Supporting Information

CPPred-RF: a sequence-based predictor for identifying cell-penetrating peptides and their uptake efficiency

Leyi Wei[1], PengWei Xing[1], Ran Su[2], Gaotao Shi[1], Zhanshan (Sam) Ma[3*] and Quan Zou[1,3*]

1. School of Computer Science and Technology, Tianjin University, Tianjin, China
2. School of Software, Tianjin University, Tianjin, China
3. State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China

*Corresponding author: samma@uidaho.edu and zouquan@tju.edu.cn

Table of Contents

Supplementary method S-1. Algorithmic details of feature descriptors

For convenience of discussion, a given peptide sequence is denoted as **P**=$P_1P_2\ldots P_L$, where P_i represents the *i*-th amino acid in **P** and L represents the length of **P**. In the following subsections, we describe how to use the above four feature descriptors to represent the sequence **P**, respectively.

1) Features based on PC-PseAAC

For given a peptide sequence **P,** the PC-PseAAC feature vector is represented by,

$$FV = [fv_1, fv_2, \ldots, fv_{20}, fv_{20+1}, \ldots, fv_{20+\lambda}]^T \qquad (1)$$

where

$$fv_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{\lambda}\theta_j}, & 1 \le u \le 20 \\ \dfrac{w\theta_{u-20}}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{\lambda}\theta_j}, & 20+1 \le u \le 20+\lambda \end{cases} \qquad (2)$$

where u is an integer; fv_u ($1 \le u \le 20$) represents the normalized appearance frequency of the 20 amino acids in **P**; λ represents the highest tier of the correlation along **P**; θ_j ($j = 1,2,\ldots,\lambda$) is the correlation function that measures the *j*-tier sequence-order correlation between all the *j*-th most contiguous residues along **P**. θ_j is subject to the following formula,

$$\theta_j = \frac{1}{L}\sum_{i=1}^{L-j}\frac{1}{3}\sum_{m=1}^{3}[H_m(P_{i+j}) - H_m(P_i)]^2 \qquad (3)$$

where $H_m(P_i)$ (m=1,2,3) represents the normalized hydrophobicity value, the hydrophilicity value, and the side-chain mass value corresponding to the *i*-th amino acid P_i in the peptide sequence **P**, respectively. They are calculated by the following formula,

$$H_m(P_i) = \frac{H'_m(P_i) - \sum_{j=1}^{20}\frac{H'_m(A_j)}{20}}{\sqrt{\frac{\sum_{l=1}^{20}[H'_m(A_l) - \sum_{j=1}^{20}\frac{H'_m(A_j)}{20}]^2}{20}}} \qquad (4)$$

where $H_m'(P_i)$ is the original value of $H_m(P_i)$.

2) *Features based on SC-PseAAC*

For given a peptide sequence **P**, the SC-PseAAC feature vector is represented by,

$$FV = [fv_1, fv_2, \dots, fv_{20}, fv_{20+1}, \dots, fv_{20+2\lambda}]^T \qquad (5)$$

where

$$fv_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + w \sum_{j=1}^{2\lambda} \tau_j}, & 1 \le u \le 20 \\ \dfrac{w\tau_{u-20}}{\sum_{i=1}^{20} f_i + w \sum_{j=1}^{2\lambda} \tau_j}, & 20+1 \le u \le 20+2\lambda \end{cases} \qquad (6)$$

where fv_u $(1 \le u \le 20)$ represents the normalized appearance frequency of the 20 amino acids in **P**; λ is the highest tier of the correlation along **P**; τ_j $(j = 1,2,\dots,\lambda)$ is the correlation function, which measures the *j*-tier sequence-correlation between all the *j*-th most contiguous residues along **P** and is defined by,

$$\begin{cases} \tau_{2k-1} = \dfrac{1}{L-k} \displaystyle\sum_{j=1}^{L-k} H_1(P_i) H_1(P_{i+k}) \\ \tau_{2k} = \dfrac{1}{L-k} \displaystyle\sum_{j=1}^{L-k} H_2(P_i) H_2(P_{i+k}) \end{cases} \qquad (7)$$

where $H_1(P_i)$ and $H_2(P_i)$ are the standardized values of hydrophobicity and hydrophilicity, respectively.

3) *Features based on adaptive skip dipeptide composition*

For given a peptide sequence **P**, dipeptide composition is defined as the fraction of any two adjacent residues $(P_i P_{i+1})$ [11]. It measures the correlation of any two adjacent residues in a sequence. However, it is obvious that the correlation information of those intervening (non-adjacent) two residues $(P_i P_j; j - i > 1)$ would be lost. To address this problem, we present a modified dipeptide composition, called adaptive skip dipeptide composition, which computes frequencies of any two residues in the sequence **P**. The proposed adaptive skip dipeptide composition sufficiently considers the correlation not only between adjacent residues but also between intervening residues. For given a peptide sequence **P**, the feature vector for adaptive skip dipeptide composition is represented by,

$$FV = [fv_1, fv_2, \ldots, fv_{400}]^T \quad (8)$$

where

$$fv_i = \frac{O^i}{\sum_{k=1}^{L} n(k)} \quad (9)$$

where O^i represents the observed number of *i*-th two-residue pair and $n(k)$ represents the number of all possible two residues with $\leq k$ intervening residues. If k=1, the feature vector is exactly the dipeptide composition.

4) Features based on physicochemical properties

Physicochemical properties have proven to provide insights into the differences between the classes of CPPs and non-CPPs [10,13]. To capture the physicochemical information, we adopted a powerful feature descriptor that uses protein physicochemical properties to represent peptide sequences [18,19]. This feature descriptor considers the following eight physicochemical properties: (1) normalized van der waals volume, (2) secondary structure, (3) solvent accessibility, (4) polarizability, (5) polarity, (6) hydrophobicity, (7) charge, and (8) surface tension. For each property, 20 standard amino acids {A,N,C,Q,G,H,I,L,M,F,P,S,T,W,Y,V,D,E,K,R} are divided into three groups, e.g., {ANCQGHILMFPSTWYV, DE, KR} for the charge property. To quantize the physicochemical information, the sequence **P** is encoded from the following three aspects: content, distribution and bivalent frequency for each physicochemical property. The details of encoding procedure can be referred to [18,19]. To this end, the peptide sequence **P** is subsequently represented with a 188-dimension feature vector.

Supplementary method S-2. Algorithmic details of performance metrics

Five commonly used metrics for a binary classification task are employed in present study, which include Sensitivity (SE), Specificity (SP), Accuracy (ACC), Mathew's correlation coefficient (MCC), and area under the receiver operating characteristic curve (AUC). Of these metrics, AUC is to calculate the area under the receiver operating characteristic curve; while the other metrics are respectively calculated as,

$$SE = \frac{TP}{TP + FN} * 100\%$$

$$SP = \frac{TN}{TN + FP} * 100\%$$

$$ACC = \frac{TP + TN}{TP + FN + TN + FP} * 100$$

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

where TP, TN, FP, and FN are the number of true positive, true negative, false positive, and false negative, respectively.

Table S-1. Predictive results of 1st layer and 2nd layer models varying the number of features.

1st layer model			**2nd layer model**		
No. of Features	***ACC***	***MCC***	***No. of Features***	***ACC***	***MCC***
10	0.742	0.485	10	0.607	0.214
20	0.789	0.578	20	0.61	0.219
30	0.819	0.64	30	0.634	0.268
40	0.842	0.684	40	0.631	0.263
50	0.845	0.691	50	0.658	0.317
60	0.854	0.708	60	0.663	0.328
70	0.859	0.72	70	0.65	0.3
80	0.855	0.71	80	0.663	0.327
90	0.861	0.723	90	0.644	0.291
100	0.866	0.732	100	0.65	0.3
110	0.883	0.766	110	0.671	0.342
120	0.876	0.752	120	0.682	0.364
130	0.881	0.762	130	0.668	0.338
140	0.879	0.758	140	0.676	0.354
150	0.883	0.767	150	0.674	0.348
160	0.886	0.774	160	0.66	0.321
170	0.886	0.773	170	0.684	0.369
180	0.896	0.793	180	0.69	0.38
190	0.9	0.801	190	0.693	0.385
200	0.9	0.801	200	0.687	0.375
210	0.905	0.81	210	0.674	0.348
220	0.903	0.806	220	0.687	0.374
230	0.904	0.808	230	0.679	0.359
240	0.896	0.793	240	0.69	0.38
250	0.906	0.812	250	0.701	0.402
260	0.9	0.802	260	0.684	0.369
270	0.907	0.815	270	0.703	0.407
280	0.9	0.802	280	0.693	0.385
290	**0.916**	**0.831**	290	0.695	0.391
300	0.895	0.79	300	0.701	0.401
310	0.903	0.805	310	0.69	0.38
320	0.909	0.819	320	0.684	0.369
330	0.906	0.812	330	0.703	0.407
340	0.907	0.815	**340**	**0.711**	**0.423**
350	0.906	0.813	350	0.666	0.332
360	0.907	0.814	360	0.687	0.375

370	0.909	0.818	370	0.711	0.423
380	0.906	0.812	380	0.687	0.374
390	0.911	0.823	390	0.69	0.38
400	0.9	0.802	400	0.687	0.374
410	0.903	0.806	410	0.687	0.375
420	0.906	0.812	420	0.69	0.38
430	0.902	0.803	430	0.668	0.337
440	0.911	0.823	440	0.693	0.385
450	0.903	0.806	450	0.693	0.385
460	0.91	0.822	460	0.676	0.353
470	0.909	0.819	470	0.682	0.364
480	0.912	0.825	480	0.69	0.38
490	0.906	0.812	490	0.668	0.337
500	0.909	0.819	500	0.687	0.375
510	0.904	0.809	510	0.671	0.342
520	0.909	0.819	520	0.698	0.396
530	0.909	0.819	530	0.698	0.396
540	0.908	0.816	540	0.698	0.396
550	0.91	0.821	550	0.674	0.348
560	0.902	0.804	560	0.682	0.364
570	0.91	0.821	570	0.679	0.359
580	0.912	0.826	580	0.684	0.369
590	0.906	0.813	590	0.701	0.401
600	0.907	0.815	600	0.703	0.407
610	0.906	0.812	610	0.684	0.369
620	0.905	0.81	620	0.698	0.396
630	0.911	0.823	630	0.69	0.38