Amino Acids (2008) 35: 321–327
DOI 10.1007/s00726-007-0623-z
Printed in The Netherlands

Amino Acids

Using pseudo amino acid composition to predict protein subcellular location: approached with amino acid composition distribution

J.-Y. Shi[1], **S.-W. Zhang**[2], **Q. Pan**[2], and **G.-P. Zhou**[3]

[1] School of Computer Science and Engneering, Northwestern Polytechnical University, Xi'an, China
[2] School of Automation, Northwestern Polytechnical University, Xi'an, China
[3] Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA, USA

Received June 25, 2007
Accepted October 15, 2007
Published online January 22, 2008;

Summary. In the Post Genome Age, there is an urgent need to develop the reliable and effective computational methods to predict the subcellular localization for the explosion of newly found proteins. Here, a novel method of pseudo amino acid (PseAA) composition, the so-called ''amino acid composition distribution'' (AACD), is introduced. First, a protein sequence is divided equally into multiple segments. Then, amino acid composition of each segment is calculated in series. After that, each protein sequence can be represented by a feature vector. Finally, the feature vectors of all sequences thus obtained are further input into the multi-class support vector machines to predict the subcellular localization. The results show that AACD is quite effective in representing protein sequences for the purpose of predicting protein subcellular localization.

Keywords: Protein subcellular localization – Amino acid composition distribution – Pseudo amino acid composition – Support vector machines

Abbreviations: AAC, amino acid composition; AACD, amino acid composition distribution; 5CV, 5-fold cross validation; DAG, directed acyclic graph; DPC, dipeptide composition; KNN, k-nearest neighbor; OVO, one-versus-one; OVR, one-versus-rest; PPC, polypeptide composition; PseAA, pseudo amino acid composition; RBF, radial basis function; SVM, support vector machines

1. Introduction

As one of the most important areas in post-genome era, proteome aims to understand proteins' potential roles, elucidate their interaction in a cellular context, and further make the corresponding functional annotation. Determination of subcellular location of proteins is very essential for annotating their biological functions. However, the traditional way to determine the localization of a protein in a cell is by biochemical experiments, which are hard to meet the demands due to both time-consuming and expensive. Therefore, to bridge this gap, there is a need to develop more effective prediction methods.

During the last decade, many theoretical and computational methods were developed in an attempt to predict subcellular localization of protein. These methods can be broadly classified into four types (Guda and Subramaniam, 2005): (1) Methods based on the sorting signals which rely on the presence of protein targeting or signal peptides (Nakai and Horton, 1999). (2) Methods based on lexical analysis of keywords (LOCkey) from the functional annotation of proteins (Nair and Rost, 2002). (3) Methods based on the uses of phylogenetic profiles (Marcotte et al., 2000), domain projection (Mott et al., 2002) or a combination of evolutionary and structural information. (4) Methods based on the differences in the amino acid composition or amino acid properties of proteins from different subcellular locations (Nakashima and Nishikawa, 1994; Chou, 2001; Park and Kanehisa, 2003; Cui et al., 2004; Huang and Li, 2004; Guda and Subramaniam, 2005; Chou and Shen, 2006a–c; 2007a, b; Höglund et al., 2006; Shen and Chou, 2007a, b, e; Shen et al., 2007). In this paper, our interest is focused on the researches about the last type.

Because this kind of prediction methods always involves the theories and methods of statistical pattern recognition, the representation of protein sequence which is also referred to as feature, plays naturally the key role in the prediction of subcellular location.

Originally, Nakashima and Nishikawa represented protein sequence with amino acid composition (AAC) and indicated that intracellular and extracellular proteins are significantly different in this representation (Nakashima and Nishikawa, 1994). The subsequent studies showed that AAC is closely related to protein subcellular localizations (Chou and Elrod, 1999; Hua and Sun, 2001). Although AAC can represent the major information of protein sequence, it always ignores the sequence-order and structure information of protein. Hence, two sequences with different functions and localizations but similar AAC, may be predicted in the same localization. To represent protein sequence better, some improved representations have been proposed and can be classified into the following two categories: One focuses on the combination of AAC and with physicochemical properties of amino acids (Chou, 2001, 2005; Chou and Cai, 2002, 2005; Pan et al., 2003; Gao et al., 2005b; Shi et al., 2007). The other one makes the direct extension of AAC (Park and Kanehisa, 2003; Bhasin and Raghava, 2004; Cui et al., 2004; Shi et al., 2006). Among these methods, the pseudo amino acid (PseAA) composition (Chou, 2001) is a popular method. According to its original definition (Chou, 2001), the PseAA composition of a given protein sample is represented by a set of greater than 20 discrete factors, where the first 20 factors represent the components of its conventional AA (amino acid) composition while the additional factors incorporate some of its sequence-order information via various modes. Since the concept of Chou's PseAA composition was introduced, it has been widely used to improve the prediction quality for various different protein attributes. Meanwhile, various kinds of PseAA composition have been proposed (Pan et al., 2003; Gao et al., 2005b; Xiao et al., 2005a, b; 2006a, b; Mondal et al., 2006; Mundra et al., 2007; Shi et al., 2007; Xiao and Chou, 2007; Zhang and Ding, 2007; Zhou et al., 2007). Owing to the wide usage of PseAA composition, recently a web-server called PseAA was established at http://chou.med.harvard.edu/bioinf/PseAA/ (Shen and Chou, 2007c), by which users can generate various kinds of PseAA composition for a given protein sequence to best fit their need. However, the current PseAA server cannot cover the specific PseAA composition as will be formulated in this paper.

Here, we introduce a novel representation of protein sequence, amino acid composition distribution (AACD), to predict subcellular location by a direct extension of AAC. Then, multi-class SVM is applied to validate the effectiveness of AACD with several datasets of protein subcellular localization.

2. Materials and methods

2.1 Datasets

In this paper, we use several datasets from 6 papers which are presented by Chou (2001), Park and Kanehisa (2003), Cui et al. (2004), Huang and Li (2004), Höglund et al. (2006), Guda and Subramaniam (2005), respectively. These datasets vary with the version of SWISS-PROT, the similarity between sequences, the type of locations, the count of subcellular localization and the total number of samples. Remarkably, Chou's datasets are composed of a training set and a testing set, and the latter one is only for the independent test (Chou, 2001); Huang and Li built two datasets with the sequence similarities which are lower than 50 and 80% (Huang and Li, 2004), respectively; Höglund et al. (2006) constructed three datasets in which many locations and sequences overlap (Höglund et al., 2006), therefore only the plant datasets is used here.

Consequently, the eight datasets were used in these papers, which are shortly denoted by CH01-T, CH01-E, PK03, CU04, HL04-1, HL04-2, HO06-P and GS05, and listed in Table 1.

2.2 Representation methods of protein sequence

Without loss of generality, we assume that there are N protein sequences in the dataset, let L_k be the length of the k-th sequence p_k, and α_i be the i-th element of 20 natural amino acids represented by English letters A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, and Y, respectively.

2.2.1 Amino acid composition

According to amino acid composition (AAC), the protein sequence p_k can be represented as a 20-D feature vector:

$$\mathrm{AAC}_k = [c_1^k, \ldots, c_i^k, \ldots, c_{20}^k], \quad k = 1, \ldots, N \tag{1}$$

where $c_i^k = t_i^k / L_k$ is the normalized occurrence frequency of amino acid α_i, and t_i^k is the count of α_i appearing in sequence p_k.

However, it is not sufficient to represent a specific protein sequence only based on AAC. For example, suppose we have two protein sequences p_1: AAADDD and p_2: DDAA. According to AAC, the feature vectors of p_1 and p_2 are represented as follows:

$$\mathrm{AAC}_1 = [0.5, 0, 0.5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]$$
$$AAC_2 = [0.5, 0, 0.5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]$$

Obviously, AAC_1 and AAC_2 are identical so that we cannot distinguish protein p_1 from protein p_2 only based on AAC. Consequently, there is a need to develop other representations of protein sequence to treat such case.

2.2.2 Polypeptide composition

Considering the sequence order information, polypeptide composition (PPC) is an improved representation which is used to calculate the nor-

Table 1. The datasets used in this paper

Datasets	Version	Similarity (%)	Location	Sample number
CH01-T	35.0	≤80	12	2191
CH01-E	35.0	≤80	11	2494
PK03	39.0	≤80	12	7589
CU04	41.0	≤90	8	8305
HL04-1	41.0	≤50	11	3572
HL04-2	41.0	≤80	11	7203
HO06-P	42.0	≤80	10	5856
GS05	45.0	≤95	9	18317

malized occurrence frequency of n residues instead of single residue in a sequence. So the protein sequence p_k can be represented as:

$$\mathrm{PPC}_k = \left\{ \begin{array}{ccc} c^k(A\overbrace{A\cdots A}^{n-1}) & \cdots & c^k(A\overbrace{V\cdots V}^{n-1}) \\ \cdots & c^k(\beta_1\beta_2\cdots\beta_j\cdots\beta_{n-1}\beta_n) & \cdots \\ c^k(V\overbrace{A\cdots A}^{n-1}) & \cdots & c^k(V\overbrace{V\cdots V}^{n-1}) \end{array} \right\}_{20\times 20^{n-1}} \quad (2)$$

where $c^k(\beta_1\beta_2\cdots\beta_j\cdots\beta_{n-1}\beta_n)$ is the normalized occurrence frequency of sub-sequence $\beta_1\beta_2\cdots\beta_j\cdots\beta_{n-1}\beta_n$ in sequence p_k, $\beta_j \in \{\alpha_i\}$, $j = 1,\ldots,n$, and $k = 1,\ldots,N$.

Because of the huge dimension of PPC, dipeptide composition (DPC) is always used to represent protein sequence (Bhasin and Raghava, 2004), and it is also called as amino acid pair composition (Park and Kanehisa, 2003) for some cases. But DPC still has 400 dimensions and is defined as the following:

$$\mathrm{DPC}_k = \left\{ \begin{array}{ccc} c^k(AA) & \cdots & c^k(AV) \\ \cdots & c^k(\beta_1\beta_2) & \cdots \\ c^k(VA) & \cdots & c^k(VV) \end{array} \right\}_{20\times 20^{n-1}} \quad (3)$$

2.2.3 Amino acid composition distribution

A protein sequence p_k can be equally divided into multiple segments, and then calculate AAC of each segment in series. So the sequence p_k can be represented as the following formula:

$$\mathrm{AACD}_n^k = \left\{ \begin{array}{ccccc} c_{1,1}^k & \cdots & c_{1,m}^k & \cdots & c_{1,n}^k \\ \cdots & \cdots & c_{i,m}^k & \cdots & \cdots \\ c_{20,1}^k & \cdots & c_{20,m}^k & \cdots & c_{20,n}^k \end{array} \right\}_{20\times n} \quad (4)$$

where n is the count of segments, $[c_{1,m}^k,\ldots,c_{i,m}^k,\ldots c_{20,m}^k]^T$ is the AAC of the m-th segment of p_k, and $c_{i,m}^k$ is defined as:

$$c_{i,m}^k = n\cdot t_{i,m}^k/L_k, \quad m = 1,\ldots,n, \quad i = 1,\ldots,20 \quad (5)$$

where $t_{i,m}^k$ is the count of α_i appearing in the m-th segment of sequence p_k.

We define this representation as amino acid composition distribution (AACD), which has lower dimension than PPC generally.

2.3 Support vector machines

Once the representation of protein sequence is fixed, the next step is to choose a classifier to predict the protein subcellular localization. Many types of classifiers, such as neural network, covariant discriminant algorithm (Chou, 2001), fuzzy KNN (Huang and Li, 2004) and support vector machines (SVM) (Chou and Cai, 2002; Park and Kanehisa, 2003; Cui et al., 2004; Höglund et al., 2006) have been applied to prediction of protein subcellular location. In these classifiers, SVM has been more broadly applied to such prediction due to its good performance of classification.

SVM was originally designed for binary classification (Vapnik, 1998), while prediction of subcellular localization is M-class classification. Usually, we can construct M-class SVMs to solve such problem based on the binary class SVM. There are mainly two kinds of approaches for multi-class SVM. The one can processes directly all data in one optimization formulation (Crammer and Singer, 2001). The other one decomposes multi-class into a series of binary SVMs, including "one-versus-rest" (OVR) (Vapnik, 1998), "one-versus-one" (OVO) (Kreßel, 1999), and "directed acyclic graph" (DAG) (Platt et al., 2000). The extensive experiments have shown that OVR, OVO and DAG are more practical (Hsu and Lin, 2002; Rifin and Klautau, 2004; Shi et al., 2006, 2007).

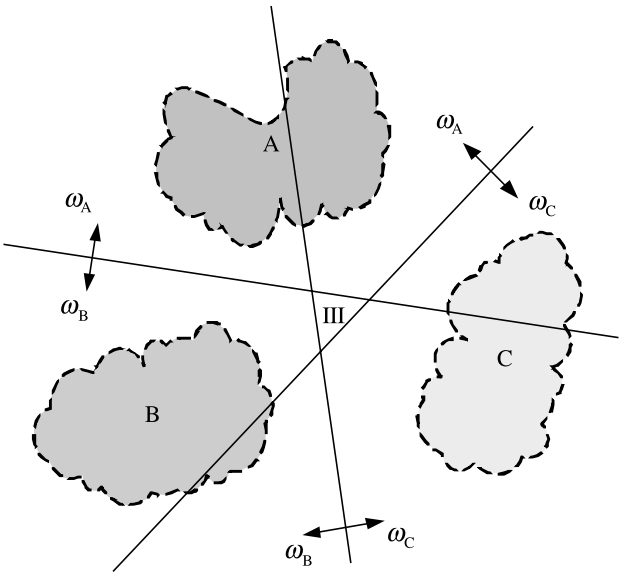

Fig. 1. The binary SVM decomposition of 3-class problem with OVO approach, where label ω_A, ω_B and ω_C denote class A, B, and C, respectively

Because of OVO approach convenient usage, it is used in this paper. For an M-class problem, OVO constructs $M(M-1)/2$ binary SVMs. During the evaluation, each of the $M(M-1)/2$ SVMs casts one vote for its most favored class, and finally the class with the most votes wins. Figure 1 shows that OVO decomposes 3-class problem into a series of binary SVMs.

The SVM software, LibSVM, is used in this paper, which can be freely downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm/ for academic research (Hsu and Lin, 2002). In addition, we do the training only with the RBF kernel in all experiments.

2.4 Prediction assessment

Among the independent dataset test, sub-sampling (e.g., 5 or 10-fold sub-sampling) test, and jackknife test, which are often used for examining the accuracy of a statistical prediction method, the jackknife test is deemed the most rigorous and objective (Chou and Zhang, 1995) as demonstrated by a penetrating analysis in a recent review (Chou and Shen, 2007e) and has been increasingly adopted by investigators to test the power of various prediction methods (see, e.g., Zhou, 1998; Zhou and Assa-Munt, 2001; Du et al., 2003; Pan et al., 2003; Zhou and Doctor, 2003; Huang and Li, 2004; Wang et al., 2004, 2006; Gao et al., 2005a, b; Liu et al., 2005a, b, 2007; Shen and Chou, 2005a, b, 2006, 2007a, b, d, e; Wang et al., 2005; Cao et al., 2006; Chen et al., 2006a, b; Chou and Shen, 2006a, d, 2007a, c, d, f; Du and Li, 2006; Du et al., 2006; Gao and Wang, 2006; Guo et al., 2006a, b; Kedarisetti et al., 2006; Mondal et al., 2006; Niu et al., 2006; Shen et al., 2006, 2007; Sun and Huang, 2006; Wen et al., 2006; Xiao et al., 2006a, b; Zhang et al., 2006a, b, c; Chen and Li, 2007; Chen et al., 2007; Ding et al., 2007; Jahandideh et al., 2007; Lin and Li, 2007a, b; Shi et al., 2007). However, since jackknife test will take a lot of computational time for SVM to complete, as a compromise, in this study we adopted the 5-fold cross-validation to examine the performance of our approach.

According to 5CV procedure, the dataset is randomly split into 5 equal subsets. In turn, we take each subset as the testing set to evaluate the prediction, and use the rest subsets to build classification model, in other words, to do the training. The average and the standard deviation of the

accuracies of all evaluations can indicate the performance of prediction, and are defined, respectively as:

$$\overline{Q} = \sum_{i=1}^{k} Q_i/k, \quad i = 1, \ldots, k \tag{6}$$

$$S = \sqrt{\sum_{i=1}^{k} (Q_i - \overline{Q})^2/(k-1)}, \quad k = 5 \tag{7}$$

where Q_i is the accuracy of the i-th evaluation and k is the count of cross-validation.

3. Results and discussion

3.1 Representations with different segments

In order to find the impact of the numbers of split segments on the classification performance, classifiers are performed on all selected datasets with various and continuous count to split protein sequences, respectively. The

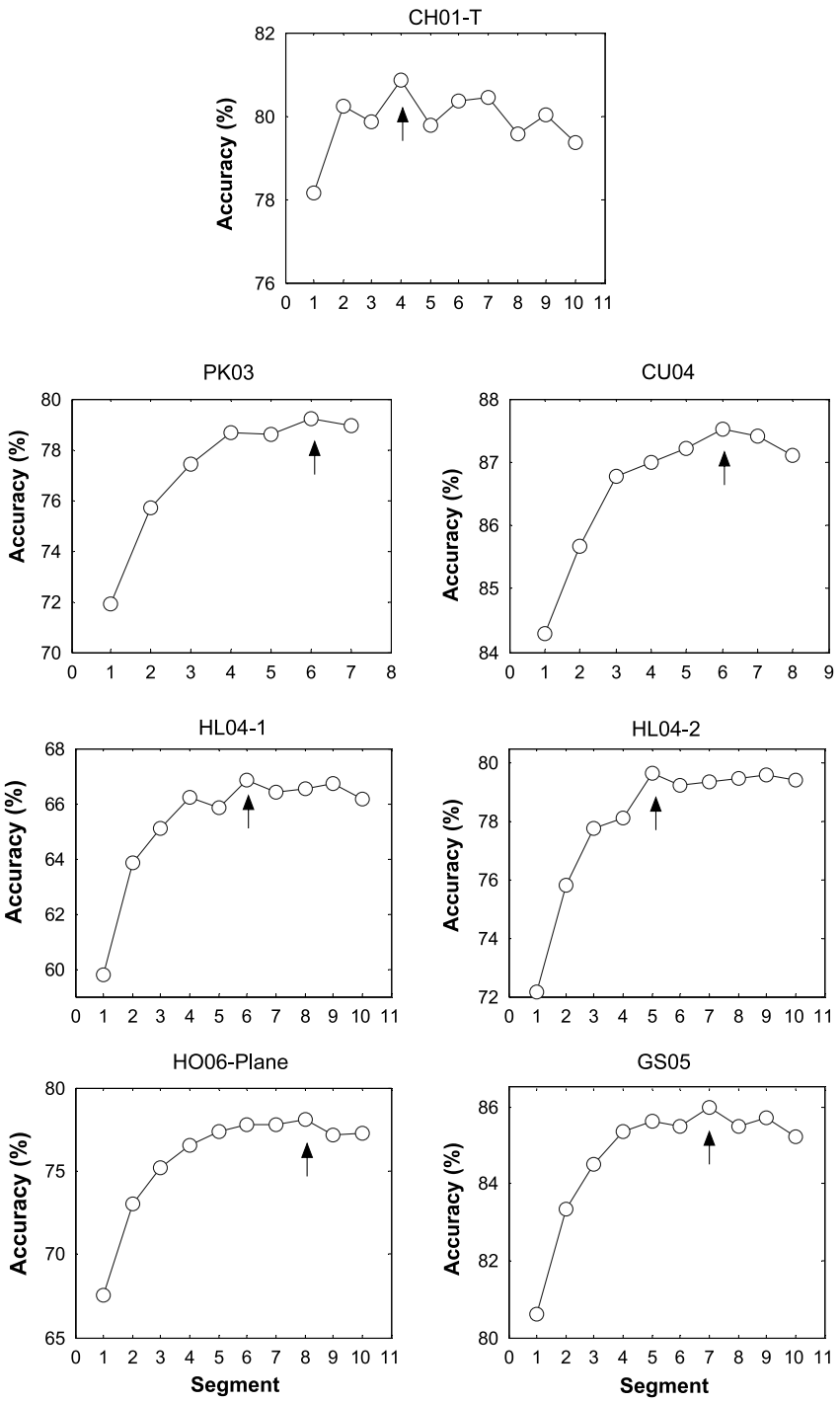

Fig. 2. The accuracies of different segments of protein sequence

up arrows in Fig. 2 are marked as the best predicted accuracies. From Fig. 2, we can see that the count of split segments has an important impact on the prediction. Firstly, for all datasets, the best accuracies achieved by AACD are remarkably higher than that of AAC (the first data point in all subfigures). This indicates that AACD can represent protein sequence better than ACC. Secondly, along with the increasing of the count of split segment, the accuracies increase at the beginning, and then decrease slowly. Hence, the optimal split corresponding to the best prediction exists for a specific dataset. The properties of dataset such as its version, the similarity between sequences, the type of locations and the count of samples, always affect the optimal splits.

3.2 Comparison with other sequence representations

In order to assess the quality of AACD representation, the results of AAC, DPC and AACD are listed in Table 2.

Table 2 shows that AACD achieves the best accuracies and always has the lower standard deviations for all datasets. It means that AACD is more effective to represent protein sequence and is better robust than both AAC and DPC for the prediction of subcellular localization.

In addition, it is worth to note that the accuracy and standard deviation of AACD are 66.84% and 0.52, respectively, which are about 7% higher than that of AAC and DPC for dataset HL04-1(sequence similarity $\leq$50%). DPC gets the lowest accuracy and has the biggest standard deviations. The results show that AACD maybe is non-sensitive to sequence similarity.

Table 2. The results (in percentage) of AAC, DPC, and AACD in 5CV test

Datasets	AAC	DPC	AACD
CH01-T	78.14 ± 0.72	76.40 ± 0.75	80.87 ± 0.84
PK03	71.91 ± 0.66	74.64 ± 0.55	79.19 ± 0.54
CU04	84.30 ± 0.56	86.48 ± 1.02	87.51 ± 0.68
HL04-1	59.80 ± 0.38	58.23 ± 0.83	66.84 ± 0.52
HL04-2	72.17 ± 0.75	72.62 ± 0.84	79.66 ± 0.59
HO06-P	67.54 ± 1.07	71.23 ± 1.44	78.09 ± 1.01
GS05	80.59 ± 0.57	83.89 ± 0.69	85.99 ± 0.81

Table 3. The results (in percentage) of the comparison with other methods

Datasets	Assessment	Original result	Our result
CH01-T	Jackknife	73.03	80.87 ± 0.84
PK03	5CV	78.2 ± 0.9	79.19 ± 0.54
CU04	Jackknife	87	87.51 ± 0.68
HL04-1	Jackknife	58.1	66.84 ± 0.52
HL04-2	Jackknife	80.1	79.66 ± 0.59
HO06-P	5CV	74.60 ± 0.80	78.09 ± 1.01
GS05	unknown	79.43	85.99 ± 0.81

Table 4. The results (in percentage) of the comparison with other PseAA methods

Datasets	Assessment	Chou (2001)	Pan et al. (2003)	Xiao et al. (2005a)	Our method
CH01-T	Jackknife	73.03	67.68	73.57	82.15
CH01-E	independent	80.87	73.86	79.79	85.49

3.3 Comparison with other methods

In order to further validate the presented methods, the results of several methods are listed in Table 3.

Table 3 shows that AACD obtains almost the best prediction results for the seven selected datasets. Although some prediction methods used jackknife test to assess their results, 5-CV and jackknife test will achieve similar estimation if the number of samples (protein sequences) is not too much less (Jain et al., 2000).

3.4 Comparison with other PseAA methods

We choose datasets CH01 to compare the performance of several PseAA methods (Chou, 2001; Pan et al., 2003; Xiao et al., 2005a) in jackknife and independent tests respectively. The results are listed in Table 4. Table 4 shows that the corresponding accuracies achieved by AACD are the highest in both jackknife and independent tests.

4. Conclusion

In this paper, we have developed a novel method of pseudo amino acid composition, amino acid composition distribution (AACD). Instead of calculating AAC of the whole sequence, the presented method is used to calculate AACs of multiple segments which are derived from the equal interval split of the whole sequence.

With this AACD representation, Multi-class SVMs are applied to predict the subcellular localization. Compared with other methods, the results show that AACD is an effective representation of protein sequence and non-sensitive to sequence similarity because of the better ability to reflect the information of protein subcellular localization.

Acknowledgements

This paper was supported in part by the National Natural Science Foundation of China (No. 60775012 and 60634030) and the Technological Innovation Foundation of Northwestern Polytechnical University (No. KC02).

References

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

Cao Y, Liu S, Zhang L, Qin J, Wang J, Tang K (2006) Prediction of protein structural class with Rough Sets. BMC Bioinformatics 7: doi:10.1186/1471-2105-7-20

Chen C, Tian YX, Zou XY, Cai PX, Mo JY (2006a) Using pseudo-amino acid composition and support vector machine to predict protein structural class. J Theor Biol 243: 444–448

Chen C, Zhou X, Tian Y, Zou X, Cai P (2006b) Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network. Anal Biochem 357: 116–121

Chen J, Liu H, Yang J, Chou KC (2007) Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino Acids 33: 423–428

Chen YL, Li QZ (2007) Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo amino acid composition. J Theor Biol doi: 10.1016/j.jtbi.2007.05.019

Chou KC (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins Struct Funct Genet 43: 246–255

Chou KC (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21: 10–19

Chou KC, Cai YD (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. J Biol Chem 277: 45765–45769

Chou KC, Cai YD (2005) Prediction of membrane protein types by incorporating amphipathic effects. J Chem Inf Model 45: 407–413

Chou KC, Elrod D (1999) Protein subcellular localization prediction. Protein Eng 12: 107–118

Chou KC, Shen HB (2006a) Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. Biochem Biophys Res Commun 347: 150–157

Chou KC, Shen HB (2006b) Large-scale predictions of Gram-negative bacterial protein subcellular locations. J Proteome Res 5: 3420–3428

Chou KC, Shen HB (2006c) Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. J Proteome Res 5: 1888–1897

Chou KC, Shen HB (2006d) Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. J Proteome Res 5: 1888–1897

Chou KC, Shen HB (2007a) Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. J Proteome Res 6: 1728–1734

Chou KC, Shen HB (2007b) Large-scale plant protein subcellular location prediction. J Cell Biochem 100: 665–678

Chou KC, Shen HB (2007c) Large-scale plant protein subcellular location prediction. J Cell Biochem 100: 665–678

Chou KC, Shen HB (2007d) MemType-2L: a Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochem Biophys Res Commun 360: 339–345

Chou KC, Shen HB (2007e) Review: recent progresses in protein subcellular location prediction. Anal Biochem 370: 1–16

Chou KC, Shen HB (2007f) Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. Biochem Biophys Res Commun 357: 633–640

Chou KC, Zhang CT (1995) Prediction of protein structural classes. Crit Rev Biochem Mol Biol 30: 275–349

Crammer K, Singer Y (2001) On the algorithmic implementation of multiclass kernel-based vector machines. J Machine Learning Res 2: 265–292

Cui Q, Jiang T, Liu B, Ma S (2004) Esub8: A novel tool to predict protein subcellular localizations in eukaryotic organisms. BMC Bioinformatics 5: 66–72

Ding YS, Zhang TL, Chou KC (2007) Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network. Protein Peptide Lett 14: 811–815

Du P, Li Y (2006) Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence. BMC Bioinformatics 7: 518

Du QS, Jiang ZQ, He WZ, Li DP, Chou KC (2006) Amino acid principal component analysis (AAPCA) and its applications in protein structural class prediction. J Biomol Struct Dyn 23: 635–640

Du QS, Wei DQ, Chou KC (2003) Correlation of amino acids in proteins. Peptides 24: 1863–1869

Gao QB, Wang ZZ (2006) Classification of G-protein coupled receptors at four levels. Protein Eng Des Sel 19: 511–516

Gao QB, Wang ZZ, Yan C, Du YH (2005a) Prediction of protein subcellular location using a combined feature of sequence. FEBS Lett 579: 3444–3448

Gao Y, Shao SH, Xiao X, Ding YS, Huang YS, Huang ZD, Chou KC (2005b) Using pseudo amino acid composition to predict protein subcellular location: approached with Lyapunov index, Bessel function, and Chebyshev filter. Amino Acids 28: 373–376

Guda C, Subramaniam S (2005) pTARGET: a new method for predicting protein subcellular localization in eukaryotes. Bioinformatics 21: 3963–3969

Guo J, Lin Y, Liu X (2006a) GNBSL: a new integrative system to predict the subcellular location for Gram-negative bacteria proteins. Proteomics 6: 5099–5105

Guo YZ, Li M, Lu M, Wen Z, Wang K, Li G, Wu J (2006b) Classifying G protein-coupled receptors and nuclear receptors based on protein power spectrum from fast Fourier transform. Amino Acids 30: 397–402

Höglund A, Dönnes P, Blum T, Adolph H-W, Kohlbacher O (2006) MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. Bioinformatics 22: 1158–1165

Hsu C, Lin CJ (2002) A comparison of methods for multi-class support vector machines. IEEE T Neural Networ 13: 415–425

Hua SJ, Sun ZR (2001) Support vector machine approach for protein subcellular localization prediction. Bioinformatics 17: 721–728

Huang Y, Li YD (2004) Prediction of protein subcellular locations using fuzzy k-NN method. Bioinformatics 20: 21–28

Jahandideh S, Abdolmaleki P, Jahandideh M, Asadabadi EB (2007) Novel two-stage hybrid neural discriminant model for predicting proteins structural classes. Biophys Chem 128: 87–93

Jain AK, Duin RPW, Mao J (2000) Statistical pattern recognition: a review. IEEE T Pattern Anal 22: 4–37

Kedarisetti KD, Kurgan LA, Dick S (2006) Classifier ensembles for protein structural class prediction with varying homology. Biochem Biophys Res Commun 348: 981–988

Kreßel UH (1999) Pairwise classification and support vector machines. In: Schölkopf B, Burges CJ, Smola AJ (eds) Advances in Kernel methods: support vector learning. MIT Press, Cambridge, MA

Lin H, Li QZ (2007a) Predicting conotoxin superfamily and family by using pseudo amino acid composition and modified Mahalanobis discriminant. Biochem Biophys Res Commun 354: 548–551

Lin H, Li QZ (2007b) Using pseudo amino acid composition to predict protein structural class: approached by incorporating 400 dipeptide components. J Comput Chem 28: 1463–1466

Liu DQ, Liu H, Shen HB, Yang J, Chou KC (2007) Predicting secretory protein signal sequence cleavage sites by fusing the marks of global alignments. Amino Acids 32: 493–496

Liu H, Wang M, Chou KC (2005a) Low-frequency Fourier spectrum for predicting membrane protein types. Biochem Biophys Res Commun 336: 737–739

Liu H, Yang J, Wang M, Xue L, Chou KC (2005b) Using Fourier spectrum analysis and pseudo amino acid composition for prediction of membrane protein types. Protein J 24: 385–389

Marcotte EM, Xenarios I, van Der Bliek A, Eisenberg D (2000) Localizing proteins in the cell from their phylogenetic profiles. Proc Natl Acad Sci USA 97: 12115–12120

Mondal S, Bhavna R, Mohan Babu R, Ramakumar S (2006) Pseudo amino acid composition and multi-class support vector machines approach for conotoxin superfamily classification. J Theor Biol 243: 252–260

Mundra P, Kumar M, Kumar KK, Jayaraman VK, Kulkarni BD (2007) Using pseudo amino acid composition to predict protein subnuclear localization: approached with PSSM. Pattern Recogn Lett 28: 1610–1615

Mott R, Schultz J, Bork P, Ponting CP (2002) Predicting protein cellular localization using a domain projection method. Genome Res 12: 1168–1174

Nair R, Rost B (2002) Inferring sub-cellular localization through automated lexical analysis. Bioinformatics 18: S78–S86

Nakai K, Horton P (1999) PSORT: a program for detecting the sorting signals of proteins and predicting their subcellular localization. Trends Biochem Sci 24: 34–36

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

Niu B, Cai YD, Lu WC, Zheng GY, Chou KC (2006) Predicting protein structural class with AdaBoost learner. Protein Peptide Lett 13: 489–492

Pan YX, Zhang ZZ, Guo ZM, Feng GY, Huang Z, He L (2003) Application of pseudo amino acid composition for predicting protein subcellular location: Stochastic signal processing approach. J Protein Chem 22: 395–402

Park KJ, Kanehisa M (2003) Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. Bioinformatics 19: 1656–1663

Platt J, Cristianini N, Shawe-Taylor J (2000) Large margin DAGs for multiclass classification. In: Solla SA, Leen TK, Müller KR (eds) Adv Neural Inform Proc Syst 12: 547–555

Rifin R, Klautau A (2004) In defense of one-vs-all classification. J Machine Learn Res 5: 101–141

Shen HB, Chou KC (2005a) Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition. Biochem Biophys Res Commun 337: 752–756

Shen HB, Chou KC (2005b) Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo amino acid composition to predict membrane protein types. Biochem Biophys Res Comm 334: 288–292

Shen HB, Chou KC (2006) Ensemble classifier for protein fold pattern recognition. Bioinformatics 22: 1717–1722

Shen HB, Chou KC (2007a) Gpos-PLoc: an ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins. Protein Eng Design and Selection 20: 39–46

Shen HB, Chou KC (2007b) Hum-mPLoc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. Biochem Biophys Res Commun 355: 1006–1011

Shen HB, Chou KC (2007c) PseAC: a flexible web-server for generating various kinds of protein pseudo amino acid composition. Anal Biochem doi: 10.10.1016/j.ab.2007.10.012

Shen HB, Chou KC (2007d) Using ensemble classifier to identify membrane protein types. Amino Acids 32: 483–488

Shen HB, Chou KC (2007e) Virus-PLoc: a fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells. Biopolymers 85: 233–240

Shen HB, Yang J, Chou KC (2006) Fuzzy KNN for predicting membrane protein types from pseudo amino acid composition. J Theor Biol 240: 9–13

Shen HB, Yang J, Chou KC (2007) Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction. Amino Acids 33: 57–67

Shi JY, Zhang SW, Liang Y, Pan Q (2006) Prediction of protein subcellular localizations using moment descriptors and support vector machine. In: Ragapakse JC, Wong L, Acharya R (eds) PRIB, Hong Kong, China. Springer, Berlin, Heidelberg

Shi JY, Zhang SW, Pan Q, Cheng YM, Xie J (2007) SVM-based method for subcellular localization of protein using multi-scale energy and pseudo amino acid composition. Amino Acids 33: 69–74

Sun XD, Huang RB (2006) Prediction of protein structural classes using support vector machines. Amino Acids 30: 469–475

Vapnik V (1998) Statistical learning theory. Wiley, New York

Wang M, Yang J, Chou KC (2005) Using string kernel to predict signal peptide cleavage site based on subsite coupling model. Amino Acids (Erratum, ibid. 2005, 29: 301) 28: 395–402

Wang M, Yang J, Liu GP, Xu ZJ, Chou KC (2004) Weighted-support vector machines for predicting membrane protein types based on pseudo amino acid composition. Protein Eng Des Sel 17: 509–516

Wang SQ, Yang J, Chou KC (2006) Using stacked generalization to predict membrane protein types based on pseudo amino acid composition. J Theor Biol 242: 941–946

Wen Z, Li M, Li Y, Guo Y, Wang K (2006) Delaunay triangulation with partial least squares projection to latent structures: a model for G-protein coupled receptors classification and fast structure recognition. Amino Acids 32: 277–283

Xiao X, Chou KC (2007) Digital coding of amino acids based on hydrophobic index. Protein Peptide Lett 14: doi: 0929-8665/07

Xiao X, Shao SH, Ding YS, Huang ZD, Huang Y, Chou KC (2005a) Using complexity measure factor to predict protein subcellular location. Amino Acids 28: 57–61

Xiao X, Shao S, Ding Y, Huang Z, Chen X, Chou KC (2005b) Using cellular automata to generate Image representation for biological sequences. Amino Acids 28: 29–35

Xiao X, Shao SH, Ding YS, Huang ZD, Chou KC (2006a) Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. Amino Acids 30: 49–54

Xiao X, Shao SH, Huang ZD, Chou KC (2006b) Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor. J Comput Chem 27: 478–482

Zhang SW, Pan Q, Zhang HC, Shao ZC, Shi JY (2006a) Prediction protein homo-oligomer types by pseudo amino acid composition: approached with an improved feature extraction and naive bayes feature fusion. Amino Acids 30: 461–468

Zhang T, Ding Y, Chou KC (2006b) Prediction of protein subcellular location using hydrophobic patterns of amino acid sequence. Comput Biol Chem 30: 367–371

Zhang TL, Ding YS (2007) Using pseudo amino acid composition and binary-tree support vector machines to predict protein structural classes. Amino Acids 10.1007/s00726-007-0496-1

Zhang ZH, Wang ZH, Zhang ZR, Wang YX (2006c) A novel method for apoptosis protein subcellular localization prediction combining encoding based on grouped weight and support vector machine. FEBS Lett 580: 6169–6174

Zhou GP (1998) An intriguing controversy over protein structural class prediction. J Protein Chem 17: 729–738

Zhou GP, Assa-Munt N (2001) Some insights into protein structural class prediction. Proteins 44: 57–59

Zhou GP, Doctor K (2003) Subcellular location prediction of apoptosis proteins. Proteins 50: 44–48

Zhou XB, Chen C, Li ZC, Zou XY (2007) Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. J Theor Biol 248: 546–551

Authors' address: Shao-Wu Zhang, College of Automation control, Northwestern Polytechnical University, 127 YouYi West Rd., Xi'an 710072, Shaanxi, China,
Fax: +86-29-88494352, E-mail: Zhangsw@nwpu.edu.cn