11. Commonly Used Feature Descriptors and Parameters for Protein or Peptide Sequences

AAC (Amino Acid Composition)

The Amino Acid Composition (AAC) encoding (2) calculates the frequency of each amino acid type in a protein or peptide sequence.

EAAC (Enhanced Amino Acid Composition)

The Enhanced Amino Acid Composition (EAAC) feature calculates the AAC based on the sequence window of fixed length that continuously slides from the N- to C-terminus of each peptide and can be usually applied to encode the peptides with an equal length.

Parameters:

- *Sliding window size:* the length of the sliding window, default is 5.

EGAAC (Enhanced GAAC)

The Enhanced GAAC (EGAAC) is also for the first time proposed in this work. It calculates GAAC in windows of fixed length continuously sliding from the N- to C-terminal of each peptide and is usually applied to peptides with an equal length.

Parameters:

- *Sliding window size:* the length of the sliding window, default is 5.

CKSAAP type 1 (Composition of k-spaced Amino Acid Pairs type 1)

The CKSAAP type 1 feature encoding calculates the frequency of amino acid pairs separated by any *k* residues (*k* = 0, 1, 2, … , 5. The default maximum value of *k* is 5) (5,6,89,90).

Parameters:

- *Gap value: the length of k-spaced amino acids, default is 3.*

CKSAAP type 2 (Composition of k-spaced Amino Acid Pairs type 2)

The CKSAAP type 2 feature encoding calculates the raw count of amino acid pairs separated by any k residues.

Parameters:

- *Gap value: the length of k-spaced amino acids, default is 3.*

CKSAAGP type 1 (Composition of k-Spaced Amino Acid Group Pairs type 1)

CKSAAGP type 1 is a variation of the CKSAAP descriptor, which is our own proposal. It calculates the frequency of amino acid group pairs separated by any *k* residues.

Parameters:

- *Gap value: the length of k-spaced amino acids, default is 3.*

CKSAAGP type 2 (Composition of k-Spaced Amino Acid Group Pairs type 2)

It calculates the raw count of amino acid group pairs separated by any *k* residues.

Parameters:

- *Gap value: the length of k-spaced amino acids, default is 3.*

DPC type 1 (Di-Peptide Composition type 1)

The Dipeptide Composition (7) gives 400 descriptors. It is defined as:

$$D(r,s) = \frac{N_{rs}}{N-1}, \qquad r,s \in \{A,C,D,\ldots Y\}$$

where N_{rs} is the number of dipeptides represented by amino acid types r and s.

DPC type 2 (Di-Peptide Composition type 2)

DPC type 2 calculate the raw count of the 400 di-peptides in a sequences.

GDPC type 1 (Grouped Di-Peptide Composition type 1)

The Grouped Di-Peptide Composition encoding is another variation of the DPC descriptor. It is

composed of a total of 25 descriptors that are defined as:

$$f(r,s)=\frac{N_{rs}}{N-1}, \quad r,s\in\{g1,g2,g3,g4,g5\}$$

where N_{rs} is the number of tripeptides represented by amino acid type groups r and s, N is the length of a protein or peptide sequence.

GDPC type 2 (Grouped Di-Peptide Composition type 2)

GDPC type 2 calculates the raw count of the 25 grouped amino acid pairs.

TPC type 1 (Tri-Peptide Composition type 1)

The Tripeptide Composition (TPC) (2) gives 8000 descriptors, defined as:

$$f(r,s,t)=\frac{N_{rst}}{N-2}, \quad r,s,t\in\{A,C,D,...,Y\}$$

where N_{rst} is the number of tripeptides represented by amino acid types r, s and t.

TPC type 2 (Tri-Peptide Composition type 2)

TPCP type 2 calculate the raw count of the 8000 tripeptides.

GTPC type 1 (Gouped Tri-Peptide Composition type 1)

The Grouped Tri-Peptide Composition encoding is also a variation of TPC descriptor, which generates 125 descriptors, defined as:

$$f(r,s,t)=\frac{N_{rst}}{N-2}, \quad r,s,t\in\{g1,g2,g3,g4,g5\}$$

where N_{rst} is the number of tripeptides represented by amino acid type groups r, s and t. $\underline{N}$ is the length of a protein or peptide sequence.

GTPC type 2 (Gouped Tri-Peptide Composition type 2)

GTPC type 2 calculates the raw count of the 125 grouped amino acid tripeptides.

DDE (Dipeptide Deviation from Expected Mean)

The Dipeptide Deviation from Expected Mean feature vector (7) is constructed by computing three parameters, i.e. dipeptide composition (D_c), theoretical mean (T_m), and theoretical variance (T_v). The above three parameters and the DDE are computed as follows. $D_c(r,s)$, the dipeptide composition measure for the dipeptide '*rs'*, is given as

$$D_c(r,s) = \frac{N_{rs}}{N-1}, \qquad r,s \in \{A,C,D,\ldots Y\}$$

where N_{rs} is the number of dipeptides represented by amino acid types r and s and N is the length of the protein or peptide. $T_m(r,s)$, the theoretical mean, is given by:

$$T_m(r,s) = \frac{C_r}{C_N} \times \frac{C_s}{C_N}$$

where C_r is the number of codons that code for the first amino acid and C_s is the number of codons that code for the second amino acid in the given dipeptide '*rs*'. C_N is the total number of possible codons, excluding the three stop codons (i.e., 61). T_v (*r,s*), the theoretical variance of the dipeptide '*rs*', is given by:

$$T_v(r,s) = \frac{T_m(r,s)(1-T_m(r,s))}{N-1}$$

Finally, *DDE(r,s)* is calculated as:

$$DDE(r,s) = \frac{D_c(r,s) - T_m(r,s)}{\sqrt{T_v(r,s)}}$$

Binary

In the Binary encoding (31,91), each amino acid is represented by a 20-dimensional binary vector, e.g.

A is encoded by (10000000000000000000), C is encoded by (01000000000000000000), …, Y is encoded by (00000000000000000001), respectively. This encoding scheme is often used to encode peptides with an equal length.

Binary_6bit

In this descriptor, the 6-letter exchange group {e_1, e_2, e_3, e_4, e_5, e_6} is adopted to represent a protein sequence (16,32), where $e_1 \in$ {H, R, K}, $e_2 \in$ {D, E, N, D}, $e_3 \in$ {C}, $e_4 \in$ {S, T, P, A, G}, $e_5 \in$ {M, I, L, V}, $e_6 \in$ {F, Y, W}. Exchange groups represent conservative replacements through evolution. These exchange groups are effectively equivalence classes of amino acids and are derived from PAM. For example, the protein sequence PVKTNVK can be represented as $e_4e_5e_1e_4e_2e_5e_1$. Then, each group is represented by a 6-dimensional binary vector, e.g. e_1 is encoded by (100000), e_2 is encoded by (010000), …, e_6 is encoded by (000001).

Binary_5bit_type 1

Like binary_6bit descriptor, the 5-letter amino acid group {e_1, e_2, e_3, e_4, e_5} is adopted to represent a protein sequence, and each group is represented by a 5-dimensional binary vector (16,33). $e_1 \in$ {G, A, V, L, M, I}, $e_2 \in$ {F, Y, W}, $e_3 \in$ {K, R, H}, $e_4 \in$ {D, E}, $e_5 \in$ {S, T, C, P, N, Q}. Then, each group is represented by a 5-dimensional binary vector, e.g. e_1 is encoded by (10000), e_2 is encoded by (01000), …, e_5 is encoded by (00001).

Binary_5bit_type 2

For this descriptor, it is based on all the possible ways that ones and zeros can be combined in a five bit unit. There are 32 possible ways to represent 20 amino acids. When the representations with no or all ones and those with 1 or 4 ones are removed there are exactly twenty representations. And A is encoded by (00011), C (00101), D (00110), E (00111), F(01001), G (01010), H (01011), I (01100), K (01101), L (01110), M (10001), N (10010), P (10011), Q (10100), R (10101), S (10110), T (11000), V (11001), W (11010), Y (11100).

Binary_3bit_type 1-7

For this descriptor, the 3-letter amino acid group {e_1, e_2, e_3} is adopted to represent a protein sequence, and each group is represented by a 3-dimensional binary vector (80). Then, each group is represented by a 5-dimensional binary vector, e.g. e_1 is encoded by (100), e_2 is encoded by (010), e_3 is encoded by (001). The division of the amino acids based on physicochemical properties (PRAM900101) in **Table 12**. The difference of the type 1 to type 7 subtypes is the different division of amino acids. The division of amino acids for type 2 to 7 is based on “Normalized van der Waals volume”, “Polarity”, “Polarizability”, “Charge”, “Secondary structure” and “Solvent accessibility” in **Table 12**.

Table 12. Amino acid physicochemical attributes and the division of the amino acids into three groups according to each attribute.

Attribute	Division		
Hydrophobicity_PRAM900101	Polar: RKEDQN	Neutral: GASTPHY	Hydrophobicity: CLVIMFW
Hydrophobicity_ARGP820101	Polar: QSTNGDE	Neutral: RAHCKMV	Hydrophobicity: LYPFIW
Hydrophobicity_ZIMJ680101	Polar: QNGSWTDERA	Neutral: HMCKV	Hydrophobicity: LPFYI
Hydrophobicity_PONP930101	Polar: KPDESNQT	Neutral: GRHA	Hydrophobicity: YMFWLCVI
Hydrophobicity_CASG920101	Polar: KDEQPSRNTG	Neutral: AHYMLV	Hydrophobicity: FIWC
Hydrophobicity_ENGD860101	Polar: RDKENQHYP	Neutral :SGTAW	Hydrophobicity: CVLIMF
Hydrophobicity_FASG890101	Polar: KERSQD	Neutral: NTPG	Hydrophobicity: AYHWVMFLIC
Normalized van der Waals volume	Volume range: 0-2.78 GASTPD	Volume range: 2.95-94.0 NVEQIL	Volume range: 4.03-8.08 MHKFRYW
Polarity	Polarity value: 4.9-6.2 LIFWCMVY	Polarity value: 8.0-9.2 PATGS	Polarity value: 10.4-13.0 HQRKNED
Polarizability	Polarizability value: 0-1.08 GASDT	Polarizability value: 0.128-120.186 GPNVEQIL	Polarizability value: 0.219-0.409 KMHFRYW
Charge	Positive: KR	Neutral: ANCQGHILMFPSTWYV	Negative: DE
Secondary structure	Helix: EALMQKRH	Strand: VIYCWFT	Coil: GNPSD
Solvent accessibility	Buried: ALFCGIVW	Exposed: PKQEND	Intermediate: MPSTHY

AESNN3 (Learn from alignments)

For this descriptor, each amino acid type is described using a three-dimensional vector. Values are taken from the three hidden units from the neural network trained on structure alignments (16,34). The values are listed in **Table 13**.

Table 13. AESNN3 values learning from alignments.

Amino acids	AESNN3 values		
A	-0.99	-0.61	0.00
R	0.28	-0.99	-0.22
N	0.77	-0.24	0.59
D	0.74	-0.72	-0.35
C	0.34	0.88	0.35
Q	0.12	-0.99	-0.99
E	0.59	-0.55	-0.99
G	-0.79	-0.99	0.10
H	0.08	-0.71	0.68
I	-0.77	0.67	-0.37
L	-0.92	0.31	-0.99
K	-0.63	0.25	0.50
M	-0.80	0.44	-0.71
F	0.87	0.65	-0.53
P	-0.99	-0.99	-0.99
S	0.99	0.40	0.37
T	0.42	0.21	0.97
W	-0.13	0.77	-0.90
Y	0.59	0.33	-0.99
V	-0.99	0.27	-0.52

Moran

The autocorrelation descriptors are defined based on the distribution of amino acid properties along the sequence (17,19,20). The amino acid properties used here are different types of amino acids index, which is retrieved from the AAindex Database (83) available at http://www.genome.jp/dbget/aaindex.html/. All the amino acid indices are centralized and standardized prior to the calculation:

$$P_r = \frac{P_r - \overline{P}}{\sigma},$$

where $\overline{P}$ is the average of the properties of the 20 amino acids and σ is the standard deviation of the properties of the 20 amino acids. $\overline{P}$ and σ can be calculated as follows:

$$\overline{P} = \frac{\sum_{r=1}^{20} P_r}{20}, \quad \sigma = \sqrt{\frac{1}{20}\sum_{r=1}^{20}(P_r - \overline{P})^2}.$$

The Moran autocorrelation descriptors (17,18) can thus be defined as:

$$I(d)=\frac{\frac{1}{N-d}\sum_{i=1}^{N-d}(P_i-\overline{P}')(P_{i+d}-\overline{P}')}{\frac{1}{N}\sum_{i=1}^{N}(P_i-\overline{P}')^2}, \qquad d=1,2,3...,nlag,$$

where d is the lag of the autocorrelation, $nlag$ is the maximum value of the lag, P_i and P_{i+d} are the properties of the amino acids at positions i and $i + d$, respectively. $\overline{P}'$ is the average of the considered property P over the entire sequence of length N and is calculated as:

$$\overline{P}'=\frac{\sum_{i=1}^{N}P_i}{N}.$$

The Moran descriptor has been successfully applied to membrane protein type prediction (17) and protein secondary structural content prediction (18).

Parameters:

- *Physicochemical properties for proteins: the names of used physicochemical amino acids indices.*

Geary

The Geary autocorrelation descriptors for a protein or peptide sequence are defined as:

$$C(d)=\frac{\frac{1}{2(N-d)}\sum_{i=1}^{N-d}(P_i-P_{i+d})^2}{\frac{1}{N-1}\sum_{i=1}^{N}(P_i-\overline{P}')^2}, \qquad d=1,2,...,nlag,$$

where d, P, P_i and P_{i+d}, $nlag$ have the same definitions as described above. The Geary descriptor has been successfully applied to population structure inferring (19).

Parameters:

- *Physicochemical properties for proteins: the names of used physicochemical amino acids indices.*

NMBroto (Normalized moreau-broto autocorrelation)

The Moreau-Broto autocorrelation descriptors are defined as follows:

$$AC(d) = \sum_{i=1}^{N-d} P_i \times P_{i+d}, \qquad d = 1, 2, ..., nlag .$$

The normalized Moreau-Broto autocorrelation descriptors are thus defined as:

$$ATS(d) = \frac{AC(d)}{N-d}, \qquad d = 1, 2, ..., nlag .$$

The NMBroto descriptor has been successfully applied to protein helix content prediction (20).

Parameters:

- *Physicochemical properties for proteins: the names of used physicochemical amino acids indices.*

CTD (Composition/Transition/Distribution)

The Composition, Transition and Distribution (CTD) features represent the amino acid distribution patterns of a specific structural or physicochemical property in a protein or peptide sequence (8-12). 13 types of physicochemical properties have been previously used for computing these features (**Table 12**). These include hydrophobicity, normalized Van der Waals Volume, polarity, polarizability, charge, secondary structures and solvent accessibility. These descriptors are calculated according to the following procedures: (i) The sequence of amino acids is transformed into a sequence of certain structural or physicochemical properties of residues; (ii) Twenty amino acids are divided into three groups for each of the seven different physicochemical attributes based on the main clusters of the amino acid indices of Tomii and Kanehisa (92). The groups of amino acids are listed in **Table 12**. The Composition/Transition/Distribution descriptors has been successfully applied to protein folding class prediction (10,11), enzyme family classification (9), RNA-binding protein prediction (12), protein structural prediction (92) and anti-cancer peptide prediction (80).

CTDC

Taking the hydrophobicity attribute as an example, all amino acids are divided into three groups: polar, neutral and hydrophobic (**Table 12**). The Composition descriptor consists of three values: the

global compositions (percentage) of polar, neutral and hydrophobic residues of the protein. An illustrated example of this encoding scheme is provided in the following **Figure 15**. The Composition descriptor can be calculated as follows:

$$C(r) = \frac{N(r)}{N}, \quad r \in \{polar, neutral, hydrophobic\},$$

where *N(r)* is the number of amino acid type *r* in the encoded sequence and *N* is the length of the sequence.

CTDT

The Transition descriptor T also consists of three values (10,11): A transition from the polar group to the neutral group is the percentage frequency with which a polar residue is followed by a neutral residue or a neutral residue by a polar residue. Transitions between the neutral group and the hydrophobic group and those between the hydrophobic group and the polar group are defined in a similar way. The transition descriptor can then be calculated as:

$$T(r,s) = \frac{N(r,s) + N(s,r)}{N-1}, \quad r,s \in \{(polar, neutral), (neutral, hydrophobic), (hydrophobic, polar)\},$$

where *N(r,s)* and *N(s,r)* are the numbers of dipeptides encoded as "*rs*" and "*sr*" respectively in the sequence, while *N* is the length of the sequence. An illustrated example of this encoding scheme is provided in the following **Figure 15.**

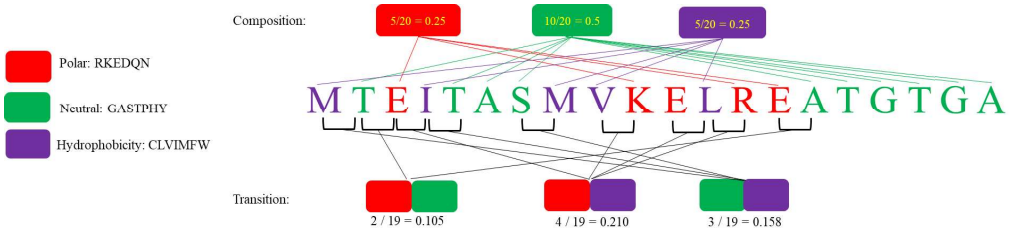

Figure 15. An example of the calculation of composition and transition descriptors. This example uses the hydrophobicity attribute.

CTDD

The Distribution descriptor consists of five values for each of the three groups (polar, neutral and hydrophobic) (10,11), namely the corresponding fraction of the entire sequence, where the first residue of a given group is located, and where 25, 50, 75 and 100% of occurrences are contained.

For example, we start with the first residue up to and including the residue that marks 25/50/75/100% of occurrences for residues of any given group and then we simply divide the position of this residue by the length of the entire sequence.

CTriad (Conjoint triad)

The Conjoint Triad descriptor (CTriad) considers the properties of one amino acid and its vicinal amino acids by regarding any three continuous amino acids as a single unit (13). First, the protein sequence is represented by a binary space (*V, F*), where *V* denotes the vector space of the sequence features, and each feature (V_i) represents a sort of triad type; *F* is the number vector corresponding to *V*, where f_i, the value of the *i*-th dimension of *F*, is the number of type V_i appearing in the protein sequence. For the amino acids that have been catalogued into seven classes, the size of *V* should be equal to 7 x 7 x 7=343. Accordingly, i = 1, 2, 3, …, 343. An illustrated example of this encoding scheme is provided in the following **Figure 16**. In principle, the longer a protein sequence, the higher the probability to have larger values of f_i, confounding the comparison of proteins with different lengths. Thus, we define a new parameter, d_i, by normalizing f_i with the following equation:

$$d_i = \frac{f_i - \min\{f_1, f_2, ..., f_{343}\}}{\max\{f_1, f_2, ..., f_{343}\}}.$$

The CTriad descriptors has been successfully applied to protein-protein interaction prediction (13).

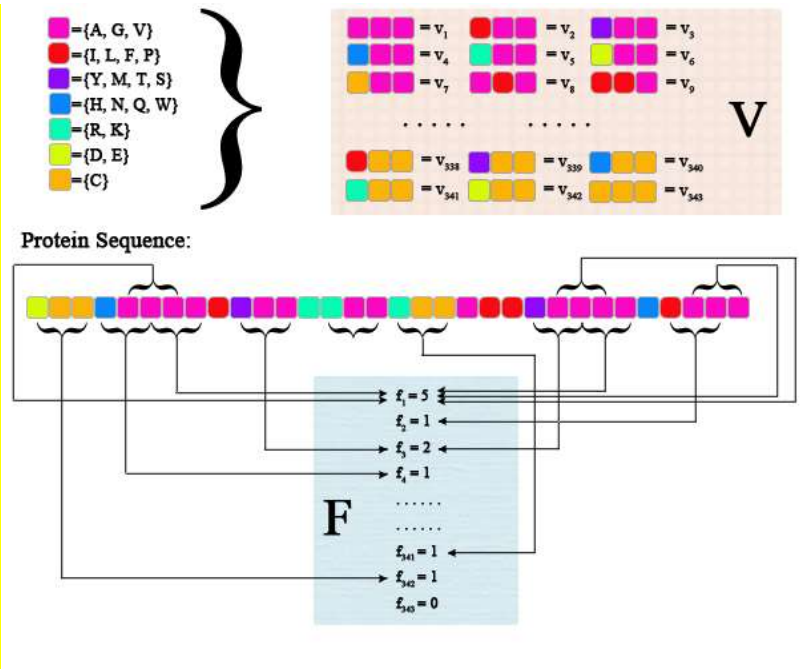

Figure 16. Schematic diagram for constructing the vector space (*V, F*) of a given protein sequence. *V* is the vector space of the sequence features; each feature (V_i) represents a triad composed of three consecutive amino acids; *F* is the number vector corresponding to *V*, and the value of the *i*-th entry of *F*, denoted f_i, is the number of occurrences that the triad associated with V_i appearing in the protein sequence. The figure was adapted from the Supplementary Figure in (13).

KSCTriad (Conjoint k-spaced Triad)

The *k*-Spaced Conjoint Triad (KSCTriad) descriptor is based on the Conjoint CTriad descriptor, which not only calculates the numbers of three continuous amino acid units, but also considers the continuous amino acid units that are separated by any *k* residues (The default maximum value of *k* is set to 5). For example, AxRxT is a 1-spaced triad. Thus, the dimensionality of the KSCTriad encoded feature vector is 343×(*k*+1). An illustrated example of this encoding scheme is provided in **Figure 17**.

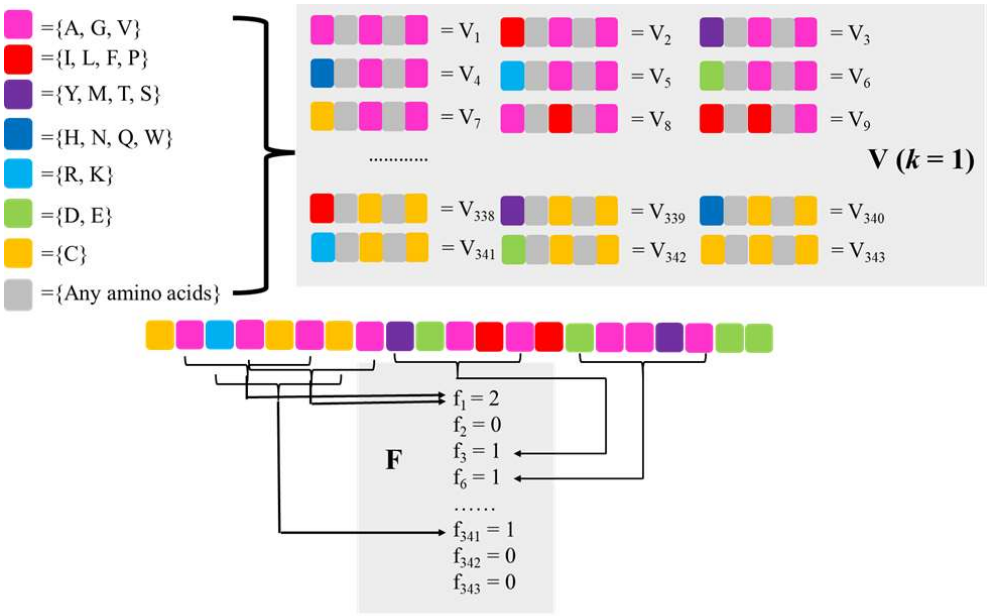

Figure 17. A schematic diagram for constructing the vector space (V, F) of protein sequence (*k*=1).

Parameters:

- *Gap value: the length of k-spaced amino acids, default is 3.*

SOCNumber (Sequence-Order-Coupling Number)

The *d*-th rank sequence-order-coupling number is defined as:

$$\tau_d = \sum_{i=1}^{N-d} (d_{i,i+d})^2, \qquad d = 1, 2, 3, ..., nlag,$$

where $d_{i,i+d}$ is the entry in a given distance matrix describing a distance between the two amino acids at position *i* and *i* + *d*, *nlag* denotes the maximum value of the lag (default value: 30) and *N* is the length of a protein or peptide sequence. As distance matrix both the Schneider-Wrede physicochemical distance matrix (26) used by Kuo-Chen Chou, and the chemical distance matrix by Grantham (93) are used. Accordingly, the descriptor dimension will be *nlag* x 2. The quasi-sequence-order descriptors described next also utilizes the two matrices. An illustrated example of this encoding scheme is provided in the following **Figure 18**.

Note: the length of the protein must be not less than the maximum value of *nlag*.

Parameter(s):

- *Lag value: the maximum value of the lag, default value is 3.*

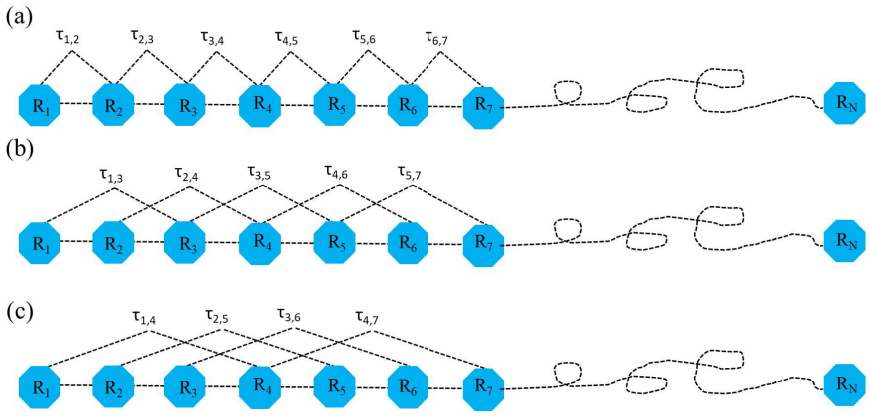

Figure 18. A schematic drawing to show (a) the 1st-rank, (b) the 2nd-rank, and (3) the 3rd-rank sequence-order-coupling mode along a protein sequence. (a) reflects the coupling mode between all the most adjacent residues, (b) shows the coupling between the adjacent plus one residue, and (c) shows the coupling between the adjacent plus two residues. This figure is adapted from (24).

QSOrder (Quasi-sequence-order)

For each amino acid type, a quasi-sequence-order descriptor can be defined as:

$$X_r = \frac{f_r}{\sum_{r=1}^{20} f_r + w\sum_{d=1}^{nlag} \tau_d}, \qquad r = 1, 2, ..., 20$$

where f_r is the normalized occurrence of amino acid type r and w is a weighting factor (w = 0.1), $nlag$ and τ_d have the same definitions as described above. These are the first 20 quasi-sequence-order descriptors. The other 30 quasi-sequence-order descriptors are defined as:

$$X_d = \frac{w\tau_d - 20}{\sum_{r=1}^{20} f_r + w\sum_{d=1}^{nlag} \tau_d}, \qquad d = 21, 22, ..., 20 + nlag\ .$$

The SOCNumber and QSOrder descriptors have been successfully applied to protein subcellular location prediction (25,26).

Parameter(s):

- *Lag value: the maximum value of the lag, default value is 3.*

- *Weight factor: the weight factor value, range from 0 to 1, and default is 0.05.*

PAAC (Pseudo-Amino Acid Composition)

This group of descriptors has been proposed in (27,28). Let $H_1^o(i)$, $H_2^o(i)$, $M^o(i)$ for $i = 1, 2, 3, \ldots 20$ be the original hydrophobicity values, the original hydrophilicity values and the original side chain masses of the 20 natural amino acids, respectively. They are converted to the following quantities by a standard conversion:

$$H_1(i) = \frac{H_1^o(i) - \frac{1}{20}\sum_{i=1}^{20} H_1^o(i)}{\sqrt{\frac{\sum_{i=1}^{20}[H_1^o(i) - \frac{1}{20}\sum_{i=1}^{20} H_1^o(i)]^2}{20}}},$$

where $H_2^o(i)$ and $M^o(i)$ are normalized as $H_2(i)$ and $M(i)$ in the same manner. An example of the correlation function is provided in the following **Figure 19**.

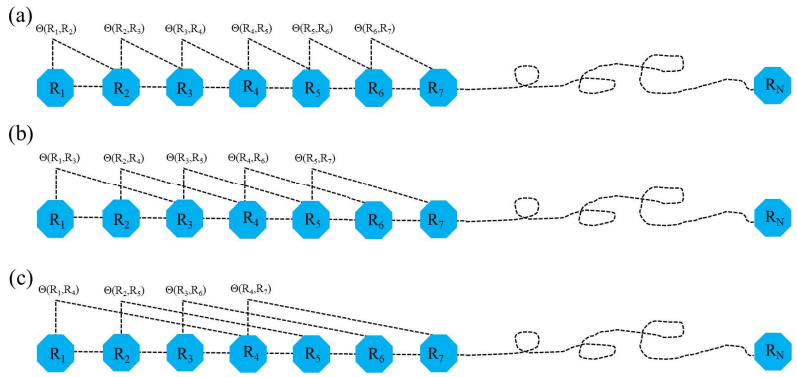

Figure 19. A schematic illustration showing (a) the first-tier, (b) the second-tier, and (3) the third-tier sequence order correlation mode along a protein sequence. (a) reflects the coupling mode between all the most adjacent residues, (b) shows the coupling between the adjacent plus one residue, and (c) shows the coupling between the adjacent plus two residues. This figure is adapted from (27) for illustration purposes.

Next, a correlation function can be defined as:

$$\Theta(R_i, R_j) = \frac{1}{3}\{[H_1(R_i) - H_1(R_j)]^2 + [H_2(R_i) - H_2(R_j)]^2 + [M(R_i) - M(R_j)]^2\}.$$

This correlation function is actually an averaged value for the three amino acid properties: hydrophobicity value, hydrophilicity value and side chain mass. Therefore, we can extend this definition of correlation function for one amino acid property or for a set of *n* amino acid properties. For one amino acid property, the correlation can be defined as:

$$\Theta(R_i, R_j) = [H_1(R_i) - H_1(R_j)]^2 ,$$

where $H(R_i)$ is the amino acid property of amino acid R_i after standardization.

For a set of *n* amino acid properties, it can be defined as:

$$\Theta(R_i, R_j) = \frac{1}{n}\sum_{n=1}^{n}[H_k(R_i) - H_k(R_j)]^2 ,$$

where $H_k(R_i)$ is the *k*-th property in the amino acid property set for amino acid R_i.

A set of descriptors called sequence order-correlated factors are defined as:

$$\theta_1 = \frac{1}{N-1}\sum_{i=1}^{N-1}\Theta(R_i, R_{i+1}) ,$$

$$\theta_2 = \frac{1}{N-2}\sum_{i=1}^{N-2}\Theta(R_i, R_{i+2}) ,$$

$$\theta_3 = \frac{1}{N-3}\sum_{i=1}^{N-3}\Theta(R_i, R_{i+3}) ,$$

$$\ldots,$$

$$\theta_\lambda = \frac{1}{N-\lambda}\sum_{i=1}^{N-\lambda}\Theta(R_i, R_{i+\lambda}) ,$$

where λ ($\lambda < N$) is an integer parameter to be chosen. Let f_i be the normalized occurrence frequency of amino acid *i* in the protein sequence. Then, a set of $20 + \lambda$ descriptors called the pseudo-amino acid composition for a protein sequence can be defines as:

$$X_c = \frac{f_c}{\sum_{r=1}^{20} f_r + w\sum_{j=1}^{\lambda}\theta_j}, \qquad (1 < c < 20) ,$$

$$X_c = \frac{w\theta_{c-20}}{\sum_{r=1}^{20} f_r + w\sum_{j=1}^{\lambda}\theta_j}, \qquad (21 < c < 20 + \lambda) ,$$

where *w* is the weighting factor for the sequence-order effect and is set to $w = 0.05$ as suggested by

Chou *et al*. (27).

Parameter(s):

- *Lambda value: integer, should be small than sequence length, default is 2.*
- *Weight value: weight factor, ranged from 0 to 1, default is 0.05.*

APAAC (Amphiphilic Pseudo-Amino Acid Composition)

Amphiphilic Pseudo-Amino Acid Composition (APAAC) was proposed in (27,28). The definition of this set of features is similar to the PAAC descriptors. Using $H_1(i)$ and $H_2(j)$ as previously defined, the hydrophobicity and hydrophilicity correlation functions are defined as:

$$H_{i,j}^1 = H_1(i)H_1(j)$$
$$H_{i,j}^2 = H_2(i)H_2(j)$$,

respectively. An illustrated example of the correlation functions is provided in the following **Figure 20**. Thus, sequence order factors can be defined as:

$$\tau_1 = \frac{1}{N-1}\sum_{i=1}^{N-1} H_{i,i+1}^1$$
$$\tau_2 = \frac{1}{N-1}\sum_{i=1}^{N-1} H_{i,i+1}^2$$
$$\tau_3 = \frac{1}{N-2}\sum_{i=1}^{N-2} H_{i,i+2}^1$$
$$\tau_4 = \frac{1}{N-2}\sum_{i=1}^{N-2} H_{i,i+2}^2$$
...
$$\tau_{2\lambda-1} = \frac{1}{N-\lambda}\sum_{i=1}^{N-\lambda} H_{i,i+\lambda}^1$$
$$\tau_{2\lambda} = \frac{1}{N-\lambda}\sum_{i=1}^{N-\lambda} H_{i,i+\lambda}^2$$

Then, Amphiphilic Pseudo-Amino Acid Composition (APAAC) is defined as:

$$P_c = \frac{f_c}{\sum_{r=1}^{20} f_r + w\sum_{j=1}^{2\lambda} \tau_j}, \quad (1 < c < 20)$$

$$P_c = \frac{\omega\tau_u}{\sum_{r=1}^{20} f_r + w\sum_{j=1}^{2\lambda} \tau_j}, \quad (21 < u < 20 + 2\lambda)$$

where *w* is the weighting factor. In *iLearnPlus* this factor is set to *w* = 0.5 as described in Chou's work (27). The PAAC and APAAC have been successfully applied to protein cellular attributes prediction (27) and enzyme subfamily classes prediction (28).

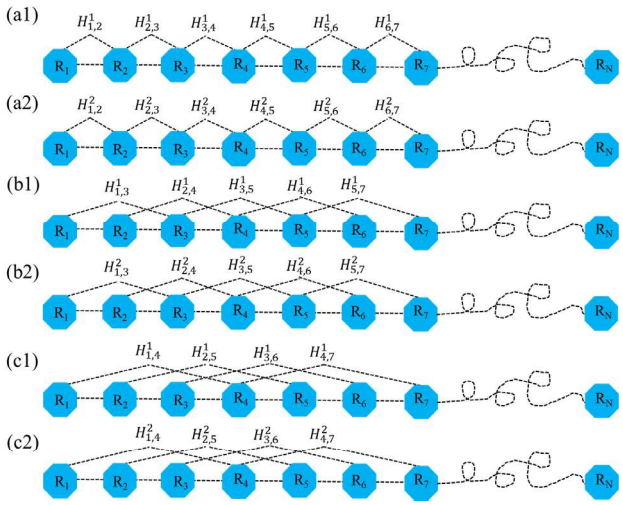

Figure 20. A schematic diagram to show (a1/a2) the first-rank, (b1/b2) the second-rank and (c1/c2) the third-rank sequence-order-coupling mode along a protein sequence through a hydrophobicity/hydrophilicity correlation function, where $H_{i,j}^1$ and $H_{i,j}^2$ are given by the aforementioned equation. Panels (a1/a2) reflects the coupling mode between the most adjacent residues, panels (b1/b2) shows the coupling between the adjacent plus one residue, and panels (c1/c2) shows the coupling between the adjacent plus two residues. This figure is adapted from (28) for illustration purposes.

Parameter(s):

- *Lambda value: integer, should be small than sequence length, default is 2.*
- *Weight value: weight factor, ranged from 0 to 1, default is 0.05.*

AAindex

Physicochemical properties of amino acids are the most intuitive features for representing biochemical reactions and have been extensively applied in bioinformatics research. The amino acid indices (AAindex) database (83) collects many published indices representing physicochemical properties of amino acids. For each physicochemical property, there is a set of 20 numerical values for all amino acids. Currently, 544 physicochemical properties can be retrieved from the AAindex

database. After removing physicochemical properties with value 'NA' for any of the amino acids, 531 physicochemical properties were left. In contrast to the residue-based encoding methods of amino acid identity and evolutionary information, a vector of 531 mean values is used to represent a sample for various window sizes. The AAINDEX descriptor (35) can be applied to encode peptides of equal length.

Parameters:

- *Physicochemical properties for proteins: the names of used physicochemical amino acids indices.*

BLOSUM62

In this descriptor, the BLOSUM62 matrix is employed to represent the protein primary sequence information as the basic feature set. A matrix comprising of $m \times n$ elements is used to represent each residue in a training dataset, where n denotes the peptide length and $m = 20$, which elements comprise 20 amino acids. Each row in the BLOSUM62 matrix is adopted to encode one of 20 amino acids. The BLOSUM62 descriptor can be applied to encode peptides of equal length. The BLOSUM62 descriptor has been successfully applied to ubiquitination site prediction (36).

ZScale

For this descriptor, each amino acid is characterized by five physicochemical descriptor variables (cf. **Table 14**), which were developed by Sandberg *et al.* in 1998 (94). The ZSCALE descriptor can be applied to encode peptides with equal length. The descriptor has been successfully applied to sumoylation site prediction (37).

Table 14. Z-scales for the 20 amino acids.

Amino acid	Z1	Z2	Z3	Z4	Z5	Amino Acid	Z1	Z2	Z3	Z4	Z5
A	0.24	-2.32	0.60	-0.14	1.30	M	-2.85	-0.22	0.47	1.94	-0.98
C	0.84	-1.67	3.71	0.18	-2.65	N	3.05	1.60	1.04	-1.15	1.61
D	3.98	0.93	1.93	-2.46	0.75	P	-1.66	0.27	1.84	0.70	2.00
E	3.11	0.26	-0.11	-3.04	-0.25	Q	1.75	0.50	-1.44	-1.34	0.66
F	-4.22	1.94	1.06	0.54	-0.62	R	3.52	2.50	-3.50	1.99	-0.17
G	2.05	4.06	0.36	-0.82	-0.38	S	2.39	-1.07	1.15	-1.39	0.67
H	2.47	1.95	0.26	3.90	0.09	T	0.75	-2.18	-1.12	-1.46	-0.40
I	-3.89	-1.73	-1.71	-0.84	0.26	V	-2.59	-2.64	-1.54	-0.85	-0.02
K	2.29	0.89	-2.49	1.49	0.31	W	-4.36	3.94	0.59	3.44	-1.59
L	-4.28	-1.30	-1.49	-0.72	0.84	Y	-2.54	2.44	0.43	0.04	-1.47

OPF_10bit

For this descriptor, the amino acids are classified into 10 groups based their physicochemical properties (**Table 15**). Note that different groups might overlap, since a specific amino acid type may have two or more physicochemical properties. To reflect the correlation of different properties, a 10-bit vector were calculated to represent each amino acid of the N-terminus of peptide. Similarly, the position of the bit of this 10-bit vector is set to 1, if the amino acid belongs to a corresponding group and 0 otherwise. The OPF_10bit descriptor has been successfully applied to anti-cancer peptides prediction (80).

Table 15. Details of the division of the standard amino acid alphabet based on ten physicochemical properties.

Rank	Physicochemical properties	Amino acid group
1	Aromatic	{F, Y, W, H}
2	Negative	{D, E}
3	Positive	{K, H, R}
4	Polar	{N, Q, S, D, E, C, T, K, R, H, Y, W}
5	Hydrophobic	{A, G, C, T, I, V, L, K, H, F, Y, W, M}
6	Aliphatic	{I, V, L}
7	Tiny	{A, S, G, C}
8	Charged	{K, H, R, D, E}
9	Small	{P, N, D, T, C, A, G, S, V}
10	Proline	{P}

OPF_7bit type 1

Similar to OPF_10bit descriptor, for this descriptor, the amino acids are classified into 7 groups. There are three subtypes of OPF_7bit descriptor (i.e. type 1, type 2 and type 3), due to different division of the amino acids. The difference of the type 1 to type 3 subtypes is the different division of amino acids. The division of amino acids for type 1 to 3 is listed in **Table 16**.

KNN (K-Nearest Neighbors)

The *K*-Nearest Neighbor for nucleotide (KNN) descriptor requires an extra training file and a label file. The training file is used to calculate the top *K*-Nearest Neighbor peptides by calculating the similarity score of two nucleotide sequences.

ASDC (Adaptive skip dinucleotide composition)

The adaptive skip dipeptide composition is a modified dipeptide composition, which sufficiently considers the correlation information present not only between adjacent residues but also between intervening residues (80). For given a sequence, the feature vector for ASDC is represented by:

$$ASDC = (f_{v1}, f_{v1}, \dots, f_{v400}),$$

where f_{vi} is calculated by

$$f_{vi} = \frac{\sum_{g=1}^{L-1} O_i^g}{\sum_{i=1}^{400} \sum_{g=1}^{L-1} O_i^g},$$

where f_{vi} denotes the occurrence frequency of all possible dipeptide with $\leq L$-1 intervening nucleotides. The ASDC descriptor has been successfully applied to anti-cancer peptide prediction

(80) and cell-penetrating peptide prediction (81).

Table 16. Details of the division of the standard amino acid alphabet based on seven physicochemical properties.

Rank	Physicochemical properties	Type 1	Type 2	Type 3
1	Hydrophobicity	{A, C, F, G, H, I, L, M, N, P, Q, S, T, V, W, Y}	{D, E}	{K, R}
2	Normalized Van der Waals volume	{C, F, I, L, M, V, W}	{A, G, H, P, S, T, Y}	{D, E, K, N, Q, R}
3	Polarity	{A, C, D, G, P, S, T}	{E, I, L, N, Q, V}	{F, H, K, M, R, W, Y}
4	Polarizibility	{C, F, I, L, M, V, W, Y}	{A, G, P, S, T}	{D, E, H, K, N, Q, R}
5	Charge	{A, D, G, S, T}	{C, E, I, L, N, P, Q, V}	{F, H, K, M, R, W, Y}
6	Secondary structures	{D, G, N, P, S}	{A, E, H, K, L, M, Q, R}	{C, F, I, T, V, W, Y}
7	Solvent accessibility	{A, C, F, G, I, L, V, W}	{H, M, P, S, T, Y}	{D, E, K, N, R, Q}

DistancePair (PseAAC of distance-pair and reduced alphabet)

The descriptor incorporates the amino acid distance pair coupling information and the amino acid reduced alphabet profile into the general pseudo amino acid composition vector. For the reduced alphabet profile, they are cp(13), cp(14), and cp(15) as defined below:

cp(13) = {MF; IL;V;A;C;WYQHP;G;T; S;N;RK;D; E}

cp(14) = {EIMV; L;F;WY;G; P;C;A; S; T;N;HRKQ; E;D}

sp(15) = {P;G; E;K;R;Q;D; S;N; T;H;C; I;V;W;YF;A; L;M}

where the single letters without a semicolon (;) to separate them mean belonging to a same cluster. The DistancePair descriptor has been successfully applied to DNA-binding protein identification (16).

Parameter(s):

- *Maximum distance: the maximum distance value, default is 0.*
- *Reduced alphabet scheme: i.e. cp(13), cp(14), cp(15) or cp(20).*

PseKRAAC (pseudo K-tuple reduced amino acids composition)

Previous studies indicate that certain residues are similar in their physicochemical features, and can be clustered into groups because they play similar structural or functional roles in proteins (95). By implementing reduced amino acid alphabets, the protein complexity can be significantly simplified, which reduces information redundancy and decreases the risk of overfitting. The Pseudo *K*-tuple Reduced Amino Acids Composition (PseKRAAC) descriptor (29) includes two different feature types for protein sequence analysis: *g-gap* and λ-correlation PseKRAAC (27).

The *g-gap* PseKRAAC is used to represent a protein sequence with a vector containing $RAAC^K$ components, where *g* represents the gap between each *K*-tuple peptides (21,52,96,97). A *g-gap* of *n* reflects the sequence-order information for all *K*-tuple peptides with the starting residues separated by *n* residues. An illustrated example of this encoding scheme ($K = 2$) is provided in the following **Figure 21A**.

The λ-correlation PseKRAAC is used to represent a protein sequence with a vector containing $RAAC^K$ components, where λ is an integer that represents the correlation tier and is less than N-K, where N is the sequence length. The *n*-th-tier correlation factor ($\lambda = n$) reflects the sequence-order correlation between the *n*-th nearest residues. An illustrated example of this encoding scheme ($K = 2$) is provided in the following **Figure 21B**.

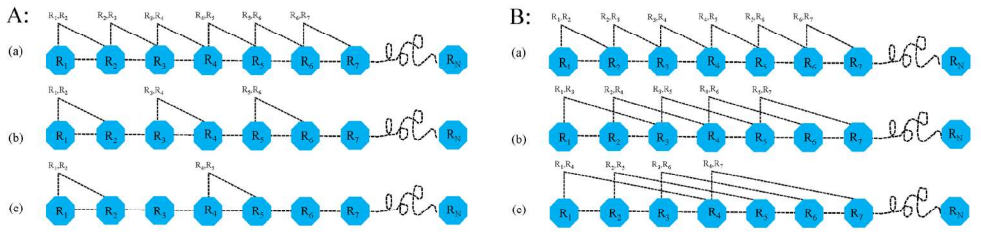

Figure 21. A schematic diagram showing: (A) *g-gap* definition of dipeptide, and (B) λ-correlation definition of dipeptide A: (a) *g-gap* of 0 reflects the sequence-order information between all adjacent dipeptides, i.e. separated by zero residues, (b) *g-gap* of 1 reflects the sequence-order information for all dipeptides with the starting residues separated by one residue, and (c) *g-gap* of 2 reflects the sequence-order information for all dipeptides with the starting residues separated by two residues; B: (a) the first-tier correlation factor reflects the sequence-order correlation between the nearest residues along a protein chain, (b) the second-tier correlation factor reflects the sequence-order correlation between the second nearest residues, (c) the third-tier correlation factor reflects the sequence-order correlation between the 3rd nearest residues, and so forth. The figure is adapted from (29).

The 16 types of reduced amino acid alphabets with different clustering approaches can be used to generate different versions of pseudo reduced amino acid compositions (PseRAACs) (**Table17**).

Table 17. A list of 16 types of reduced amino acid alphabets for proteins (29).

Type	Description	Cluster	Reference
1	RedPSSM	2-19	(98)
2	BLOSUM 62 matrix	2-6, 8, 15	(99)
3	PAM matrix (3A) and WAG matrix (3B)	2-19	(100)
4	Protein Blocks	5,8,9,11,13	(101)
5	BLOSUM50 matrix	3,4,8,10,15	(101)
6	Multiple cluster	4,5A,5B,5C	(102)
7	Metric multi-dimensional scaling	2-19	(103)
8	Grantham Distance Matrix	2-19	(104)
9	Grantham Distance Matrix	2-19	(104)
10	BLOSUM matrix for SWISS-PROT	2-19	(105)
11	BLOSUM matrix for SWISS-PROT	2-19	(105)
12	BLOSUM matrix for DAPS	2-18	(106)
13	Coarse-graining substitution matrices	4,12,17	(107)
14	Alphabet Simplifer	2-19	(108)
15	MJ matrix	2-16	(109)
16	BLOSUM50 matrix	2-16	(109)

Parameter(s):

- *RAAC subtype model: RAAC subtype model, i.e. g-gap or lambda-correlation.*
- *Gap value: the gap value for between two amino acids.*
- *Lambda value: the lambda value in lambda-correlation model.*
- *K-tuple: k-tuple value.*
- *RAAC cluster: the RAAC cluster type for each type of feature descriptor.*

12. Commonly Used Feature Descriptors and Parameters for Ligands

Basak

The feature descriptor calculates some commonly used basak information index based on its topological structure, and 21 molecular connectivity can be obtained for this feature descriptor.

Burden

The feature descriptor calculates 64 burden eigvenvalue descriptors.

Phamacophore