Amino Acids (2015) 47:1485–1493
DOI 10.1007/s00726-015-1974-5

ORIGINAL ARTICLE

Prediction and analysis of cell-penetrating peptides using pseudo-amino acid composition and random forest models

Lei Chen[1,2] · Chen Chu[3] · Tao Huang[4] · Xiangyin Kong[4] · Yu-Dong Cai[1]

Received: 19 February 2015 / Accepted: 27 March 2015 / Published online: 18 April 2015

Abstract Cell-penetrating peptides, a group of short peptides, can traverse cell membranes to enter cells and thus facilitate the uptake of various molecular cargoes. Thus, they have the potential to become powerful drug delivery systems. The correct identification of peptides as cell-penetrating or non-cell-penetrating would accelerate this application. In this study, we determined which features were important for a peptide to be cell-penetrating or non-cell-penetrating and built a predictive model based on the key features extracted from this analysis. The investigated peptides were retrieved from a previous study, and each was encoded as a numeric vector according to six properties of amino acids—amino acid frequency, codon diversity, electrostatic charge, molecular volume, polarity, and secondary structure—by the pseudo-amino acid composition method. Methods of minimum redundancy maximum relevance and incremental feature selection were then employed to analyze these features, and some were found to be key determinants of cell penetration. In parallel, an optimal random forest prediction model was built. We hope that our findings will provide new resources for the study of cell-penetrating peptides.

Keywords Cell-penetrating peptide · Pseudo-amino acid composition · Minimum redundancy maximum relevance · Incremental feature selection · Random forest

Introduction

Cell-penetrating peptides (CPPs) are a group of short peptides that are able to penetrate the cell membrane at low micromolar concentrations both in vivo and in vitro without using chiral receptors or causing significant membrane damage (Madani et al. 2011). Thus, CPPs are ideal candidates for development as a powerful drug delivery system (El-Andaloussi et al. 2005). Many biologically active compounds, such as doxorubicin and cytarabine, are often ineffective due to their inaccessibility to the target cell, thereby necessitating use of specific carriers. Like other drug delivery systems (e.g., liposomes and micelles), CPPs are

Handling Editor: F. Albericio.

L. Chen and C. Chu have contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s00726-015-1974-5) contains supplementary material, which is available to authorized users.

✉ Lei Chen
chen_lei1@163.com

✉ Yu-Dong Cai
cai_yud@126.com

Chen Chu
chuchen@sibcb.ac.cn

Tao Huang
tohuangtao@126.com

Xiangyin Kong
xykong@sibs.ac.cn

[1] College of Life Science, Shanghai University, Shanghai 200444, People's Republic of China

[2] College of Information Engineering, Shanghai Maritime University, Shanghai 201306, People's Republic of China

[3] Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, People's Republic of China

[4] Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, People's Republic of China

hypothesized to bind 'cargo' compounds either electrostatically or covalently, pulling the cargo through the cell membrane in the complex and leaving it at target sites within the cell, thereby greatly improving the effect of the drug (Heitz et al. 2009; Vives et al. 2008).

CPPs have been shown promise in this capacity since their discovery in the 1980s, and many studies have focused on their definition, categorization and mechanisms of cellular uptake (Holm et al. 2006; Madani et al. 2011; Mueller et al. 2008; Montrose et al. 2013; Lindberg et al. 2013; Al-Soraj et al. 2010). Traditionally, potential CPPs were examined and screened individually for cell-penetrating capacity in vitro in cell lines (Gao et al. 2011; Lee et al. 2012), which is labor-intensive and not suitable for high-throughput studies. Although plenty of efforts have been made to improve the prediction, selection and investigation of CPPs, the development of more time- and cost-efficient methods for screening and designing effective CPPs is desirable. With the advancement of computer-based algorithms and software, computational methods have greatly improved the efficiency of many aspects of biology and pharmacy, including protein structure simulation (Marks et al. 2012; Shah et al. 2003), drug design (Basak 2013; Chen et al. 2014b; Ou-Yang et al. 2012; Chen et al. 2014a), and the prediction of biologically active molecules (Li et al. 2014a; Wang et al. 2011).

Diverse computational prediction and design algorithms tailored to CPPs have been proposed by researchers in this field (Sanders et al. 2011; Gautam et al. 2013). Sanders et al. (2011) complied a benchmark dataset that includes 111 known CPPs and 34 known non-CPPs from published literature. They represented the biochemical properties of the peptide with 61 features and built a support vector machine (SVM)-based classifier whose true-positive rate, false-positive rate and total accuracy were 0.759, 0.768 and 0.7586, respectively. Gautam et al. (2013) also proposed an SVM-based model to predict CPPs using various features, including amino acid and dipeptide composition, binary pattern profile, and physicochemical properties. Their maximum accuracy obtained via analysis of an independent dataset was 81.31 %.

The prediction performance of previous methods has been relatively low. To improve prediction performance, we proposed a new method for predicting the cell-penetrating properties of CPPs based on different features and a different machine learning model. In our method, each investigated CPP or non-CPP was encoded by pseudo-amino acid composition (PseAAC) to obtain a 270-D (dimension) vector (i.e., each CPP or non-CPP was represented by 270 features). Several feature selection methods, including minimum redundancy maximum relevance (mRMR) (Peng et al. 2005) and incremental feature selection (IFS), were employed to analyze these features and to build an optimal prediction model. Our method outperformed previous methods in analysis of the same dataset (Sanders et al. 2011): the sensitivity (SN), specificity (SP), and prediction accuracy (ACC) of our method were 0.9550, 0.4412 and 0.8345, higher respective 0.759, 0.232 and 0.7586 reported by Sanders et al. (2011).

In addition to building the prediction model, we determined the key features predicted to result in the cell-penetrating or non-cell-penetrating character of a peptide. This feature is discussed at the end of the study. We hope that the findings presented in this study will provide a strong basis for the categorization and characterization of CPPs as well as broaden the biochemical and medical applications of CPPs.

Materials and methods

Materials

The investigated CPPs and non-CPPs were retrieved from published literature authored by Sanders et al. (Sanders et al. 2011), in which 111 known CPPs and 34 known non-CPPs were compiled from other published literature (Hällbrink et al. 2005; Hansen et al. 2008; Anaspec 2010). Detailed information regarding these 111 CPPs and 34 non-CPPs is available in Supplementary Material I.

Encoding method

To build an effective prediction method, the first step is to encode each sample into a numeric vector that encompasses its essential properties and can be processed by computers. For a protein sequence, the most classic encoding method is amino acid composition, which consists of 20 discrete numbers defined as the occurrence frequency of amino acid in the protein sequence. However, this encoding method does not consider the sequence information of the protein (i.e., the obtained numeric vector loses the sequence information of the original protein). To address this issue, Chou proposed a generalized encoding method, called PseAAC, in 2001 (Chou 2001), which partly takes into account the sequence information. Until now, PseAAC has been applied in the investigation of many protein and protein-related problems (Huang et al. 2014; Lin 2008; Chen et al. 2009; Kong et al. 2014; Zhou and Cai 2006; Hajisharifi et al. 2014; Ding et al. 2011; Nanni et al. 2012; Hayat and Khan 2010; Zou et al. 2011). In this study, we employed this method to encode each protein sequence and other selected key features for the identification of CPPs and non-CPPs. Here, we provided a brief description of this method, but readers can refer to Chou's paper (Chou 2001) for a more detailed description.

Let $A_1A_2\cdots A_L$ be a protein sequence consisting of L amino acid residues. Its sequence information can be measured by the following λ discrete correlation factors:

$$\theta_j = \frac{1}{L-j}\sum_{i=1}^{L-j}[F(A_{i+j}) - F(A_i)]^2 \quad j = 1, 2, \ldots, \lambda, \lambda < L \tag{1}$$

where $F(A_i)$ is a value of the amino acid A_i for a certain property (e.g., hydrophilicity) of amino acids, which can be computed by

$$F(A_i) = \frac{F_o(A_i) - \sum_{X\in AA}\frac{F_o(X)}{20}}{\sqrt{\frac{\sum_{X\in AA}\left[F_o(A_i) - \sum_{X\in AA}\frac{F_o(X)}{20}\right]^2}{20}}} \tag{2}$$

where AA is a set consisting of 20 types of amino acids. If the original values (i.e., $F_o(X)$) of all types of amino acids for a certain property of amino acids are defined in advance, then $F(X)$ for all $X \in AA$ can be computed by Eq. 2, thereby calculating λ discrete correlation factors by Eq. 1. According to the discrete correlation factors calculated by Eq. 1, some discrete numbers can be computed by

$$v_i = \begin{cases} \frac{f_i}{\sum_{k=1}^{20} f_k + \varpi\sum_{j=1}^{\lambda}\theta_j} & 1 \le i \le 20 \\ \frac{\varpi\theta_{i-20}}{\sum_{k=1}^{20} f_k + \varpi\sum_{j=1}^{\lambda}\theta_j} & 21 \le i \le 20+\lambda \end{cases} \tag{3}$$

where ϖ is the weight of the sequence information and f_k ($1 \le k \le 20$) is the occurrence frequency of the 20 amino acids. Then, the PseAAC of a protein sequence can be represented by a vector as below:

$$V = [v_1, v_2, \ldots, v_{20}, v_{21}, \ldots, v_{20+\lambda}]^{\mathrm{T}} \tag{4}$$

In this study, ϖ and λ were set to be 0.15 and 50, respectively, and the following five properties of amino acids—(I) Codon diversity; (II) Electrostatic charge; (III) Molecular volume; (IV) Polarity; and (V) Secondary structure—were employed when taking into account the sequence information of protein sequence. The original values of all amino acids, which were retrieved from previous studies (Atchley et al. 2005; Rubinstein et al. 2009; Huang et al. 2010), are listed in Table 1. Each of the properties considered required 50 components to contain the sequence information. Thus, each protein sequence can be encoded into a $20 + 50 \times 5 = 270$-D (dimension) vector. Accordingly, each CPP or non-CPP was represented by 270 features in this study, and the distribution of these features is listed in Table 2.

mRMR method

As mentioned in "Encoding method", each investigated CPP or non-CPP was encoded into a 270-D (dimension) vector (i.e., each sample was represented by 270 features). However, they may not contribute equally to determining

Table 1 Original values of five properties for all amino acids

Amino acid	Polarity	Secondary structure	Molecular volume	Codon diversity	Electrostatic charge
A	−0.591	−1.302	−0.733	1.57	−0.146
C	−1.343	0.465	−0.862	−1.02	−0.255
D	1.05	0.302	−3.656	−0.259	−3.242
E	1.357	−1.453	1.477	0.113	−0.837
F	−1.006	−0.59	1.891	−0.397	0.412
G	−0.384	1.652	1.33	1.045	2.064
H	0.336	−0.417	−1.673	−1.474	−0.078
I	−1.239	−0.547	2.131	0.393	0.816
K	1.831	−0.561	0.533	−0.277	1.648
L	−1.019	−0.987	−1.505	1.266	−0.912
M	−0.663	−1.524	2.219	−1.005	1.212
N	0.945	0.828	1.299	−0.169	0.933
P	0.189	2.081	−1.628	0.421	−1.392
Q	0.931	−0.179	−3.005	−0.503	−1.853
R	1.538	−0.055	1.502	0.44	2.897
S	−0.228	1.399	−4.76	0.67	−2.647
T	−0.032	0.326	2.213	0.908	1.313
V	−1.337	−0.279	−0.544	1.242	−1.262
W	−0.595	0.009	0.672	−2.128	−0.184
Y	0.26	0.83	3.097	−0.838	1.512

Table 2 Distribution of the features used to encode each CPP and non-CPP

Feature category	Number of features
Amino acid frequency	20
Polarity	50
Second structure	50
Molecular volume	50
Codon diversity	50
Electrostatic charge	50

the CPP or non-CPP status of a peptide. To address this, we adopted a popular feature selection method—minimum redundancy maximum relevance—to distinguish them. This method was first introduced by Peng et al. (2005) and was executed based on the criteria of Max-Relevance and Min-Redundancy. Accordingly, two feature lists—MaxRel feature list and mRMR feature list—can be obtained, where the former is constructed by the criterion of Max-Relevance (i.e., sorts each feature according to its contributions to classification) and the latter list is constructed by considering criteria of both Max-Relevance and Min-Redundancy (i.e., sorts each feature according to its contribution to classification and redundancy with features listed before it). In this study, we formulated these two lists as below:

$$\begin{cases} \text{MaxRel features list: } F_{\text{MaxRel}} = [f_1^M, f_2^M, \ldots, f_n^M] \\ \text{mRMR features list: } F_{\text{mRMR}} = [f_1^m, f_2^m, \ldots, f_n^m] \end{cases} \quad (5)$$

where n is the total number of analyzed features. Readers can refer to Peng et al.'s paper (Peng et al. 2005) for a detailed description of this method. The mRMR method has previously been applied to the investigation of many complicated biological systems (Chen et al. 2013; Huang et al. 2011; Zhang et al. 2008; Mohabatkar et al. 2011; Han et al. 2013; Xu et al. 2014; Li et al. 2014b; Song et al. 2014).

Ten-fold cross-validation

Ten-fold cross-validation is a type of cross-validation method that is often used to evaluate the performance of classification and prediction models (Kohavi 1995). In this method, original samples are randomly and equally divided into ten parts. Accordingly, this method contains a procedure that is executed in ten rounds. For the *i*th round, samples in the *i*th part are singled out as testing samples, while samples in the other nine parts are used to train the classification or prediction methods. Thus, each sample is tested exactly once. Compared to another cross-validation method, the jackknife test (Chen et al. 2012), ten-fold cross-validation takes less time and provides similar results. Thus, it was adopted to evaluate the method used in this study.

Random forest

Random forest is an ensemble classifier that integrates multiple decision trees (Breiman 2001). Each decision tree is constructed as follows:

1. Let S be a dataset consisting of N samples. Randomly take N samples from S, with replacement, to comprise the dataset S' that is used to construct the decision tree, while non-selected samples are used as testing samples to evaluate its performance.
2. Suppose that samples in S were represented by M features. m is set to be a positive integer much less than M. The tree is grown at each node by randomly selecting m features from M features and optimally splitting the node on these m features.
3. The decision tree is fully grown and not pruned.

For any query sample, decision trees in the random forest make predictions, and the predicted result of the random forest is decided by a majority vote. The random forest can always provide good performance; it has been used to tackle some biological prediction problems (Kandaswamy et al. 2011; Li et al. 2012a, b; Lin et al. 2011; Pugalenthi et al. 2012; Shameer et al. 2011; Trost and Kusalik 2013).

Weka (Witten and Frank 2005) is a popular suite of machine learning software that collects many widely used state-of-art machine learning algorithms. It has a classifier called RandomForest that implements the procedure described above. For convenience, we used this classifier without modification as the classification model with default parameters in this study. In detail, each random forest constructs ten decision trees, and m is set to be $[\log_2 M + 1]$.

Accuracy measurement

In a two-class classification problem, the predicted results are always counted as four values: true positive (TP), true negative (TN), false positive (FP) and false negative (FN) (Baldi et al. 2000; Chen et al. 2010). Accordingly, the quality of the predicted results is evaluated by calculating sensitivity (SN), specificity (SP), prediction accuracy (ACC) and Matthews's correlation coefficient (MCC) (Matthews 1975) as follows:

$$\begin{cases} \text{SN} = \dfrac{\text{TP}}{\text{TP} + \text{FN}} \\ \text{SP} = \dfrac{\text{TN}}{\text{TN} + \text{FP}} \\ \text{ACC} = \dfrac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \\ \text{MCC} = \dfrac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TN} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TP} + \text{FN}) \times (\text{TP} + \text{FP})}} \end{cases} \quad (6)$$

It can be seen from Eq. 6 that SN and SP only use some of the TP, TN, FP and FN values, while ACC and MCC use all of them. Thus, SN and SP are not appropriate for a complete evaluation of the quality of the predicted results. Furthermore, although ACC uses all the TP, TN, FP and FN values, it cannot completely reflect the quality of the predicted results if the classes in the dataset are of very different sizes. In view of the number of investigated CPPs and non-CPPs, ACC is not a good choice. MCC is generally regarded as a balanced measure even if the classes are of very different sizes; it is, therefore, employed as a key measure for evaluating the quality of predicted results of the current method.

IFS method

As mentioned in "mRMR method", two feature lists can be produced by the mRMR method. The MaxRel feature list was used to analyze the importance of features (i.e., their contributions to classification), while the mRMR feature list was used to build an optimal classification model. The IFS method attempts to construct an optimal classification model based on the mRMR feature list and a machine learning algorithm (here, we selected random forest). First, according to the mRMR feature list $F_{\mathrm{mRMR}} = [f_1^m, f_2^m, \ldots, f_n^m]$, n feature sets are constructed: $F_{\mathrm{mRMR}}^i = [f_1^m, f_2^m, \ldots, f_i^m](1 \le i \le n)$. Second, for each F_{mRMR}^i $(1 \le i \le n)$, RandomForest was executed in Weka on a dataset in which samples were represented by features in F_{mRMR}^i evaluated by ten-fold cross-validation. Finally, SNs, SPs, ACCs and MCCs were calculated according to Eq. 6, and an IFS curve with the superscript i of F_{mRMR}^i as its X axis and MCC value as its Y axis was plotted, thereby obtaining a feature set with the highest MCC. Features in this set are called optimal features.

Results and discussion

Results of mRMR

As mentioned in "Encoding method", each CPP or non-CPP was represented by 270 features. To analyze these features, the mRMR method was employed (this program can be downloaded at the website http://research.janelia.org/peng/proj/mRMR/). For convenience, it was executed using default parameters. From this analysis, we obtained two feature lists—MaxRel feature list and mRMR feature list—which are provided in Supplementary Material II. Features were ranked by their values for determining whether a peptide was cell-penetrating or non-cell-penetrating in the MaxRel features list (i.e., features with a high rank contributed more, whereas those with low rank gave little or no contribution). Here, we investigated the highest 10 % of features on the MaxRel features list (refer to the first 27 features of the MaxRel features list in Supplementary Material II). Figure 1 shows the distribution of these 27 features, from which we can see that the amino acid frequency and five properties of amino acids were included at almost the same level of importance.

Results of IFS

To build an optimal prediction method and extract the optimal combination of features, the IFS method and random forest were employed. Figure 2 shows the IFS curve with MCC on the Y axis and the number of features participating in the model on the X axis. The SN, SP, ACC and MCC values obtained via the IFS method are available in Supplementary Material III. The highest MCC value was 0.4867, which was obtained using the first 32 features in the mRMR features list and the random forest classification method. The SN, SP and ACC were 0.9550, 0.4412 and 0.8345, respectively. This indicates that the optimal model has good discriminating power for determining CPPs and non-CPPs.

Comparison with other methods

To indicate the effectiveness of the optimal prediction method obtained in "Results of IFS", we compared our results with those reported in the study by Sanders et al. (2011), which were obtained from the same dataset and are listed in Table 3. The SN, SP and ACC in (Sanders et al. 2011) were 0.759, 0.232 and 0.7586, respectively, from which we can infer that the MCC was about −0.0089. All values are much lower than the corresponding values obtained by our method. Based on the fact that the MCC obtained by the optimal prediction method is not very low, it is not necessary to use a more complicated dataset compilation strategy, as described in (Sanders et al. 2011), to improve the predicted results.

Analysis of the optimal features

As described in "Results of IFS", the optimal prediction model used the first 32 features in the mRMR features list to encode each sample and used the random forest prediction method. These 32 features were deemed to be optimal features for determination of CPPs and non-CPPs. The distribution of these optimal features is shown in Fig. 3; we can see that the numbers of features relating to polarity, codon diversity and amino acid frequency were eight, eight, and seven, respectively, implying that these three amino acid properties are important factors for determination of CPPs and non-CPPs. Furthermore, four features were related to secondary structure and molecular volume.

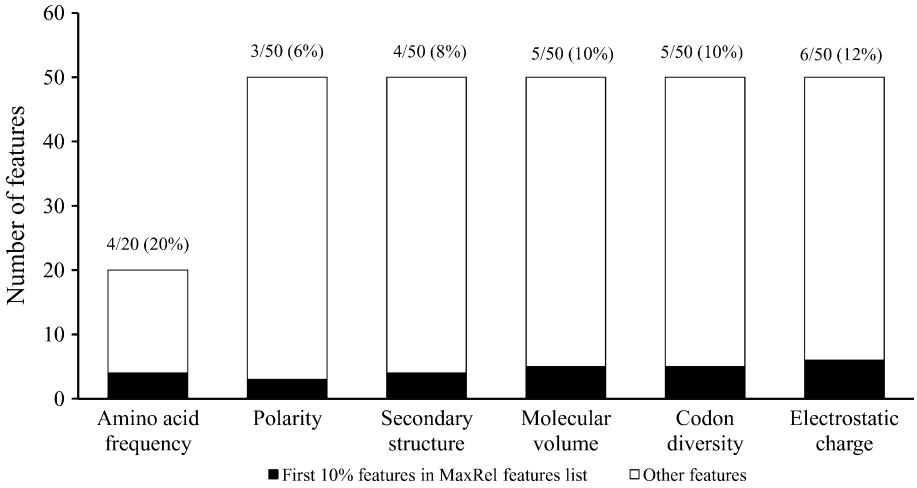

Fig. 1 Distribution of the first 27 features in the MaxRel features list

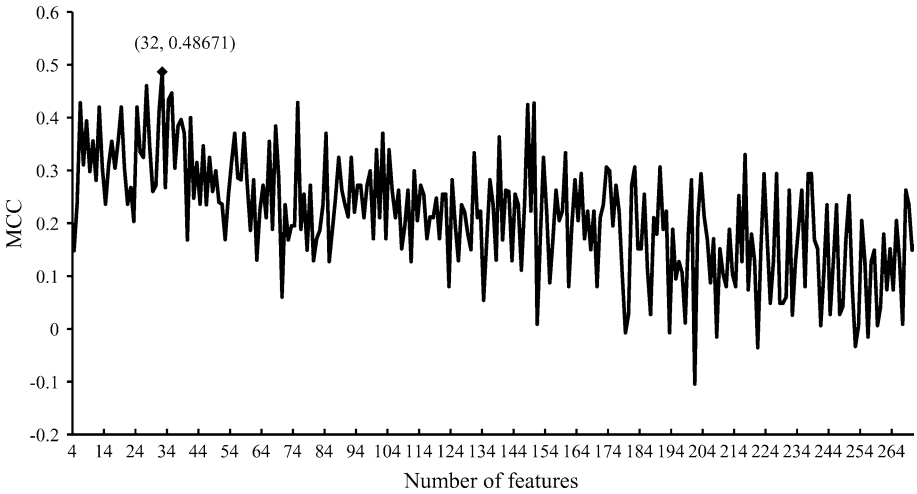

Fig. 2 The IFS curve. The *X* axis represents the number of features participating in the classification model. The *Y* axis represents the Matthews's correlation coefficient (MCC) value evaluated by the classification model and ten-fold cross-validation. The highest MCC value of IFS is 0.48671 using 32 features

Table 3 Comparison with the method presented by Sanders et al.

Measurement	Our method	Sanders et al.'s method
SN	0.9550	0.759
SP	0.4412	0.232
ACC	0.8345	0.7586
MCC	0.4867	−0.0089

Only one feature was related to electrostatic charge, implying that this property provided only a minor contribution for determination of CPPs and non-CPPs.

The amino acid frequency and molecular volume directly and indirectly refer to the selectivity of amino acids in composing CPPs. Previous studies have indicated that CPPs are water soluble and are polybasic peptides with a net positive charge at physiological pH (Jarver and Langel 2006). These properties enable CPPs to penetrate the cell membrane at low micromolar concentrations both in vivo and in vitro and with or without membrane receptors, all without causing significant membrane damage (Madani et al. 2011; Richard et al. 2005; Murriel and Dowdy 2006). Usually, positively charged arginine and lysine residues are specifically enriched in CPPs (Su et al. 2009), indicating a high amino acid selectivity and agreeing with our results. All amino acids can be divided into two main classes based on polarity: the polar and non-polar amino acids. Polar amino acids are those with side chains that prefer to reside in an aqueous (i.e., water) environment. By this definition, polar amino acids include aspartate, glutamate, asparagine, glutamine and the CPP-enriched arginine and lysine, consistent with our optimal features. This selectivity in amino acid composition closely correlates with the secondary structure of proteins and polypeptides (Eisenhaber et al. 1996; Malkov et al. 2009). For CPPs, secondary structure determines their interactions and insertion in the cell membrane and thus plays a crucial role (Eiriksdottir et al. 2010; Ye et al. 2010). It has been shown that several CPPs adopt a specific secondary structure, which functions to stabilize their interactions with the membrane and promote cellular uptake (Eiriksdottir et al. 2010).

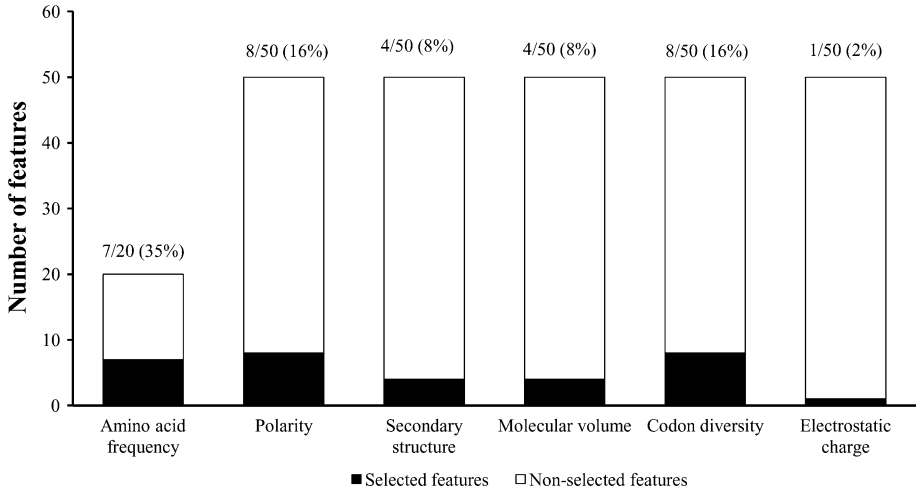

Fig. 3 Distribution of the 32 selected optimal features

Codon diversity implies a codon usage bias. It has been reported that codon usage bias can be used to predict gene expression levels (Henry and Sharp 2007; Roymondal et al. 2009). However, in our study, we show that codon diversity also strongly correlates with the ability of CCPs to penetrate the cell, revealing a previously unknown role of codon bias in CCP function.

As a result, the precise prediction of CCPs could be dependent on the traits of polarity, frequency, codon diversity and the molecular volume of amino acids, as well as the secondary structure of the peptide itself.

Conclusion

This study analyzed the contributing features of a peptide that determine its cell-penetrating capacity. Several feature selection methods, including minimum redundancy maximum relevance and incremental feature selection, were employed to analyze features extracted from protein sequence, including amino acid frequency, codon diversity, electrostatic charge, molecular volume, polarity, and secondary structure. As a result, key features were selected, and an optimal prediction method was constructed based on these features. We hope that this contribution will aid in uncovering the mechanism by which peptides penetrate cells.

Acknowledgments This study was supported by the National Basic Research Program of China (2011CB510101, 2011CB510102), the National Natural Science Foundation of China (61202021, 31371335, 61373028), the Innovation Program of Shanghai Municipal Education Commission (12YZ120, 12ZZ087), and the Shanghai Educational Development Foundation (12CG55).

Conflict of interest The authors declare that they have no conflict of interest.

References

Al-Soraj MH, Watkins CL, Vercauteren D, De Smedt SC, Braeckmans K, Jones AT (2010) siRNA versus pharmacological inhibition of endocytic pathways for studying cellular uptake of cell penetrating peptides. J Control Release 148(1):e86–87

Anaspec I (2010) Cell permeable peptides (CPP)/drug delivery peptides. In: Anaspec I (ed) Anaspec's catalog listing of cell permeable peptides (CPP)

Atchley WR, Zhao J, Fernandes AD, Drüke T (2005) Solving the protein sequence metric problem. Proc Natl Acad Sci USA 102(18):6395–6400

Baldi P, Brunak S, Chauvin Y, Andersen C, Nielsen H (2000) Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16(5):412–424

Basak SC (2013) Recent developments and future directions at current computer aided drug design. Curr Comput Aided Drug Des 9(1):1

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Chen C, Chen L, Zou X, Cai P (2009) Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine. Protein Pept Lett 16(1):27–31

Chen L, Feng KY, Cai YD, Chou KC, Li HP (2010) Predicting the network of substrate-enzyme-product triads by combining compound similarity and functional domain composition. BMC Bioinform 11:293

Chen L, Zeng WM, Cai YD, Feng KY, Chou KC (2012) Predicting anatomical therapeutic chemical (ATC) classification of drugs by integrating chemical-chemical interactions and similarities. PLoS One 7(4):e35254

Chen L, Zeng W-M, Cai Y-D, Huang T (2013) Prediction of metabolic pathway using graph property, chemical functional group and chemical structural set. Curr Bioinform 8(2):200–207

Chen L, Lu J, Huang T, Yin J, Wei L, Cai Y-D (2014a) Finding candidate drugs for hepatitis C based on chemical-chemical and chemical-protein interactions. PLoS One 9(9):e107767

Chen L, Lu J, Zhang N, Huang T, Cai Y-D (2014b) A hybrid method for prediction and repositioning of drug anatomical therapeutic chemical classes. Mol Bio Syst 10(4):868–877

Chou KC (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins 43(3):246–255

Ding H, Liu L, Guo F-B, Huang J, Lin H (2011) Identify Golgi protein types with modified mahalanobis discriminant algorithm and pseudo amino acid composition. Protein Pept Lett 18(1):58–63

Eiriksdottir E, Konate K, Langel U, Divita G, Deshayes S (2010) Secondary structure of cell-penetrating peptides controls membrane interaction and insertion. Biochim Biophys Acta 1798(6):1119–1128

Eisenhaber F, Imperiale F, Argos P, Frommel C (1996) Prediction of secondary structural content of proteins from their amino acid composition alone I: new analytic vector decomposition methods. Proteins 25(2):157–168

El-Andaloussi S, Holm T, Langel U (2005) Cell-penetrating peptides: mechanisms and applications. Curr Pharm Des 11(28):3597–3611

Gao S, Simon MJ, Hue CD, Morrison B 3rd, Banta S (2011) An unusual cell penetrating peptide identified using a plasmid display-based functional selection platform. ACS Chem Biol 6(5):484–491

Gautam A, Chaudhary K, Kumar R, Sharma A, Kapoor P, Tyagi A, Raghava GP (2013) In silico approaches for designing highly effective cell penetrating peptides. J Transl Med 11:74

Hajisharifi Z, Piryaiee M, Mohammad Beigi M, Behbahani M, Mohabatkar H (2014) Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test. J Theor Biol 341:34–40

Hällbrink M, Kilk K, Elmquist A, Lundberg P, Lindgren M, Jiang Y, Pooga M, Soomets U, Langel Ü (2005) Prediction of cell-penetrating peptides. Int J Pept Res Ther 11(4):249–259

Han GS, Anh V, Krishnajith AP, Tian Y-C (2013) An ensemble method for predicting subnuclear localizations from primary protein structures. PLoS One 8(2):e57225

Hansen M, Kilk K, Langel Ü (2008) Predicting cell-penetrating peptides. Adv Drug Deliv Rev 60(4):572–579

Hayat M, Khan A (2010) Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition. J Theor Biol 271(1):10–17

Heitz F, Morris MC, Divita G (2009) Twenty years of cell-penetrating peptides: from molecular mechanisms to therapeutics. Br J Pharmacol 157(2):195–206

Henry I, Sharp PM (2007) Predicting gene expression level from codon usage bias. Mol Biol Evol 24(1):10–12

Holm T, Johansson H, Lundberg P, Pooga M, Lindgren M, Langel U (2006) Studying the uptake of cell-penetrating peptides. Nat Protoc 1(2):1001–1005

Huang T, Shi XH, Wang P, He Z, Feng KY, Hu L, Kong X, Li YX, Cai YD, Chou KC (2010) Analysis and prediction of the metabolic stability of proteins based on their sequential features, subcellular locations and interaction networks. PLoS One 5(6):e10972

Huang T, Chen L, Cai Y, Chou C (2011) Classification and analysis of regulatory pathways using graph property, biochemical and physicochemical property, and functional property. PLoS One 6(9):e25297

Huang G, Zhang Y, Chen L, Zhang N, Huang T, Cai Y-D (2014) Prediction of multi-type membrane proteins in human by an integrated approach. PLoS One 9(3):e93553

Jarver P, Langel U (2006) Cell-penetrating peptides: a brief introduction. Biochim Biophys Acta 1758(3):260–263

Kandaswamy KK, Chou KC, Martinetz T, Moller S, Suganthan PN, Sridharan S, Pugalenthi G (2011) AFP-Pred: a random forest approach for predicting antifreeze proteins from sequence-derived properties. J Theor Biol 270:56–62

Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of international joint conference on artificial intelligence, 1995. Lawrence Erlbaum Associates Ltd, pp 1137–1145

Kong L, Zhang L, Lv J (2014) Accurate prediction of protein structural classes by incorporating predicted secondary structure information into the general form of Chou's pseudo amino acid composition. J Theor Biol 344:12–18

Lee JH, Song HS, Park TH, Lee SG, Kim BG (2012) Screening of cell-penetrating peptides using mRNA display. Biotechnol J 7(3):387–396

Li BQ, Feng KY, Chen L, Huang T, Cai YD (2012a) Prediction of protein-protein interaction sites by Random Forest algorithm with mRMR and IFS. PLoS One 7(8):e43927

Li BQ, Hu LL, Chen L, Feng KY, Cai YD, Chou KC (2012b) Prediction of protein domain with mRMR feature selection and analysis. PLoS One 7(6):e39308

Li BQ, Zhang YC, Huang GH, Cui WR, Zhang N, Cai YD (2014a) Prediction of aptamer-target interacting pairs with pseudo-amino acid composition. PLoS One 9(1):e86729

Li Z, Chen L, Lai Y, Dai Z, Zou X (2014b) The prediction of methylation states in human DNA sequences based on hexanucleotide composition and feature selection. Anal Methods 6(6):1897–1904

Lin H (2008) The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. J Theor Biol 252(2):350–356

Lin WZ, Fang JA, Xiao X, Chou KC (2011) iDNA-Prot: identification of DNA binding proteins using random forest with grey model. PLoS One 6:e24756

Lindberg S, Munoz-Alarcon A, Helmfors H, Mosqueira D, Gyllborg D, Tudoran O, Langel U (2013) PepFect15, a novel endosomolytic cell-penetrating peptide for oligonucleotide delivery via scavenger receptors. Int J Pharm 441(1–2):242–247

Madani F, Lindberg S, Langel U, Futaki S, Graslund A (2011) Mechanisms of cellular uptake of cell-penetrating peptides. J Biophys 2011:414729

Malkov SN, Zivkovic MV, Beljanski MV, Stojanovic SD, Zaric SD (2009) A reexamination of correlations of amino acids with particular secondary structures. Protein J 28(2):74–86

Marks DS, Hopf TA, Sander C (2012) Protein structure prediction from sequence variation. Nat Biotechnol 30(11):1072–1080

Matthews B (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA)-Protein. Structure 405(2):442–451

Mohabatkar H, Mohammad Beigi M, Esmaeili A (2011) Prediction of GABAA receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine. J Theor Biol 281(1):18–23

Montrose K, Yang Y, Sun X, Wiles S, Krissansen GW (2013) Xentry, a new class of cell-penetrating peptide uniquely equipped for delivery of drugs. Sci Rep 3:1661

Mueller J, Kretzschmar I, Volkmer R, Boisguerin P (2008) Comparison of cellular uptake using 22 CPPs in 4 different cell lines. Bioconjug Chem 19(12):2363–2374

Murriel CL, Dowdy SF (2006) Influence of protein transduction domains on intracellular delivery of macromolecules. Expert Opin Drug Deliv 3(6):739–746

Nanni L, Lumini A, Gupta D, Garg A (2012) Identifying bacterial virulent proteins by fusing a set of classifiers based on variants of Chou's pseudo amino acid composition and on evolutionary information. IEEE/ACM Trans Comput Biol Bioinform (TCBB) 9(2):467–475

Ou-Yang SS, Lu JY, Kong XQ, Liang ZJ, Luo C, Jiang H (2012) Computational drug discovery. Acta Pharmacol Sin 33(9):1131–1140

Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 1226–1238

Pugalenthi G, Kandaswamy KK, Chou KC, Vivekanandan S, Kolatkar P (2012) RSARF: prediction of residue solvent accessibility from protein sequence using random forest method. Protein Pept Lett 19:50–56

Richard JP, Melikov K, Brooks H, Prevot P, Lebleu B, Chernomordik LV (2005) Cellular uptake of unconjugated TAT peptide involves

clathrin-dependent endocytosis and heparan sulfate receptors. J Biol Chem 280(15):15300–15306

Roymondal U, Das S, Sahoo S (2009) Predicting gene expression level from relative codon usage bias: an application to Escherichia coli genome. DNA Res 16(1):13–30

Rubinstein ND, Mayrose I, Pupko T (2009) A machine-learning approach for predicting B-cell epitopes. Mol Immunol 46(5):840–847

Sanders WS, Johnston CI, Bridges SM, Burgess SC, Willeford KO (2011) Prediction of cell penetrating peptides by support vector machines. PLoS Comput Biol 7(7):e1002101

Shah M, Passovets S, Kim D, Ellrott K, Wang L, Vokler I, LoCascio P, Xu D, Xu Y (2003) A computational pipeline for protein structure prediction and analysis at genome scale. Bioinformatics 19(15):1985–1996

Shameer K, Pugalenthi G, Kandaswamy KK, Sowdhamini R (2011) 3dswap-pred: prediction of 3D domain swapping from protein sequence using random forest approach. Protein Pept Lett 18:1010–1020

Song L, Li D, Zeng X, Wu Y, Guo L, Zou Q (2014) nDNA-prot: identification of DNA-binding proteins based on unbalanced classification. BMC Bioinform 15(1):298

Su Y, Doherty T, Waring AJ, Ruchala P, Hong M (2009) Roles of arginine and lysine residues in the translocation of a cell-penetrating peptide from (13)C, (31)P, and (19)F solid-state NMR. Biochemistry 48(21):4587–4595

Trost B, Kusalik A (2013) Computational phosphorylation site prediction in plants using random forests and organism-specific instance weights. Bioinformatics 29(6):686–694

Vives E, Schmidt J, Pelegrin A (2008) Cell-penetrating and cell-targeting peptides in drug delivery. Biochim Biophys Acta 1786(2):126–138

Wang P, Hu L, Liu G, Jiang N, Chen X, Xu J, Zheng W, Li L, Tan M, Chen Z, Song H, Cai YD, Chou KC (2011) Prediction of antimicrobial peptides based on sequence alignment and feature selection methods. PLoS ONE 6(4):e18476

Witten IH, Frank E (2005) Data Mining: practical machine learning tools and techniques. Morgan Kaufmann Pub, San Francisco

Xu Y, Deng Y, Ji Z, Liu H, Liu Y, Peng H, Wu J, Fan J (2014) Identification of thyroid carcinoma related genes with mRMR and shortest path approaches. PLoS One 9(4):e94022

Ye J, Fox SA, Cudic M, Rezler EM, Lauer JL, Fields GB, Terentis AC (2010) Determination of penetratin secondary structure in live cells with Raman microscopy. J Am Chem Soc 132(3):980–988

Zhang Y, Ding C, Li T (2008) Gene selection algorithm by combining reliefF and mRMR. BMC Genom 9(Suppl 2):S27

Zhou GP, Cai YD (2006) Predicting protease types by hybridizing gene ontology and pseudo amino acid composition. Proteins Struct Funct Bioinf 63(3):681–684

Zou D, He Z, He J, Xia Y (2011) Supersecondary structure prediction using Chou's pseudo amino acid composition. J Comput Chem 32(2):271–278