Bio-Medical Materials and Engineering 26 (2015) S2003–S2009
DOI 10.3233/BME-151504
IOS Press

Multisite protein subcellular localization prediction based on entropy density

Qing Zhao, Dong Wang*, Yuehui Chen* and Xumi Qu
School of Information Science and Engineering, University of Jinan, Jinan, P.R. China

Abstract. Protein subcellular localization prediction is currently receiving much attention in the field of protein research. Many researchers make great efforts to study single-site protein subcellular localization, but the experimental data shows that many proteins can be found in two or more sub-cellular locations, prompting the study of multisite protein sub-cellular localization. This study utilized a Gpos-mPLOC data set and pseudo amino acid compositions, physicochemical properties of amino acid composition, and entropy density as three effective feature extraction methods. Then, these features were then placed in a multi-label k nearest neighbor classifier to predict subcellular protein locations. Experimental results verified that this approach provides a localization precision of 66.73% through the Jack-knife test.

Keywords: Multisite, pseudo amino acid composition, physicochemical propertiess, entropy density, multi-label k nearest neighbor

1. Introduction

Proteins play an important role in life activities and are closely related to protein function. However, in order for proteins to correctly perform their corresponding functions, correct subcellular location is necessary, otherwise the protein may harm the body or even create serious dangerous effects [1]. In light of this, protein subcellular localization has great realistic significance. Predicting protein subcellular localization has been a hot topic for many years, and researchers have made great efforts to predict protein subcellular localization and established numerous effective prediction algorithms recent years [2].

However, most current research focuses on single-site subcellular protein localization. Experimental data on the study of proteins has shown that some proteins can be located in two or more subcellular locations, thus expanding the field of protein subcellular localization [3]. Multisite protein datasets have been proposed, leading many scholars begin to study multisite protein subcellular localization. In much research surrounding multisite protein subcellular localization, many feature extraction models have been proposed, such as pseudo amino acid composition (Pse-AAC), physicochemical properties of amino acid composition, and gene ontology, among other methods [4]. Many prediction methods

* Address for correspondence: Dong Wang, School of Information Science and Engineering, University of Jinan, Jinan, P.R. China. Tel.: 15376418126; Fax: 0531-89736503; E-mail: ise_wangd@ujn.edu.cn.
Yuehui Chen, School of Information Science and Engineering, University of Jinan, Jinan, P.R. China. Tel.: 13969187693; Fax: 0531-89736503; E-mail:yhchen@ujn.edu.cn.

0959-2989/15/$35.00 © 2015 – IOS Press and the authors.

This article is published with Open Access and distributed under the terms of the Creative Commons Attribution and Non-Commercial License.

have also been proposed including multi-label k nearest neighbor classification (ML-KNN), and rank support vector machine (rank-SVM).

In this paper, we do research onwe investigate a Gpos-mPloc data set in terms of. We choose pseudo amino acid composition, the physicochemical properties of amino acid composition, and entropy density. The use of these three feature extraction methods allows us to obtain efficient features, and to use a multi-label k nearest neighbor classifier to predict the sub-cellular localization.

2. Dataset

The Gpos-mPLoc dataset was developed to identify the subcellular localization of Gram-positive bacterial proteins by combining gene ontology with functional domain and sequential evolution information. Compared to the old Gpos-PLoc dataset, the new predictor is much more powerful and flexible. Additionally, it has the particular capacity to deal with multiple-location proteins, as indicated by the character "m" in front of "PLoc" of its name.

This benchmark dataset includes 523 Gram-positive bacterial protein sequences, classified into four subcellular locations [5]. Among the 519 different proteins, 515 are single-location proteins and four proteins belong to two locations, as shown in Figure 1.

3. Feature extraction

Feature extraction is an important component of multisite protein subcellular localization prediction, and its effective features can greatly increase prediction precision. In order to obtain efficient features, the protein composition information and amino acid location information must be exclusively considered. In this article, we choose three effective feature extraction methods: pseudo amino acid composition [6], the physicochemical properties of amino acid composition, and entropy density three effective feature extraction methods. Pseudo amino acid composition and entropy density reflect both composition and information and location information, but are not sufficient to express proteins.

As amino acid have hundreds of physicochemical properties, physiochemical information can also include location information, so the physicochemical properties of amino acid composition was determined to be most effective for efficient feature extraction.

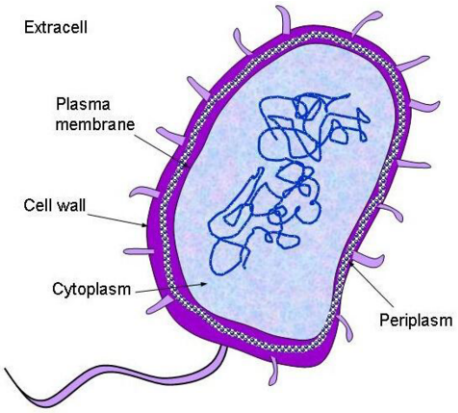

Fig. 1. Four subcellular locations of Gram-positive bacterial proteins.

3.1. Pseudo amino acid composition

Pseudo amino acid composition is the most commonly used method of protein feature extraction, and offers several advantages. It is not only a reflection of protein composition information, but also it takes amino acid position information into account [7]. A protein sequence S of length L can be expressed by the following equation:

$$\mathrm{S} = \left[\mathrm{s}_1, \mathrm{s}_2, \cdots, \mathrm{s}_{20}, \mathrm{s}_{20+1}, \cdots, \mathrm{s}_{20+\mathrm{n}}\right]^{\mathrm{T}} \left(\mathrm{n} < \mathrm{L}\right) \tag{1}$$

The first 20 dimensions are the frequency at which each amino acid appears in protein sequence S, and represents the composition information. The other n dimension reflects the relationship among amino acids, and expresses the position information.

The relationship between amino acids can be calculated by the following equation:

$$\tau_n = \frac{1}{L-n}\sum_{i=1}^{L-n} J_{i,i+n} (n < L) \tag{2}$$

$J_{i,i+n}$ can be obtained from the following equation:

$$J_{i,i+n} = \frac{1}{3}\left\{\left[H_1(R_{i+n}) - H_1(R_i)\right]^2 + \left[H_2(R_{i+n}) - H_2(R_i)\right]^2 + \left[M(R_{i+n}) - M(R_i)\right]^2\right\} \tag{3}$$

$H_1(R_i), H_2(R_i), M(R_i)$ represent the hydrophobicity, the hydrophilic and the side chain molecular weight of amino acid residue R_i, respectively.

Next, the feature vector is calculated from the following equation:

$$P_j = \begin{cases} \dfrac{f_j}{\sum_{i=1}^{20} f_i + \omega\sum_{n=1}^{\lambda}\tau_n} \left(1 \le j \le 20\right) \\ \dfrac{\omega\tau_{j-20}}{\sum_{i=1}^{20} f_i + \omega\sum_{n=1}^{\lambda}\tau_n} \left(20+1 \le j \le 20+\lambda\right) \end{cases} \tag{4}$$

ω is the weight factor, which is typically 0.05. f_i represents the frequency of amino acid residue i in the protein sequence. According to this method, if the value of λ is 20 or 30, then the dimension of feature vector is 40 or 50 [6].

3.2. Physiochemical properties of amino acid composition

All proteins can be divided into three groups on the basis of seven physicochemical properties: hydrophobicity, normalized van der Waals volume, polarity, polarizaibility, charge, secondary

Table 1

Three groups based on seven physicochemical properties

	polar	neutral	hydrophobic
hydrophobicity	RKEDQN	GASTPHY	CLVIMFW
Normalized vander Waals	GASCTPD	NVEQIL	MHKFRYW
polarity	LIFWCMVY	PATGS	HQRKNED
polarizibility	GASDT	CPNVEQIL	KMHFRYW
charge	KR	ANCQGHILMFPSTWYV	DE
secondary structures	EALMQKRH	VIYCWFT	GNPSD
solvent accessibility	ALFCGIVW	RKQEND	MPSTHY

structures and solvent accessibility [8]. The three groups are displayed in Table 1.

ForIn each kind ofof the seven physiochemical properties, counting frequency of polar acid, neutral acid and hydrophobic acid in every protein sequence, thenthere are can get 21 dimensional feature vectors.

3.3. Entropy density

According to the entropy of proteins, a protein sequence can be expressed by 20 dimensional numerical vectors [9]. The information entropy information of a protein can be calculated by the following equation:

$$H(S) = -\sum_{i=1}^{20} f_i \log f_i \tag{5}$$

f_i represents the frequency with which an amino acid i appears in protein S. Therefore, the entropy density of acid *i* can calculated as follows:

$$s_i(S) = -\frac{1}{H(S)} f_i \log f_i \tag{6}$$

A protein sequence can be expressed by vector s(S), as follows:

$$s(S) = (s_1(S), s_2(S), s_3(S), \cdots s_{20}(S))^T \tag{7}$$

4. Classification algorithm

A key step in the localization prediction is to select an appropriate classification algorithm. With the progressive study of multisite protein subcellular location, researchers have proposed many multi-label algorithms, including the Multi-label k Nearest Neighbor (ML-kNN) algorithm [10] and the Back Propagation Multi-label (BP-MLL) algorithm, among others. In this article, the multi-label k nearest neighbor algorithm is used to achieve multisite protein subcellular location classification.

The multi-label k nearest neighbor algorithm is derived from the k nearest neighbor algorithm; the

unknown sample obtains its label information from the label information of its k neighbors. The basic principle is to first calculate the prior probability of every label. Then, ensure the k neighbors of the unknown sample according to the k nearest neighbor method, and note the label information of the k neighbors. Then, calculate the probability that the unknown sample contains or doesn't contain each label while its k neighbors are under certain circumstances. Finally, determine whether the unknown sample contains the label, as confirmed by the Bayes decision theory [11].

Supposing that Y represents the binary label vector and X_0^i indicates the training sample that does not contain label i, X_1^i indicates the training sample that does contain label i. C_m^i represents the number of k neighbors to unknown sample m for each label i. Thus, we obtain prior probability $P(X_0^i)$ and $P(X_1^i)$. For sample set t, conditional probability $P(E_j^i / X_0^i)$ and $P(E_j^i / X_1^i)$, E_j^i represent j samples which contain label i within sample set t.

According to the principle of maximum posterior probability:

$$y_m(i) = \arg\max P(X_b^i / E_{C_m^i}^i)(b = 0 or 1) \tag{8}$$

Finally, the Bayes equation determines the final result:

$$\begin{aligned} y_m(i) &= \arg\max \frac{P(X_b^i)P(E_{C_m^i}^i / X_b^i)}{P(E_{C_m^i}^i)} \\ &= \arg\max P(X_b^i)P(E_{C_m^i}^i / X_b^i) \end{aligned} \tag{9}$$

5. Evaluation and conclusion

The ML-kNN algorithm is a typical multi-label learning algorithm [12]. In the multi-label learning system, one sample typically has multiple labels. The evaluating metrics, such as precision and recall, as used in single-label learning system cannot reflect a multi-label learning system. The most common measures used to estimate the performance of an algorithm in multi-label learning are Hamming Loss, One Error, Coverage, Ranking Loss and Average Precision methods.

Given multi-label classifier y (.) and the test set $N = \{(x_i, Y_i) | 1 < i < m\}$, Y_i represents the related label set of x_i. There are m samples and q labels in the training set.

1). Hamming Loss

$$Ham = \frac{1}{m}\sum_{i=1}^{t}\frac{1}{q}|y(x_i)\Delta Y_i| \tag{10}$$

Δ represents the difference between two symmetric sets. This index which judges the misclassification of individual labels prefers smaller values.

2). One Error

$$One-E=\frac{1}{m}\sum_{i=1}^{m}\left[\left[\arg\max f(x_i,y)\right]\notin Y_i\right] \tag{11}$$

f (x_i, y) represents the real valued function, which tests the condition that the label with the highest degree of membership value is not included in the label set; smaller values are preferred in the system.

3). Coverage

$$Cov=\frac{1}{m}\sum_{i=1}^{m}\max rankf(x_i,y)-1 \tag{12}$$

The rank f (x_i, y) represents a sorting function. This parameter evaluates the depth of search if all related labels are covered in the collating sequence; smaller values are more beneficial.

4). Ranking Loss

$$Rank=\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\left|Y_i\right|\left|\overline{Y_i}\right|}\left|\left\{(y',y'')\middle|f(x_i,y')\le f(x_i,y''),(y',y'')\in Y_i\times\overline{Y_i}\right\}\right| \tag{13}$$

$\overline{Y_i}$ is the complementary set of Y_i. This value reflects the condition of an incorrect sort in the sort sequence; smaller value proves the superior performance of the system.

5) Average Precision

$$Ave=\frac{1}{m}\sum_{i=1}^{m}\frac{1}{Y_i}\sum\frac{\left|\left\{y'\middle|rankf(x_i,y')\le rankf(x_i,y)y'\in Y_i\right\}\right|}{rankf(x_i,y)} \tag{14}$$

For the k nearest neighbor algorithm, different k values can obtain different results, as displayed in Table 2.

As shown in Table 2, best results are obtained when k is equal to 2, when the prediction precision reaches 66.7%.

This paper emphasizes a particular feature combination [13], allowing the entropy density to predict multisite protein subcellular localization and holding great promise to improve the precision of prediction. The prediction precision of the proposed method is shown in Table 3.

In the prediction process, feature extraction and the classification algorithm are extremely important components. For feature extraction, features should fully express protein information [14]. In addition,

Table 2

Results according to various k values

k	Ham	One-E	Cov	Rank	Ave
1	0.162	0.311	0.498	0.165	0.812
2	0.168	0.282	**0.386**	**0.119**	**0.854**
3	**0.143**	0.285	0.415	0.132	0.841
4	0.156	**0.273**	0.423	0.138	0.845
5	0.157	0.287	0.427	0.141	0.833

Table 3

Prediction precision of various methods

Feature Extraction	PseAAC	PseAAC+PC	PseAAC+PC+ED
Precision	47.99%	62.33%	66.73%

choosing an appropriate classification algorithm is integral in order to achieve higher prediction precision.

Acknowledgments

This research was partially supported by the Program for Scientific research innovation team in colleges and universities of Shandong Province 2012-2015, the Key Project of Natural Science Foundation of Shandong Province (ZR2011FZ001), the Natural Science Foundation of Shandong Province (ZR2011FL022 and ZR2013FL002), the Key Subject Research Foundation of Shandong Province and the Shandong Provincial Key Laboratory of Network Based Intelligent Computing. This work was also supported by the National Natural Science Foundation of China (Grant Nos. 61302128, 61201428 and 61203105), and the scientific research foundation of University of Jinan (xky1109).

References

[1] X. Xiao and Z.C. Wu and K.C. Chou, iLoc-Virus: A multi-label learning classifier for identifying the sub-cellular localization of virus proteins with both single and multiple sites, Journal of Theoretical Biology **3** (2011), 42-51.
[2] S.P. Qiao and B.Q. Yan, Review of protein sub-cellular localization prediction, Application Research of Computers **2** (2014), 321-327.
[3] K.C. Chou and H.B. Shen, Recent progress in protein subcellular location prediction, Analytical Biochemistry **2** (2007), 1-6.
[4] P.F. Du and C Xu, Predicting multisite protein sub-cellular locations: progress and challenges, Expert Review of Proteomics **8** (2013), 227-237.
[5] H.B. Shen and K.C. Chou, Virus-mPLoc: A fusion classifier for viral protein sub-cellular location prediction by incorporating multiple sites, Molecular BioSystems **11** (2010), 175-186.
[6] L.Q. Li, S.J. Yu and W.D. Xiao, Prediction of bacterial protein sub-cellular localization by incorporating various features into Chou's PseAAC and a backward feature selection approach, Biochimie **3** (2012), 100-107.
[7] S.Y. Mei, Multi-kernel transfer learning based on Chou's PseAAC formulation for protein submito chondria localization, Journal of Theoretical Biology **3** (2012), 121-130.
[8] G.L. Fan and Q.Z. Li, Predict mycobacterial proteins sub-cellular locations by incorporating pseudo-average chemical shift into the general form of Chou's pseudo amino acid composition, Journal of Theoretical Biology **9** (2012), 88-95.
[9] C. Huang and J.Q. Yuan, Predicting protein subchloroplast locations with both single and multiple sites via three different modes of Chou's pseudo amino acid compositions, Journal of Theoretical Biology **2** (2013), 205-212.
[10] Q.M. Hu, The research on protein sequence feature extraction and its application on protein subcellular location, Hunan University, 2010.
[11] S. Zhang and H.X. Zhang, Modified KNN algorithm for multi-label learning, Application Research of Computers **2** (2011), 4445-4450.
[12] Z.X. Li, Y.Q. Zhou, C.L. Zhang and S.M. Zhou, Survey on multi-label learning, Application Research of Computers **4** (2014), 1601-1605.
[13] P. Zakeri, B. Moshiri and M. Sadeghi, Prediction of protein submitochondria locations based on data fusion of various features of sequences, Journal of Theoretical Biology **9** (2011), 208-216.
[14] K.C. Chou, Some remarks on predicting multi-label attributes in molecular biosystems, Molecular Biosystems **10** (2013), 1092-1100.