**Pse-in-One 2.0**: a web server for generating comprehensive modes of pseudo components of DNA, RNA, and protein sequences (2017 update)

2017-01-08

**Home-page**: [http://bioinformatics.hitsz.edu.cn/Pse-in-One2.0/](http://bioinformatics.hitsz.edu.cn/Pse-in-One2.0/)
Content

1. DNA ........................................................................................................................... 4
  1.1 Deoxyribonucleic acid composition ........................................................................ 4
     1.1.1 Basic kmer (Kmer)........................................................................................ 4
     1.1.2 Reverse complementary kmer (RevKmer)................................................... 4
     1.1.3 Increment of diversity (IDKmer).................................................................. 4
     1.1.4 Mismatch....................................................................................................... 5
     1.1.5 Subsequence.................................................................................................. 5
  1.2 Autocorrelation ........................................................................................................ 5
     1.2.1 Dinucleotide-based auto covariance (DAC) ................................................. 5
     1.2.2 Dinucleotide-based cross covariance (DCC).............................................. 6
     1.2.3 Dinucleotide-based auto-cross covariance (DACC) ..................................... 6
     1.2.4 Trinucleotide-based auto covariance (TAC) ................................................. 6
     1.2.5 Trinucleotide-based cross covariance (TCC) ................................................ 7
     1.2.6 Trinucleotide-based auto-cross covariance (TACC) ..................................... 7
     1.2.7 Moran autocorrelation (MAC) ...................................................................... 7
     1.2.8 Geary autocorrelation (GAC) ....................................................................... 8
     1.2.9 Normalized Moreau–Broto autocorrelation (NMBAC) .............................. 8
  1.3 Pseudo deoxyribonucleic acid composition ............................................................. 9
     1.3.1 Pseudo dinucleotide composition (PseDNC) ................................................ 9
     1.3.2 Pseudo k-tuple nucleotide composition (PseKNC) ........................................ 10
     1.3.3 General parallel correlation pseudo dinucleotide composition (PC-PseDNC-General) ......................................................................................... 10
     1.3.4 General parallel correlation pseudo trinucleotide composition (PC-PseTNC-General) ......................................................................................... 11
     1.3.5 General series correlation pseudo dinucleotide composition (SC-PseDNC-General) ......................................................................................... 12
     1.3.6 General series correlation pseudo trinucleotide composition (SC-PseTNC-General) ......................................................................................... 13
  2. RNA ......................................................................................................................... 14
  2.1 Ribonucleic acid composition ................................................................................ 14
     2.1.1 Basic kmer (Kmer)...................................................................................... 14
     2.1.2 Mismatch..................................................................................................... 14
     2.1.3 Subsequence................................................................................................ 15
  2.2 Autocorrelation ...................................................................................................... 15
     2.2.1 Dinucleotide-based auto covariance (DAC) ................................................. 15
     2.2.2 Dinucleotide-based cross covariance (DCC).............................................. 15
     2.2.3 Dinucleotide-based auto-cross covariance (DACC) ..................................... 16
     2.2.4 Moran autocorrelation (MAC) ...................................................................... 16
     2.2.5 Geary autocorrelation (GAC) ....................................................................... 16
     2.2.6 Normalized Moreau–Broto autocorrelation (NMBAC) .............................. 17
  2.3 Pseudo ribonucleic acid composition ..................................................................... 17
     2.3.1 General parallel correlation pseudo dinucleotide composition (PC-PseDNC-General) ......................................................................................... 17
     2.3.2 General series correlation pseudo dinucleotide composition (SC-PseDNC-General) ......................................................................................... 18
  2.4 Predicted structure composition ............................................................................. 19
     2.4.1 Local structure-sequence triplet element (Triplet)....................................... 19
2.4.2 Pseudo-structure status composition (PseSSC) ................................................. 20
2.4.3 Pseudo-distance structure status pair composition (PseDPC) ......................... 21
3. Protein ...................................................................................................................... 22
3.1 Amino acid composition ....................................................................................... 22
3.1.1 Basic kmer (Kmer) ....................................................................................... 22
3.1.2 Distance-based Residue (DR) ....................................................................... 22
3.1.3 PseAAC of Distance-Pairs and reduced alphabet scheme (Distance Pair) ... 22
3.2 Autocorrelation .................................................................................................. 23
3.2.1 Auto covariance (AC) .................................................................................. 23
3.2.2 Cross covariance (CC) .................................................................................. 23
3.2.3 Auto-cross covariance (ACC) ....................................................................... 24
3.2.4 Physicochemical distance transformation (PDT) ......................................... 24
3.3 Pseudo amino acid composition ......................................................................... 24
3.3.1 Parallel correlation pseudo amino acid composition (PC-PseAAC) ....... 24
3.3.2 Series correlation pseudo amino acid composition (SC-PseAAC) ........... 26
3.3.3 General parallel correlation pseudo amino acid composition
             (PC-PseAAC-General) ............................................................................... 28
3.3.4 General series correlation pseudo amino acid composition
             (SC-PseAAC-General) ............................................................................... 29
3.4 Frequency Profile ................................................................................................ 30
3.4.1 Top-n-gram ................................................................................................ 30
3.4.2 Profile-based physicochemical distance transformation (PDT-Profile) .... 30
3.4.3 Distance-based Top-n-gram (DT) ............................................................... 30
3.4.4 Profile-based Auto covariance (AC-PSSM) .............................................. 31
3.4.5 Profile-based Cross covariance (CC-PSSM) .............................................. 31
3.4.6 Profile-based Auto-cross covariance (ACC-PSSM) .................................. 31
Table 1. The names of the 148 physicochemical indices for dinucleotides (DNA) .... 32
Table 2. The names of the 12 physicochemical indices for trinucleotides (DNA) .... 33
Table 3. The names of the 90 physicochemical indices for dinucleotides (DNA) .... 33
Table 4. The names of the 6 physicochemical indices for dinucleotides (DNA) ..... 35
Table 5. The names of the 22 physicochemical indices for dinucleotides (RNA) ... 35
Table 6. The names of the 11 physicochemical indices for dinucleotides (RNA) ... 35
Table 7. The names of the 547 physicochemical indices for amino acids .......... 35
Table 8. The names of the 3 physicochemical indices for amino acids ............... 40
Table 9. The names of the 2 physicochemical indices for amino acids ............... 40
References ............................................................................................................... 40
1. DNA

1.1 Deoxyribonucleic acid composition

1.1.1 Basic kmer (Kmer)

Basic kmer (1) is the simplest approach to represent the DNAs, in which the DNA sequences are represented as the occurrence frequencies of $k$ neighboring nucleic acids. This approach has been successfully applied to human gene regulatory sequence prediction (2,3), enhancer identification (1), etc.

1.1.2 Reverse complementary kmer (RevKmer)

The reverse complementary kmer (2,3) is a variant of the basic kmer, in which the kmers are not expected to be strand-specific, so reverse complements are collapsed into a single feature. For example, if $k=2$, there are totally 16 basic kmers (‘AA’, ’AC’, ’AG’, ’AT’, ’CA’, ’CC’, ’CG’, ’CT’, ’GA’, ’GC’, ’GG’, ’GT’, ’TA’, ’TC’, ’TG’, ’TT’), but by removing the reverse complementary kmers, there are only 10 distinct kmers in the reverse complementary kmer approach (‘AA’, ’AC’, ’AG’, ’AT’, ’CA’, ’CC’, ’CG’, ’GA’, ’GC’, ’TA’). For more information of this approach, please refer to (2,3).

1.1.3 Increment of diversity (IDKmer)

Suppose a DNA sequence $D$ with $L$ nucleic acid residues; i.e.

$$D = R_1R_2R_3R_4R_5R_6\cdots R_L$$

(1)

where $R_1$ represents the nucleic acid residue at the sequence position 1, $R_2$ the nucleic acid residue at position 2 and so forth.

The increment of diversity has been successfully applied in the prediction of exonintron splice sites for several model genomes (4), transcription start site prediction, and studying the organization of nucleosomes around splice sites (4). In this method, the sequence features are converted into the increment of diversity (ID), defined by the relation of sequence $X$ with standard source $S$:

$$ID = \text{Diversity}(X+S) - \text{Diversity}(S) - \text{Diversity}(X)$$

(2)

We obtain an $r$-dimensional feature vector. The feature vector $R$ is designed by the following considerations. The kmers are responsible for the discrimination between positive samples and negative samples, and therefore they construct the diversity sources. Based on this, 2 kmer-based increments of diversities $ID_1$ ($ID_2$) between sequence $D$ and the standard source in positive (negative) training set can be easily introduced as the feature vectors. For more information of this approach, please refer to (5), (6) and (7).
1.1.4 Mismatch

Mismatch (8-10) calculates the occurrences of a k-length neighboring nucleic acids that differ by at most m mismatches \((m < k)\). For a 3-length subsequence “AAC”, and max one mismatch, we need to consider 3 cases, “-AC”, ”A-C” and “AA-”, “-” can be replaced by any nucleic acid residue. The mismatch feature vector of sequence \(D\) (Eq. 1) is defined:

\[
f_{k,m}(D) = \left( \sum_{j=0}^{m} c_{1,j} + \sum_{j=0}^{m} c_{2,j} + \cdots + \sum_{j=0}^{m} c_{4^k,j} \right)
\]

where \(c_{i,j}\) represents the occurrences of \(i\)-th k-mer type in \(D\), with \(j\) mismatches, \(i = 1, 2, ..., 4^k; j = 0, 1, ..., m\).

1.1.5 Subsequence

Subsequence (8,10,11) is an approach that allows non-contiguous matching. For a 3-mer “AAC” in a sequence \(D\) (Eq. 1), we need to consider a pattern, “A*A*C”, “*” can be replaced by 0 or more letters which represents nucleic acid residues, and when “**” represents 0, it represents an exact matching, or represents non-contiguous matching. For each subsequence, there is a dimension of the feature vector and the value of such coordinate depends on its occurrences, length \(l\) and a decay factor \(\delta \in [0,1]\). The subsequence feature vector of sequence \(D\) is defined:

\[
f_{k,m}(D) = \left( \sum_{l(a_j) \text{ is exact matching}}^{k-\text{mer in } x} \delta^{l(a_j)}, \sum_{l(a_j) \text{ is non-contiguous matching}}^{k-\text{mer in } x} \delta^{l(a_j)} \right)
\]

where

\[
l(a_j) = \begin{cases} 0, & a_j \text{ is exact matching;} \\ |a_j|, & a_j \text{ is non-contiguous matching.} \end{cases}
\]

\(|a_j|\) represents the length of \(a_j\), \(i = 1, 2, ..., 4^k\).

1.2 Autocorrelation

1.2.1 Dinucleotide-based auto covariance (DAC)

The DAC (12-14) measures the correlation of the same physicochemical index between two dinucleotides separated by a distance of \(lag\) along the sequence, which can be calculated as:

\[
DAC(u, lag) = \frac{\sum_{i=1}^{L-lag-1} (P_u(R_i,R_{i+lag}) - \bar{P}_u)(P_u(R_{i+lag},R_{i+lag+1}) - \bar{P}_u)}{(L-lag-1)}
\]

where \(u\) is a physicochemical index, \(L\) is the length of the DNA sequence \(D\), \(P_u(R_i,R_{i+lag})\) means the numerical value of the physicochemical index \(u\) for the dinucleotide \(R_iR_{i+1}\) at position \(i\), \(\bar{P}_u\) is the average value for physicochemical index \(u\) along the whole sequence:
In such a way, the length of DAC feature vector is $N \times LAG$, where $N$ is the number of physicochemical indices (Table 1) extracted from two papers (14,15), and LAG is the maximum of lag ($lag = 1, 2, \ldots, LAG$).

1.2.2 Dinucleotide-based cross covariance (DCC)

Given a DNA sequence $D$ (Eq. 1), the DCC (12,14) approach measures the correlation of two different physicochemical indices between two dinucleotides separated by lag nucleic acids along the sequence, which can be calculated by:

$$DCC(u_1, u_2, lag) = \sum_{j=1}^{L-\text{lag}-1} \left( P_{u_1}(R_{i+j}) - \overline{P}_{u_1} \right) \left( P_{u_2}(R_{i+j\text{lag}}) - \overline{P}_{u_2} \right) / (L - \text{lag} - 1)$$

where $u_1, u_2$ are two different physicochemical indices, $L$ is the length of the DNA sequence, $P_{u_1}(R_{i+j})$ is the numerical value of the physicochemical index $u_1$ for the dinucleotide $R_iR_{i+j}$ at position $i$, $\overline{P}_{u_1}$ is the average value for physicochemical index value $u_1$ along the whole sequence:

$$\overline{P}_u = \sum_{j=1}^{L-1} P_u(R_{i+j}) / (L - 1)$$

In such a way, the length of the DCC feature vector is $N \times (N-1) \times LAG$, where LAG is the maximum of lag ($lag = 1, 2, \ldots, LAG$); $N$ is the number of physicochemical indices (Table 1).

1.2.3 Dinucleotide-based auto-cross covariance (DACC)

DACC(12,14) is a combination of DAC and DCC. Therefore, the length of the DACC feature vector is $N \times N \times LAG$, where $N$ is the number of physicochemical indices (Table 1) and LAG is the maximum of lag ($lag = 1, 2, \ldots, LAG$).

1.2.4 Trinucleotide-based auto covariance (TAC)

Given a DNA sequence $D$ (Eq. 1), the TAC approach (12-14) measures the correlation of the same physicochemical index between two trinucleotides separated by lag nucleic acids along the sequence, which can be calculated as:

$$\text{TAC}(\text{lag}, u) = \sum_{i=1}^{L-\text{lag}-2} \left( P_u(R_{i+j1}R_{i+j2}) - \overline{P}_u \right) \left( P_u(R_{i+j\text{lag}1}R_{i+j\text{lag}2}) - \overline{P}_u \right) / (L - \text{lag} - 2)$$

where $u$ is a physicochemical index, $L$ is the length of the DNA sequence, $P_u(R_{i+j1}R_{i+j2})$ represents the numerical value of the physicochemical index $u$ for the trinucleotide $R_{i+j1}R_{i+j2}$ at position $i$, $\overline{P}_u$ is the average value for physicochemical index $u$ along the whole sequence:
\[
\overline{P}_u = \sum_{j=1}^{L} P_u (R_j \cdot R_{j+1} \cdot R_{j+2}) / (L - 2)
\]

In such a way, the length of TAC feature vector is \(N \times \text{LAG}\), where \(N\) is the number of physicochemical indices (Table 2) extracted from (14), and LAG is the maximum of \(\text{lag} (\text{lag}=1, 2, \ldots, \text{LAG})\).

### 1.2.5 Trinucleotide-based cross covariance (TCC)

Given a DNA sequence \(D\) (Eq. 1), the TCC(12,14) approach measures the correlation of two different physicochemical indices between two trinucleotides separated by \(\text{lag}\) nucleic acids along the sequence, which can be calculated by:

\[
\text{TCC}(u_1, u_2, \text{lag}) = \sum_{i=1}^{L-\text{lag}+2} \left( P_{u_1}(R_i \cdot R_{i+1} \cdot R_{i+2}) - \overline{P}_{u_1} \right) \left( P_{u_2}(R_{i+\text{lag}} \cdot R_{i+\text{lag}+1} \cdot R_{i+\text{lag}+2}) - \overline{P}_{u_2} \right) / (L-\text{lag}-2)
\]

where \(u_1, u_2\) are two physicochemical indices; \(L\) is the length of the DNA sequence; \(P_{u_i}(R_i \cdot R_{i+1} \cdot R_{i+2})\) represents the numerical value of the physicochemical index \(u_1\) (\(u_2\)) for the trinucleotide \(R_i \cdot R_{i+1} \cdot R_{i+2}\) at position \(i\); \(\overline{P}_{u_i}\) (\(\overline{P}_{u_2}\)) is the average value for physicochemical index \(u_1\) (\(u_2\)) along the whole sequence:

\[
\overline{P}_u = \sum_{j=1}^{L-2} P_u (R_j \cdot R_{j+1} \cdot R_{j+2}) / (L - 2)
\]

In such a way, the length of TCC feature vector is \(N \times (N-1) \times \text{LAG}\), where \(N\) is the number of physicochemical index (Table 2) extracted from (14), and LAG is the maximum of \(\text{lag} (\text{lag}=1, 2, \ldots, \text{LAG})\).

### 1.2.6 Trinucleotide-based auto-cross covariance (TACC)

TACC (12,14) is a combination of TAC and TCC. Therefore, the length of the TACC feature vector is \(N \times N \times \text{LAG}\), where \(N\) is the number of physicochemical indices (Table 2) extracted from (14), and LAG is the maximum of \(\text{lag} (\text{lag}=1, 2, \ldots, \text{LAG})\).

### 1.2.7 Moran autocorrelation (MAC)

Given a DNA sequence \(D\) (Eq. 1), the MAC (14,16) approach measures the correlation of the same properties between two residues separated by a distance of \(\text{lag}\) along the sequence, which can be calculated by:

\[
\text{MAC}(u, k, \text{lag}) = \left[ 1 / (L-\text{lag}-k+1) \right] \sum_{i=1}^{L-\text{lag}-k+1} \left( P_u(x_i) - \overline{P}_u(x) \right) \left( P_u(x_{i+k}) - \overline{P}_u(x) \right) / \left( 1 / L-k+1 \right) \sum_{i=1}^{L-k+1} \left( P_u(x_i) - \overline{P}_u(x) \right)^2
\]

where \(u\) is a physicochemical index, \(L\) is the length of the DNA sequence, \(x\) represents trinucleotide or dinucleotide. When \(x\) represents dinucleotide, the value of \(k\) is 2, its
corresponding physicochemical indices are listed in Table 3. When \( x \) represents trinucleotide, the value of \( k \) is 3, its corresponding physicochemical indices are listed in Table 2. \( P_u(x) \) represents the numerical value of the physicochemical index \( u \) for \( x \) at position \( i \), \( \bar{P}_u(x) \) is the average value for physicochemical index \( u \) along the whole sequence. When \( \text{lag} = 1 \), the nearest neighbor correlations at a physicochemical property \( u \) are measured; When \( \text{lag} = 2 \), next second nearest neighbor correlation are considered, and so on.

### 1.2.8 Geary autocorrelation (GAC)

Given a DNA sequence \( D \) (Eq. 1), the GAC (14,17) approach measures the correlation of the same properties between two residues separated by a distance of \( \text{lag} \) along the sequence, which can be calculated by:

\[
\text{GAC}(u, k, \text{lag}) = \frac{\left[1/(L-\text{lag}-k+1)\right] \sum_{i=1}^{L-\text{lag}-k+1} \left(P_u(x_i) - P_u(x_{i+\text{lag}})\right)^2}{(1/L-k+1) \sum_{i=1}^{L-k+1} \left(P_u(x_i) - \bar{P}_u(x)\right)^2}
\]

where \( u \) is a physicochemical index, \( L \) is the length of the DNA sequence, \( x \) represents trinucleotide or dinucleotide. When \( x \) represents dinucleotide, the value of \( k \) is 2, its corresponding physicochemical indices are listed in Table 3. When \( x \) represents trinucleotide, the value of \( k \) is 3, its corresponding physicochemical indices are listed in Table 2. \( P_u(x) \) means the numerical value of the physicochemical index \( u \) for \( x \) at position \( i \), \( \bar{P}_u(x) \) is the average value for physicochemical index \( u \) along the whole sequence. When \( \text{lag} = 1 \), the nearest neighbor correlations at a physicochemical property \( u \) are measured; When \( \text{lag} = 2 \), next second nearest neighbor correlations are considered, and so on.

### 1.2.9 Normalized Moreau–Broto autocorrelation (NMBAC)

Given a DNA sequence \( D \) (Eq. 1), the NMBAC (14,18) approach measures the correlation of the same properties between two residues separated by a distance of \( \text{lag} \) along the sequence, which can be calculated by:

\[
\text{NMBAC}(u, k, \text{lag}) = \frac{\sum_{i=1}^{L-\text{lag}-k+1} \left(P_u(x_i) \times P_u(x_{i+\text{lag}})\right)^2}{L-k-\text{lag}+1}
\]

where \( u \) is a physicochemical index, \( L \) is the length of the DNA sequence, \( x \) represents trinucleotide or dinucleotide. When \( x \) represents dinucleotide, the value of \( k \) is 2, its corresponding physicochemical indices are listed in Table 3. When \( x \) represents trinucleotide, the value of \( k \) is 3, its corresponding physicochemical indices are listed in Table 2. \( P_u(x) \) means the numerical value of the physicochemical index \( u \) for \( x \) at position \( i \). When \( \text{lag} = 1 \), the nearest neighbor correlations at a physicochemical property \( u \) are measured; When \( \text{lag} = 2 \), next second nearest neighbor correlations are considered, and so on.
1.3 Pseudo deoxyribonucleic acid composition

1.3.1 Pseudo dinucleotide composition (PseDNC)

PseDNC (19) is an approach incorporating the contiguous local sequence-order information and the global sequence-order information into the feature vector of the DNA sequence.

Given a DNA sequence $D$ (Eq. 1), the PseDNC feature vector of $D$ is defined:

$$D = [d_1 \ d_2 \ \cdots \ d_{16} \ d_{16+1} \ \cdots \ d_{16+\lambda}]^T$$

where

$$d_k = \left\{ \begin{array}{ll}
\frac{f_k}{\sum_{i=1}^{16} f_i + w \sum_{j=1}^{\lambda} \theta_j} & (1 \leq k \leq 16) \\
\frac{w \theta_{k-16}}{\sum_{i=1}^{16} f_i + w \sum_{j=1}^{\lambda} \theta_j} & (17 \leq k \leq 16 + \lambda)
\end{array} \right.$$

where $f_k (k=1,2,\cdots,16)$ is the normalized occurrence frequency of dinucleotides in the DNA sequence; the parameter $\lambda$ is an integer, representing the highest counted rank (or tier) of the correlation along a DNA sequence; $w$ is the weight factor ranged from 0 to 1; $\theta_j (j=1,2,\cdots,\lambda)$ is called the j-tier correlation factor that reflects the sequence-order correlation between all the most contiguous dinucleotides along a DNA sequence, which is defined:

$$\theta_j = \left\{ \begin{array}{ll}
\frac{1}{L-2} \sum_{i=1}^{L-2} \Theta(R_{j_i}, R_{j_{i+1}}, R_{j_{i+2}}) & (\lambda < L) \\
\frac{1}{L-3} \sum_{i=1}^{L-3} \Theta(R_{j_i}, R_{j_{i+1}}, R_{j_{i+2}}, R_{j_{i+3}}) & \\
\frac{1}{L-4} \sum_{i=1}^{L-4} \Theta(R_{j_i}, R_{j_{i+1}}, R_{j_{i+2}}, R_{j_{i+3}}, R_{j_{i+4}}) & \\
\cdots \\
\frac{1}{L-1-\lambda} \sum_{i=1}^{L-1-\lambda} \Theta(R_{j_i}, R_{j_{i+1}}, R_{j_{i+2}}, R_{j_{i+3}}, \cdots, R_{j_{i+\lambda+1}})
\end{array} \right.$$

where the correlation function is given by

$$\Theta(R_{j_i}, R_{j_{i+1}}, R_{j_{i+2}}, R_{j_{i+3}}) = \frac{1}{\mu} \sum_{u=1}^{6} [P_u(R_{j_i}, R_{j_{i+1}}) - P_u(R_{j_i}, R_{j_{i+1}})]^2$$

where $\mu$ is the number of physicochemical indices, in this approach, 6 indices reflecting the local DNA structural properties (19) (Table 4) are employed to generate the PseDNC feature vector; $P_u(R_{j_i}, R_{j_{i+1}})$ represents the numerical value of the $u$-th ($u=1,2,\cdots,6$) physicochemical index of the dinucleotide $R_{j_i} R_{j_{i+1}}$ at position $i$ ($j_i$).
1.3.2 Pseudo k-tuple nucleotide composition (PseKNC)

PseKNC (20,21) extends the PseDNC approach by incorporating k-tuple nucleotide composition.

Given a DNA sequence \(D\) (Eq. 1), the feature vector of \(D\) is defined:

\[
D = \begin{bmatrix}
    d_1 & d_2 & \cdots & d_4 & d_{4+1} & \cdots & d_{4+k}
\end{bmatrix}^T
\]  

(21)

where

\[
d_u = \frac{\sum_{i=1}^{d_u} f_u + w \sum_{j=1}^{\lambda} \theta_j}{\sum_{i=1}^{d_u} f_u + w \sum_{j=1}^{\lambda} \theta_j}
\]  

(22)

where \(\lambda\) is the number of the total counted ranks (or tiers) of the correlations along a DNA sequence; \(f_u\) (\(u=1,2,\cdots,4^k\)) is the frequency of oligonucleotide that is normalized to \(\sum_{i=1}^{d_u} f_u = 1\); \(w\) is a weight factor; \(\theta_j\) is given by

\[
\theta_j = \frac{1}{L-j-1} \sum_{i=1}^{L-j-1} \Theta(R_{i+j}, R_{i+j+1}) \quad (j = 1, 2, \cdots, \lambda; \lambda < L)
\]  

(23)

which represents the \(j\)-tier structural correlation factor between all the \(j\)-th most contiguous dinucleotides. The correlation function \(\Theta(R_{i+j}, R_{i+j+1})\) is defined by

\[
\Theta(R_{i+j}, R_{i+j+1}) = \frac{1}{\mu} \sum_{v=1}^{\mu} \left( P_v(R_{i+j}) - P_v(R_{i+j+1}) \right)^2
\]  

(24)

where \(\mu\) is the number of physicochemical indices, in this study, 6 indices reflecting the local DNA structural properties (19) (Table 4) are employed to generate the PseKNC feature vector; \(P_v(R_{i+j})\) \((P_v(R_{i+j}, R_{i+j+1}))\) represents the numerical value of the \(v\)-th \((v = 1,2,\cdots,\mu)\) physicochemical index for the dinucleotide \(R_{i+j}, R_{i+j+1}\) at position \(i+j\).

For more information about this approach, please refer to (20,21).

1.3.3 General parallel correlation pseudo dinucleotide composition (PC-PseDNC-General)

In PC-PseDNC-General (22) approach, the users cannot only select the 148 built-in physiochemical indices (Table 1), but also can upload their own indices to generate the PC-PseDNC-General feature vector.

Given a DNA sequence \(D\) (Eq. 1), the PC-PseDNC-General feature vector of \(D\) is defined:

\[
D = \begin{bmatrix}
    d_1 & d_2 & \cdots & d_{16} & d_{16+1} & \cdots & d_{16+2\lambda}
\end{bmatrix}^T
\]  

(25)

where
where $f_k$ ($k=1,2,\cdots,16$) is the normalized occurrence frequency of dinucleotides in the DNA sequence; the parameter $\lambda$ is an integer, representing the highest counted rank (or tier) of the correlation along a DNA sequence; $w$ is the weight factor ranging from 0 to 1; $\theta_j$ ($j=1, 2, \cdots, \lambda$) is called the $j$-tier correlation factor that reflects the sequence-order correlation between all the most contiguous dinucleotides along a DNA sequence, which is defined:

$$
\theta_j = \begin{cases} 
\frac{1}{L-2} \sum_{i=1}^{L-2} \Theta(R_{i+1}, R_{i+1}R_{i+2}) \\
\frac{1}{L-3} \sum_{i=1}^{L-3} \Theta(R_{i+2}, R_{i+2}R_{i+3}) \\
\frac{1}{L-4} \sum_{i=1}^{L-4} \Theta(R_{i+3}, R_{i+3}R_{i+4}) \\
\cdots \\
\frac{1}{L-\lambda} \sum_{i=1}^{L-\lambda} \Theta(R_{i+\lambda}, R_{i+\lambda}R_{i+\lambda+1}) 
\end{cases}$$

(27)

where the correlation function is given by

$$
\Theta(R_{i+u}, R_{i+u}R_{i+1}) = \frac{1}{\mu} \sum_{u=1}^{\mu} (P_u(R_{i+u}) - P_u(R_{i+1}))^2
$$

(28)

where $\mu$ is the number of physicochemical indices listed in the Table 1; $P_u(R_{i+u})$ ($P_u(R_{i+1})$) represents the numerical value of the $u$-th ($u=1,2,\cdots,\mu$) physicochemical index for the dinucleotide $R_{i+u}$ ($R_{i+1}$) at position $i$ ($j$).

### 1.3.4 General parallel correlation pseudo trinucleotide composition (PC-PseTNC-General)

In PC-PseTNC-General (22) approach, the users cannot only select the 12 built-in physiochemical indices (Table 2), but also can upload their own indices to generate the PC-PseTNC-General feature vector.

Given a DNA sequence $D$ (Eq. 1), the PC-PseTNC-General feature vector of $D$ is defined:

$$
D = [d_1, d_2, \cdots, d_{64}, d_{64+1}, \cdots, d_{64+\lambda}]^T
$$

(29)

where

$$
d_k = \begin{cases} 
\frac{f_k}{\sum_{i=1}^{64} f_i + w \sum_{j=1}^{k} \theta_j} & (1 \leq k \leq 64) \\
\frac{w \theta_{k-64}}{\sum_{i=1}^{64} f_i + w \sum_{j=1}^{k} \theta_j} & (64 + 1 \leq k \leq 64 + \lambda)
\end{cases}
$$

(30)
where $f_k$ ($k=1,2,\cdots,16$) is the normalized occurrence frequency of dinucleotide in the DNA sequence; the parameter $\lambda$ is an integer, representing the highest counted rank (or tier) of the correlation along a DNA sequence; $w$ is the weight factor ranging from 0 to 1; $\theta_j$ ($j=1,2,\cdots,\lambda$) is called the $j$-tier correlation factor that reflects the sequence-order correlation between all the most contiguous dinucleotides along a DNA sequence, which is defined:

$$\theta_j = \left\{ \begin{array}{ll}
\frac{1}{L-3} \sum_{i=1}^{L-3} \Theta(R_{i+i}, R_{i+i}, R_{i+i}, R_{i+i}) \\
\frac{1}{L-4} \sum_{i=1}^{L-4} \Theta(R_{i+i}, R_{i+i}, R_{i+i}, R_{i+i}) \\
\frac{1}{L-5} \sum_{i=1}^{L-5} \Theta(R_{i+i}, R_{i+i}, R_{i+i}, R_{i+i}) \\
\ldots \\
\frac{1}{L-2+\lambda} \sum_{i=1}^{L-2+\lambda} \Theta(R_{i+i}, R_{i+i}, R_{i+i}, R_{i+i}) \\
\end{array} \right. 
$$

(31)

where the correlation function is given by

$$\Theta(R_{i+i}, R_{i+i}, R_{i+i}, R_{i+i}) = \frac{1}{\mu} \sum_{u=1}^{\mu} \left[ P_u(R_{i+i}, R_{i+i}, R_{i+i}) - P_u(R_{j+i}, R_{j+i}) \right]^2 
$$

(32)

where $\mu$ is the number of physicochemical indices considered that are listed in the Table 2; $P_u(R_{i+i}, R_{i+i}, R_{i+i}, R_{i+i})$ represents the numerical value of the $u$-th ($u=1,2,\cdots,\mu$) physicochemical index for the tri-nucleotide $R_{i+i}, R_{i+i}, R_{i+i}$ at position $i (j)$.

### 1.3.5 General series correlation pseudo dinucleotide composition (SC-PseDNC-General)

SC-PseDNC-General (22) is a variant of PC-PseDNC-General, which differs in the equations of calculating the correlation factors reflecting the sequence-order correlation between all the most contiguous dinucleotides along a DNA sequence. Given a DNA sequence $D$ (Eq. 1), the SC-PseDNC-General feature vector of $D$ is defined:

$$D = [d_1 \ d_2 \ \cdots \ d_16 \ d_{16+1} \ \cdots \ d_{16+\lambda} \ \cdots \ d_{16+\lambda+1} \ \cdots \ d_{16+\lambda+\Lambda}]^T 
$$

(33)

where

$$d_k = \begin{cases}
\frac{f_k}{\sum_{i=1}^{16} f_i + w \sum_{i=1}^{\Lambda} \theta_j} & (1 \leq k \leq 16) \\
\frac{w \theta_k}{\sum_{i=1}^{16} f_i + w \sum_{i=1}^{\Lambda} \theta_j} & (17 \leq k \leq 16 + \lambda \Lambda) 
\end{cases} 
$$

(34)

where $f_k$ ($k=1,2,\cdots,16$) is the normalized occurrence frequency of dinucleotide in the DNA sequence; the parameter $\lambda$ is an integer, representing the highest counted rank (or tier) of the correlation along a DNA sequence; $w$ is the weight factor ranging from 0 to 1; $\Lambda$ is the number of physicochemical indices (Table 1); $\theta_j$ ($j=1,2,\cdots,\lambda$) is
called the *j*-tier correlation factor that reflects the sequence-order correlation between all the most contiguous dinucleotides along a DNA sequence, which is defined:

\[
\begin{align*}
\theta_1 &= \frac{1}{L-3} \sum_{i=1}^{L-3} J_{i,i+1} \\
\theta_2 &= \frac{1}{L-3} \sum_{i=1}^{L-3} J_{i+1,i+2} \\
\vdots & \\
\theta_{\lambda} &= \frac{1}{L-3} \sum_{i=1}^{L-3} J_{i,i+\lambda} \\
\theta_{\lambda+1} &= \frac{1}{L-\lambda-2} \sum_{i=1}^{L-\lambda-2} J_{i,i+\lambda+1} \\
\theta_{\lambda+2} &= \frac{1}{L-\lambda-2} \sum_{i=1}^{L-\lambda-2} J_{i,i+\lambda+2} \\
\end{align*}
\]

(35)

The correlation function is given by

\[
J_{i,i+m}^{u,\lambda} = P_u(R_{i+1}, R_{i+2}, \ldots, R_{i+m}) 
\prod_{p=1}^{\lambda} P_p(R_{i+p}, R_{i+m+p})
\]

(36)

where \( P_u(R_{i}, R_{i+1}) \) represents the numerical value of the *u*-th physicochemical index for the dinucleotide \( R_{i} R_{i+1} \) at position \( i \) \((i+m)\).

### 1.3.6 General series correlation pseudo trinucleotide composition (SC-PseTNC-General)

SC-PseTNC-General (22) is a variant of PC-PseTNC-General, which differs in the equations of calculating the correlation factors reflecting the sequence-order correlation between all the most contiguous dinucleotides along a DNA sequence. Given a DNA sequence \( D \) (Eq. 1), the SC-PseTNC-General feature vector of \( D \) is defined:

\[
D = [d_1 \ d_2 \ \cdots \ d_{64} \ d_{64+1} \ \cdots \ d_{64+\lambda} \ d_{64+\lambda+1} \ \cdots \ d_{64+\lambda\Lambda}]^T
\]

(37)

where

\[
d_k = \begin{cases} 
\frac{f_k}{\sum_{i=1}^{64} f_i + \sum_{j=1}^{\lambda\Lambda} \theta_j} & (1 \leq k \leq 64) \\
\frac{w \theta_k}{\sum_{i=1}^{64} f_i + \sum_{j=1}^{\lambda\Lambda} \theta_j} & (64 + 1 \leq k \leq 64 + \lambda\Lambda) 
\end{cases}
\]

(38)

where \( f_k \) \((k=1, 2, \cdots, 64)\) is the normalized occurrence frequency of trinucleotide in the DNA sequence; the parameter \( \lambda \) is an integer, representing the highest counted rank (or tier) of the correlation along a DNA sequence; \( w \) is the weight factor ranging from 0 to 1; \( \Lambda \) is the number of physicochemical indices (Table 2); \( \theta_j \) \((j = 1, 2, \cdots, \lambda)\) is called the *j*-tier correlation factor reflecting the sequence-order correlation between all the most contiguous trinucleotides along a DNA sequence, which is defined:
The correlation function is given by

\[
\theta_i = \frac{1}{L-4} \sum_{j=1}^{L-4} J_{i,j+1}^1 \\
\theta_2 = \frac{1}{L-4} \sum_{j=1}^{L-4} J_{i,j+1}^2 \\
\ldots \\
\theta_{\lambda} = \frac{1}{L-4} \sum_{j=1}^{L-4} J_{i,j+1}^\lambda \\
\theta_{\lambda+1} = \frac{1}{L-4} \sum_{j=1}^{L-4} J_{i,j+1}^{\lambda+1} \\
\theta_{\lambda-1} = \frac{1}{L-4} \sum_{j=1}^{L-4} J_{i,j+1}^{\lambda-1} \\
\theta_{\lambda} = \frac{1}{L-4} \sum_{j=1}^{L-4} J_{i,j+1}^{\lambda} \\
\theta_{\lambda=1} = \frac{1}{L-4} \sum_{j=1}^{L-4} J_{i,j+1}^{\lambda=1} \\
\theta_{\lambda} = \frac{1}{L-4} \sum_{j=1}^{L-4} J_{i,j+1}^{\lambda} \\
\theta_{\lambda} = \frac{1}{L-4} \sum_{j=1}^{L-4} J_{i,j+1}^{\lambda}(39)
\]

The correlation function is given by

\[
I^u_{i,j+1} = P_i(R, R_{i+1}, R_{i+2}) \cdot P_m(R_{i+m}, R_{i+m+1}, R_{i+m+2}) \\
( u = 1, 2, \ldots, \lambda; \; m=1, 2, \ldots, \lambda; \; i=1, 2, \ldots, L-m-2)
\]

where \( P_i(R, R_{i+1}, R_{i+2}) \) (\( P_m(R_{i+m}, R_{i+m+1}, R_{i+m+2}) \)) represents the numerical value of the \( u \)-th \(( u = 1, 2, \ldots, \mu)\) physiochemical index for the tri-nucleotide \( R_{i,m}R_{i,m+1}R_{i,m+2} \) at position \( i (i+m) \).

2. RNA

2.1 Ribonucleic acid composition

2.1.1 Basic kmer (Kmer)

Basic kmer (23) is the simplest approach to represent the RNAs, in which the RNA sequences are represented as the occurrence frequencies of \( k \) neighboring nucleic acids.

2.1.2 Mismatch

Suppose an RNA sequence \( R \) with \( L \) nucleic acid residues; i.e.

\[
R = R_1R_2R_3R_4R_5R_6R_7\ldots R_L
\]

where \( R_1 \) represents the nucleic acid residue at the sequence position 1, \( R_2 \) the nucleic acid residue at position 2, and so forth. Mismatch (8-10) calculates the occurrences of a \( k \)-length neighboring nucleic acids that differ by at most \( m \) mismatches \(( m < k) \). For a 3-length subsequence “AAC”, and max one mismatch, we need to consider 3 cases, “-AC”, ”A-C” and “AA-”, “-” can be replaced by any nucleic acid residue. The mismatch feature vector of sequence \( R \) is defined:
\[ f_{k,m}(\mathbf{R}) = \left\{ \sum_{j=0}^{m} c_{1,j} + \sum_{j=0}^{m} c_{2,j} + \ldots + \sum_{j=0}^{m} c_{d,x,j} \right\} \]  \hspace{2cm} (42)

where \( c_{i,j} \) represents the occurrences of \( i \)-th \( k \)-mer type in \( \mathbf{R} \), with \( j \) mismatches, \( i = 1, 2, \ldots, 4^k; j = 0, 1, \ldots, m \).

### 2.1.3 Subsequence

Subsequence (8, 10, 11) is an approach that allows non-contiguous matching. For a 3-mer “AAC” in a sequence \( \mathbf{R} \) (Eq. 41), we need to consider a pattern, “A*A*C”, “*” can be replaced by 0 or more letters which represents nucleic acid residues, and when “*” represents 0, it represents an exact matching, or represents non-contiguous matching. For each subsequence, there is a dimension of the feature vector and the value of such coordinate depends on its occurrences, length \( l \) and a decay factor \( \delta \in [0, 1] \). The subsequence feature vector of sequence \( \mathbf{R} \) is defined:

\[ f_{k,m}(x) = \left( \sum_{k \text{-mer } a_i \text{ in } x} \delta^{l(a_i)} \sum_{k \text{-mer } a_j \text{ in } x} \delta^{l(a_j)} \ldots \sum_{k \text{-mer } a_{\alpha_i} \text{ in } x} \delta^{l(a_{\alpha_i})} \right) \]  \hspace{2cm} (43)

where

\[ l(a_i) = \begin{cases} 0, & a_i \text{ is exact matching;} \\ |a_i|, & a_i \text{ is non-contiguous matching}. \end{cases} \]  \hspace{2cm} (44)

\[ |a_i| \text{ represents the length of } a_i, i = 1, 2, \ldots, 4^k. \]

### 2.2 Autocorrelation

#### 2.2.1 Dinucleotide-based auto covariance (DAC)

The DAC(12-14) measures the correlation of the same physicochemical index between two dinucleotides separated by a distance of \( \text{lag} \) along the sequence, which can be calculated as:

\[ \text{DAC}(u, \text{lag}) = \sum_{i=1}^{L-\text{lag}+1} (P_u(R_i R_{i+1}) - \overline{P}_u)(P_u(R_{i+\text{lag} R_{i+\text{lag}+1}}) - \overline{P}_u) / (L - \text{lag} - 1) \]  \hspace{2cm} (45)

where \( u \) is a physicochemical index; \( L \) is the length of the RNA sequence \( \mathbf{R} \) (Eq. 41), \( P_u(R_i R_{i+1}) \) (\( P_u(R_{i+\text{lag} R_{i+\text{lag}+1}}) \)) means the numerical value of the physicochemical index \( u \) for the dinucleotide \( R_i R_{i+1} \) (\( R_{i+\text{lag}} R_{i+\text{lag}+1} \)) at position \( i \) (\( j \)), \( \overline{P}_u \) is the average value for physicochemical index \( u \) along the whole sequence:

\[ \overline{P}_u = \sum_{j=1}^{L-1} P_u(R_j R_{j+1}) / (L - 1) \]  \hspace{2cm} (46)

In such a way, the length of DAC feature vector is \( N \times \text{LAG} \), where \( N \) is the number of physicochemical indices (Table 5), which are extracted from (14,15), and LAG is the maximum of \( \text{lag} \) (\( \text{lag} = 1, 2, \ldots, \text{LAG} \)).

#### 2.2.2 Dinucleotide-based cross covariance (DCC)
Given an RNA sequence $R$ (Eq. 41), the DCC (12,14) approach measures the correlation of two different physicochemical indices between two dinucleotides separated by $\text{lag}$ nucleic acids along the sequence, which can be calculated by:

$$\text{DCC}(u_1, u_2, \text{lag}) = \sum_{i=1}^{L-\text{lag}-1} (P_{u_1}(R_iR_{i+1}) - \overline{P}_{u_1})(P_{u_2}(R_{i+\text{lag}}R_{i+\text{lag}+1}) - \overline{P}_{u_2}) / (L - \text{lag} - 1) \quad (47)$$

where $u_1$, $u_2$ are two different physicochemical indices, $L$ is the length of the RNA sequence, $P_{u_i}(R_iR_{i+1})$ is the numerical value of the physicochemical index $u_i$ for the dinucleotide $R_iR_{i+1}$ at position $i$, $\overline{P}_{u_i}$ is the average value for physicochemical index value $u_i$ along the whole sequence:

$$\overline{P}_{u_i} = \frac{1}{L} \sum_{j=1}^{L-1} P_{u_i}(R_jR_{j+1}) / (L - 1) \quad (48)$$

In such a way, the length of the DCC feature vector is $N^* (N-1)^* \text{LAG}$, where $N$ is the number of physicochemical indices (Table 5) and LAG is the maximum of $\text{lag}$ ($\text{lag} = 1, 2, \ldots, \text{LAG}$).

### 2.2.3 Dinucleotide-based auto-cross covariance (DACC)

DACC (12,14) is a combination of DAC and DCC. Therefore, the length of the DACC feature vector is $N^* N^* \text{LAG}$, where $N$ is the number of physicochemical indices (Table 5) and LAG is the maximum of $\text{lag}$ ($\text{lag} = 1, 2, \ldots, \text{LAG}$).

### 2.2.4 Moran autocorrelation (MAC)

Given a RNA sequence $R$ (Eq. 41), the MAC (14,16) approach measures the correlation of the same properties between two residues separated by a distance of $\text{lag}$ along the sequence, which can be calculated by:

$$\text{MAC}(u, k, \text{lag}) = \frac{1 / (L - \text{lag} - k + 1)}{1 / k + 1} \sum_{i=1}^{L-\text{lag}-k+1} (P_u(x_i) - \overline{P}_u(x))(P_u(x_{i+k}) - \overline{P}_u(x)) \quad (49)$$

where $u$ is a physicochemical index, $L$ is the length of the RNA sequence, $x$ represents dinucleotide, its corresponding physicochemical indices are listed in Table 6. $P_u(x)$ means the numerical value of the physicochemical index $u$ for $x$ at position $i$, $\overline{P}_u(x)$ is the average value for physicochemical index $u$ along the whole sequence. When $\text{lag} = 1$, the nearest neighbor correlations at a physicochemical property $u$ are measured; When $\text{lag} = 2$, next second nearest neighbor correlations are considered, and so on.

### 2.2.5 Geary autocorrelation (GAC)

Given a RNA sequence $R$ (Eq. 41), the GAC (14,17) approach measures the correlation of the same properties between two residues separated by a distance of $\text{lag}$
along the sequence, which can be calculated by:

$$\text{GAC}(u,k,\text{lag}) = \frac{1}{(L-\text{lag}-k+1)} \sum_{i=1}^{L-\text{lag}-k+1} \left( P_u(x_i) - P_u(x_{i+\text{lag}}) \right)^2$$

(50)

where $u$ is a physicochemical index, $L$ is the length of the RNA sequence, $x$ represents dinucleotide, its corresponding physicochemical indices are listed in Table 6. $P_u(x)$ means the numerical value of the physicochemical index $u$ for $x$ at position $i$. $\bar{P}_u(x)$ is the average value for physicochemical index $u$ along the whole sequence. When $\text{lag} = 1$, the nearest neighbor correlations at a physicochemical property $u$ are measured; When $\text{lag} = 2$, next second nearest neighbor correlations are considered, and so on.

2.2.6 Normalized Moreau–Broto autocorrelation (NMBAC)

Given a RNA sequence $R$ (Eq. 41), the NMBAC (14,18) approach measures the correlation of the same properties between two residues separated by a distance of $\text{lag}$ along the sequence, which can be calculated by:

$$\text{NMBAC}(u,\text{lag}) = \sum_{i=1}^{L-\text{lag}-1} \left( P_u(x_i) \times P_u(x_{i+\text{lag}}) \right)^2$$

(51)

where $u$ is a physicochemical index, $L$ is the length of the RNA sequence, $x$ represents dinucleotide, its corresponding physicochemical indices are listed in Table 6. $P_u(x)$ means the numerical value of the physicochemical index $u$ for $x$ at position $i$. When $\text{lag} = 1$, the nearest neighbor correlations at a physicochemical property $u$ are measured; When $\text{lag} = 2$, next second nearest neighbor correlations are considered, and so on.

2.3 Pseudo ribonucleic acid composition

2.3.1 General parallel correlation pseudo dinucleotide composition

(PC-PseDNC-General)

In PC-PseDNC-General (14) approach, the users cannot only select the 22 built-in physiochemical indices (Table 5), but also can upload their own indices to generate the PC-PseDNC-General feature vector.

Given an RNA sequence $R$ (Eq. 41), the PC-PseDNC-General feature vector of $R$ is defined:

$$R = [d_1 \ d_2 \ \cdots \ d_{16} \ d_{16+1} \ \cdots \ d_{16+\text{lag}}]^T$$

(52)

where
where $f_k$ ($k=1,2,\cdots,16$) is the normalized occurrence frequency of dinucleotide in the RNA sequence; the parameter $\lambda$ is an integer, representing the highest counted rank (or tier) of the correlation along a RNA sequence; $w$ is the weight factor ranging from 0 to 1; $\theta_j$ ($j=1,2,\cdots,\lambda$) is called the $j$-tier correlation factor reflecting the sequence-order correlation between all the most contiguous dinucleotides along an RNA sequence, which is defined:

\[
\begin{align*}
\theta_1 &= \frac{1}{L-2} \sum_{i=1}^{L-3} \Theta(R, R_{i+1}, R_{i+3}, R_{i+5}) \\
\theta_2 &= \frac{1}{L-3} \sum_{i=1}^{L-4} \Theta(R, R_{i+1}, R_{i+2}, R_{i+5}) \\
\theta_3 &= \frac{1}{L-4} \sum_{i=1}^{L-5} \Theta(R, R_{i+1}, R_{i+3}, R_{i+5}) \\
& \quad \vdots \\
\theta_\lambda &= \frac{1}{L-\lambda-1} \sum_{i=1}^{L-\lambda-2} \Theta(R, R_{i+1}, R_{i+3}, R_{i+5})
\end{align*}
\]

where the correlation function is given by

\[
\Theta(R, R_{i+1}, R_j R_{j+1}) = \frac{1}{\mu} \sum_{u=1}^{\mu} \left[ P_u(R_{i+1}) - P_u(R_j R_{j+1}) \right]^2
\]

where $\mu$ is the number of physicochemical indices considered that are listed in the Table 5: $P_u(R_{i+1})$ ($P_u(R_j R_{j+1})$) represents the numerical value of the $u$-th ($u=1,2,\cdots,\mu$) physicochemical index for the dinucleotide $R_{i+1}$ ($R_j R_{j+1}$) at position $i$ ($j$).

**2.3.2 General series correlation pseudo dinucleotide composition (SC-PseDNC-General)**

SC-PseDNC-General (14) is a variant of PC-PseDNC-General, which differs in the equations of calculating the correlation factors reflecting the sequence-order correlation between all the most contiguous dinucleotides along an RNA sequence. Given an RNA sequence $R$ (Eq. 41), the SC-PseDNC-General feature vector of $R$ is defined:

\[
R = \begin{bmatrix}
  d_1 & d_2 & \cdots & d_{16} & d_{16+1} & \cdots & d_{16+\lambda} & d_{16+\lambda+1} & \cdots & d_{16+\lambda+\Lambda}
\end{bmatrix}^T
\]
where \( f_k \) \((k=1, 2, \ldots, 16)\) is the normalized occurrence frequency of dinucleotides in the RNA sequence; the parameter \( \lambda \) is an integer, representing the highest counted rank (or tier) of the correlation along an RNA sequence; \( w \) is the weight factor ranging from 0 to 1; \( \Lambda \) is the number of physicochemical indices (Table 5); \( \theta_j \) \((j = 1, 2, \ldots, \lambda)\) is called the \( j \)-tier correlation factor reflecting the sequence-order correlation between all the most contiguous dinucleotides along an RNA sequence, which is defined:

\[
\begin{align*}
\theta_1 & = \frac{1}{L-3} \sum_{i=1}^{L-3} J_{i,i+1}^1 \\
\theta_2 & = \frac{1}{L-3} \sum_{i=1}^{L-3} J_{i,i+1}^2 \\
& \\
& \\
\theta_{\lambda} & = \frac{1}{L-3} \sum_{i=1}^{L-3} J_{i,i+1}^\lambda \\
& \\
\theta_{\lambda-1} & = \frac{1}{L-\lambda-2} \sum_{i=1}^{L-\lambda-2} J_{i,i+1}^{\lambda-1} \\
\theta_{\lambda} & = \frac{1}{L-\lambda-2} \sum_{i=1}^{L-\lambda-2} J_{i,i+1}^\lambda
\end{align*}
\]

The correlation function is given by

\[
\begin{align*}
J_{i,i+m}^u &= P_u(R_{i}, R_{i+1}) \cdot P_u(R_{i+m}, R_{i+m+1}) \\
u &= 1, 2, \ldots, \Lambda; \ m &= 1, 2, \ldots, \lambda; \ i &= 1, 2, \ldots, L-\lambda-2
\end{align*}
\]

\( P_u(R_{i}, R_{i+1}) \cdot P_u(R_{i+m}, R_{i+m+1}) \) represents the numerical value of the \( u \)-th \((u = 1, 2, \ldots, \mu)\) physiochemical index for the dinucleotide \( R_i, R_{i+1} \) \((R_{i+m}, R_{i+m+1})\) at position \( i \) \((i+m)\).

### 2.4 Predicted structure composition

#### 2.4.1 Local structure-sequence triplet element (Triplet)

The Triplet(24) is an early approach to use the structure information of RNA sequences, and showed better performance for microRNA identification compared with other sequence-based methods.

Given an RNA sequence \( R \) (Eq. 41), formulating it according to its secondary structure derived from the Vienna RNA software package (25) \((\text{released 2.1.6})\), we have

\[
R = \Psi_1 \Psi_2 \Psi_3 \Psi_4 \Psi_5 \cdots \Psi_\mu
\]
where \( \Psi_1 \) denotes the structure status of \( R_1 \), \( \Psi_2 \) the structure status of \( R_2 \), and so forth.

In the predicted secondary structure, there are only two statuses for each nucleotide, paired or unpaired, indicated by brackets "(" or ")" and dots ".", respectively. The left bracket "(" means that the paired nucleotide is located near the 5'-end and can be paired with another nucleotide at the 3'-end, which is indicated by a right bracket ")". We don't distinguish these two situations and use "(" for both situations. For any 3 adjacent nucleotides, there are 8 \( 2^3 \) possible structure compositions: "((", "((", "(...)", "(", ".(", ".(", "(", "...", "..."). Considering the middle nucleotide among the 3 adjacent nucleotides, there are 32 \( 4 \times 8 \) possible structure-sequence combinations, which we denote as \( f_A^{("((")} \), \( f_C^{("((")} \), etc.

Therefore, Triplet approach formulates a feature vector containing 32 \( 4 \times 8 \) components as given by

\[
\mathbf{D} = [ f_A^{("((")}, f_A^{(('.)",} \cdots f_A^{("...")}, f_C^{("((")}, \cdots f_U^{("...")}]^T \tag{61}
\]

where \( f \) represents the normalized occurrence frequency of the structure-sequence compositions.

### 2.4.2 Pseudo-structure status composition (PseSSC)

Given an RNA sequence \( R \) \((\text{Eq. 41})\), we can formulate its secondary structure as \( \text{Eq. 60} \). They can be any of the 10 structure statuses; i.e.,

\[
\Psi_i = \{ \text{A, C, G, U, A-U, U-A, G-C, C-G, G-U, U-G} \} \quad i = 1, 2, \cdots, L \tag{62}
\]


The PseSSC \((26)\) approach formulates a feature vector containing \( 10^n + \lambda \) components as given by

\[
\mathbf{R} = \begin{bmatrix} f_1^* \ f_2^* \ f_3^* \cdots f_{10^n}^* \ f_{10^n+1}^* \cdots f_{10^n+\lambda}^* \end{bmatrix}^T \tag{63}
\]

where

\[
f^* = \begin{cases} 
\frac{f_u}{\sum_{i=1}^{10^n} f_i + w \sum_{j=1}^{\lambda} 0_j} & (1 \leq u \leq 10^n) \\
\frac{w 0_{u-10^n}}{\sum_{i=1}^{10^n} f_i + w \sum_{j=1}^{\lambda} 0_j} & (10^n + 1 \leq u \leq 10^n + \lambda)
\end{cases} \tag{64}
\]

where \( f_i \) \((i = 1, 2, \cdots, 10^n)\) represents the normalized occurrence frequency of the structure status combination of \( n \) adjacent nucleobases, \( w \) is the weight factor used to adjust the effect of the correlation factors, and \( 0_j \) is the \( j \)-tier sequence correlation factor given by
where $\lambda$ is an integer, representing the highest counted rank (or tier) of the structural correlation along an RNA chain; $\theta_i$ is the $i$th-tier correlation factor reflecting the structure-order information between all the $i$th most contiguous bases along an RNA chain, and the correlation function $\Theta(\Psi_i, \Psi_j)$ is given by

\[
\Theta(\Psi_i, \Psi_j) = \left[ F(\Psi_i) - F(\Psi_j) \right]^2
\]  

(66)

where $F(\Psi_i)$ is the free energy of the structure status $\Psi_i$ of the nucleobase at position $i$, and $F(\Psi_j)$ is the free energy of the structure status $\Psi_j$ of the nucleobase at position $j$.

### 2.4.3 Pseudo-distance structure status pair composition (PseDPC)

Given an RNA sequence $R$ (Eq. 41), its feature vector (Eq. 60) can also be formulated as follows. In order to capture the structure-order information of the RNA sequence $R$, a concept called the occurrences of “distance structure status pair” or just “distance-pair” has been proposed, as formulated by

\[
D(\Psi_i, \Psi_j | k) = \begin{cases} 
\theta_1 & \text{if } k = 0 \text{ then } i = j \\
\theta_2 & \text{if } k = 1 \\
\theta_3 & \text{if } k = 2 \\
\vdots & \\
\theta_L & \text{if } k = L - 1 
\end{cases}
\]

(67)

where $\Psi_i$ and $\Psi_j$ can be any of the 10 structure statuses of an RNA chain $R$ (cf. Eq. 62), and $k$ ($0 \leq k \leq L - 1$) represents the distance between structure statuses $\Psi_i$ and $\Psi_j$ along the RNA chain $R$. Suppose $\Psi_i$ is A–U, $\Psi_j$ is U–G, and $k = 3$, then $D$(A–U, U–G|3) means the structure status pair (A–U, U–G) with its two counterparts separated by two nucleotides along the RNA chain $R$.

The approach PseDPC (27) formulates a feature vector as below:

\[
[d_1 d_2 d_3 \ldots d_{\Omega} d_{\Omega+1} d_{\Omega+2} \ldots d_{\Omega+\lambda}]^T
\]

(68)
where

\[
d_u = \begin{cases} 
\frac{f_u}{1 + w \sum_{j=1}^{\lambda} \theta_j} & (1 \leq u \leq \Omega) \\
\frac{w \theta_j \Omega}{1 + w \sum_{j=1}^{\lambda} \theta_j} & (\Omega + 1 \leq u \leq \Omega + \lambda)
\end{cases}
\]  \hspace{1cm} (69)

where \(\theta_j\) is the \(j\)-tier sequence correlation factor computed by Eq. 65, \(w\) is the weight factor used to adjust the effect of the correlation factors, \(\Omega = 10 + 100n\), where \(n\) represents the maximum distance between two structure statuses, and \(f_u\) is the occurrences of the distance-pairs \(D(\Psi_i, \Psi_j | k)\) calculated by

\[
f_u = \begin{cases} 
f(D(\Psi_i, \Psi_j | 0)) & \text{if } 1 \leq u \leq 10 \\
f(D(\Psi_i, \Psi_j | 1)) & \text{if } 11 \leq u \leq 110 \\
\vdots & \vdots \\
f(D(\Psi_i, \Psi_j | n)) & \text{if } 10 + 100(n-1) \leq u \leq 10 + 100n
\end{cases}
\]  \hspace{1cm} (70)

3. Protein

3.1 Amino acid composition

3.1.1 Basic kmer (Kmer)

Basic kmer (28) is the simplest approach to represent the proteins, in which the protein sequences are represented as the occurrence frequencies of \(k\) neighboring amino acids.

3.1.2 Distance-based Residue (DR)

Distance-based Residue (29) is a sequence-based method, in which the feature vector representation for protein is based on the distance between residue pairs. The proposed feature vectors was calculated by counting the occurrences of all possible residue pairs within a certain distance threshold. The dimension of the feature vector is \(20 + 20 \times 20 \times d_{\text{MAX}}\), where 20 is the size of the alphabet of amino acids and \(d_{\text{MAX}}\) is the distance threshold which representing the maximum distance between residue pairs. For more information of this approach, please refer to (29).

3.1.3 PseAAC of Distance-Pairs and reduced alphabet scheme (Distance Pair)
PseAAC of Distance-Pairs and reduced alphabet scheme (30) is a sequenced-based method, in which the feature vector representation for protein is based on reduced alphabet scheme and the distance between residue pairs. The proposed reduced alphabet approach can significantly cut down the dimension of the PseAAC vector and improve the predictive performance. The dimension of the feature vector is $n + dn^2$, where $n$ represents the number of clusters for a given profile, $d$ is the distance threshold which representing the maximum distance between residue pairs.

The reduced alphabet used here is as follows:

\[
\begin{align*}
\text{cp}(13) &= \{\text{MF; IL; V; A; C; WYQHP; G; T; S; N; RK; D; E}\} \\
\text{cp}(14) &= \{\text{IMV; L; F; WY; G; P; C; A; S; T; N; HRKQ; E; D}\} \\
\text{cp}(19) &= \{\text{P; G; E; K; R; Q; D; S; N; T; H; C; I; V; W; YF; A; L; M}\} \\
\text{cp}(20) &= \{\text{A; C; D; E; F; G; H; I; K; L; M; N; P; Q; R; S; T; V; W; Y}\}
\end{align*}
\]  

(71)

For more information of this approach, please refer to (30).

### 3.2 Autocorrelation

#### 3.2.1 Auto covariance (AC)

Suppose a protein sequence $P$ with $L$ amino acid residues; i.e.

\[
P = R_1R_2R_3R_4R_5R_6...R_L
\]  

(72)

where $R_1$ represents the amino acid residue at the sequence position 1, $R_2$ the amino acid residue at position 2 and so forth.

The AC (12,13,31) approach measures the correlation of the same property between two residues separated by a distance of lag along the sequence, which can be calculated as:

\[
AC(i,\text{lag}) = \frac{\sum_{i=1}^{L-\text{lag}} (P_u(R_i) - \bar{P}_u)(P_u(R_{i+\text{lag}}) - \bar{P}_u)}{(L-\text{lag})}
\]  

(73)

where $u$ is a physicochemical index, $L$ is the length of the protein sequence, $P_u(R_i)$ means the numerical value of the physicochemical index $u$ for the amino acid $R_i$ at position $i$, $\bar{P}_u$ is the average value for physicochemical index $u$ along the whole sequence:

\[
\bar{P}_u = \frac{\sum_{j=1}^{L} P_u(R_j)}{L}
\]  

(74)

In such a way, the length of AC feature vector is $N\times\text{LAG}$, where $N$ is the number of physicochemical indices (Table 7) extracted from AAindex (32); LG is the maximum of lag ($\text{lag}=1,2,...,\text{LG}$).

For more information of this approach, please refer to (12,13).

#### 3.2.2 Cross covariance (CC)
Given a protein sequence $P$ (Eq. 72), the CC (12,13,31) approach measures the correlation of two different properties between two residues separated by a distance of $\text{lag}$ along the sequence, which can be calculated by:

$$
\text{CC}(u_1, u_2, \text{lag}) = \sum_{j=1}^{L-\text{lag}} (P_{u_1}(R_j) - \overline{P}_{u_1})(P_{u_2}(R_{j+\text{lag}}) - \overline{P}_{u_2}) / (L-\text{lag}) \tag{75}
$$

where $u_1, u_2$ are two different physicochemical indices, $L$ is the length of the protein sequence, $P_{u_1}(R_j)$ ($P_{u_2}(R_{j+\text{lag}})$) is the numerical value of the physicochemical index $u_1$ ($u_2$) for the amino acid $R_j$ ($R_{j+\text{lag}}$) at position $i$ ($i+\text{lag}$), $\overline{P}_{u_1}$ ($\overline{P}_{u_2}$) is the average value for physicochemical index value $u_1$ ($u_2$) along the whole sequence:

$$
\overline{P}_{u} = \frac{1}{L} \sum_{j=1}^{L} P_{u}(R_j) \tag{76}
$$

In such a way, the length of the CC feature vector is $N*(N-1)*\text{LAG}$, where $N$ is the number of physicochemical indices (Table 7) and LAG is the maximum of $\text{lag}$ ($\text{lag}=1, 2, \ldots, \text{LAG}$).

For more information of this approach, please refer to (12,13).

### 3.2.3 Auto-cross covariance (ACC)

ACC (12,13,31) is a combination of AC and CC. Therefore, the length of the ACC feature vector is $N*N*\text{LAG}$, where $N$ is the number of physicochemical indices (Table 7) and LAG is the maximum of $\text{lag}$ ($\text{lag}=1, 2, \ldots, \text{LAG}$).

### 3.2.4 Physicochemical distance transformation (PDT)

Physicochemical distance transformation (PDT) (33) is able to incorporate the sequence-order effects into prediction. Each protein sequence is converted into a series of numbers by using physicochemical property scores in the amino acid index (AAIndex) (34), and then the sequence is converted into a fixed length vector by PDT. 547 different physicochemical properties were used in this approach as shown in Table 7. For more information of this approach, please refer to (33).

### 3.3 Pseudo amino acid composition

#### 3.3.1 Parallel correlation pseudo amino acid composition (PC-PseAAC)

PC-PseAAC (35) is an approach incorporating the contiguous local sequence-order information and the global sequence-order information into the feature vector of the protein sequence.

Given a Protein sequence $P$ (Eq. 72), the PC-PseAAC feature vector of $P$ is defined:

$$
P = [x_1 \ x_2 \ \cdots \ x_{20} \ x_{20+1} \ \cdots \ x_{20+\lambda}]^T \tag{77}
$$

where
\[
x_u = \begin{cases} 
  \frac{f_u}{\sum_{i=1}^{u} f_i + w \sum_{j=1}^{\lambda} \theta_j} & (1 \leq u \leq 20) \\
  \frac{w \theta_{u-20}}{\sum_{i=1}^{20} f_i + w \sum_{j=1}^{\lambda} \theta_j} & (20 + 1 \leq u \leq 20 + \lambda)
\end{cases} 
\]  
(78)

where \(f_i (i=1,2,\ldots,20)\) is the normalized occurrence frequency of the 20 amino acids in the protein \(P\); the parameter \(\lambda\) is an integer, representing the highest counted rank (or tier) of the correlation along a protein sequence; \(w\) is the weight factor ranging from 0 to 1; \(\theta_j (j=1,2,\ldots,\lambda)\) is called the \(j\)-tier correlation factor reflecting the sequence-order correlation between all the \(j\)-th most contiguous residues along a protein chain, which is defined:

\[
\begin{align*}
\theta_1 &= \frac{1}{L-1} \sum_{i=1}^{L-1} \Theta(R_i, R_{i+1}) \\
\theta_2 &= \frac{1}{L-2} \sum_{i=2}^{L-2} \Theta(R_i, R_{i+2}) \\
\theta_3 &= \frac{1}{L-3} \sum_{i=3}^{L-3} \Theta(R_i, R_{i+3}) \\
& \quad \ldots \ldots \ldots \\
\theta_\lambda &= \frac{1}{L-\lambda} \sum_{i=\lambda}^{L-\lambda} \Theta(R_i, R_{i+\lambda}) 
\end{align*}
\]  
(79)

where the correlation function is given by

\[
\Theta(R_i, R_j) = \frac{1}{3} \left[ \left( H_1(R_j) - H_1(R_i) \right)^2 + \left( H_2(R_j) - H_2(R_i) \right)^2 + \left( M(R_j) - M(R_i) \right)^2 \right] 
\]  
(80)

where \(H_1(R_i), H_2(R_i),\) and \(M(R_i)\) are, respectively, the hydrophobicity value, hydrophilicity value, and side-chain mass (Table 8) of the amino acid \(R_i\); Note that before substituting the values of hydrophobicity, hydrophilicity, and side-chain mass into Eq. 80, they are all subjected to a standard conversion as described by the following equation:
where $H_1^o(i)$ is the original hydrophobicity value of the $i$-th amino acid; $H_2^o(i)$ the corresponding original hydrophilicity value; $M^o(i)$ the mass of the $i$-th amino acid side chain.

### 3.3.2 Series correlation pseudo amino acid composition (SC-PseAAC)

SC-PseAAC (36) is a variant of PC-PseAAC. Given a protein sequence $P$ (Eq. 72), the SC-PseAAC feature vector of $P$ is defined:

$$
P = \left[ p_1, p_2, \cdots, p_{20}, p_{20+\lambda}, \cdots, p_{20+2\lambda} \right]^T$$

where

$$p_u = \begin{cases} 
  \frac{f_u}{\sum_{i=1}^{20} f_i + w \sum_{j=1}^{2\lambda} \tau_j} & (1 \leq u \leq 20) \\
  \frac{w \tau_u - 20}{\sum_{i=1}^{20} f_i + w \sum_{j=1}^{2\lambda} \tau_j} & (20 + 1 \leq u \leq 20 + 2\lambda) 
\end{cases}$$

where $f_i (i = 1, 2, \ldots, 20)$ is the normalized occurrence frequency of the 20 native amino acids in the protein $P$; the parameter $\lambda$ is an integer, representing the highest counted rank (or tier) of the correlation along a protein sequence; $w$ is the weight factor ranging from 0 to 1; $\tau_j$ the $j$-tier sequence-correlation factor that reflects the sequence-order correlation between all the most contiguous residues along a protein sequence, which is defined:
\[
\begin{align*}
\tau_1 &= \frac{1}{L-1} \sum_{i=1}^{L-1} H_{i,i+1}^1 \\
\tau_2 &= \frac{1}{L-1} \sum_{i=1}^{L-1} H_{i,i+1}^2 \\
\tau_3 &= \frac{1}{L-2} \sum_{i=1}^{L-2} H_{i,i+2}^1 \\
\tau_4 &= \frac{1}{L-2} \sum_{i=1}^{L-2} H_{i,i+2}^2 \\
\quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \ quad
3.3.3 General parallel correlation pseudo amino acid composition (PC-PseAAC-General)

The PC-PseAAC-General approach (31) cannot only incorporate comprehensive built-in indices (Table 7) extracted from AAindex (32), but also allow the users to upload their own indices to generate the PC-PseAAC-General feature vector. Given a protein sequence \( P \) (Eq. 72), the PC-PseAAC-General feature vector of \( P \) is defined:

\[
P = \left[ x_1 \ x_2 \ \cdots \ x_{20} \ x_{20+1} \ \cdots \ x_{20+\lambda} \right]^T
\]  

(87)

where

\[
x_u = \begin{cases} 
\frac{f_u}{\sum_{j=1}^{20} f_j + w \sum_{j=1}^{\lambda} \theta_j} & (1 \leq u \leq 20) \\
\frac{w \theta_u - 20}{\sum_{j=1}^{20} f_j + w \sum_{j=1}^{\lambda} \theta_j} & (20 + 1 \leq u \leq 20 + \lambda)
\end{cases}
\]  

(88)

where \( f_i (i=1,2,\cdots,20) \) is the normalized occurrence frequency of the 20 amino acids in the protein \( P \); the parameter \( \lambda \) is an integer, representing the highest counted rank (or tier) of the correlation along a protein sequence; \( w \) is the weight factor ranging from 0 to 1; \( \theta_j (j=1,2,\cdots,\lambda) \) is called the \( j \)-tier correlation factor reflecting the sequence-order correlation between all the \( j \)-th most contiguous residues along a protein chain, which is defined:

\[
\begin{align*}
\theta_1 &= \frac{1}{L-1} \sum_{i=1}^{L-1} \Theta(R_i, R_{i+1}) \\
\theta_2 &= \frac{1}{L-2} \sum_{i=1}^{L-2} \Theta(R_i, R_{i+2}) \\
\theta_3 &= \frac{1}{L-3} \sum_{i=1}^{L-3} \Theta(R_i, R_{i+3}) \\
& \quad \cdots \cdots \\
\theta_\lambda &= \frac{1}{L-\lambda} \sum_{i=1}^{L-\lambda} \Theta(R_i, R_{i+\lambda})
\end{align*}
\]  

(89)

where the correlation function is given by

\[
\Theta(R_i, R_j) = \frac{1}{\mu} \sum_{u=1}^{\mu} [H_u(R_i) - H_u(R_j)]^2
\]  

(90)

where \( \mu \) is the number of physicochemical indices considered that listed in the Table 7; \( H_u(R_i) \) is the \( u \)-th physicochemical index value of the amino acid \( R_i \); \( H_u(R_j) \), the \( u \)-th physicochemical index value for the amino acid \( R_j \). Note that before substituting the physicochemical indices values into Eq. 90, they are all subjected to a standard conversion as described by the following equation:
where \( H_u^0(i) \) is the \( u\)-th original physicochemical value of the \( i\)-th amino acid.

### 3.3.4 General series correlation pseudo amino acid composition (SC-PseAAC-General)

The SC-PseAAC-General approach (31) cannot only incorporate comprehensive built-in indices (Table 7) extracted from AAindex (32), but also allow the users to upload their own indices to generate the SC-PseAAC-General feature vector.

Given a protein sequence \( P \) (Eq. 72), the SC-PseAAC-General feature vector of \( P \) is defined:

\[
P = \left[ p_1, p_2, \ldots, p_{20}, p_{20+1}, \ldots, p_{20+\lambda}, p_{20+\lambda+1}, \ldots, p_{20+\lambda+\Lambda} \right]^T
\]

where

\[
p_u = \begin{cases} 
\frac{f_u}{\sum_{i=1}^{20} f_i + w \sum_{j=1}^{\lambda \Lambda} \tau_j} & (1 \leq u \leq 20) \\
\frac{w \tau_{u-20}}{\sum_{i=1}^{20} f_i + w \sum_{j=1}^{\lambda \Lambda} \tau_j} & (20 + 1 \leq u \leq 20 + \lambda \Lambda)
\end{cases}
\]

where \( f_i \) (\( i = 1, 2, \ldots, 20 \)) is the normalized occurrence frequency of the 20 native amino acids in the protein \( P \), the parameter \( \lambda \) is an integer, representing the highest counted rank (or tier) of the correlation along a protein sequence; \( w \) is the weight factor ranging from 0 to 1; \( \Lambda \) is the number of physicochemical indices (Table 7); \( \tau_j \) the \( j \)-tier sequence-correlation factor reflecting the sequence-order correlation between all the most contiguous residues along a protein sequence, which is defined:

\[
\tau_j = \begin{cases} 
\frac{1}{L-1} \sum_{i=1}^{L-1} H_{ij+1}^j & \lambda < (L-1) \\
\frac{1}{L-1} \sum_{i=1}^{L-1} H_{ij+1}^\lambda & \lambda = (L-1) \\
\end{cases}
\]

where \( H_{ij+m}^q \) is the correlation function given by
where \( h^p(R_i) \) is the \( \zeta \)-th physicochemical value for the \( i \)-th (\( i = 1, 2, \ldots, L \)) amino acid in Eq. 72, and the dot (\( \cdot \)) means the multiplication sign.

Note that before substituting the physicochemical values into Eq. 95, they are all subjected to a standard conversion as described by the following equation:

\[
\begin{align*}
\hat{h}_\zeta(R_i) &= \sum_{k=1}^{20} h_\zeta^2(R_k) - \frac{1}{20} \sum_{k=1}^{20} h_\zeta^2(R_k)^2 \\
&= \sqrt{\left[ \sum_{k=1}^{20} h_\zeta^2(R_k) - \frac{1}{20} \sum_{k=1}^{20} h_\zeta^2(R_k)^2 \right]^2}
\end{align*}
\]  

where we use the \( R_i, (i = 1, 2, \ldots, 20) \) to represent the 20 native amino acids. The symbols \( h_\zeta \) represent the \( \zeta \)-th original physicochemical value of the amino acid in the brackets right after the symbols.

### 3.4 Frequency Profile

#### 3.4.1 Top-n-gram

Top-n-gram (37) can be viewed as a novel profile-based building blocks of proteins, containing the evolutionary information extracted from the frequency profiles. The frequency profiles calculated from the multiple sequence alignments outputted by PSI-BLAST (38) are converted into Top-n-grams by combining the n most frequent amino acids in each amino acid frequency profile. The protein sequences are transformed into fixed dimension feature vectors by the occurrence times of each Top-n-gram. For more information of this approach, please refer to (37).

#### 3.4.2 Profile-based physicochemical distance transformation (PDT-Profile)

The process of profile-based PDT (33) is similar as that of sequence-based PDT (33). Except that there is an additional step of extracting the evolutionary information from the frequency profiles. The target frequencies in the frequency profiles reflect the probabilities of the corresponding amino acids appearing in the specific sequence positions. The higher the frequency is, the more likely the corresponding amino acid occurs. It is reasonable to use the \( n \)-th most frequent amino acids in the frequency profiles to represent the protein sequences. Each amino acid in a protein sequence is replaced by its corresponding \( n \)-th most frequent amino acid in the frequency profile. Therefore, the resulting protein sequence takes the evolutionary information in the frequency profile into consideration. For more information of this approach, please refer to (33).

#### 3.4.3 Distance-based Top-n-gram (DT)
Distance-based Top-n-gram (29) is a profile-based method which considers the distances between Top-n-gram (37) pairs. Replacing all the amino acids in a protein sequence can be represented as a sequence of Top-n-grams instead of a sequence of amino acids. Distance-based Top-n-gram was proposed, which extends the original Top-n-gram-based feature vector by considering the relative position information of Top-n-gram pairs in protein sequences. In this study, the Top-1-gram was selected to construct the Distance-based Top-n-gram feature vector in order to reduce the dimension of the feature vectors and reduce the computational cost. The proposed feature vectors was calculated by counting the occurrences of all possible Top-n-gram pairs within a certain distance threshold. The dimension of the feature vector is $20 + 20 \times 20 \times d_{\text{MAX}}$, where 20 is the size of the alphabet of amino acids and $d_{\text{MAX}}$ is the distance threshold which representing the maximum distance between Top-1-gram pairs. For more information of this approach, please refer to (29).

3.4.4 Profile-based Auto covariance (AC-PSSM)

AC-PSSM (12) can transform the PSSMs of different lengths into fixed-length vector. The AC variable measures the correlation of the same property between two residues separated by a distance of $\text{lag}$ along the sequence, which can be calculated as:

$$AC(i, \text{lag}) = \sum_{j=1}^{L-\text{lag}} \left( S_{i,j} - \bar{S}_i \right) \left( S_{i,j+\text{lag}} - \bar{S}_i \right) / (L - \text{lag})$$  \hspace{1cm} (97)

where $i$ is one of the residues, $L$ is the length of the protein sequence, $S_{i,j}$ is the PSSM score of amino acid $i$ at position $j$, $\bar{S}_i$ is the average score for amino acid $i$ along the whole sequence:

$$\bar{S}_i = \sum_{j=1}^{L} S_{i,j} / L$$  \hspace{1cm} (98)

In such a way, the number of AC variables can be calculated as $20\times L\text{AG}$, where $L\text{AG}$ is the maximum of $\text{lag}$ ($\text{lag}=1, 2, ..., L\text{AG}$).

3.4.5 Profile-based Cross covariance (CC-PSSM)

CC-PSSM (12) can transform the PSSMs of different lengths into fixed-length vectors. The CC variable measures the correlation of two different properties between two residues separated by $\text{lag}$ along the sequence, which can be calculated by:

$$CC(i1, i2, \text{lag}) = \sum_{j=1}^{L-\text{lag}} \left( S_{i1,j} - \bar{S}_{i1} \right) \left( S_{i2,j+\text{lag}} - \bar{S}_{i2} \right) / (L - \text{lag})$$  \hspace{1cm} (99)

where $i1, i2$ are two different amino acids and $\bar{S}_{i1}$ ($\bar{S}_{i2}$) is the average score for amino acid $i1$ ($i2$) along the sequence. Since the CC variables are not symmetric, the total number of CC variables is $380\times L\text{AG}$.

3.4.6 Profile-based Auto-cross covariance (ACC-PSSM)

ACC-PSSM (12), as one of the multivariate modeling tools, can transform the PSSMs of different lengths into fixed-length vectors by measuring the correlation between any two properties. ACC results in two kinds of variables: AC between the same
property, and cross-covariance (CC) between two different properties. Each protein sequence is represented as a vector of either AC variable or ACC variable that is a combination of AC and CC.

**Table 1.** The names of the 148 physicochemical indices for dinucleotides (DNA).

<table>
<thead>
<tr>
<th>Base stacking</th>
<th>Protein induced deformability</th>
<th>B-DNA twist</th>
</tr>
</thead>
<tbody>
<tr>
<td>Propeller twist</td>
<td>Duplex stability:(freeenergy)</td>
<td>Duplex tability(disruptenergy)</td>
</tr>
<tr>
<td>Protein DNA twist</td>
<td>Stabilising energy of Z-DNA</td>
<td>Aida_BA_transition</td>
</tr>
<tr>
<td>Breslauer dS</td>
<td>Electron interaction</td>
<td>Hartman_trans_free_energy</td>
</tr>
<tr>
<td>Lisser_BZ_transition</td>
<td>Polar_interaction</td>
<td>SantaLucia_dG</td>
</tr>
<tr>
<td>Sarai_flexibility</td>
<td>Stability</td>
<td>Stacking_energy</td>
</tr>
<tr>
<td>Sugimoto_dS</td>
<td>Watson-Crick_interaction</td>
<td>Twist</td>
</tr>
<tr>
<td>Shift</td>
<td>Slide</td>
<td>Rise</td>
</tr>
<tr>
<td>Twist stiffness</td>
<td>Tilt stiffness</td>
<td>Shift rise</td>
</tr>
<tr>
<td>Twist_shift</td>
<td>Enthalpy1</td>
<td>Twist_twist</td>
</tr>
<tr>
<td>Shift2</td>
<td>Tilt3</td>
<td>Tilt1</td>
</tr>
<tr>
<td>Slide (DNA-protein complex)1</td>
<td>Tilt_shift</td>
<td>Twist_tilt</td>
</tr>
<tr>
<td>Roll_rise</td>
<td>Stacking energy</td>
<td>Stacking energy1</td>
</tr>
<tr>
<td>Propeller Twist</td>
<td>Roll11</td>
<td>Rise (DNA-protein complex)</td>
</tr>
<tr>
<td>Roll2</td>
<td>Roll3</td>
<td>Roll1</td>
</tr>
<tr>
<td>Slide_slide</td>
<td>Enthalpy</td>
<td>Shift_shift</td>
</tr>
<tr>
<td>Flexibility_slide</td>
<td>Minor Groove Distance</td>
<td>Rise (DNA-protein complex)1</td>
</tr>
<tr>
<td>Roll (DNA-protein complex)1</td>
<td>Entropy</td>
<td>Cytosine content</td>
</tr>
<tr>
<td>Major Groove Distance</td>
<td>Twist (DNA-protein complex)</td>
<td>Purine (AG) content</td>
</tr>
<tr>
<td>Tilt_slide</td>
<td>Major Groove Width</td>
<td>Major Groove Depth</td>
</tr>
<tr>
<td>Free energy6</td>
<td>Free energy7</td>
<td>Free energy4</td>
</tr>
<tr>
<td>Free energy3</td>
<td>Free energy1</td>
<td>Twist_roll</td>
</tr>
<tr>
<td>Flexibility_shift</td>
<td>Shift (DNA-protein complex)1</td>
<td>Thymine content</td>
</tr>
<tr>
<td>Tip</td>
<td>Keto (GT) content</td>
<td>Roll stiffness</td>
</tr>
<tr>
<td>Entropy1</td>
<td>Roll_slide</td>
<td>Slide (DNA-protein complex)</td>
</tr>
<tr>
<td>Twist2</td>
<td>Twist5</td>
<td>Twist4</td>
</tr>
<tr>
<td>Tilt (DNA-protein complex)1</td>
<td>Twist_slide</td>
<td>Minor Groove Depth</td>
</tr>
<tr>
<td>Persistance Length</td>
<td>Rise3</td>
<td>Shift stiffness</td>
</tr>
<tr>
<td>Slide3</td>
<td>Slide2</td>
<td>Slide1</td>
</tr>
<tr>
<td>Rise1</td>
<td>Rise stiffness</td>
<td>Mobility to bend towards minor groove</td>
</tr>
<tr>
<td>----------------</td>
<td>----------------</td>
<td>---------------------------------------</td>
</tr>
<tr>
<td>Dinucleotide GC Content</td>
<td>A-philicity</td>
<td>Wedge</td>
</tr>
<tr>
<td>DNA denaturation</td>
<td>Bending stiffness</td>
<td>Free energy5</td>
</tr>
<tr>
<td>Breslauer_dG</td>
<td>Breslauer_dH</td>
<td>Shift (DNA-protein complex)</td>
</tr>
<tr>
<td>Helix-Coil transition</td>
<td>Ivanov BA transition</td>
<td>Slide rise</td>
</tr>
<tr>
<td>SantaLucia_dH</td>
<td>SantaLucia_dS</td>
<td>Minor Groove Width</td>
</tr>
<tr>
<td>Sugimoto_dG</td>
<td>Sugimoto_dH</td>
<td>Twist</td>
</tr>
<tr>
<td>Tilt</td>
<td>Roll</td>
<td>Twist7</td>
</tr>
<tr>
<td>Clash Strength</td>
<td>Roll_roll</td>
<td>Roll (DNA-protein complex)</td>
</tr>
<tr>
<td>Adenine content</td>
<td>Direction</td>
<td>Probability contacting nucleosome core</td>
</tr>
<tr>
<td>Roll_shift</td>
<td>Shift_slide</td>
<td>Shift1</td>
</tr>
<tr>
<td>Tilt4</td>
<td>Tilt2</td>
<td>Free energy8</td>
</tr>
<tr>
<td>Twist (DNA-protein complex)_1</td>
<td>Tilt_rise</td>
<td>Free energy2</td>
</tr>
<tr>
<td>Stacking energy2</td>
<td>Stacking energy3</td>
<td>Rise_rise</td>
</tr>
<tr>
<td>Tilt_tilt</td>
<td>Roll4</td>
<td>Tilt_roll</td>
</tr>
<tr>
<td>Minor Groove Size</td>
<td>GC content</td>
<td>Inclination</td>
</tr>
<tr>
<td>Slide stiffness</td>
<td>Melting Temperature</td>
<td>Twist3</td>
</tr>
<tr>
<td>Tilt (DNA-protein complex)</td>
<td>Guanine content</td>
<td>Twist6</td>
</tr>
<tr>
<td>Major Groove Size</td>
<td>Twist_rise</td>
<td>Rise2</td>
</tr>
<tr>
<td>Melting Temperature</td>
<td>Free energy</td>
<td>Mobility to bend towards major groove</td>
</tr>
<tr>
<td>Bend</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 2.** The names of the 12 physicochemical indices for trinucleotides (DNA).

<table>
<thead>
<tr>
<th>Bendability (DNAse)</th>
<th>Bendability (consensus)</th>
<th>Trinucleotide GC Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consensus_roll</td>
<td>Consensus-Rigid</td>
<td>Dnase I</td>
</tr>
<tr>
<td>MW-Daltons</td>
<td>MW-kg</td>
<td>Nucleosome</td>
</tr>
<tr>
<td>Nucleosome positioning</td>
<td>Dnase I-Rigid</td>
<td>Nucleosome-Rigid</td>
</tr>
</tbody>
</table>

**Table 3.** The names of the 90 physicochemical indices for dinucleotides (DNA).

<table>
<thead>
<tr>
<th>Base stacking</th>
<th>Protein induced deformability</th>
<th>B-DNA twist</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dinucleotide GC Content</td>
<td>A-philicity</td>
<td>Propeller twist</td>
</tr>
<tr>
<td>Duplex stability-free energy</td>
<td>Duplex stability-disrupt energy</td>
<td>DNA denaturation</td>
</tr>
<tr>
<td>Bending stiffness</td>
<td>Protein DNA twist</td>
<td>Stabilising energy of Z-DNA</td>
</tr>
<tr>
<td>-------------------</td>
<td>-------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>Aida_{BA}_transition</td>
<td>Breslauer_{dG}</td>
<td>Breslauer_{dH}</td>
</tr>
<tr>
<td>Breslauer_{dS}</td>
<td>Electron_interaction</td>
<td>Hartman_{trans_free_energy}</td>
</tr>
<tr>
<td>Helix-Coil_transition</td>
<td>Ivanov_{BA}_transition</td>
<td>Lisser_{BZ}_transition</td>
</tr>
<tr>
<td>Polar_interaction</td>
<td>SantaLucia_{dG}</td>
<td>SantaLucia_{dH}</td>
</tr>
<tr>
<td>SantaLucia_{dS}</td>
<td>Sarai_{flexibility}</td>
<td>Stability</td>
</tr>
<tr>
<td>Stacking_energy</td>
<td>Sugimoto_{dG}</td>
<td>Sugimoto_{dH}</td>
</tr>
<tr>
<td>Sugimoto_{dS}</td>
<td>Watson-Crick_interaction</td>
<td>Twist</td>
</tr>
<tr>
<td>Tilt</td>
<td>Roll</td>
<td>Shift</td>
</tr>
<tr>
<td>Slide</td>
<td>Rise</td>
<td>Stacking energy</td>
</tr>
<tr>
<td>Bend</td>
<td>Tip</td>
<td>Inclination</td>
</tr>
<tr>
<td>Major Groove Width</td>
<td>Major Groove Depth</td>
<td>Major Groove Size</td>
</tr>
<tr>
<td>Major Groove Distance</td>
<td>Minor Groove Width</td>
<td>Minor Groove Depth</td>
</tr>
<tr>
<td>Minor Groove Size</td>
<td>Minor Groove Distance</td>
<td>Persistance Length</td>
</tr>
<tr>
<td>Melting Temperature</td>
<td>Mobility to bend towards major groove</td>
<td>Mobility to bend towards minor groove</td>
</tr>
<tr>
<td>Propeller Twist</td>
<td>Clash Strength</td>
<td>Enthalpy</td>
</tr>
<tr>
<td>Free energy</td>
<td>Twist_{twist}</td>
<td>Tilt_{tilt}</td>
</tr>
<tr>
<td>Roll_{roll}</td>
<td>Twist_{tilt}</td>
<td>Twist_{roll}</td>
</tr>
<tr>
<td>Tilt_{roll}</td>
<td>Shift_{shift}</td>
<td>Slide_{slide}</td>
</tr>
<tr>
<td>Rise_{rise}</td>
<td>Shift_{slide}</td>
<td>Shift_{rise}</td>
</tr>
<tr>
<td>Slide_{rise}</td>
<td>Twist_{shift}</td>
<td>Twist_{slide}</td>
</tr>
<tr>
<td>Twist_{rise}</td>
<td>Tilt_{shift}</td>
<td>Tilt_{slide}</td>
</tr>
<tr>
<td>Tilt_{rise}</td>
<td>Roll_{shift}</td>
<td>Roll_{slide}</td>
</tr>
<tr>
<td>Roll_{rise}</td>
<td>Slide stiffness</td>
<td>Shift stiffness</td>
</tr>
<tr>
<td>Roll stiffness</td>
<td>Rise stiffness</td>
<td>Tilt stiffness</td>
</tr>
<tr>
<td>Twist stiffness</td>
<td>Wedge</td>
<td>Direction</td>
</tr>
<tr>
<td>Flexibility_{slide}</td>
<td>Flexibility_{shift}</td>
<td>Entropy</td>
</tr>
</tbody>
</table>
Table 4. The names of the 6 physicochemical indices for dinucleotides (DNA).

<table>
<thead>
<tr>
<th>Twist(DNA)</th>
<th>Tilt(DNA)</th>
<th>Roll(DNA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shift(DNA)</td>
<td>Slide(DNA)</td>
<td>Rise(DNA)</td>
</tr>
</tbody>
</table>

Table 5. The names of the 22 physicochemical indices for dinucleotides (RNA).

<table>
<thead>
<tr>
<th>Shift (RNA)</th>
<th>Hydrophilicity (RNA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hydrophilicity (RNA)</td>
<td>GC content</td>
</tr>
<tr>
<td>Purine (AG) content</td>
<td>Keto (GT) content</td>
</tr>
<tr>
<td>Adenine content</td>
<td>Guanine content</td>
</tr>
<tr>
<td>Cytosine content</td>
<td>Thymine content</td>
</tr>
<tr>
<td>Slide (RNA)</td>
<td>Rise (RNA)</td>
</tr>
<tr>
<td>Tilt (RNA)</td>
<td>Roll (RNA)</td>
</tr>
<tr>
<td>Twist (RNA)</td>
<td>Stacking energy (RNA)</td>
</tr>
<tr>
<td>Enthalpy (RNA)</td>
<td>Entropy (RNA)</td>
</tr>
<tr>
<td>Free energy (RNA)</td>
<td>Free energy (RNA)</td>
</tr>
<tr>
<td>Enthalpy (RNA)</td>
<td>Entropy (RNA)</td>
</tr>
</tbody>
</table>

Table 6. The names of the 11 physicochemical indices for dinucleotides (RNA).

<table>
<thead>
<tr>
<th>Shift</th>
<th>Slide</th>
<th>Rise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tilt</td>
<td>Roll</td>
<td>Twist</td>
</tr>
<tr>
<td>Stacking energy</td>
<td>Enthalpy</td>
<td>Entropy</td>
</tr>
<tr>
<td>Free energy</td>
<td>Hydrophilicity</td>
<td></td>
</tr>
</tbody>
</table>

Table 7. The names of the 547 physicochemical indices for amino acids.

<table>
<thead>
<tr>
<th>Hydrophobicity</th>
<th>Hydrophilicity</th>
<th>Mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARGP820102</td>
<td>ARGP820103</td>
<td>BEGF750101</td>
</tr>
<tr>
<td>BHAR880101</td>
<td>BIGC670101</td>
<td>BIOV880101</td>
</tr>
<tr>
<td>BROC820102</td>
<td>BULH740101</td>
<td>BULH740102</td>
</tr>
<tr>
<td>BUNA790103</td>
<td>BURA740101</td>
<td>BURA740102</td>
</tr>
<tr>
<td>CHAM820102</td>
<td>CHAM830101</td>
<td>CHAM830102</td>
</tr>
<tr>
<td>CHAM830105</td>
<td>CHAM830106</td>
<td>CHAM830107</td>
</tr>
<tr>
<td>CHOC760101</td>
<td>CHOC760102</td>
<td>CHOC760103</td>
</tr>
<tr>
<td>------------</td>
<td>------------</td>
<td>------------</td>
</tr>
<tr>
<td>CHOP780201</td>
<td>CHOP780202</td>
<td>CHOP780203</td>
</tr>
<tr>
<td>CHOP780206</td>
<td>CHOP780207</td>
<td>CHOP780208</td>
</tr>
<tr>
<td>CHOP780211</td>
<td>CHOP780212</td>
<td>CHOP780213</td>
</tr>
<tr>
<td>CHOP780216</td>
<td>CIDH920101</td>
<td>CIDH920102</td>
</tr>
<tr>
<td>CIDH920105</td>
<td>COHE430101</td>
<td>CRAJ730101</td>
</tr>
<tr>
<td>DAWD720101</td>
<td>DAYM780101</td>
<td>DAYM780201</td>
</tr>
<tr>
<td>EISD840101</td>
<td>EISD860101</td>
<td>EISD860102</td>
</tr>
<tr>
<td>FASG760102</td>
<td>FASG760103</td>
<td>FASG760104</td>
</tr>
<tr>
<td>FAUJ880101</td>
<td>FAUJ880102</td>
<td>FAUJ880103</td>
</tr>
<tr>
<td>FAUJ880106</td>
<td>FAUJ880107</td>
<td>FAUJ880108</td>
</tr>
<tr>
<td>FAUJ880111</td>
<td>FAUJ880112</td>
<td>FAUJ880113</td>
</tr>
<tr>
<td>FINA910102</td>
<td>FINA910103</td>
<td>FINA910104</td>
</tr>
<tr>
<td>GEIM800102</td>
<td>GEIM800103</td>
<td>GEIM800104</td>
</tr>
<tr>
<td>GEIM800107</td>
<td>GEIM800108</td>
<td>GEIM800109</td>
</tr>
<tr>
<td>GOLD730101</td>
<td>GOLD730102</td>
<td>GRAR740101</td>
</tr>
<tr>
<td>GUYH850101</td>
<td>HOPA770101</td>
<td>HOPT810101</td>
</tr>
<tr>
<td>HUTJ700103</td>
<td>ISOY800101</td>
<td>ISOY800102</td>
</tr>
<tr>
<td>ISOY800105</td>
<td>ISOY800106</td>
<td>ISOY800107</td>
</tr>
<tr>
<td>JANJ780102</td>
<td>JANJ780103</td>
<td>JANJ780104</td>
</tr>
<tr>
<td>JOND750102</td>
<td>JOND920101</td>
<td>JOND920102</td>
</tr>
<tr>
<td>KANM800101</td>
<td>KANM800102</td>
<td>KANM800103</td>
</tr>
<tr>
<td>KARP850102</td>
<td>KARP850103</td>
<td>KHAG800101</td>
</tr>
<tr>
<td>KRIW790101</td>
<td>KRIW790102</td>
<td>KRIW790103</td>
</tr>
<tr>
<td>LEVM760101</td>
<td>LEVM760102</td>
<td>LEVM760103</td>
</tr>
<tr>
<td>LEVM760106</td>
<td>LEVM760107</td>
<td>LEVM780101</td>
</tr>
<tr>
<td>LEVM780104</td>
<td>LEVM780105</td>
<td>LEVM780106</td>
</tr>
<tr>
<td>LIFS790102</td>
<td>LIFS790103</td>
<td>MANP780101</td>
</tr>
<tr>
<td>MAXF760103</td>
<td>MAXF760104</td>
<td>MAXF760105</td>
</tr>
<tr>
<td>MEEJ800101</td>
<td>MEEJ800102</td>
<td>MEEJ810101</td>
</tr>
<tr>
<td>MEIH800102</td>
<td>MEIH800103</td>
<td>MIYS850101</td>
</tr>
<tr>
<td>NAGK730103</td>
<td>NAKH900101</td>
<td>NAKH900102</td>
</tr>
<tr>
<td>NAKH900105</td>
<td>NAKH900106</td>
<td>NAKH900107</td>
</tr>
<tr>
<td>NAKH900110</td>
<td>NAKH900111</td>
<td>NAKH900112</td>
</tr>
<tr>
<td>NAKH920102</td>
<td>NAKH920103</td>
<td>NAKH920104</td>
</tr>
<tr>
<td>NAKH920107</td>
<td>NAKH920108</td>
<td>NISK800101</td>
</tr>
<tr>
<td>OOBM770101</td>
<td>OOBM770102</td>
<td>OOBM770103</td>
</tr>
<tr>
<td>OOBM850101</td>
<td>OOBM850102</td>
<td>OOBM850103</td>
</tr>
<tr>
<td>PALJ810101</td>
<td>PALJ810102</td>
<td>PALJ810103</td>
</tr>
<tr>
<td>PALJ810106</td>
<td>PALJ810107</td>
<td>PALJ810108</td>
</tr>
<tr>
<td>PALJ810111</td>
<td>PALJ810112</td>
<td>PALJ810113</td>
</tr>
<tr>
<td>PALJ810116</td>
<td>PARJ860101</td>
<td>PLIV810101</td>
</tr>
<tr>
<td>PONP800103</td>
<td>PONP800104</td>
<td>PONP800105</td>
</tr>
<tr>
<td>PONP800108</td>
<td>PRAM820101</td>
<td>PRAM820102</td>
</tr>
<tr>
<td>PRAM900102</td>
<td>PRAM900103</td>
<td>PRAM900104</td>
</tr>
<tr>
<td>QIAN880101</td>
<td>QIAN880102</td>
<td>QIAN880103</td>
</tr>
<tr>
<td>QIAN880106</td>
<td>QIAN880107</td>
<td>QIAN880108</td>
</tr>
<tr>
<td>GEOR030103</td>
<td>GEOR030104</td>
<td>GEOR030105</td>
</tr>
<tr>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
</tr>
<tr>
<td>GEOR030108</td>
<td>GEOR030109</td>
<td>ZHOH040101</td>
</tr>
<tr>
<td>BAEK050101</td>
<td>HARY940101</td>
<td>PONJ960101</td>
</tr>
<tr>
<td>OLSK800101</td>
<td>KIDA850101</td>
<td>GUYH850102</td>
</tr>
<tr>
<td>GUYH850105</td>
<td>ROSM880104</td>
<td>ROSM880105</td>
</tr>
<tr>
<td>BLAS910101</td>
<td>CASG920101</td>
<td>CORJ870101</td>
</tr>
<tr>
<td>CORJ870104</td>
<td>CORJ870105</td>
<td>CORJ870106</td>
</tr>
<tr>
<td>MIYS990101</td>
<td>MIYS990102</td>
<td>MIYS990103</td>
</tr>
<tr>
<td>ENG0860101</td>
<td>FASG890101</td>
<td>TANS770101</td>
</tr>
<tr>
<td>ANDN920101</td>
<td>ARG0820101</td>
<td>TANS770106</td>
</tr>
<tr>
<td>BEGF750102</td>
<td>BEGF750103</td>
<td>VASM830101</td>
</tr>
<tr>
<td>BIOV880102</td>
<td>BRO0820101</td>
<td>VHEG790101</td>
</tr>
<tr>
<td>BUNA790101</td>
<td>BUNA790102</td>
<td>WERD780103</td>
</tr>
<tr>
<td>CHAM810101</td>
<td>CHAM820101</td>
<td>WOLS870102</td>
</tr>
<tr>
<td>CHAM830103</td>
<td>CHAM830104</td>
<td>YUTK870104</td>
</tr>
<tr>
<td>CHAM830108</td>
<td>CHOC750101</td>
<td>ZIMJ80104</td>
</tr>
<tr>
<td>CHOC760104</td>
<td>CHOP800101</td>
<td>AURR80104</td>
</tr>
<tr>
<td>CHOP780204</td>
<td>CHOP780205</td>
<td>AURR80109</td>
</tr>
<tr>
<td>CHOP780209</td>
<td>CHOP780210</td>
<td>AURR80114</td>
</tr>
<tr>
<td>CHOP780214</td>
<td>CHOP780215</td>
<td>AURR80119</td>
</tr>
<tr>
<td>CIDH920103</td>
<td>CIDH920104</td>
<td>VINC940102</td>
</tr>
<tr>
<td>CRAJ730102</td>
<td>CRAJ730103</td>
<td>MUNV940103</td>
</tr>
<tr>
<td>DESM900101</td>
<td>DESM900102</td>
<td>MONM990101</td>
</tr>
<tr>
<td>EISD860103</td>
<td>FASG760101</td>
<td>KUMS000102</td>
</tr>
<tr>
<td>FASG760105</td>
<td>FAUJ830101</td>
<td>NAD010101</td>
</tr>
<tr>
<td>FAUJ880104</td>
<td>FAUJ880105</td>
<td>NAD010106</td>
</tr>
<tr>
<td>FAUJ880109</td>
<td>FAUJ880110</td>
<td>CEDJ970101</td>
</tr>
<tr>
<td>FINA770101</td>
<td>FINA910101</td>
<td>FUKS010101</td>
</tr>
<tr>
<td>GAR0730101</td>
<td>GEIM800101</td>
<td>FUKS010106</td>
</tr>
<tr>
<td>GEIM800105</td>
<td>GEIM800106</td>
<td>FUKS010111</td>
</tr>
<tr>
<td>GEIM800110</td>
<td>GEIM800111</td>
<td>AVBF000104</td>
</tr>
<tr>
<td>GRAR740102</td>
<td>GRAR740103</td>
<td>AVBF000109</td>
</tr>
<tr>
<td>HUTT700101</td>
<td>HUTT700102</td>
<td>COS940101</td>
</tr>
<tr>
<td>ISOY800103</td>
<td>ISOY800104</td>
<td>WILM950104</td>
</tr>
<tr>
<td>ISOY800108</td>
<td>JANJ870101</td>
<td>BASU050102</td>
</tr>
<tr>
<td>JANJ890102</td>
<td>JOND750101</td>
<td>GEOR030101</td>
</tr>
<tr>
<td>JUKT750101</td>
<td>JUNT780101</td>
<td>GEOR030106</td>
</tr>
<tr>
<td>KANM800104</td>
<td>KARP850101</td>
<td>ZHOH040102</td>
</tr>
<tr>
<td>KLEP840101</td>
<td>KRIW710101</td>
<td>DIGM050101</td>
</tr>
<tr>
<td>KYT820101</td>
<td>LAWE840101</td>
<td>GUYH850103</td>
</tr>
<tr>
<td>LEVM760104</td>
<td>LEVM760105</td>
<td>JACR890101</td>
</tr>
<tr>
<td>LEVM780102</td>
<td>LEVM780103</td>
<td>CORJ870102</td>
</tr>
<tr>
<td>LEWP710101</td>
<td>LEWP790101</td>
<td>CORJ870107</td>
</tr>
<tr>
<td>MAXF760101</td>
<td>MAXF760102</td>
<td>MIYS990104</td>
</tr>
<tr>
<td>MAXF760106</td>
<td>MCM0640101</td>
<td>TANS770102</td>
</tr>
<tr>
<td>MEEJ810102</td>
<td>MEIH800101</td>
<td>TANS770107</td>
</tr>
<tr>
<td>NAGK730101</td>
<td>NAGK730102</td>
<td>VASM830102</td>
</tr>
<tr>
<td>NAKH900103</td>
<td>NAKH900104</td>
<td>WARP780101</td>
</tr>
<tr>
<td>------------</td>
<td>------------</td>
<td>------------</td>
</tr>
<tr>
<td>NAKH900108</td>
<td>NAKH900109</td>
<td>WERD780104</td>
</tr>
<tr>
<td>NAKH920113</td>
<td>NAKH920101</td>
<td>WOLS870103</td>
</tr>
<tr>
<td>NAKH920105</td>
<td>NAKH920106</td>
<td>ZASB820101</td>
</tr>
<tr>
<td>NISK860101</td>
<td>NOZY710101</td>
<td>ZIMJ680105</td>
</tr>
<tr>
<td>OOBM770104</td>
<td>OOBM770105</td>
<td>AURR980105</td>
</tr>
<tr>
<td>OOBM850104</td>
<td>OOBM850105</td>
<td>AURR980110</td>
</tr>
<tr>
<td>PALJ810104</td>
<td>PALJ810105</td>
<td>AURR980115</td>
</tr>
<tr>
<td>PALJ810109</td>
<td>PALJ810110</td>
<td>AURR980120</td>
</tr>
<tr>
<td>PALJ810114</td>
<td>PALJ810115</td>
<td>VINM940103</td>
</tr>
<tr>
<td>PONP800101</td>
<td>PONP800102</td>
<td>MUNV940104</td>
</tr>
<tr>
<td>PONP800106</td>
<td>PONP800107</td>
<td>BLAM930101</td>
</tr>
<tr>
<td>PRAM820103</td>
<td>PRAM900101</td>
<td>KUMS000103</td>
</tr>
<tr>
<td>PTIO830101</td>
<td>PTIO830102</td>
<td>NADH010102</td>
</tr>
<tr>
<td>QIAN880104</td>
<td>QIAN880105</td>
<td>NADH010107</td>
</tr>
<tr>
<td>QIAN880109</td>
<td>QIAN880110</td>
<td>CEDJ970102</td>
</tr>
<tr>
<td>QIAN880114</td>
<td>QIAN880115</td>
<td>FUSS010102</td>
</tr>
<tr>
<td>QIAN880119</td>
<td>QIAN880120</td>
<td>FUSS010107</td>
</tr>
<tr>
<td>QIAN880124</td>
<td>QIAN880125</td>
<td>FUSS010112</td>
</tr>
<tr>
<td>QIAN880129</td>
<td>QIAN880130</td>
<td>AVFM000105</td>
</tr>
<tr>
<td>QIAN880134</td>
<td>QIAN880135</td>
<td>YANJ020101</td>
</tr>
<tr>
<td>QIAN880139</td>
<td>RACS770101</td>
<td>PONP930101</td>
</tr>
<tr>
<td>RACS820102</td>
<td>RACS820103</td>
<td>KUHL950101</td>
</tr>
<tr>
<td>RACS820107</td>
<td>RACS820108</td>
<td>BASU050103</td>
</tr>
<tr>
<td>RACS820112</td>
<td>RACS820113</td>
<td>GEOJ030102</td>
</tr>
<tr>
<td>RADA880103</td>
<td>RADA880104</td>
<td>GEOJ030107</td>
</tr>
<tr>
<td>RADA880108</td>
<td>RICJ880101</td>
<td>ZOHJ040103</td>
</tr>
<tr>
<td>RICJ880105</td>
<td>RICJ880106</td>
<td>WOLR790101</td>
</tr>
<tr>
<td>RICJ880110</td>
<td>RICJ880111</td>
<td>GUYH080104</td>
</tr>
<tr>
<td>RICJ880115</td>
<td>RICJ880116</td>
<td>COWJ900101</td>
</tr>
<tr>
<td>ROBB760103</td>
<td>ROBB760104</td>
<td>CORJ870103</td>
</tr>
<tr>
<td>ROBB760108</td>
<td>ROBB760109</td>
<td>CORJ870108</td>
</tr>
<tr>
<td>ROBB760113</td>
<td>ROBB790101</td>
<td>MIYS990105</td>
</tr>
<tr>
<td>ROSM880102</td>
<td>ROSM880103</td>
<td>SNEP660104</td>
</tr>
<tr>
<td>SNEP660103</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 8. The names of the 3 physicochemical indices for amino acids.

<table>
<thead>
<tr>
<th>Hydrophobicity</th>
<th>hydrophilicity</th>
<th>mass</th>
</tr>
</thead>
</table>

Table 9. The names of the 2 physicochemical indices for amino acids.

<table>
<thead>
<tr>
<th>hydrophobicity</th>
<th>hydrophilicity</th>
</tr>
</thead>
</table>

References


