BioSeq-Analysis: a platform for DNA, RNA, and protein sequence analysis based on machine learning approaches

Feature description of BioSeq-Analysis

2017-10-21

Home-page: http://bioinformatics.hitsz.edu.cn/BioSeq-Analysis/

Content

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 [4], 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.

$$\mathbf{D} = \mathrm{R}_1\mathrm{R}_2\mathrm{R}_3\mathrm{R}_4\mathrm{R}_5\mathrm{R}_6\mathrm{R}_7\cdots\mathrm{R}_L \quad (1)$$

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 [5], transcription start site prediction, and studying the organization of nucleosomes around splice sites [5]. 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*:

$$\mathrm{ID} = \mathrm{Diversity}(X + S) - \mathrm{Diversity}(S) - \mathrm{Diversity}(X) \quad (2)$$

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 [6], [7] and [8].

1.1.4 Mismatch

Mismatch [9-11] 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:

$$f_{k,m}(\mathbf{D}) = \left(\sum_{j=0}^{m} c_{1,j}, \sum_{j=0}^{m} c_{2,j}, ..., \sum_{j=0}^{m} c_{4^k,j} \right) \tag{3}$$

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 [9, 11, 12] 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:

$$f_{k,m}(\mathbf{D}) = \left(\sum_{k-mer\ \alpha_1\ in\ x} \delta^{l(\alpha_1)}, \sum_{k-mer\ \alpha_2\ in\ x} \delta^{l(\alpha_2)}, ..., \sum_{k-mer\ \alpha_{4^k}\ in\ x} \delta^{l(\alpha_{4^k})} \right) \tag{4}$$

where

$$l(\alpha_i) = \begin{cases} 0, & \alpha_i \text{ is exact matching;} \\ |\alpha_i|, & \alpha_i \text{ is non-contiguous matching.} \end{cases} \tag{5}$$

$|\alpha_i|$ represents the length of α_i, i =1, 2,…, 4^k.

1.2 Autocorrelation

1.2.1 Dinucleotide-based auto covariance (DAC)

The DAC [13-15] 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:

$$\mathrm{DAC}(u, lag) = \sum_{i=1}^{L-lag-1} (P_u(\mathrm{R}_i\mathrm{R}_{i+1}) - \overline{P}_u)(P_u(\mathrm{R}_{i+lag}\mathrm{R}_{i+lag+1}) - \overline{P}_u) / (L - lag - 1) \tag{6}$$

where *u* is a physicochemical index, *L* is the length of the DNA sequence **D**, $P_u(\mathrm{R}_i\mathrm{R}_{i+1})$ means the numerical value of the physicochemical index *u* for the dinucleotide $\mathrm{R}_i\mathrm{R}_{i+1}$ at position *i*, $\overline{P}_u$ is the average value for physicochemical index *u* along the whole sequence:

$$\overline{P_u} = \sum_{j=1}^{L-1} P_u(\mathrm{R}_j\mathrm{R}_{j+1}) / (L-1) \quad (7)$$

In such a way, the length of DAC feature vector is *N*∗LAG, where *N* is the number of physicochemical indices (**Table 1**) extracted from two papers [15, 16], and LAG is the maximum of *lag* (*lag* = 1, 2, … , LAG).

1.2.2 Dinucleotide-based cross covariance (DCC)

Given a DNA sequence **D** (**Eq. 1**), the DCC [13, 15] 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:

$$\mathrm{DCC}(u_1,u_2,lag) = \sum_{i=1}^{L-lag-1} (P_{u_1}(R_iR_{i+1}) - \overline{P}_{u_1})(P_{u_2}(R_{i+lag}R_{i+lag+1}) - \overline{P}_{u_2}) / (L-lag-1) \quad (8)$$

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

$$\overline{P_u} = \sum_{j=1}^{L-1} P_u(\mathrm{R}_j\mathrm{R}_{j+1}) / (L-1) \quad (9)$$

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

1.2.3 Dinucleotide-based auto-cross covariance (DACC)

DACC[13, 15] is a combination of DAC and DCC. Therefore, the length of the DACC feature vector is *N***N**LAG, where *N* is the number of physicochemical indices (**Table 1**) and LAG is the maximum of *lag* (*lag* = 1, 2, …, LAG).

1.2.4 Trinucleotide-based auto covariance (TAC)

Given a DNA sequence **D** (**Eq. 1**), the TAC approach [13-15] measures the correlation of the same physicochemical index between two trinucleotides separated by *lag* nucleic acids along the sequence, which can be calculated as:

$$\mathrm{TAC}(lag,u) = \sum_{i=1}^{L-lag-2} (P_u(\mathrm{R}_i\mathrm{R}_{i+1}\mathrm{R}_{i+2}) - \overline{P}_u)(P_u(\mathrm{R}_{i+lag}\mathrm{R}_{i+lag+1}\mathrm{R}_{i+lag+2}) - \overline{P}_u) / (L-lag-2) \quad (10)$$

where *u* is a physicochemical index, *L* is the length of the DNA sequence, $P_u(\mathrm{R}_i\mathrm{R}_{i+1}\mathrm{R}_{i+2})$ represents the numerical value of the physicochemical index *u* for the trinucleotide $\mathrm{R}_i\mathrm{R}_{i+1}\mathrm{R}_{i+2}$ at position *i*, $\overline{P}_u$ is the average value for physicochemical index *u* along the whole sequence:

$$\overline{P_u} = \sum_{j=1}^{L-2} P_u(\mathrm{R}_j\mathrm{R}_{j+1}\mathrm{R}_{j+2}) / (L-2) \quad (11)$$

In such a way, the length of TAC feature vector is $N*$LAG, where N is the number of physicochemical indices (**Table 2**) extracted from [15], and LAG is the maximum of *lag* (lag=1, 2, …, LAG).

1.2.5 Trinucleotide-based cross covariance (TCC)

Given a DNA sequence **D** (**Eq. 1**), the TCC [13, 15] approach measures the correlation of two different physicochemical indices between two trinucleotides separated by *lag* nucleic acids along the sequence, which can be calculated by:

$$\mathrm{TCC}(u_1,u_2,\mathrm{lag}) = \sum_{i=1}^{L-lag-2} (P_{u_1}(\mathrm{R}_i\mathrm{R}_{i+1}\mathrm{R}_{i+2}) - \overline{P}_{u_1})(P_{u_2}(\mathrm{R}_{i+lag}\mathrm{R}_{i+lag+1}\mathrm{R}_{i+lag+2}) - \overline{P}_{u_2}) / (L\text{-}lag\text{-}2) \quad (12)$$

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

$$\overline{P_u} = \sum_{j=1}^{L-2} P_u(\mathrm{R}_j\mathrm{R}_{j+1}\mathrm{R}_{j+2}) / (L-2) \quad (13)$$

In such a way, the length of TCC feature vector is $N*(N\text{-}1)*$LAG, where N is the number of physicochemical index (**Table 2**) extracted from [15], and LAG is the maximum of *lag* (*lag* = 1, 2, …, LAG).

1.2.6 Trinucleotide-based auto-cross covariance (TACC)

TACC [13, 15] is a combination of TAC and TCC. Therefore, the length of the TACC feature vector is $N*N*$LAG, where N is the number of physicochemical indices (**Table 2**) extracted from [15], and LAG is the maximum of *lag* (*lag* = 1, 2, …, LAG).

1.2.7 Moran autocorrelation (MAC)

Given a DNA sequence **D** (**Eq. 1**), the MAC [15, 17] approach measures the correlation of the same properties between two residues separated by a distance of *lag* along the sequence, which can be calculated by:

$$\mathrm{MAC}(u,k,lag) = \frac{\left[1/\left(L-lag\text{-}k+1\right)\right] \sum_{i=1}^{L-lag-k+1} \left(P_u\left(x_i\right) - \overline{P}_u\left(x\right)\right)\left(P_u(x_{i+lag}) - \overline{P}_u\left(x\right)\right)}{\left(1/L\text{-k+1}\right) \sum_{i=1}^{L\text{-k+1}} \left(P_u\left(x_i\right) - \overline{P}_u\left(x\right)\right)^2} \quad (14)$$

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*, $\overline{P}_u(x)$ is the average value for physicochemical index *u* along the whole sequence. When *lag* = 1, the nearest neighbor correlations at a physicochemical property *u* are measured; When *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 [15, 18] approach measures the correlation of the same properties between two residues separated by a distance of *lag* along the sequence, which can be calculated by:

$$\mathrm{GAC}(u,k,lag)=\frac{\left[1/\left(L-lag\text{-}k+1\right)\right]\sum_{i=1}^{L-lag\text{-}k+1}\left(P_u(x_i)-P_u(x_{i+lag})\right)^2}{\left(1/L\text{-k+1}\right)\sum_{i=1}^{L\text{-k+1}}\left(P_u(x_i)-\overline{P}_u(x)\right)^2} \quad (15)$$

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*, $\overline{P}_u(x)$ is the average value for physicochemical index *u* along the whole sequence. When *lag* = 1, the nearest neighbor correlations at a physicochemical property *u* are measured; When *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 [15, 19] approach measures the correlation of the same properties between two residues separated by a distance of *lag* along the sequence, which can be calculated by:

$$\mathrm{NMBAC}(u,k,lag)=\frac{\sum_{i=1}^{L-lag\text{-}k+1}\left(P_u(x_i)\times P_u(x_{i+lag})\right)^2}{L\text{-}k-lag+1} \quad (16)$$

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 *lag* = 1, the nearest neighbor correlations at a physicochemical property *u* are measured; When *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 [20] 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:

$$\mathbf{D} = \begin{bmatrix} d_1 & d_2 & \cdots & d_{16} & d_{16+1} & \cdots & d_{16+\lambda} \end{bmatrix}^{\mathbf{T}} \tag{17}$$

where

$$d_k = \begin{cases} \dfrac{f_k}{\sum_{i=1}^{16} f_i + w\sum_{j=1}^{\lambda} \theta_j} & (1 \le k \le 16) \\ \dfrac{w\theta_{k-16}}{\sum_{i=1}^{16} f_i + w\sum_{j=1}^{\lambda} \theta_j} & (17 \le k \le 16+\lambda) \end{cases} \tag{18}$$

where f_k (k=1,2,⋯,16) is the normalized occurrence frequency of dinucleotides in the DNA sequence; the parameter λ 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; θ_j (j=1,2,⋯,λ) 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:

$$\begin{cases} \theta_1 = \dfrac{1}{L\text{-}2}\sum_{i=1}^{L-2} \Theta(\mathrm{R}_i\mathrm{R}_{i+1}, \mathrm{R}_{i+1}\mathrm{R}_{i+2}) \\ \theta_2 = \dfrac{1}{L\text{-}3}\sum_{i=1}^{L-3} \Theta(\mathrm{R}_i\mathrm{R}_{i+1}, \mathrm{R}_{i+2}\mathrm{R}_{i+3}) \\ \theta_3 = \dfrac{1}{L\text{-}4}\sum_{i=1}^{L-4} \Theta(\mathrm{R}_i\mathrm{R}_{i+1}, \mathrm{R}_{i+3}\mathrm{R}_{i+4}) \qquad (\lambda < L) \\ \cdots\cdots \\ \theta_\lambda = \dfrac{1}{L\text{-}1\text{-}\lambda}\sum_{i=1}^{L-1-\lambda} \Theta(\mathrm{R}_i\mathrm{R}_{i+1}, \mathrm{R}_{i+\lambda}\mathrm{R}_{i+\lambda+1}) \end{cases} \tag{19}$$

where the correlation function is given by

$$\Theta(\mathrm{R}_i\mathrm{R}_{i+1}, \mathrm{R}_j\mathrm{R}_{j+1}) = \frac{1}{\mu}\sum_{u=1}^{\mu} [P_u(\mathrm{R}_i\mathrm{R}_{i+1}) - P_u(\mathrm{R}_j\mathrm{R}_{j+1})]^2 \tag{20}$$

where μ is the number of physicochemical indices, in this approach, 6 indices reflecting the local DNA structural properties [20] (**Table 4**) are employed to generate the PseDNC feature vector; $P_u(\mathrm{R}_i\mathrm{R}_{i+1})$ ($P_u(\mathrm{R}_j\mathrm{R}_{j+1})$) represents the numerical value of the u-th ($u = 1,2,\cdots,\mu$) physicochemical index of the dinucleotide $\mathrm{R}_i\mathrm{R}_{i+1}$ ($\mathrm{R}_j\mathrm{R}_{j+1}$) at position i (j).

1.3.2 Pseudo k-tuple nucleotide composition (PseKNC)

PseKNC [21, 22] extends the PseDNC approach by incorporating *k*-tuple nucleotide composition.

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

$$\mathbf{D} = \begin{bmatrix} d_1 & d_2 & \cdots & d_{4^k} & d_{4^k+1} & \cdots & d_{4^k+\lambda} \end{bmatrix}^{\mathbf{T}} \tag{21}$$

where

$$d_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{4^k} f_i + w\sum_{j=1}^{\lambda}\theta_j} & (1 \le u \le 4^k) \\ \dfrac{w\theta_{u-4^k}}{\sum_{i=1}^{4^k} f_i + w\sum_{j=1}^{\lambda}\theta_j} & (4^k \le u \le 4^k + \lambda) \end{cases} \tag{22}$$

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

$$\theta_j = \frac{1}{L\text{-}j\text{-}1}\sum_{i=1}^{L-j-1}\Theta(\mathrm{R}_i\mathrm{R}_{i+1}, \mathrm{R}_{i+j}\mathrm{R}_{i+j+1}) \quad (j = 1, 2, \cdots, \lambda; \lambda < L) \tag{23}$$

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

$$\Theta(\mathrm{R}_i\mathrm{R}_{i+1}, \mathrm{R}_{i+j}\mathrm{R}_{i+j+1}) = \frac{1}{\mu}\sum_{v=1}^{\mu}[P_v(\mathrm{R}_i\mathrm{R}_{i+1}) - P_v(\mathrm{R}_{i+j}\mathrm{R}_{i+j+1})]^2 \tag{24}$$

where μ is the number of physicochemical indices, in this study, 6 indices reflecting the local DNA structural properties [20] (**Table 4**) are employed to generate the PseKNC feature vector; $P_v(\mathrm{R}_i\mathrm{R}_{i+1})$ ($P_v(\mathrm{R}_{i+j}\mathrm{R}_{i+j+1})$) represents the numerical value of the v-th ($v = 1, 2, \cdots, \mu$) physicochemical index for the dinucleotide $\mathrm{R}_i\mathrm{R}_{i+1}$ ($\mathrm{R}_{i+j}\mathrm{R}_{i+j+1}$) at position i (i+j).

For more information about this approach, please refer to [21, 22].

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

In PC-PseDNC-General [23] 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:

$$\mathbf{D} = \begin{bmatrix} d_1 & d_2 & \cdots & d_{16} & d_{16+1} & \cdots & d_{16+\lambda} \end{bmatrix}^{\mathbf{T}} \tag{25}$$

where

$$d_k = \begin{cases} \dfrac{f_k}{\sum_{i=1}^{16} f_i + w\sum_{j=1}^{\lambda}\theta_j} & (1 \le k \le 16) \\ \dfrac{w\theta_{k-16}}{\sum_{i=1}^{16} f_i + w\sum_{j=1}^{\lambda}\theta_j} & (16+1 \le k \le 16+\lambda) \end{cases} \tag{26}$$

where f_k（k=1,2,⋯,16) is the normalized occurrence frequency of dinucleotides in the DNA sequence; the parameter λ 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; θ_j (j=1, 2, ⋯, λ) 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:

$$\begin{cases} \theta_1 = \dfrac{1}{L\text{-}2}\sum_{i=1}^{L-2}\Theta(\mathrm{R}_i\mathrm{R}_{i+1}, \mathrm{R}_{i+1}\mathrm{R}_{i+2}) \\ \theta_2 = \dfrac{1}{L\text{-}3}\sum_{i=1}^{L-3}\Theta(\mathrm{R}_i\mathrm{R}_{i+1}, \mathrm{R}_{i+2}\mathrm{R}_{i+3}) \\ \theta_3 = \dfrac{1}{L\text{-}4}\sum_{i=1}^{L-4}\Theta(\mathrm{R}_i\mathrm{R}_{i+1}, \mathrm{R}_{i+3}\mathrm{R}_{i+4}) \qquad (\lambda < L-1) \\ \cdots\cdots \\ \theta_\lambda = \dfrac{1}{L\text{-}1\text{-}\lambda}\sum_{i=1}^{L-1-\lambda}\Theta(\mathrm{R}_i\mathrm{R}_{i+1}, \mathrm{R}_{i+\lambda}\mathrm{R}_{i+\lambda+1}) \end{cases} \tag{27}$$

where the correlation function is given by

$$\Theta(\mathrm{R}_i\mathrm{R}_{i+1}, \mathrm{R}_j\mathrm{R}_{j+1}) = \frac{1}{\mu}\sum_{u=1}^{\mu}[P_u(\mathrm{R}_i\mathrm{R}_{i+1}) - P_u(\mathrm{R}_j\mathrm{R}_{j+1})]^2 \tag{28}$$

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

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

In PC-PseTNC-General [23] 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:

$$\mathbf{D} = \begin{bmatrix} d_1 & d_2 & \cdots & d_{64} & d_{64+1} & \cdots & d_{64+\lambda} \end{bmatrix}^{\mathrm{T}} \tag{29}$$

where

$$d_k = \begin{cases} \dfrac{f_k}{\sum_{i=1}^{64} f_i + w\sum_{j=1}^{\lambda}\theta_j} & (1 \le k \le 64) \\ \dfrac{w\theta_{k-64}}{\sum_{i=1}^{64} f_i + w\sum_{j=1}^{\lambda}\theta_j} & (64+1 \le k \le 64+\lambda) \end{cases} \tag{30}$$

where f_k（k=1,2,⋯,64) is the normalized occurrence frequency of trinucleotide in the DNA sequence; the parameter λ 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; θ_j (j=1, 2, ⋯, λ) is called the j-tier correlation factor that reflects the sequence-order correlation between all the most contiguous trinucleotides along a DNA sequence, which is defined:

$$\begin{cases} \theta_1 = \dfrac{1}{L\text{-}3}\sum_{i=1}^{L-3}\Theta(\mathrm{R}_i\mathrm{R}_{i+1}\mathrm{R}_{i+2},\mathrm{R}_{i+1}\mathrm{R}_{i+2}\mathrm{R}_{i+3}) \\ \theta_2 = \dfrac{1}{L\text{-}4}\sum_{i=1}^{L-4}\Theta(\mathrm{R}_i\mathrm{R}_{i+1}\mathrm{R}_{i+2},\mathrm{R}_{i+2}\mathrm{R}_{i+3}\mathrm{R}_{i+4}) \\ \theta_3 = \dfrac{1}{L\text{-}5}\sum_{i=1}^{L-5}\Theta(\mathrm{R}_i\mathrm{R}_{i+1}\mathrm{R}_{i+2},\mathrm{R}_{i+3}\mathrm{R}_{i+4}\mathrm{R}_{i+5}) \\ \cdots\cdots \\ \theta_\lambda = \dfrac{1}{L\text{-}2\text{-}\lambda}\sum_{i=1}^{L-2-\lambda}\Theta(\mathrm{R}_i\mathrm{R}_{i+1}\mathrm{R}_{i+2},\mathrm{R}_{i+\lambda}\mathrm{R}_{i+\lambda+1}\mathrm{R}_{i+\lambda+2}) \end{cases} \quad (\lambda < L-2) \tag{31}$$

where the correlation function is given by

$$\Theta(\mathrm{R}_i\mathrm{R}_{i+1}\mathrm{R}_{i+2},\mathrm{R}_j\mathrm{R}_{j+1}\mathrm{R}_{j+2}) = \frac{1}{\mu}\sum_{u=1}^{\mu}[P_u(\mathrm{R}_i\mathrm{R}_{i+1}\mathrm{R}_{i+2}) - P_u(\mathrm{R}_j\mathrm{R}_{j+1}\mathrm{R}_{j+2})]^2 \tag{32}$$

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

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

SC-PseDNC-General [23] 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:

$$\mathbf{D} = \begin{bmatrix} d_1 & d_2 & \cdots & d_{16} & d_{16+1} & \cdots & d_{16+\lambda} & d_{16+\lambda+1} & \cdots & d_{16+\lambda\Lambda} \end{bmatrix}^{\mathbf{T}} \tag{33}$$

where

$$d_k = \begin{cases} \dfrac{f_k}{\sum_{i=1}^{16} f_i + w\sum_{j=1}^{\lambda\Lambda}\theta_j} & (1 \le k \le 16) \\ \dfrac{w\theta_{k-16}}{\sum_{i=1}^{16} f_i + w\sum_{j=1}^{\lambda\Lambda}\theta_j} & (17 \le k \le 16+\lambda\Lambda) \end{cases} \tag{34}$$

where f_k（k=1, 2, ⋯, 16) is the normalized occurrence frequency of dinucleotide in the DNA sequence; the parameter λ 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; Λ is the number of physicochemical indices (**Table 1**); θ_j (j = 1, 2, ⋯, λ) 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:

$$\begin{cases} \theta_1 = \dfrac{1}{L\text{-}3}\sum_{i=1}^{L-3} \mathrm{J}_{i,i+1}^{1} \\ \theta_2 = \dfrac{1}{L\text{-}3}\sum_{i=1}^{L-3} \mathrm{J}_{i,i+1}^{2} \\ \cdots\cdots \\ \theta_\Lambda = \dfrac{1}{L\text{-}3}\sum_{i=1}^{L-3} \mathrm{J}_{i,i+1}^{\Lambda} & \lambda < (L-2) \\ \cdots\cdots \\ \theta_{\lambda\Lambda-1} = \dfrac{1}{L\text{-}\lambda\text{-}2}\sum_{i=1}^{L-\lambda-2} \mathrm{J}_{i,i+\lambda}^{\Lambda-1} \\ \theta_{\lambda\Lambda} = \dfrac{1}{L\text{-}\lambda\text{-}2}\sum_{i=1}^{L-\lambda-2} \mathrm{J}_{i,i+\lambda}^{\Lambda} \end{cases} \tag{35}$$

The correlation function is given by

$$\mathrm{J}_{i,i+m}^{u} = P_u(\mathrm{R}_i\mathrm{R}_{i+1}) \cdot P_u(\mathrm{R}_{i+m}\mathrm{R}_{i+m+1}) \quad (u = 1,2,\cdots,\Lambda;\ m=1,2,\cdots,\lambda;\ i=1,2,\cdots,L\text{-}m\text{-}1) \tag{36}$$

where $P_u(\mathrm{R}_i\mathrm{R}_{i+1})$ ($P_u(\mathrm{R}_{i+m}\mathrm{R}_{i+m+1})$) represents the numerical value of the *u*-th ($u = 1, 2, \cdots, \mu$) physiochemical index for the dinucleotide $\mathrm{R}_i\mathrm{R}_{i+1}$ ($\mathrm{R}_{i+m}\mathrm{R}_{i+m+1}$) at position *i* (*i*+*m*).

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

SC-PseTNC-General [23] 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:

$$\mathbf{D} = \begin{bmatrix} d_1 & d_2 & \cdots & d_{64} & d_{64+1} & \cdots & d_{64+\lambda} & d_{64+\lambda+1} & \cdots & d_{64+\lambda\Lambda} \end{bmatrix}^{\mathbf{T}} \tag{37}$$

where

$$d_k = \begin{cases} \dfrac{f_k}{\sum_{i=1}^{64} f_i + w\sum_{j=1}^{\lambda\Lambda}\theta_j} & (1 \le k \le 64) \\ \dfrac{w\theta_{k-64}}{\sum_{i=1}^{64} f_i + w\sum_{j=1}^{\lambda\Lambda}\theta_j} & (64+1 \le k \le 64+\lambda\Lambda) \end{cases} \tag{38}$$

where f_k（k=1, 2, ⋯, 64) is the normalized occurrence frequency of trinucleotide in the DNA sequence; the parameter λ 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; Λ is the number of physicochemical indices (**Table 2**); θ_j (j = 1, 2, ⋯, λ) 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:

$$\begin{cases} \theta_1 = \dfrac{1}{L\text{-}4}\sum_{i=1}^{L-4} \mathrm{J}_{i,i+1}^{1} \\ \theta_2 = \dfrac{1}{L\text{-}4}\sum_{i=1}^{L-4} \mathrm{J}_{i,i+1}^{2} \\ \cdots\cdots \\ \theta_\Lambda = \dfrac{1}{L\text{-}4}\sum_{i=1}^{L-4} \mathrm{J}_{i,i+1}^{\Lambda} \qquad \lambda < (L-3) \\ \cdots\cdots \\ \theta_{\lambda\Lambda-1} = \dfrac{1}{L\text{-}\lambda\text{-}3}\sum_{i=1}^{L-\lambda-3} \mathrm{J}_{i,i+\lambda}^{\Lambda-1} \\ \theta_{\lambda\Lambda} = \dfrac{1}{L\text{-}\lambda\text{-}3}\sum_{i=1}^{L-\lambda-3} \mathrm{J}_{i,i+\lambda}^{\Lambda} \end{cases} \tag{39}$$

The correlation function is given by

$$\begin{cases} \mathrm{J}_{i,i+m}^{u} = P_u(\mathrm{R}_i\mathrm{R}_{i+1}\mathrm{R}_{i+2}) \cdot P_u(\mathrm{R}_{i+m}\mathrm{R}_{i+m+1}\mathrm{R}_{i+m+2}) \\ u = 1, 2, \cdots, \Lambda;\ m = 1,2,\cdots,\lambda;\ i = 1,2,\cdots,L\text{-}m\text{-}2 \end{cases} \tag{40}$$

where $P_u(\mathrm{R}_i\mathrm{R}_{i+1}\mathrm{R}_{i+2})$ ($P_u(\mathrm{R}_{i+m}\mathrm{R}_{i+m+1}\mathrm{R}_{i+m+2})$) represents the numerical value of the *u*-th ($u = 1,\ 2,\ \cdots,\ \mu$) physiochemical index for the tri-nucleotide $\mathrm{R}_i\mathrm{R}_{i+1}\mathrm{R}_{i+2}$ ($\mathrm{R}_{i+m}\mathrm{R}_{i+m+1}\mathrm{R}_{i+m+2}$) at position *i* (*i*+*m*).

2. RNA

2.1 Ribonucleic acid composition

2.1.1 Basic kmer (Kmer)

Basic kmer [24] 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.

$$\mathbf{R} = \mathrm{R}_1\mathrm{R}_2\mathrm{R}_3\mathrm{R}_4\mathrm{R}_5\mathrm{R}_6\mathrm{R}_7\cdots\mathrm{R}_L \tag{41}$$

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 [9-11] 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:

$$f_{k,m}(\mathbf{R}) = \left(\sum_{j=0}^{m} c_{1,j}, \sum_{j=0}^{m} c_{2,j}, ..., \sum_{j=0}^{m} c_{4^k,j} \right) \quad (42)$$

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

2.1.3 Subsequence

Subsequence [9, 11, 12] is an approach that allows non-contiguous matching. For a 3-mer "*AAC*" in a sequence **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 **R** is defined:

$$f_{k,m}(x) = \left(\sum_{k-mer\ \alpha_1\ in\ x} \delta^{l(\alpha_1)}, \sum_{k-mer\ \alpha_2\ in\ x} \delta^{l(\alpha_2)}, ..., \sum_{k-mer\ \alpha_{4^k}\ in\ x} \delta^{l(\alpha_{4^k})} \right) \quad (43)$$

where

$$l(\alpha_i) = \begin{cases} 0, & \alpha_i \text{ is exact matching;} \\ |\alpha_i|, & \alpha_i \text{ is non-contiguous matching.} \end{cases} \quad (44)$$

$|\alpha_i|$ represents the length of α_i, i=1, 2,…, 4^k.

2.2 Autocorrelation

2.2.1 Dinucleotide-based auto covariance (DAC)

The DAC[13-15] 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:

$$\mathrm{DAC}(u, lag) = \sum_{i=1}^{L-lag-1} (P_u(\mathrm{R}_i\mathrm{R}_{i+1}) - \overline{P}_u)(P_u(\mathrm{R}_{i+lag}\mathrm{R}_{i+lag+1}) - \overline{P}_u) / (L - lag - 1) \quad (45)$$

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

$$\overline{P_u} = \sum_{j=1}^{L-1} P_u(\mathrm{R}_j\mathrm{R}_{j+1}) / (L-1) \quad (46)$$

In such a way, the length of DAC feature vector is *N*∗LAG, where *N* is the number of physicochemical indices (**Table 5**), which are extracted from [15, 16], and LAG is the maximum of *lag* (*lag* = 1, 2, … , LAG).

2.2.2 Dinucleotide-based cross covariance (DCC)

Given an RNA sequence **R** (**Eq. 41**), the DCC [13, 15] 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:

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

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

$$\overline{P_u} = \sum_{j=1}^{L-1} P_u(\mathrm{R}_j\mathrm{R}_{j+1}) / (L-1) \quad (48)$$

In such a way, the length of the DCC feature vector is N*(N-1)*LAG, where N is the number of physicochemical indices (**Table 5**) and LAG is the maximum of *lag* (*lag*=1, 2, …, LAG).

2.2.3 Dinucleotide-based auto-cross covariance (DACC)

DACC [13, 15] is a combination of DAC and DCC. Therefore, the length of the DACC feature vector is N*N*LAG, where N is the number of physicochemical indices (**Table 5**) and LAG is the maximum of *lag* (*lag* = 1, 2, …, LAG).

2.2.4 Moran autocorrelation (MAC)

Given a RNA sequence **R** (**Eq. 41**), the MAC [15, 17] approach measures the correlation of the same properties between two residues separated by a distance of *lag* along the sequence, which can be calculated by:

$$\mathrm{MAC}(u,k,lag) = \frac{\left[1/\left(L-lag\text{-}k+1\right)\right]\sum_{i=1}^{L-lag-k+1}\left(P_u\left(x_i\right)-\overline{P}_u\left(x\right)\right)\left(P_u(x_{i+lag})-\overline{P}_u\left(x\right)\right)}{\left(1/L\text{-k+1}\right)\sum_{i=1}^{L\text{-k+1}}\left(P_u\left(x_i\right)-\overline{P}_u\left(x\right)\right)^2} \quad (49)$$

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**. $P_u\left(x\right)$ means the numerical value of the physicochemical index u for x at position i, $\overline{P}_u\left(x\right)$ is the average value for physicochemical index u along the whole sequence. When *lag* = 1, the nearest neighbor correlations at a physicochemical property u are measured; When *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 [15, 18] approach measures the correlation of the same properties between two residues separated by a distance of *lag*

along the sequence, which can be calculated by:

$$\mathrm{GAC}(u,k,lag)=\frac{\left[1/\left(L-lag\text{-}k\text{+}1\right)\right]\sum_{i=1}^{L-lag\text{-}k+1}\left(P_u\left(x_i\right)-P_u\left(x_{i+lag}\right)\right)^2}{\left(1/L\text{-k+}1\right)\sum_{i=1}^{L\text{-k+}1}\left(P_u\left(x_i\right)-\overline{P}_u\left(x\right)\right)^2} \tag{50}$$

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**. $P_u\left(x\right)$ 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 lag = 1, the nearest neighbor correlations at a physicochemical property u are measured; When 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 [15, 19] approach measures the correlation of the same properties between two residues separated by a distance of lag along the sequence, which can be calculated by:

$$\mathrm{NMBAC}(u,lag)=\frac{\sum_{i=1}^{L-lag\text{-}1}\left(P_u\left(x_i\right)\times P_u\left(x_{i+lag}\right)\right)^2}{L-lag\text{-}1} \tag{51}$$

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**. $P_u\left(x\right)$ means the numerical value of the physicochemical index u for x at position i. When lag = 1, the nearest neighbor correlations at a physicochemical property u are measured; When 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 [15] 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:

$$\mathbf{R}=\left[d_1 \quad d_2 \quad \cdots \quad d_{16} \quad d_{16+1} \quad \cdots \quad d_{16+\lambda}\right]^{\mathbf{T}} \tag{52}$$

where

$$d_k = \begin{cases} \dfrac{f_k}{\sum_{i=1}^{16} f_i + w\sum_{j=1}^{\lambda}\theta_j} & (1 \le k \le 16) \\ \dfrac{w\theta_{k-16}}{\sum_{i=1}^{16} f_i + w\sum_{j=1}^{\lambda}\theta_j} & (16+1 \le k \le 16+\lambda) \end{cases} \tag{53}$$

where f_k（k=1,2,⋯,16) is the normalized occurrence frequency of dinucleotide in the RNA sequence; the parameter λ 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; θ_j (j=1, 2, ⋯, λ) 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:

$$\begin{cases} \theta_1 = \dfrac{1}{L\text{-}2}\sum_{i=1}^{L-2}\Theta(\mathrm{R}_i\mathrm{R}_{i+1}, \mathrm{R}_{i+1}\mathrm{R}_{i+2}) \\ \theta_2 = \dfrac{1}{L\text{-}3}\sum_{i=1}^{L-3}\Theta(\mathrm{R}_i\mathrm{R}_{i+1}, \mathrm{R}_{i+2}\mathrm{R}_{i+3}) \\ \theta_3 = \dfrac{1}{L\text{-}4}\sum_{i=1}^{L-4}\Theta(\mathrm{R}_i\mathrm{R}_{i+1}, \mathrm{R}_{i+3}\mathrm{R}_{i+4}) \qquad (\lambda < L) \\ \cdots\cdots \\ \theta_\lambda = \dfrac{1}{L\text{-}1\text{-}\lambda}\sum_{i=1}^{L-1-\lambda}\Theta(\mathrm{R}_i\mathrm{R}_{i+1}, \mathrm{R}_{i+\lambda}\mathrm{R}_{i+\lambda+1}) \end{cases} \tag{54}$$

where the correlation function is given by

$$\Theta(\mathrm{R}_i\mathrm{R}_{i+1}, \mathrm{R}_j\mathrm{R}_{j+1}) = \frac{1}{\mu}\sum_{u=1}^{\mu}[P_u(\mathrm{R}_i\mathrm{R}_{i+1}) - P_u(\mathrm{R}_j\mathrm{R}_{j+1})]^2 \tag{55}$$

where μ is the number of physicochemical indices considered that are listed in the **Table 5**; $P_u(\mathrm{R}_i\mathrm{R}_{i+1})$ ($P_u(\mathrm{R}_j\mathrm{R}_{j+1})$) represents the numerical value of the u-th ($u = 1, 2, \cdots, \mu$) physicochemical index for the dinucleotide $\mathrm{R}_i\mathrm{R}_{i+1}$($\mathrm{R}_j\mathrm{R}_{j+1}$) at position i (j).

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

SC-PseDNC-General [15] 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:

$$\mathbf{R} = \begin{bmatrix} d_1 & d_2 & \cdots & d_{16} & d_{16+1} & \cdots & d_{16+\lambda} & d_{16+\lambda+1} & \cdots & d_{16+\lambda\Lambda} \end{bmatrix}^{\mathbf{T}} \tag{56}$$

where

$$d_k = \begin{cases} \dfrac{f_k}{\sum_{i=1}^{16} f_i + w\sum_{j=1}^{\lambda\Lambda}\theta_j} & (1 \le k \le 16) \\ \dfrac{w\theta_{k-16}}{\sum_{i=1}^{16} f_i + w\sum_{j=1}^{\lambda\Lambda}\theta_j} & (16+1 \le k \le 16+\lambda\Lambda) \end{cases} \tag{57}$$

where f_k（k=1, 2, ⋯, 16) is the normalized occurrence frequency of dinucleotides in the RNA sequence; the parameter λ 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; Λ is the number of physicochemical indices (**Table 5**); θ_j (j = 1, 2, ⋯, λ) 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:

$$\begin{cases} \theta_1 = \dfrac{1}{L\text{-}3}\sum_{i=1}^{L-3} \mathrm{J}_{i,i+1}^{1} \\ \theta_2 = \dfrac{1}{L\text{-}3}\sum_{i=1}^{L-3} \mathrm{J}_{i,i+1}^{2} \\ \cdots\cdots \\ \theta_\Lambda = \dfrac{1}{L\text{-}3}\sum_{i=1}^{L-3} \mathrm{J}_{i,i+1}^{\Lambda} \qquad \lambda < (L-2) \\ \cdots\cdots \\ \theta_{\lambda\Lambda-1} = \dfrac{1}{L\text{-}\lambda\text{-}2}\sum_{i=1}^{L-\lambda-2} \mathrm{J}_{i,i+\lambda}^{\Lambda-1} \\ \theta_{\lambda\Lambda} = \dfrac{1}{L\text{-}\lambda\text{-}2}\sum_{i=1}^{L-\lambda-2} \mathrm{J}_{i,i+\lambda}^{\Lambda} \end{cases} \tag{58}$$

The correlation function is given by

$$\begin{cases} \mathrm{J}_{i,i+m}^{u} = P_u(\mathrm{R}_i\mathrm{R}_{i+1}) \cdot P_u(\mathrm{R}_{i+m}\mathrm{R}_{i+m+1}) \\ u = 1, 2, \cdots, \Lambda;\ m = 1, 2, \cdots, \lambda;\ i = 1, 2, \cdots, L\text{-}\lambda\text{-}2 \end{cases} \tag{59}$$

$P_u(\mathrm{R}_i\mathrm{R}_{i+1})$ ($P_u(\mathrm{R}_{i+m}\mathrm{R}_{i+m+1})$) represents the numerical value of the u-th ($u = 1, 2, \cdots, \mu$) physiochemical index for the dinucleotide $\mathrm{R}_i\mathrm{R}_{i+1}$ ($\mathrm{R}_{i+m}\mathrm{R}_{i+m+1}$) at position i (i+m).

2.4 Predicted structure composition

2.4.1 Local structure-sequence triplet element (Triplet)

The Triplet[25] 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 [26] (released 2.1.6), we have

$$\mathbf{R} = \Psi_1\Psi_2\Psi_3\Psi_4\Psi_5\cdots\Psi_L \tag{60}$$

where Ψ_1 denotes the structure status of R_1, Ψ_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: "(((", "((.", "(..", "(.(", ".((", ".(.", "..(" and "...". Considering the middle nucleotide among the 3 adjacent nucleotides, there are 32 (4 × 8) possible structure-sequence combinations, which we denote as $f_{\mathrm{A}}(\text{"(((")$, $f_{\mathrm{C}}(\text{"(((")$, etc.

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

$$\mathbf{D} = [\ f_{\mathrm{A}}(\text{"(((")} \quad f_{\mathrm{A}}(\text{"((.")} \quad \cdots \quad f_{\mathrm{A}}(\text{"..."}) \quad f_{\mathrm{C}}(\text{"(((")} \quad \cdots \quad f_{\mathrm{U}}(\text{"..."})\ ]^{\mathrm{T}} \tag{61}$$

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** (**Eq. 41**), we can formulate its secondary structure as **Eq. 60**. They can be any of the 10 structure statuses; i.e.,

$$\begin{gathered} \Psi_i \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}, \mathrm{A\text{-}U}, \mathrm{U\text{-}A}, \mathrm{G\text{-}C}, \mathrm{C\text{-}G}, \mathrm{G\text{-}U}, \mathrm{U\text{-}G}\} \\ i = 1, 2, \cdots, L \end{gathered} \tag{62}$$

where A, C, G, U represent the structure statuses of the four unpaired nucleobases, while A–U, U–A, G–C, C–G, G–U, U–G represent the structure statuses of the six paired bases.

The PseSSC [27] approach formulates a feature vector containing $10^n + \lambda$ components as given by

$$\mathbf{R} = \left[\ f_1^* \quad f_2^* \quad f_3^* \quad \cdots \quad f_{10^n}^* \quad f_{10^n+1}^* \quad \cdots \quad f_{10^n+\lambda}^*\ \right]^{\mathrm{T}} \tag{63}$$

where

$$f^* = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{10^n} f_i + w\sum_{j=1}^{\lambda} \theta_j} & (1 \le u \le 10^n) \\ \dfrac{w\theta_{u-10^n}}{\sum_{i=1}^{10^n} f_i + w\sum_{j=1}^{\lambda} \theta_j} & (10^n + 1 \le u \le 10^n + \lambda) \end{cases} \tag{64}$$

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 θ_j is the *j*-tier sequence correlation factor given by

$$\begin{cases} \theta_1 = \dfrac{1}{L-1}\sum_{i=1}^{L-1}\Theta(\Psi_i,\Psi_{i+1}) \\ \theta_2 = \dfrac{1}{L-2}\sum_{i=1}^{L-2}\Theta(\Psi_i,\Psi_{i+2}) \\ \theta_3 = \dfrac{1}{L-3}\sum_{i=1}^{L-3}\Theta(\Psi_i,\Psi_{i+3}) \qquad (\lambda < L) \\ \cdots\cdots \\ \theta_\lambda = \dfrac{1}{L-\lambda}\sum_{i=1}^{L-\lambda}\Theta(\Psi_i,\Psi_{i+\lambda}) \end{cases} \tag{65}$$

where λ is an integer, representing the highest counted rank (or tier) of the structural correlation along an RNA chain; θ_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

$$\Theta(\Psi_i,\Psi_j) = \left[F(\Psi_i) - F(\Psi_j)\right]^2 \tag{66}$$

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

$$\begin{cases} D(\Psi_i,\Psi_j\,|\,0) & \text{if } k=0 \text{ then } i=j \\ D(\Psi_i,\Psi_j\,|\,1) & \text{if } k=1 \\ D(\Psi_i,\Psi_j\,|\,2) & \text{if } k=2 \\ \vdots & \vdots \\ D(\Psi_i,\Psi_j\,|\,L-1) & \text{if } k=L-1 \end{cases} \tag{67}$$

where Ψ_i and Ψ_j can be any of the 10 structure statuses of an RNA chain **R** (cf. **Eq. 62**), and k ($0 \le k \le L-1$) represents the distance between structure statuses Ψ_i and Ψ_j along the RNA chain **R**. Suppose Ψ_i is A–U, Ψ_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 [28] formulates a feature vector as below:

$$[d_1 d_2 d_3 \ldots d_u \ldots d_\Omega d_{\Omega+1} d_{\Omega+2} \ldots d_{\Omega+\lambda}]^T \tag{68}$$

where

$$d_u = \begin{cases} \dfrac{f_u}{1+w\sum_{j=1}^{\lambda}\theta_j} & (1 \le u \le \Omega) \\ \dfrac{w\theta_{u-\Omega}}{1+w\sum_{j=1}^{\lambda}\theta_j} & (\Omega+1 \le u \le \Omega+\lambda) \end{cases} \quad (69)$$

where θ_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

$$f_u = \begin{cases} f\left(D\left(\Psi_i, \Psi_j \mid 0\right)\right) & \text{if } 1 \le u \le 10 \\ f\left(D\left(\Psi_i, \Psi_j \mid 1\right)\right) & \text{if } 11 \le u \le 110 \\ f\left(D\left(\Psi_i, \Psi_j \mid 2\right)\right) & \text{if } 111 \le u \le 210 \\ \vdots & \vdots \\ f\left(D\left(\Psi_i, \Psi_j \mid n\right)\right) & \text{if } 10 + 100(n-1) \le u \le 10 + 100n \end{cases} \quad (70)$$

3. Protein

3.1 Amino acid composition

3.1.1 Basic kmer (Kmer)

Basic kmer [29] 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 [30] 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 * 20 * d_{MAX}, where 20 is the size of the alphabet of amino acids and d_{MAX} is the distance threshold which representing the maximum distance between residue pairs. For more information of this approach, please refer to [30].

3.1.3 PseAAC of Distance-Pairs and reduced alphabet scheme (Distance Pair)

PseAAC of Distance-Pairs and reduced alphabet scheme [31] 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:

$$\begin{cases} \text{cp(13)} = \{\text{MF; IL; V; A; C; WYQHP; G; T; S; N; RK; D; E}\} \\ \text{cp(14)} = \{\text{IMV; L; F; WY; G; P; C; A; S; T; N; HRKQ; E; D}\} \\ \text{cp(19)} = \{\text{P; G; E; K; R; Q; D; S; N; T; H; C; I; V; W; YF; A; L; M}\} \\ \text{cp(20)} = \{\text{A; C; D; E; F; G; H; I; K; L; M; N; P; Q; R; S; T; V; W; Y}\} \end{cases} \quad (71)$$

For more information of this approach, please refer to [31].

3.2 Autocorrelation

3.2.1 Auto covariance (AC)

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

$$\mathbf{P} = \mathrm{R}_1\mathrm{R}_2\mathrm{R}_3\mathrm{R}_4\mathrm{R}_5\mathrm{R}_6\mathrm{R}_7\cdots\mathrm{R}_L \quad (72)$$

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 [13, 14, 32] 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:

$$\mathrm{AC}(i, lag) = \sum_{i=1}^{L-lag} (P_u(\mathrm{R}_i) - \overline{P}_u)(P_u(\mathrm{R}_{i+lag}) - \overline{P}_u) / (L - lag) \quad (73)$$

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

$$\overline{P_u} = \sum_{j=1}^{L} P_u(\mathrm{R}_j) / L \quad (74)$$

In such a way, the length of AC feature vector is $N*\mathrm{LAG}$, where N is the number of physicochemical indices (**Table 7**) extracted from AAindex [33]; LG is the maximum of *lag* (*lag*=1,2,...,LG).
For more information of this approach, please refer to [13, 14].

3.2.2 Cross covariance (CC)

Given a protein sequence **P** (**Eq. 72**), the CC [13, 14, 32] approach measures the correlation of two different properties between two residues separated by a distance of *lag* along the sequence, which can be calculated by:

$$\mathrm{CC}(u_1,u_2,lag) = \sum_{i=1}^{L-lag} (P_{u_1}(R_i) - \overline{P}_{u_1})(P_{u_2}(R_{i+lag}) - \overline{P}_{u_2}) / (L - lag) \quad (75)$$

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

$$\overline{P_u} = \sum_{j=1}^{L} P_u(\mathrm{R}_j) / L \quad (76)$$

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

For more information of this approach, please refer to [13, 14].

3.2.3 Auto-cross covariance (ACC)

ACC [13, 14, 32] is a combination of AC and CC. Therefore, the length of the ACC feature vector is N*N*LAG, where N is the number of physicochemical indices (**Table 7**) and LAG is the maximum of *lag* (*lag* = 1, 2, …, LAG).

3.2.4 Physicochemical distance transformation (PDT)

Physicochemical distance transformation (PDT) [34] 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) [35], 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 [34].

3.3 Pseudo amino acid composition

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

PC-PseAAC [36] 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:

$$\mathbf{P} = \begin{bmatrix} x_1 & x_2 & \cdots & x_{20} & x_{20+1} & \cdots & x_{20+\lambda} \end{bmatrix}^{\mathbf{T}} \quad (77)$$

where

$$x_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{\lambda}\theta_j} & (1 \le u \le 20) \\ \dfrac{w\theta_{u-20}}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{\lambda}\theta_j} & (20+1 \le u \le 20+\lambda) \end{cases} \quad (78)$$

where f_i (i=1,2,···,20) is the normalized occurrence frequency of the 20 amino acids in the protein **P**; the parameter λ 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; θ_j (j=1,2,···,λ) 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:

$$\begin{cases} \theta_1 = \dfrac{1}{L\text{-}1}\sum_{i=1}^{L-1}\Theta(\mathrm{R}_i, \mathrm{R}_{i+1}) \\ \theta_2 = \dfrac{1}{L\text{-}2}\sum_{i=1}^{L-2}\Theta(\mathrm{R}_i, \mathrm{R}_{i+2}) \\ \theta_3 = \dfrac{1}{L\text{-}3}\sum_{i=1}^{L-3}\Theta(\mathrm{R}_i, \mathrm{R}_{i+3}) \qquad (\lambda < L) \\ \cdots\cdots \\ \theta_\lambda = \dfrac{1}{L\text{-}\lambda}\sum_{i=1}^{L-\lambda}\Theta(\mathrm{R}_i, \mathrm{R}_{i+\lambda}) \end{cases} \quad (79)$$

where the correlation function is given by

$$\Theta(\mathrm{R}_i, \mathrm{R}_j) = \frac{1}{3}\left\{\left[H_1(\mathrm{R}_j) - H_1(\mathrm{R}_i)\right]^2 + \left[H_2(\mathrm{R}_j) - H_2(\mathrm{R}_i)\right]^2 + \left[M(\mathrm{R}_j) - M(\mathrm{R}_i)\right]^2\right\} \quad (80)$$

where $H_1(\mathrm{R}_i)$, $H_2(\mathrm{R}_i)$, and $M(\mathrm{R}_i)$ 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:

$$\begin{cases} H_1(i) = \dfrac{H_1^0(i) - \sum_{i=1}^{20} \dfrac{H_1^0(i)}{20}}{\sqrt{\dfrac{\sum_{i=1}^{20} \left[H_1^0(i) - \sum_{i=1}^{20} \dfrac{H_1^0(i)}{20} \right]^2}{20}}} \\ H_2(i) = \dfrac{H_2^0(i) - \sum_{i=1}^{20} \dfrac{H_2^0(i)}{20}}{\sqrt{\dfrac{\sum_{i=1}^{20} \left[H_2^0(i) - \sum_{i=1}^{20} \dfrac{H_2^0(i)}{20} \right]^2}{20}}} \\ M(i) = \dfrac{M^0(i) - \sum_{i=1}^{20} \dfrac{M^0(i)}{20}}{\sqrt{\dfrac{\sum_{i=1}^{20} \left[M^0(i) - \sum_{i=1}^{20} \dfrac{M^0(i)}{20} \right]^2}{20}}} \end{cases} \quad (81)$$

where $H_1^0(i)$ is the original hydrophobicity value of the *i*-th amino acid; $H_2^0(i)$ the corresponding original hydrophilicity value; $M^0(i)$ the mass of the *i*-th amino acid side chain.

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

SC-PseAAC [37] is a variant of PC-PseAAC. Given a protein sequence **P** (**Eq. 72**), the SC-PseAAC feature vector of **P** is defined:

$$\mathbf{P} = \left[p_1 \quad p_2 \quad \cdots \quad p_{20} \quad p_{20+1} \quad \cdots \quad p_{20+\lambda} \quad p_{20+\lambda+1} \quad \cdots \quad p_{20+2\lambda} \right]^{\mathbf{T}} \quad (82)$$

where

$$p_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + w \sum_{j=1}^{2\lambda} \tau_j} & (1 \le u \le 20) \\ \dfrac{w\tau_{u-20}}{\sum_{i=1}^{20} f_i + w \sum_{j=1}^{2\lambda} \tau_j} & (20+1 \le u \le 20+2\lambda) \end{cases} \quad (83)$$

where f_i (i = 1, 2,..., 20) is the normalized occurrence frequency of the 20 native amino acids in the protein **P**; the parameter λ 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; τ_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:

$$\begin{cases} \tau_1 = \dfrac{1}{L\text{-}1}\sum_{i=1}^{L-1} H_{i,i+1}^1 \\ \tau_2 = \dfrac{1}{L\text{-}1}\sum_{i=1}^{L-1} H_{i,i+1}^2 \\ \tau_3 = \dfrac{1}{L\text{-}2}\sum_{i=1}^{L-2} H_{i,i+2}^1 \\ \tau_4 = \dfrac{1}{L\text{-}2}\sum_{i=1}^{L-2} H_{i,i+2}^2 \qquad (\lambda < L-1) \\ \cdots\cdots \\ \tau_{2\lambda-1} = \dfrac{1}{L\text{-}\lambda}\sum_{i=1}^{L-\lambda} H_{i,i+\lambda}^1 \\ \tau_{2\lambda} = \dfrac{1}{L\text{-}\lambda}\sum_{i=1}^{L-\lambda} H_{i,i+\lambda}^2 \end{cases} \tag{84}$$

where $H_{i,\mathrm{j}}^1$ and $H_{i,\mathrm{j}}^2$ are the hydrophobicity and hydrophilicity correlation functions given by

$$\begin{cases} H_{i,j}^1 = h^1(\mathrm{R}_i)\cdot h^1(\mathrm{R}_j) \\ H_{i,j}^2 = h^2(\mathrm{R}_i)\cdot h^2(\mathrm{R}_j) \end{cases} \tag{85}$$

where $h^1(\mathrm{R}_i)$ and $h^2(\mathrm{R}_i)$ are, respectively, the hydrophobicity and hydrophilicity values(**Table 9**) for the *i*-th (*i* = 1, 2, . . . ,*L*) amino acid in **Eq. 72**, and the dot (•) means the multiplication sign.

Note that before substituting the values of hydrophobicity and hydrophilicity into **Eq. 85**, they are all subjected to a standard conversion as described by the following equation:

$$\begin{cases} h^1(\mathrm{R}_i) = \dfrac{h_0^1(\mathrm{R}_i) - \sum_{k=1}^{20} \dfrac{h_0^1(\mathbb{R}_k)}{20}}{\sqrt{\dfrac{\sum_{u=1}^{20}\left[h_0^1(\mathbb{R}_u) - \sum_{k=1}^{20} \dfrac{h_0^1(\mathbb{R}_k)}{20}\right]^2}{20}}} \\ h^2(\mathrm{R}_i) = \dfrac{h_0^2(\mathrm{R}_i) - \sum_{k=1}^{20} \dfrac{h_0^2(\mathbb{R}_k)}{20}}{\sqrt{\dfrac{\sum_{u=1}^{20}\left[h_0^2(\mathbb{R}_u) - \sum_{k=1}^{20} \dfrac{h_0^2(\mathbb{R}_k)}{20}\right]^2}{20}}} \end{cases} \tag{86}$$

where we use the $\mathbb{R}_i$ (*i* = 1, 2, . . . , 20) to represent the 20 native amino acids. The symbols h_0^1 and h_0^2 represent the original hydrophobicity and hydrophilicity values of the amino acid in the brackets right after the symbols.

For more information of the SC-PseAAC, please refer to [37].

3.3.3 General parallel correlation pseudo amino acid composition (PC-PseAAC-General)

The PC-PseAAC-General approach [32] cannot only incorporate comprehensive built-in indices (**Table 7**) extracted from AAindex [33], 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:

$$\mathbf{P} = \begin{bmatrix} x_1 & x_2 & \cdots & x_{20} & x_{20+1} & \cdots & x_{20+\lambda} \end{bmatrix}^{\mathbf{T}} \tag{87}$$

where

$$x_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{\lambda}\theta_j} & (1 \le u \le 20) \\ \dfrac{w\theta_{u-20}}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{\lambda}\theta_j} & (20+1 \le u \le 20+\lambda) \end{cases} \tag{88}$$

where f_i (i=1,2,···,20) is the normalized occurrence frequency of the 20 amino acids in the protein **P**; the parameter λ 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; θ_j (j=1,2,···,λ) 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:

$$\begin{cases} \theta_1 = \dfrac{1}{L\text{-}1}\sum_{i=1}^{L-1}\Theta(\mathrm{R}_i, \mathrm{R}_{i+1}) \\ \theta_2 = \dfrac{1}{L\text{-}2}\sum_{i=1}^{L-2}\Theta(\mathrm{R}_i, \mathrm{R}_{i+2}) \\ \theta_3 = \dfrac{1}{L\text{-}3}\sum_{i=1}^{L-3}\Theta(\mathrm{R}_i, \mathrm{R}_{i+3}) \\ \cdots\cdots \\ \theta_\lambda = \dfrac{1}{L\text{-}\lambda}\sum_{i=1}^{L-\lambda}\Theta(\mathrm{R}_i, \mathrm{R}_{i+\lambda}) \end{cases} \quad (\lambda < L) \tag{89}$$

where the correlation function is given by

$$\Theta(\mathrm{R}_i, \mathrm{R}_j) = \frac{1}{\mu}\sum_{u=1}^{\mu}[H_u(\mathrm{R}_i) - H_u(\mathrm{R}_j)]^2 \tag{90}$$

where μ is the number of physicochemical indices considered that listed in the **Table 7**; $H_u(\mathrm{R}_i)$ is the u-th physicochemical index value of the amino acid R_i; $H_u(\mathrm{R}_j)$, 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:

$$H_u(i) = \frac{H_u^0(i) - \sum_{i=1}^{20} \frac{H_u^0(i)}{20}}{\sqrt{\frac{\sum_{i=1}^{20} \left[H_u^0(i) - \sum_{i=1}^{20} \frac{H_u^0(i)}{20} \right]^2}{20}}} \tag{91}$$

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 [32] cannot only incorporate comprehensive built-in indices (**Table 7**) extracted from AAindex [33], 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:

$$\mathbf{P} = \left[p_1 \quad p_2 \quad \cdots \quad p_{20} \quad p_{20+1} \quad \cdots \quad p_{20+\lambda} \quad p_{20+\lambda+1} \quad \cdots \quad p_{20+\lambda\Lambda} \right]^{\mathbf{T}} \tag{92}$$

where

$$p_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + w \sum_{j=1}^{\lambda\Lambda} \tau_j} & (1 \le u \le 20) \\ \dfrac{w\tau_{u-20}}{\sum_{i=1}^{20} f_i + w \sum_{j=1}^{\lambda\Lambda} \tau_j} & (20+1 \le u \le 20+\lambda\Lambda) \end{cases} \tag{93}$$

where f_i (i = 1, 2,... , 20) is the normalized occurrence frequency of the 20 native amino acids in the protein **P**, the parameter λ 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; Λ is the number of physicochemical indices (**Table 7**); τ_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:

$$\begin{cases} \tau_1 = \dfrac{1}{L\text{-}1} \sum_{i=1}^{L-1} H_{i,i+1}^1 \\ \tau_2 = \dfrac{1}{L\text{-}1} \sum_{i=1}^{L-1} H_{i,i+1}^2 \\ \cdots\cdots \\ \tau_\Lambda = \dfrac{1}{L\text{-}1} \sum_{i=1}^{L-1} H_{i,i+1}^\Lambda \qquad \lambda < (L-1) \\ \cdots\cdots \\ \tau_{\lambda\Lambda-1} = \dfrac{1}{L\text{-}\lambda} \sum_{i=1}^{L-\lambda} H_{i,i+\lambda}^{\Lambda-1} \\ \tau_{\lambda\Lambda} = \dfrac{1}{L\text{-}\lambda} \sum_{i=1}^{L-\lambda} H_{i,i+\lambda}^{\Lambda} \end{cases} \tag{94}$$

where $H_{i,i+m}^{\xi}$ is the correlation function given by

$$\begin{cases} H_{i,i+m}^{\zeta} = h^{\zeta}(\mathrm{R}_i) \cdot h^{\zeta}(\mathrm{R}_{i+m}) \\ \zeta = 1, 2, \cdots, \Lambda;\ m = 1, 2, \cdots, \lambda;\ i = 1, 2, \cdots, L - m \end{cases} \tag{95}$$

where $h^{\zeta}(\mathrm{R}_i)$ is the ζ-th physicochemical value for the i-th (i = 1, 2, . . . ,L) amino acid in **Eq. 72**, and the dot (•) 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:

$$h^{\zeta}(\mathrm{R}_i) = \frac{h_0^{\zeta}(\mathrm{R}_i) - \sum_{k=1}^{20} \frac{h_0^{\zeta}(\mathbb{R}_k)}{20}}{\sqrt{\frac{\sum_{u=1}^{20} \left[h_0^{\zeta}(\mathbb{R}_u) - \sum_{k=1}^{20} \frac{h_0^{\zeta}(\mathbb{R}_k)}{20} \right]^2}{20}}} \tag{96}$$

where we use the $\mathbb{R}_i$ (i = 1, 2, . . . , 20) to represent the 20 native amino acids. The symbols h_0^{ζ} represent the ζ-th original physicochemical value of the amino acid in the brackets right after the symbols.

3.4 Profile-based features

3.4.1 Top-n-gram

Top-n-gram [38] 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 [39] 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 [38].

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

The process of profile-based PDT [34] is similar as that of sequence-based PDT [34]. 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 [34].

3.4.3 Distance-based Top-n-gram (DT)

Distance-based Top-n-gram [30] is a profile-based method which considers the distances between Top-n-gram [38] 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 * 20 * d_{MAX}, where 20 is the size of the alphabet of amino acids and d_{MAX} is the distance threshold which representing the maximum distance between Top-1-gram pairs. For more information of this approach, please refer to [30].

3.4.4 Profile-based Auto covariance (AC-PSSM)

AC-PSSM [13] 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 *lag* along the sequence, which can be calculated as:

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

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, $\overline{S}_i$ is the average score for amino acid i along the whole sequence:

$$\overline{S}_i=\sum_{j=1}^{L}S_{i,j}/L \tag{98}$$

In such a way, the number of AC variables can be calculated as 20∗LAG, where LAG is the maximum of *lag* (*lag*=1, 2, ..., LAG).

3.4.5 Profile-based Cross covariance (CC-PSSM)

CC-PSSM [13] 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 lag along the sequence, which can be calculated by:

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

where $i1$,$i2$ are two different amino acids and $\overline{S}_{i1}$ ($\overline{S}_{i2}$) 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∗LAG.

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

ACC-PSSM [13], 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.

3.4.7 PSSM distance transformation (PSSM-DT)

PSSM-DT [40] can transform the PSSMs of different lengths into fixed-length vector. It can transform the PSSM information into uniform numeric representation by approximately measuring the occurrence probabilities of any pairs of amino acid separated by a distance along the sequence in a sequence.

3.4.8 PSSM relation transformation (PSSM-RT)

PSSM-RT [41] can transform the PSSMs of different lengths into fixed-length vector. It encodes residues by utilizing the relationships of evolutionary information between residues.

3.4.9 Sequence conservation score (CS)

Sequence conservation score (CS) [42] uses the characteristic that a sequence has been maintained by evolution despite speciation across species. We employed the software rate4site to generate the CS features [43].

3.5 Predicted structure features

3.5.1 Secondary structure (SS)

SS [44] can transform the regular structures in the polypeptide chain, α-helix, β-fold, β-turn into fixed-length vector, where 0, 1, and 3 represent the α-helix, β-fold and β-turn, respectively. We used the PSIPRED [45] package to generate the SS features.

3.5.2 Solvent accessible surface area (SASA)

SASA [46] is the surface area of a biomolecule that is accessible to a solvent. In the SASA method, the SPIDER2 [47] was used to generate the feature vectors.

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

Base stacking	Protein induced deformability	B-DNA twist
Propeller twist	Duplex stability:(freeenergy)	Duplex tability(disruptenergy)
Protein DNA twist	Stabilising energy of Z-DNA	Aida_BA_transition
Breslauer_dS	Electron_interaction	Hartman_trans_free_energy
Lisser_BZ_transition	Polar_interaction	SantaLucia_dG
Sarai_flexibility	Stability	Stacking_energy
Sugimoto_dS	Watson-Crick_interaction	Twist
Shift	Slide	Rise
Twist stiffness	Tilt stiffness	Shift_rise
Twist_shift	Enthalpy1	Twist_twist
Shift2	Tilt3	Tilt1
Slide (DNA-protein complex)1	Tilt_shift	Twist_tilt
Roll_rise	Stacking energy	Stacking energy1
Propeller Twist	Roll11	Rise (DNA-protein complex)
Roll2	Roll3	Roll1
Slide_slide	Enthalpy	Shift_shift
Flexibility_slide	Minor Groove Distance	Rise (DNA-protein complex)1
Roll (DNA-protein complex)1	Entropy	Cytosine content
Major Groove Distance	Twist (DNA-protein complex)	Purine (AG) content
Tilt_slide	Major Groove Width	Major Groove Depth
Free energy6	Free energy7	Free energy4
Free energy3	Free energy1	Twist_roll
Flexibility_shift	Shift (DNA-protein complex)1	Thymine content
Tip	Keto (GT) content	Roll stiffness
Entropy1	Roll_slide	Slide (DNA-protein complex)
Twist2	Twist5	Twist4
Tilt (DNA-protein complex)1	Twist_slide	Minor Groove Depth
Persistance Length	Rise3	Shift stiffness
Slide3	Slide2	Slide1
Rise1	Rise stiffness	Mobility to bend towards minor groove
Dinucleotide GC Content	A-philicity	Wedge
DNA denaturation	Bending stiffness	Free energy5
Breslauer_dG	Breslauer_dH	Shift (DNA-protein complex)
Helix-Coil_transition	Ivanov_BA_transition	Slide_rise

SantaLucia_dH	SantaLucia_dS	Minor Groove Width
Sugimoto_dG	Sugimoto_dH	Twist1
Tilt	Roll	Twist7
Clash Strength	Roll_roll	Roll (DNA-protein complex)
Adenine content	Direction	Probability contacting nucleosome core
Roll_shift	Shift_slide	Shift1
Tilt4	Tilt2	Free energy8
Twist (DNA-protein complex)1	Tilt_rise	Free energy2
Stacking energy2	Stacking energy3	Rise_rise
Tilt_tilt	Roll4	Tilt_roll
Minor Groove Size	GC content	Inclination
Slide stiffness	Melting Temperature1	Twist3
Tilt (DNA-protein complex)	Guanine content	Twist6
Major Groove Size	Twist_rise	Rise2
Melting Temperature	Free energy	Mobility to bend towards major groove
Bend		

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

Bendability (DNAse)	Bendability (consensus)	Trinucleotide GC Content
Consensus_roll	Consensus-Rigid	Dnase I
MW-Daltons	MW-kg	Nucleosome
Nucleosome positioning	Dnase I-Rigid	Nucleosome-Rigid

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

Base stacking	Protein induced deformability	B-DNA twist
Dinucleotide GC Content	A-philicity	Propeller twist
Duplex stability-free energy	Duplex stability-disrupt energy	DNA denaturation
Bending stiffness	Protein DNA twist	Stabilising energy of Z-DNA
Aida_BA_transition	Breslauer_dG	Breslauer_dH
Breslauer_dS	Electron_interaction	Hartman_trans_free_energy
Helix-Coil_transition	Ivanov_BA_transition	Lisser_BZ_transition

Polar_interaction	SantaLucia_dG	SantaLucia_dH
SantaLucia_dS	Sarai_flexibility	Stability
Stacking_energy	Sugimoto_dG	Sugimoto_dH
Sugimoto_dS	Watson-Crick_interaction	Twist
Tilt	Roll	Shift
Slide	Rise	Stacking energy
Bend	Tip	Inclination
Major Groove Width	Major Groove Depth	Major Groove Size
Major Groove Distance	Minor Groove Width	Minor Groove Depth
Minor Groove Size	Minor Groove Distance	Persistance Length
Melting Temperature	Mobility to bend towards major groove	Mobility to bend towards minor groove
Propeller Twist	Clash Strength	Enthalpy
Free energy	Twist_twist	Tilt_tilt
Roll_roll	Twist_tilt	Twist_roll
Tilt_roll	Shift_shift	Slide_slide
Rise_rise	Shift_slide	Shift_rise
Slide_rise	Twist_shift	Twist_slide
Twist_rise	Tilt_shift	Tilt_slide
Tilt_rise	Roll_shift	Roll_slide
Roll_rise	Slide stiffness	Shift stiffness
Roll stiffness	Rise stiffness	Tilt stiffness
Twist stiffness	Wedge	Direction
Flexibility_slide	Flexibility_shift	Entropy

Table 4. The names of the 6 physicochemical indices for dinucleotides (DNA).

Twist(DNA)	Tilt(DNA)	Roll(DNA)
Shift(DNA)	Slide(DNA)	Rise(DNA)

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

Shift (RNA)	Hydrophilicity (RNA)
Hydrophilicity (RNA)	GC content
Purine (AG) content	Keto (GT) content
Adenine content	Guanine content
Cytosine content	Thymine content
Slide (RNA)	Rise (RNA)
Tilt (RNA)	Roll (RNA)
Twist (RNA)	Stacking energy (RNA)
Enthalpy (RNA)	Entropy (RNA)
Free energy (RNA)	Free energy (RNA)
Enthalpy (RNA)	Entropy (RNA)

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

Shift	Slide	Rise
Tilt	Roll	Twist
Stacking energy	Enthalpy	Entropy
Free energy	Hydrophilicity	

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

Hydrophobicity	Hydrophilicity	Mass
ARGP820102	ARGP820103	BEGF750101
BHAR880101	BIGC670101	BIOV880101
BROC820102	BULH740101	BULH740102
BUNA790103	BURA740101	BURA740102
CHAM820102	CHAM830101	CHAM830102
CHAM830105	CHAM830106	CHAM830107
CHOC760101	CHOC760102	CHOC760103
CHOP780201	CHOP780202	CHOP780203
CHOP780206	CHOP780207	CHOP780208
CHOP780211	CHOP780212	CHOP780213
CHOP780216	CIDH920101	CIDH920102
CIDH920105	COHE430101	CRAJ730101
DAWD720101	DAYM780101	DAYM780201
EISD840101	EISD860101	EISD860102
FASG760102	FASG760103	FASG760104

FAUJ880101	FAUJ880102	FAUJ880103
FAUJ880106	FAUJ880107	FAUJ880108
FAUJ880111	FAUJ880112	FAUJ880113
FINA910102	FINA910103	FINA910104
GEIM800102	GEIM800103	GEIM800104
GEIM800107	GEIM800108	GEIM800109
GOLD730101	GOLD730102	GRAR740101
GUYH850101	HOPA770101	HOPT810101
HUTJ700103	ISOY800101	ISOY800102
ISOY800105	ISOY800106	ISOY800107
JANJ780102	JANJ780103	JANJ790101
JOND750102	JOND920101	JOND920102
KANM800101	KANM800102	KANM800103
KARP850102	KARP850103	KHAG800101
KRIW790101	KRIW790102	KRIW790103
LEVM760101	LEVM760102	LEVM760103
LEVM760106	LEVM760107	LEVM780101
LEVM780104	LEVM780105	LEVM780106
LIFS790102	LIFS790103	MANP780101
MAXF760103	MAXF760104	MAXF760105
MEEJ800101	MEEJ800102	MEEJ810101
MEIH800102	MEIH800103	MIYS850101
NAGK730103	NAKH900101	NAKH900102
NAKH900105	NAKH900106	NAKH900107
NAKH900110	NAKH900111	NAKH900112
NAKH920102	NAKH920103	NAKH920104
NAKH920107	NAKH920108	NISK800101
OOBM770101	OOBM770102	OOBM770103
OOBM850101	OOBM850102	OOBM850103
PALJ810101	PALJ810102	PALJ810103
PALJ810106	PALJ810107	PALJ810108
PALJ810111	PALJ810112	PALJ810113
PALJ810116	PARJ860101	PLIV810101
PONP800103	PONP800104	PONP800105
PONP800108	PRAM820101	PRAM820102
PRAM900102	PRAM900103	PRAM900104
QIAN880101	QIAN880102	QIAN880103
QIAN880106	QIAN880107	QIAN880108
QIAN880111	QIAN880112	QIAN880113
QIAN880116	QIAN880117	QIAN880118
QIAN880121	QIAN880122	QIAN880123
QIAN880126	QIAN880127	QIAN880128
QIAN880131	QIAN880132	QIAN880133
QIAN880136	QIAN880137	QIAN880138
RACS770102	RACS770103	RACS820101
RACS820104	RACS820105	RACS820106
RACS820109	RACS820110	RACS820111

RACS820114	RADA880101	RADA880102
RADA880105	RADA880106	RADA880107
RICJ880102	RICJ880103	RICJ880104
RICJ880107	RICJ880108	RICJ880109
RICJ880112	RICJ880113	RICJ880114
RICJ880117	ROBB760101	ROBB760102
ROBB760105	ROBB760106	ROBB760107
ROBB760110	ROBB760111	ROBB760112
ROSG850101	ROSG850102	ROSM880101
SIMZ760101	SNEP660101	SNEP660102
SUEM840101	SUEM840102	SWER830101
TANS770103	TANS770104	TANS770105
TANS770108	TANS770109	TANS770110
VASM830103	VELV850101	VENT840101
WEBA780101	WERD780101	WERD780102
WOEC730101	WOLR810101	WOLS870101
YUTK870101	YUTK870102	YUTK870103
ZIMJ680101	ZIMJ680102	ZIMJ680103
AURR980101	AURR980102	AURR980103
AURR980106	AURR980107	AURR980108
AURR980111	AURR980112	AURR980113
AURR980116	AURR980117	AURR980118
ONEK900101	ONEK900102	VINM940101
VINM940104	MUNV940101	MUNV940102
MUNV940105	WIMW960101	KIMC930101
PARS000101	PARS000102	KUMS000101
KUMS000104	TAKK010101	FODM020101
NADH010103	NADH010104	NADH010105
MONM990201	KOEP990101	KOEP990102
CEDJ970103	CEDJ970104	CEDJ970105
FUKS010103	FUKS010104	FUKS010105
FUKS010108	FUKS010109	FUKS010110
AVBF000101	AVBF000102	AVBF000103
AVBF000106	AVBF000107	AVBF000108
MITS020101	TSAJ990101	TSAJ990102
WILM950101	WILM950102	WILM950103
GUOD860101	JURD980101	BASU050101
SUYM030101	PUNT030101	PUNT030102
GEOR030103	GEOR030104	GEOR030105
GEOR030108	GEOR030109	ZHOH040101
BAEK050101	HARY940101	PONJ960101
OLSK800101	KIDA850101	GUYH850102
GUYH850105	ROSM880104	ROSM880105
BLAS910101	CASG920101	CORJ870101
CORJ870104	CORJ870105	CORJ870106
MIYS990101	MIYS990102	MIYS990103
ENGD860101	FASG890101	TANS770101

ANDN920101	ARGP820101	TANS770106
BEGF750102	BEGF750103	VASM830101
BIOV880102	BROC820101	VHEG790101
BUNA790101	BUNA790102	WERD780103
CHAM810101	CHAM820101	WOLS870102
CHAM830103	CHAM830104	YUTK870104
CHAM830108	CHOC750101	ZIMJ680104
CHOC760104	CHOP780101	AURR980104
CHOP780204	CHOP780205	AURR980109
CHOP780209	CHOP780210	AURR980114
CHOP780214	CHOP780215	AURR980119
CIDH920103	CIDH920104	VINM940102
CRAJ730102	CRAJ730103	MUNV940103
DESM900101	DESM900102	MONM990101
EISD860103	FASG760101	KUMS000102
FASG760105	FAUJ830101	NADH010101
FAUJ880104	FAUJ880105	NADH010106
FAUJ880109	FAUJ880110	CEDJ970101
FINA770101	FINA910101	FUKS010101
GARJ730101	GEIM800101	FUKS010106
GEIM800105	GEIM800106	FUKS010111
GEIM800110	GEIM800111	AVBF000104
GRAR740102	GRAR740103	AVBF000109
HUTJ700101	HUTJ700102	COSI940101
ISOY800103	ISOY800104	WILM950104
ISOY800108	JANJ780101	BASU050102
JANJ790102	JOND750101	GEOR030101
JUKT750101	JUNJ780101	GEOR030106
KANM800104	KARP850101	ZHOH040102
KLEP840101	KRIW710101	DIGM050101
KYTJ820101	LAWE840101	GUYH850103
LEVM760104	LEVM760105	JACR890101
LEVM780102	LEVM780103	CORJ870102
LEWP710101	LIFS790101	CORJ870107
MAXF760101	MAXF760102	MIYS990104
MAXF760106	MCMT640101	TANS770102
MEEJ810102	MEIH800101	TANS770107
NAGK730101	NAGK730102	VASM830102
NAKH900103	NAKH900104	WARP780101
NAKH900108	NAKH900109	WERD780104
NAKH900113	NAKH920101	WOLS870103
NAKH920105	NAKH920106	ZASB820101
NISK860101	NOZY710101	ZIMJ680105
OOBM770104	OOBM770105	AURR980105
OOBM850104	OOBM850105	AURR980110
PALJ810104	PALJ810105	AURR980115
PALJ810109	PALJ810110	AURR980120

PALJ810114	PALJ810115	VINM940103
PONP800101	PONP800102	MUNV940104
PONP800106	PONP800107	BLAM930101
PRAM820103	PRAM900101	KUMS000103
PTIO830101	PTIO830102	NADH010102
QIAN880104	QIAN880105	NADH010107
QIAN880109	QIAN880110	CEDJ970102
QIAN880114	QIAN880115	FUKS010102
QIAN880119	QIAN880120	FUKS010107
QIAN880124	QIAN880125	FUKS010112
QIAN880129	QIAN880130	AVBF000105
QIAN880134	QIAN880135	YANJ020101
QIAN880139	RACS770101	PONP930101
RACS820102	RACS820103	KUHL950101
RACS820107	RACS820108	BASU050103
RACS820112	RACS820113	GEOR030102
RADA880103	RADA880104	GEOR030107
RADA880108	RICJ880101	ZHOH040103
RICJ880105	RICJ880106	WOLR790101
RICJ880110	RICJ880111	GUYH850104
RICJ880115	RICJ880116	COWR900101
ROBB760103	ROBB760104	CORJ870103
ROBB760108	ROBB760109	CORJ870108
ROBB760113	ROBB790101	MIYS990105
ROSM880102	ROSM880103	SNEP660104
SNEP660103		

Table 8. The names of the 3 physicochemical indices for amino acids.

Hydrophobicity	hydrophilicity	mass

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

hydrophobicity	hydrophilicity

References

1. Liu B, Fang L, Wang S et al. Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy, Journal of theoretical biology 2015;385:153-159.
2. Noble WS, Kuehn S, Thurman R et al. Predicting the in vivo signature of human gene regulatory sequences, Bioinformatics 2005;21 Suppl 1:i338-343.
3. Gupta S, Dennis J, Thurman RE et al. Predicting human nucleosome occupancy from primary sequence, PLoS Comput Biol 2008;4:e1000134.
4. Lee D, Karchin R, Beer MA. Discriminative prediction of mammalian enhancers

from DNA sequence, Genome Res 2011;21:2167-2180.
5. Zhang L, Luo L. Splice site prediction with quadratic discriminant analysis using diversity measure, Nucleic acids research 2003;31:6214-6220.
6. Chen W, Luo L, Zhang L. The organization of nucleosomes around splice sites, Nucleic acids research 2010;38:2788-2798.
7. Liu G, Liu J, Cui X et al. Sequence-dependent prediction of recombination hotspots in Saccharomyces cerevisiae, Journal of theoretical biology 2012;293:49-54.
8. Liu B, Liu F, Fang L et al. repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects, Bioinformatics 2015;31:1307-1309.
9. El-Manzalawy Y, Dobbs D, Honavar V. Predicting flexible length linear B-cell epitopes, Computational Systems Bioinformatics 2008;7:121-132.
10. Leslie CS, Eskin E, Cohen A et al. Mismatch string kernels for discriminative protein classification, Bioinformatics 2004;20:467-476.
11. Luo L, Li D, Zhang W et al. Accurate Prediction of Transposon-Derived piRNAs by Integrating Various Sequential and Physicochemical Features, PloS one 2016;11:e0153268.
12. Lodhi H, Saunders C, Shawe-Taylor J et al. Text classification using string kernels, Journal of Machine Learning Research 2002;2:419-444.
13. Dong Q, Zhou S, Guan J. A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation, Bioinformatics 2009;25:2655-2662.
14. Guo Y, Yu L, Wen Z et al. Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences, Nucleic Acids Research 2008;36:3025-3030.
15. Chen W, Zhang X, Brooker J et al. PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions, Bioinformatics 2015b;31:119-120.
16. Friedel M, Nikolajewa S, Suhnel J et al. DiProDB: a database for dinucleotide properties, Nucleic Acids Res 2008;37:D37-D40.
17. Horne DS. Prediction of protein helix content from an autocorrelation analysis of sequence hydrophobicities, Biopolymers 1988;27:451-477.
18. Sokal RR, Thomson BA. Population structure inferred by local spatial autocorrelation: an example from an Amerindian tribal population, American journal of physical anthropology 2006;129:121-131.
19. Feng Z-P, Zhang C-T. Prediction of membrane protein types based on the hydrophobic index of amino acids, Journal of protein chemistry 2000;19:269-275.
20. Chen W, Feng PM, Lin H et al. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition, Nucleic Acids Research 2013:gks1450.
21. Guo S-H, Deng E-Z, Xu L-Q et al. iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition, Bioinformatics 2014:btu083.
22. Lin H, Deng E-Z, Ding H et al. iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition, Nucleic acids research 2014;42:12961-12972.

23. Liu B, Zhang D, Xu R et al. Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection, Bioinformatics 2014;30:472-479.
24. Wei L, Liao M, Gao Y et al. Improved and promising identification of human microRNAs by incorporating a high-quality negative set, IEEE/ACM Transactions on Computational Biology and Bioinformatics 2014;11:192-201.
25. Xue C, Li F, He T et al. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine, BMC Bioinformatics 2005;6:310.
26. Lorenz R, Bernhart SH, Höner zu Siederdissen C et al. ViennaRNA Package 2.0, Algorithms for Molecular Biology 2011;6:1-14.
27. Liu B, Fang L, Liu F et al. Identification of real microRNA precursors with a pseudo structure status composition approach, PLoS One 2015;10:e0121501.
28. Liu B, Fang L, Liu F et al. iMiRNA-PseDPC: microRNA precursor identification with a pseudo distance-pair composition approach, Journal of Biomolecular Structure and Dynamics 2016;34:223-235.
29. Liu B, Wang X, Lin L et al. A Discriminative Method for Protein Remote Homology Detection and Fold Recognition Combining Top-n-grams and Latent Semantic Analysis, BMC Bioinformatics 2008;9:510.
30. Liu B, Xu J, Zou Q et al. Using distances between Top-n-gram and residue pairs for protein remote homology detection, BMC bioinformatics 2014;15:1.
31. Liu B, Xu J, Lan X et al. iDNA-Prot| dis: identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition, PLoS One 2014;9:e106691.
32. Cao D-S, Xu Q-S, Liang Y-Z. propy: a tool to generate various modes of Chou's PseAAC, Bioinformatics 2013;29:960-962.
33. Kawashima S, Pokarowski P, Pokarowska M et al. AAindex: amino acid index database, progress report 2008, Nucleic Acids Res 2008;36:D202-D205.
34. Liu B, Wang X, Chen Q et al. Using amino acid physicochemical distance transformation for fast protein remote homology detection, PLoS One 2012;7:e46633.
35. Kawashima S, Kanehisa M. AAindex: amino acid index database, Nucleic Acids Research 2000;28:374-374.
36. Chou K-C. Prediction of protein cellular attributes using pseudo-amino-acid-composition, PROTEINS: Structure, Function, and Genetics 2001;43:246-255.
37. Chou K-C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, Bioinformatics 2005;21:10-19.
38. Liu B, Wang X, Lin L et al. A discriminative method for protein remote homology detection and fold recognition combining Top-n-grams and latent semantic analysis, BMC bioinformatics 2008;9:1.
39. Altschul SF, Madden TL, Schäffer AA et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic acids research 1997;25:3389-3402.
40. Xu R, Zhou J, Wang H et al. Identifying DNA-binding proteins by combining support vector machine and PSSM distance transformation, Bmc Systems Biology 2015;9:S10.
41. Zhou J, Lu Q, Xu R et al. EL_PSSM-RT: DNA-binding residue prediction by

integrating ensemble learning with PSSM Relation Transformation, BMC bioinformatics 2017;18:379.
42. Glaser F, Rosenberg YA, Pupko T et al. The ConSurf-HSSP database: the mapping of evolutionary conservation among homologs onto PDB structures, Proteins-structure Function & Bioinformatics 2005;58:610.
43. Pupko T, Bell RE, Mayrose I et al. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues, Bioinformatics 2002;18 Suppl 1:S71.
44. Cuff JA, Barton GJ. Application of multiple sequence alignment profiles to improve protein secondary structure prediction, Proteins-structure Function & Bioinformatics 2000;40:502-511.
45. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices, Journal of Molecular Biology 1999;292:195-202.
46. Heffernan R, Paliwal K, Lyons J et al. Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning, Scientific reports 2015;5:11476.
47. Yang Y, Heffernan R, Paliwal K et al. SPIDER2: A Package to Predict Secondary Structure, Accessible Surface Area, and Main-Chain Torsional Angles by Deep Neural Networks 2017.