataset can be downloaded from the website http://www.csbio.sjtu.edu.cn/bioinf/Gpos-multi/Data.htm.

Table 1. Number of proteins in different subcellular locations for the dataset S1 downloaded from Virus-mPloc.

Order	Subcellular location	Number of proteins
1	Viral capsid	8
2	Host cell membrane	33
3	Host endoplasmic reticulum	20
4	Host cytoplasm	87
5	Host nucleus	84
6	Secreted	20

Table 2. Number of proteins in different subcellular locations for the dataset S2 derived from Gpos-mPloc.

Order	Subcellular location	Number of proteins
1	Cell membrane	174
2	Cell wall	18
3	Cytoplasm	208
4	Extracell	123

New feature extraction method

The 20 amino acids can be divided into four groups (Pánek et al., 2005), which use the letters L, B, W, and P, respectively, where L includes strong hydrophilic amino acids, B includes strong hydrophobic amino acids, W includes weak hydrophilic or hydrophobic amino acids, and P, G, and C have relatively unique properties, which are divided into a class using the letter P. The information for classification is listed in Table 2. For example, a sequence of protein {MERIKELRDLMSQG} can be converted to {BLLBLLBLLBBWLP} as shown in Table 3.

Table 3. Classification of amino acids.

Hydrophilic-hydrophobic property	Abbreviation	Amino acids
Strong hydrophilic	L	D, E, H, K, N, R, Q
Strong hydrophobicity	B	A, F, I, L M, V
Weak hydrophilic or hydrophobic	W	S, T, W, Y
Special	P	C, G, P

The amino acid type is mainly determined by its side chain R. Different amino acid sequences have different structures and perform different functions. The *pK* value represents the dissociation constant of the amino acid. A smaller *pK* value is indicates greater disintegration, where pK_{a1} and pK_{a2} (Shen and Chou, 2008) represent the dissociation constants of -*COOH* and -*NH* in an α carbon atom. This study used the dissociation constant pK_{a2}.

$$pK_{a1} = -\log \frac{[H^+][R_0]}{[R^+]} \quad \text{(Equation 1)}$$

$$pK_{a2} = -\log\frac{[H^+][R^-]}{[R_0]} \qquad \text{(Equation 2)}$$

Define pK_i (i = 1, 2, …, 20) to represent the pK value of 20 amino acids according to R, D, E, N, Q, K, H, L, I, V, A, M, F, S, Y, T, W, P, G, and C in order (Chen and Li, 2007), with pK values at 25°C shown in Table 4.

Table 4. Twenty amino acids and their pK_{a2} values.

Amino acids	Abbreviation	pK_{a2}
Arginine	R	9.09
Aspartic	D	9.60
Glutamic	E	9.67
Asparagine	N	9.09
Glutamine	Q	9.13
Lysine	K	8.90
Histidine	H	8.97
Leucine	L	9.60
Isoleucine	I	9.76
Valine	V	9.74
Alanine	A	9.87
Methionine	M	9.21
Phenylalanine	F	9.24
Serine	S	9.15
Threonine	T	9.12
Tyrosine	Y	9.11
Tryptophan	W	9.39
Proline	P	10.60
Glycine	G	9.60
Cystine	C	10.78

The value of pK_i (i = L, B, P, W) can be indirectly expressed as

$$pK_L = \sum_{i=1}^{L_L} \frac{pK_i}{L_L} \qquad \text{(Equation 3)}$$

$$pK_B = \sum_{i=L_L+1}^{L_L+L_B} \frac{pK_i}{L_B} \qquad \text{(Equation 4)}$$

$$pK_W = \sum_{i=L_L+L_B+1}^{L_L+L_B+L_W} \frac{pK_i}{L_W} \qquad \text{(Equation 5)}$$

where L_L represents the number of amino acids in class L, L_B represents the number of amino acids in class B, L_W represents the number of amino acids in class W, and L_P represents the number of amino acids in class P.

Any two-letter combination, such as LL, PW, and so on, can form the 16-dimensonal feature vector, as shown in Table 5.

Table 5. Values of two-letter combinations.

Two-letter combination	pK value
LL	$pK_{11} = \frac{pK_L + pK_L}{2}$
LB or BL	$pK_{12} = pK_{21} = \frac{pK_L + pK_B}{2}$
LW or WL	$pK_{13} = pK_{31} = \frac{pK_L + pK_W}{2}$
LP or PL	$pK_{14} = pK_{41} = \frac{pK_L + pK_P}{2}$
BB	$pK_{22} = \frac{pK_B + pK_B}{2}$
BW or WB	$pK_{23} = pK_{32} = \frac{pK_B + pK_W}{2}$
BP or PB	$pK_{33} = \frac{pK_W + pK_W}{2}$
WW	$pK_{24} = pK_{42} = \frac{pK_B + pK_P}{2}$
WP or PW	$pK_{34} = pK_{43} = \frac{pK_W + pK_P}{2}$
PP	$pK_{44} = \frac{pK_P + pK_P}{2}$

We used the new vector based on the pK value and the frequencies of the 20 native amino acids to predict subcellular locations.

$$V = \{pK_1 \cdot f_1, \cdots pK_{20} \cdot f_{20}, pK_{11} \cdot f_{11}, \cdots pK_{44} \cdot f_{44}\} \quad \text{(Equation 7)}$$

$$f_i = \frac{n_i}{n} (i = 1,2,\cdots,20) \quad \text{(Equation 8)}$$

$$f_{ij} = \frac{n_{ij}}{n-1} (i, j = 1,2,3,4) \quad \text{(Equation 9)}$$

Each element of the 36 dimensional vector was composed of two parts of the product. One was based on the pK value and the other was based on the relative frequency.

Algorithm

ML-KNN (Zhang and Zhou, 2007) is a simple non-parametric multilabel classifier that uses the k-nearest neighbor algorithm for statistics of the category tag information of neighbor samples and exploits the principle of maximum posterior probability to infer the no example of label set. Let $D = \{(x_i, Y_i) \mid 1 \leq i \leq p\}$ be the multilabel training set and let x be the no example. Suppose that $N(x)$ represents the set of knn of x discerned in the training set. For the j - th category of y_i, the ML-KNN algorithm will calculate the following statistics,

$$C_j = \sum_{(x,Y)\in N(x)} \{y_j \in Y\} \quad \text{(Equation 10)}$$

where C_j expresses the number of neighbors when x belongs to the $N(x)$ class, H_j represents this event for x containing the category of y_i, and $P(H_j \mid C_j)$ denotes the posterior probability established for H_j when $N(x)$ includes the number samples of C_j with a category label y_i. A multi-label classifier is required and can be expressed as:

$$h(x) = \{y_j \mid \frac{P(H_j \mid C_j)}{P(\neg H_j \mid C_j)} > 0.5, 1 \le j \le q\} \quad \text{(Equation 11)}$$

The formula shows that the posterior probability $P(H_j \mid C_j)$ is greater than $P(\neg H_j \mid C_j)$, the mark y_i belongs to the example of x. The function based on Bayes' theorem can be rewritten as:

$$h(x) = \{y_j \mid \frac{P(H_j)P(C_j \mid H_j)}{P(\neg H_j)P(C_j \mid \neg H_j)} > 0.5, 1 \le j \le q\} \quad \text{(Equation 12)}$$

However, objectively uneven training set results in relatively low accuracy. To overcome this limitation, an attribute weighted-based improving algorithm was proposed (Shen and Bai, 2008). The weighted factor w_t is defined as:

$$w_t = \frac{\log(a + \frac{AvgNum}{Num(C_t)})}{\log(a+1)} (1 < t < N) \quad \text{(Equation 13)}$$

$$a = \frac{MaxNum}{AvgNum} \quad \text{(Equation 14)}$$

where $AvgNum$ refers to the average number of samples in different categories and $Num(C_t)$ refers to the C_t class containing the number of samples. In the ML-KNN algorithm, the prior probability is weighted and referred to as wML-KNN.

Evaluation functions

Supposing a multi-label test set χ = {(x_i, X_i) | 1 ≤ i ≤ n}, based on the notations in Section 4, the following widely used multi-label evaluation measurements (Schapire and Singer, 2000; Zhang, 2009; Huang and Yuan, 2013):

Hamming loss:

$$hamming\_loss \qquad \chi(t) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{C}\left|t(x_i)\Delta X_i\right| \qquad \text{(Equation 15)}$$

Δ represents the symmetric difference between two data sets. Hamming loss is used to evaluate how many times an instance-label pair is classified falsely. A lower hamming_loss value indicates better performance.

$$one\_error \qquad \chi(g) = \frac{1}{n}\sum_{i=1}^{n}[\![\arg\max g(x_i,c)]\!] \notin X_i] \qquad \text{(Equation 16)}$$

One-error evaluates how many times the top-level label is not in set of appropriate labels of the instance. A smaller one-error value indicates better performance.

$$\text{coverage} \qquad \chi(g) = \frac{1}{n}\sum_{i=1}^{n}\max rank_g(x_i,c)-1 \qquad \text{(Equation 17)}$$

Coverage is used to evaluate how far we must go down the list of labels in order to cover all proper labels of the instance on the average. A lower coverage value indicates better performance.

$$average\_prec \qquad \chi(g) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\left|X_i\right|}\sum_{c\in X_i}AVP(x_i) \qquad \text{(Equation 18)}$$

$$AVP(x_i) = \frac{\left|c'\middle|rank_g(x_i,c') \le rank_g(x_i,c), \qquad c' \in X_i\right|}{rank_g(x_i,c)} \qquad \text{(Equation 19)}$$

Average precision is used to evaluate the average fraction of labels ranked above a particular label $c \in X$, which actually are in *X*. A larger value of average precision value indicates better performance.

Absolute-true (Chou et al., 2012):

$$absolute\_true \qquad \chi(t) = \frac{1}{n}\sum_{i=1}^{n}\left[x_i = c\right] \qquad \text{(Equation 20)}$$

The evaluation index is used to show the proportion of predicted labels consistent with the real tag set. A large absolute-true value indicates better performance.

RESULTS

In this study, in order to enhance the accuracy of prediction of multiplex protein subcellular localization, we adopted a new feature extraction method to compare the results of ML-KNN and wML-KNN in datasets S1 and S2. A large number of experiments showed that when *k* had a value of 1, the system showed the highest accuracy. The best experimental results for each metric are shown in Tables 6-9. To demonstrate that the feature extraction method is superior to previous methods, we adopted PseAAC as a representative former method to extract features for comparison with the new method. As shown in the following tables, for each evaluation criterion, downward arrow indicates that lower values are preferred, while upward arrow indicates that larger values are preferred.

Table 6. Results of each evaluation criterion for two different algorithms on dataset S1 using our new feature extraction method.

Evaluation metric	Algorithm	
	ML-KNN	wML-KNN
Hamming loss↓	0.1362	0.1278
One-error↓	0.3810	0.3333
Coverage↓	1.0670	0.9563
Average precision↑	0.7701	0.8079

Table 7. Results of each evaluation criterion for two different algorithms on dataset S1 using PseAAC as feature extraction method.

Evaluation metric	Algorithm	
	ML-KNN	wML-KNN
Hamming loss↓	0.1409	0.1316
One-error↓	0.3968	0.3452
Coverage↓	1.0992	0.9921
Average precision↑	0.7614	0.7990

Table 8. Results of each evaluation criterion for two different algorithms on dataset S2 using our new feature extraction method.

Evaluation metric	Algorithm	
	ML-KNN	wML-KNN
Hamming loss↓	0.1707	0.0631
One-error↓	0.3480	0.1224
Coverage↓	0.5889	0.1625
Average precision↑	0.7919	0.9347

Table 9. Results of each evaluation criterion for two different algorithms on dataset S2 using PseAAC as feature extraction method.

Evaluation metric	Algorithm	
	ML-KNN	wML-KNN
Hamming loss↓	0.1769	0.0798
One-error↓	0.3595	0.1587
Coverage↓	0.5946	0.2084
Average precision↑	0.7871	0.9149

As shown in Tables 6-7, the new feature extraction method performed better on all evaluation measurements compared to when PseAAC was adopted as the feature extraction method.

As shown in Tables 8-9, the new feature extraction method performed better for every evaluation measurement compared to PseAAC as the feature extraction method.

Tables 6 and 8 show that the datasets S1 and S2 achieved the best performance for every evaluation metric. As shown in Tables 6-9, the wML-KNN algorithm showed superior performance compared to ML-KNN for four different evaluation criteria using the same dataset and same feature extraction method, and thus best performance for each criterion is presented in the rightmost section.

As shown in Table 10, the new feature extraction method showed superior performance, achieving 57.94 and 64.24% accuracy for the ML-KNN algorithm and 59.92 and 86.81% accuracy rates for the wML-KNN algorithm. The best performance for each measurement was better than those shown in Tables 6 and 8. PseAAC showed 54.76 and 62.25% accuracy for the ML-KNN algorithm and 57.14 and 83.37% accuracy rates for the wML-KNN algorithm. Thus, the novel feature extraction method is useful for extracting the characteristics of proteins to improve the accuracy of predictions of protein subcellular locations.

Table 10. Best accuracy of different datasets using different feature extraction methods.

Algorithm method	ML-KNN		wML-KNN	
	S1	S2	S1	S2
New feature extraction method	0.5794	0.6424	0.5992	0.8681
PseAAC	0.5476	0.6252	0.5714	0.8337

CONCLUSIONS

Predicting protein subcellular locations is a challenging and complex problem, particularly when the system contains some proteins with both single and multiplex site proteins. In this study, using a new feature extraction method, we obtained a higher accuracy rate for dataset S1 from the predictor of Virus-mPLoc and for dataset S2 from Gpos-mPLoc. Based on this novel feature extraction representation, when wML-KNN was adopted, better results were obtained than when the ML-KNN algorithm was used.

Overall, the new feature extraction method can extract the comprehensive characteristics of proteins, which can achieve the best performance for different datasets in every criteria compared to previous feature extraction methods and some feature fusion methods. Thus, better performance using other novel feature extraction methods should be further explored. Our future studies will focus on multisite protein subcellular localization.

ACKNOWLEDGMENTS

Research partially supported by the Youth Project of National Natural Science Fund (Grant #61302128), Youth Science and Technology Star Program of Jinan City (#201406003), Natural Science Foundation of Shandong Province (#ZR2011FL022, #ZR2013FL002), Shandong Provincial Natural Science Foundation, China, under Grant #ZR2015JL025, Program for Scientific research innovation team in Colleges and Universities of Shandong Province (#2012-2015), and Shandong Provincial Key Laboratory of Network Based Intelligent Computing.

REFERENCES

Chou KC and Cai YD (2005). Predicting protein localization in budding yeast. *Bioinformatics* 21: 944-950. http://dx.doi.org/10.1093/bioinformatics/bti104

Chou KC, Wu ZC and Xiao X (2012). iLoc-Hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. *Mol. Biosyst.* 8: 629-641. http://dx.doi.org/10.1039/C1MB05420A

Chen YL and Li QZ (2007). Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo-amino acid composition. *J. Theor. Biol.* 248: 377-381. http://dx.doi.org/10.1016/j.jtbi.2007.05.019

Du P and Xu C (2013). Predicting multisite protein subcellular locations: progress and challenges. *Expert Rev. Proteomics* 10: 227-237. http://dx.doi.org/10.1586/epr.13.16

Huang C and Yuan J (2013). Using radial basis function on the general form of Chou's pseudo amino acid composition and PSSM to predict subcellular locations of proteins with both single and multiple sites. *Biosystems* 113: 50-57. http://dx.doi.org/10.1016/j.biosystems.2013.04.005

Lin WZ, Fang JA, Xiao X and Chou KC (2013). iLoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins. *Mol. Biosyst.* 9: 634-644. http://dx.doi.org/10.1039/c3mb25466f

Pánek J, Eidhammer I and Aasland R (2005). A new method for identification of protein (sub)families in a set of proteins based on hydropathy distribution in proteins. *Proteins* 58: 923-934. http://dx.doi.org/10.1002/prot.20356

Qu XM, Chen YH, Qiao SP (2013).Predicting the Subcellular Localization of Proteins with Multiple Sites Based on N-terminal Signals. DOI: 10.1109/ISCC-C.2013.101.

Schapire RE and Singer Y (2000). BoosTexter: a boosting-based system for text categorization. *Mach. Learn.* 39: 135-168. http://dx.doi.org/10.1023/A:1007649029923

Shen HB and Chou KC (2008). PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. *Anal. Biochem.* 373: 386-388. http://dx.doi.org/10.1016/j.ab.2007.10.012

Shen ZB and Bai QY (2008). Knn text classification method based on weight modify. *Comput. Sci.* 35: 123-126.

Su CY, Lo A, Lin CC, Chang F, et al. (2005). A novel approach for prediction of multi-labeled protein subcellular localization for prokaryotic bacteria [C]//Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference Workshops, August 8-12, Stanford, California. Piscataway: IEEE, 79-80.

Zhang ML (2009). ML-RBF: RBP neural networks for multi-label learning. *Neural Process. Lett.* 29: 61-74. http://dx.doi.org/10.1007/s11063-009-9095-3

Zhang ML and Zhou ZH (2007). ML-KNN: a lazy learning approach to multi-label learning. *Pattern Recognit.* 40: 2038-2048. http://dx.doi.org/10.1016/j.patcog.2006.12.019

Zhang SW, Liu YF, Yu Y, Zhang TH, et al. (2013). MSLoc-DT: A new method for predicting the protein subcellular location of multispecies based on decision templates. *Anal. Biochem.* 449: 164-171.