**ALLFEATURE** – An integrated Python package for DNA, RNA and Protein sequence data analysis

Supplementary Material

**Content**

1. Introduction
2. Tutorial
3. DNA and RNA Feature Extraction Methods
4. Protein Feature Extraction Methods
5. Feature Selection and Dimensionality Reduction Methods
6. Predicating Models
7. Output files
8. References

**1 Introduction**

ALLFEATURE integrated 20 feature selection methods, a total of 16 dimensionality reduction methods and 13 prediction/classification models. In addition, our pipeline tool also contains the step of feature extraction to generate a total of 60 different models of features from DNA, RNA and protein sequencing data. The flowchart of the Python pipeline tool is shown in **Figure 1**. Compared with other software packages or webservers, the proposed ALLFEATURE has the following advantages: (i) sufficient feature selection methods and dimension reduction methods, including regularization-based, statistic-based, information-based, tree-based and recursive feature elimination-based approaches. (ii) Ten machine learning methods and three deep learning approach: deep neural network (DNN) [52], convolutional neural network (CNN) [53], and recurrent neural network (RNN) [54]. (iii) More abundant graphical display results, including box figure, ROC curves, etc.

G:\feature_selection_package\正文\flowchart\flowchart (第二版).tif

Figure 1. The overflow of ALLFEATURE.

**2 Download and Installation**

The tool is developed using Python 3 (Python Version 3.0 of above) and it can be run on Linux operating system. We strongly recommend user to install Anaconda Python 3.7 or above version to avoid installing other packages. Anconda can be freely downloaded from <https://www.anaconda.com/distribution/#download-section>

After installing Anaconda, the following packages need to be installed:

1. xgboost
2. skrebate
3. lightgbm

The source code is freely available at <https://github.com/ashinandjay/FeatureSelection>

To install our tool, first download the zip file manually from github, or use the code below in Unix:

cd your\_folder\_path

wget https://github.com/ashinandjay/FeatureSelection/archive/master.zip

Unzip the file:

unzip master.zip

**3 DNA and RNA Feature Extraction Methods**

**ALLFEATURE** directly extracts features from DNA, RNA or protein sequences based on a total of 60 different types of feature extraction methods. The step of feature extraction consists of 16 feature extraction methods for DNA and 12 feature extraction methods for RNA; 32 feature extraction methods for protein sequences, which can be shown in **Table 1** and **Table 2**, respectively.

Table 1. List of 16 DNA feature extraction methods and 12 RNA feature extraction methods

|  |  |  |  |
| --- | --- | --- | --- |
| **DNA Feature Extraction Methods** | **RNA Feature Extraction Methods** | **Extraction Method Description** | |
| Kmer | Kmer | | DNA or RNA sequence are represented as the occurrence frequencies of k neighboring nucleic acids [55, 56] |
| Reverse Compliment Kmer (RCKmer) | Reverse Compliment Kmer (RCKmer) | | A variant of Kmer descriptor by removing the reverse compliment Kmer [55, 57] |
| Pseudo Dinucleotide Composition (PseDNC) | Pseudo Dinucleotide Composition (PseDNC) | | Incorporating the contiguous local sequence-order and global sequence-order information [58] |
| Pseudo k-tuple Nucleotide Composition (PseKNC) | - | | Extending the PseDNC by incorporating k-tuple nucleotide composition [59] |
| Dinucleotide Based Auto Covariance (DAC) | Dinucleotide Based Auto Covariance (DAC) | | Measuring the correlation of the same physicochemical index between two dinucleotides separated by lag along the sequence [60, 61] |
| Dinucleotide Based Cross Covariance (DCC) | Dinucleotide Based Cross Covariance (DCC) | | Measuring the correlation of two different physicochemical indices between two dinucleotides separated by lag nucleic acids [60, 61] |
| Dinucleotide Based Auto-cross Covariance (DACC) | Dinucleotide Based Auto-cross Covariance (DACC) | | Combining of DAC and DCC [60-61, 43] |
| Trinucleotide Based Auto Covariance (TAC) | - | | Measuring the correlation of the same physicochemical index between trinucleotides separated by lag nucleic acids [43] |
| Trinucleotide Based Cross Covariance (TCC) | - | | Measuring the correlation of two different physicochemical indices between two trinucleotides separated by lag nucleic acids [43] |
| Trinucleotide Based Auto-Cross Covariance (TACC) | - | | Combining of TCC and TACC [43] |
| Nucleic Acid Composition (NAC) | Nucleic Acid Composition (NAC) | | Calculating the frequency of each nucleic acid type in nucleotide sequence [48] |
| Di-Nucleotide Composition (DNC) | Di-Nucleotide Composition (DNC) | | Containing 16 NAC descriptors [48] |
| Tri-Nucleotide Composition (TNC) | Tri-Nucleotide Composition (TNC) | | Containing 64 NAC descriptors [48] |
| zCurve Mathematical Formula (zCurve) | zCurve Mathematical Formula (zCurve) | | Calculating three components in three axis in genomic sequence analysis [45] |
| monoMonoKGap Theoretical Description (MonoKGap) | monoMonoKGap Theoretical Description (MonoKGap) | | Calculating features based on value of kgap [45] |
| monoDiKGap Theoretical Description (MonoDiKGap) | monoDiKGap Theoretical Description (MonoDiKGap) | | Calculating features based on value of 4 \* kgap [45] |

1. Kmer

For Kmer descriptor, the DNA sequences are represented as the occurrence frequencies of  neighboring nucleic acids. The Kmer (k=3) descriptor can be calculated as:

 (7)

where  is the number of kmer type , while  is the length of a nucleotide sequence.

1. Reverse Compliment Kmer (RCKmer)

The reverse compliment kmer is a variant of kmer descriptor, in which the kmers are not expected to be strand specific. For instance, for a DNA sequence, there are 16 types of 2-mers (i.e. 'AA', 'AC', 'AG', 'AT', 'CA', 'CC', 'CG', 'CT', 'GA', 'GC', 'GG', 'GT', 'TA', 'TC', 'TG', 'TT'), ‘TT’ is reverse compliment with ‘AA’. After removing the reverse compliment kmers, there are only 10 distinct kmers in the reverse compliment kmer approach ('AA', 'AC', 'AG', 'AT', 'CA', 'CC', 'CG', 'GA', 'GC', 'TA').

1. Pseudo dinucleotide composition (PseDNC)

PseDNC 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, the PseDNC feature vector of D is defined:



 (8)

where  is the normalized occurrence frequency of dinucleotides in the RNA sequence, the parameter  is an integer, representing the highest counted rank (or tier) of the correlation along a RNA sequence, w is the weight factor ranged from 0 to 1,  is called the correlation factor that reflects the sequence-order correlation between all the most contiguous dinucleotides along a RNA sequence, which is defined:

 (9)

where the correlation function is given by

 (10)

where  is the number of physicochemical indices, in this approach, 6 indices reflecting the local RNA structural properties are employed to generate the PseDNC feature vector,  represents the numerical value of the  physicochemical index of the dinucleotide .

1. Pseudo k-tuple nucleotide composition (PseKNC)

PseKNC extends the PseDNC approach by incorporating *k*-tuple nucleotide composition. Given a DNA sequence D, the feature vector of D is defined:



 (11)

where  is the number of the total counted ranks (or tiers) of the correlations along a DNA sequence;  is the frequency of oligonucleotide that is normalized to ; is a weight factor; is given by

 (12)

which represents the  structural correlation factor between all the  most contiguous dinucleotides. The correlation function  is defined by

 (13)

where  is the number of physicochemical indices, in this study, 6 indices reflecting the local DNA structural properties are employed to generate the PseKNC feature vector;  represents the numerical value of the  physicochemical index for the dinucleotide .

1. Dinucleotide-based auto covariance (DAC)

The DAC 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:

 (14)

where  is a physicochemical index,  is the length of the RNA sequence  means the numerical value of the physicochemical index  for the dinucleotide  at position ,  is the average value for physicochemical index  along the whole sequence:

In such a way, the length of DAC feature vector is , where  is the number of physicochemical indices, and  is the maximum of sequence.

1. Dinucleotide-based cross covariance (DCC)

Given a DNA sequence, the DCC 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:

 (15)

where  are two different physicochemical indices,  is the length of the DNA sequence,  is the numerical value of the physicochemical index  for the dinucleotide .  is the average value for physicochemical index value  along the whole sequence:

 (16)

In such a way, the length of the DCC feature vector is ,  is the maximum of sequence; N is the number of physicochemical indices.

1. Dinucleotide-based auto-cross covariance (DACC)

DACC is a combination of DAC and DCC. Therefore, the length of the DACC feature vector is , where N is the number of physicochemical indices and  is the maximum of sequence.

1. Trinucleotide-based Auto Covariance (TAC)

The Trinucleotide-based auto covariance (TAC) encoding measures the correlation of the same physicochemical index between trinucleotides separated by lag nucleic acids along the sequence, and can be calculated as:

 (17)

where  is a physicochemical index,  is the length of the DNA sequence  means the numerical value of the physicochemical index  for the dinucleotide  at position , is the average value for physicochemical index  along the whole sequence. The dimension of TAC feature vector is , where  is the number of physicochemical indices, and  is the maximum of sequence.

1. Trinucleotide-based Cross Covariance (TCC)

The trinucleotide-based cross covariance (TCC) encoding measures the correlation of two different physicochemical indices between two trinucleotides separated by nucleic acids along the sequence. The TCC encoding can be calculated as:

 (18)

where  are two different physicochemical indices,  is the length of the DNA sequence,  is the numerical value of the physicochemical index  for the dinucleotide .  is the average value for physicochemical index value  along the whole sequence:

 (19)

In such a way, the length of the TCC feature vector is ,  is the maximum of sequence; N is the number of physicochemical indices.

1. Trinucleotide-based auto-cross covariance (TACC)

The trinucleotide-based auto-cross covariance (TACC) encoding is a combination of TAC and TACC encoding. Thus, the dimension of the TACC encoding is , where N is the number of physicochemical indices and  is the maximum of DNA sequence.

1. Nucleic Acid Composition (NAC)

The Nucleic Acid Composition (NAC) encoding calculates the frequency of each nucleic acid type in a nucleotide sequence. The frequencies of all 4 natural nucleic acids (i.e. “ACGT or U”) can be calculated as:

 (20)

where *N*(*t*) is the number of nucleic acid type *t*, while *N* is the length of a nucleotide sequence.

1. Di-Nucleotide Composition (DNC)

The Di-Nucleotide Composition gives 16 descriptors. It is defined as:

 (21)

where is the number of di-nucleotide represented by nucleic acid types *r* and *s*.

1. Tri-Nucleotide Composition (TNC)

The Tri-Nucleotide Composition gives 64 descriptors. It is defined as:

 (22)

where is the number of tri-nucleotide represented by nucleic acid types *r, s* and *t*.

1. zCurve Mathematical Formula (zCurve)

Z-curve theory is often used in genomic sequence analysis. It has got three components in three axis. They are defined as following.

 (23)

Three features will generate using the zCurve method.

1. MonoKGap Theoretical Description (MonoKGap)

When -kgap=n then the  features will exist for DNA and RNA but  features will exist for protein. When -kgap=1, feature structure will be X\_X.

When -kgap=2, feature structure will be X\_X, and X\_\_X. When -kgap=3, feature structure will be X\_X. X\_\_X, and X\_\_\_X. Described with appropriate examples: When -kgap=1 then only sixteen features will exist for DNA and RNA but four hundred (400) features will exist for protein. Features will be numbers of A\_A, A\_C, A\_G, A\_T, C\_A, C\_C, C\_G, C\_T, G\_A, G\_C, G\_G, G\_T, T\_A, T\_C, T\_G, and T\_T of the whole sequence of DNA respectively. When -kgap=2 then only thirty-two features will exist for DNA and RNA but eight hundred (800) features will exist for protein. Features will be numbers of A\_A, A\_C, A\_G, A\_T, C\_A,C\_C, C\_G, C\_T, G\_A, G\_C, G\_G, G\_T, T\_A, T\_C, T\_G,T\_T, A\_\_A, A\_\_C, A\_\_G, A\_\_T, C\_\_A, C\_\_C, C\_\_G, C\_\_T, G\_\_A, G\_\_C, G\_\_G, G\_\_T, T\_\_A, T\_\_C, T\_\_G, and T\_\_T of the whole sequence of DNA respectively.

When -kgap=3 then only forty-eight features will exist for DNA and RNA, but one thousand and two hundred (1,200) features will exist for protein. Features will be numbers of A\_A, A\_C, A\_G, A\_T, C\_A,C\_C, C\_G, C\_T, G\_A, G\_C, G\_G, G\_T, T\_A, T\_C, T\_G,T\_T, A\_\_A, A\_\_C, A\_\_G,A\_\_T, C\_\_A, C\_\_C, C\_\_G, C\_\_T, G\_\_A, G\_\_C, G\_\_G, G\_\_T, T\_\_A, T\_\_C, T\_\_G, T\_\_T, A\_\_\_A, A\_\_\_C, A\_\_\_G, A\_\_\_T, C\_\_\_A, C\_\_\_C, C\_\_\_G, C\_\_\_T, G\_\_\_A, G\_\_\_C, G\_\_\_G,G\_\_\_T, T\_\_\_A, T\_\_\_C, T\_\_\_G, and T\_\_\_T of the whole sequence of DNA respectively.

1. MonoDiKGap Theoretical Description (MonoDiKGap)

When kgap=n then the  features will exist for DNA and RNA, but features will exist for protein.

When -kgap=1, feature structure will be X\_XX.

When -kgap=2, feature structure will be X\_XX, and X\_\_XX.

When -kgap=3, feature structure will be X\_XX, X\_\_XX, and X\_\_\_XX. Described with appropriate examples:

When -kgap=1 then only sixty-four (64) features will exist for DNA and RNA, but eight thousand (8,000) features will exist for protein. Features will be numbers of A\_AA, A\_AC,A\_AG, A\_AT, A\_CA, A\_CC, A\_CG, A\_CT, A\_GA, A\_GC, A\_GG, A\_GT, A\_TA, A\_TC, A\_TG,A\_TT, C\_AA, C\_AC, C\_AG, C\_AT, C\_CA, C\_CC, C\_CG, C\_CT, C\_GA, C\_GC, C\_GG, C\_GT,C\_TA, C\_TC, C\_TG, C\_TT, G\_AA, G\_AC, G\_AG, G\_AT, G\_CA, G\_CC, G\_CG, G\_CT, G\_GA,G\_GC, G\_GG, G\_GT, G\_TA, G\_TC, G\_TG, G\_TT, T\_AA, T\_AC, T\_AG, T\_AT, T\_CA, T\_CC,T\_CG, T\_CT, T\_GA, T\_GC, T\_GG, T\_GT, T\_TA, T\_TC, T\_TG, and T\_TT of the whole sequence of DNA respectively.

When -kgap=2 then only hundred and twenty-eight (128) features will exist for DNA and RNA, but sixteen thousand (16,000) features will exist for protein. Features will be numbers of A\_AA, A\_AC,A\_AG, A\_AT, A\_CA, A\_CC, A\_CG, A\_CT, A\_GA, A\_GC, A\_GG, A\_GT, A\_TA, A\_TC, A\_TG,A\_TT, C\_AA, C\_AC, C\_AG, C\_AT, C\_CA, C\_CC, C\_CG, C\_CT, C\_GA, C\_GC, C\_GG, C\_GT,C\_TA, C\_TC, C\_TG, C\_TT, G\_AA, G\_AC, G\_AG, G\_AT, G\_CA, G\_CC, G\_CG, G\_CT, G\_GA,G\_GC, G\_GG, G\_GT, G\_TA, G\_TC, G\_TG, G\_TT, T\_AA, T\_AC, T\_AG, T\_AT, T\_CA, T\_CC,T\_CG, T\_CT, T\_GA, T\_GC, T\_GG, T\_GT, T\_TA, T\_TC, T\_TG,T\_TT, A\_\_AA, A\_\_AC, A\_\_AG, A\_\_AT, A\_\_CA, A\_\_CC, A\_\_CG, A\_\_CT, A\_\_GA, A\_\_GC,A\_\_GG, A\_\_GT, A\_\_TA, A\_\_TC, A\_\_TG, A\_\_TT, C\_\_AA, C\_\_AC, C\_\_AG, C\_\_AT, C\_\_CA,C\_\_CC, C\_\_CG, C\_\_CT, C\_\_GA, C\_\_GC, C\_\_GG, C\_\_GT, C\_\_TA, C\_\_TC, C\_\_TG, C\_\_TT,G\_\_AA, G\_\_AC, G\_\_AG, G\_\_AT, G\_\_CA, G\_\_CC, G\_\_CG, G\_\_CT, G\_\_GA, G\_\_GC, G\_\_GG,G\_\_GT, G\_\_TA, G\_\_TC, G\_\_TG, G\_\_TT, T\_\_AA, T\_\_AC, T\_\_AG, T\_\_AT, T\_\_CA, T\_\_CC,T\_\_CG, T\_\_CT, T\_\_GA, T\_\_GC, T\_\_GG, T\_\_GT, T\_\_TA, T\_\_TC, T\_\_TG, and T\_\_TT of the whole sequence of DNA respectively.

**4 Protein Feature Extraction Methods**

Table 2. List of 32 Protein feature extraction methods and their description

|  |  |  |
| --- | --- | --- |
| **Protein Feature Extraction** | **Extraction Method Description** | |
| Amino Acid Composition (AAC) | | Calculating the frequencies of 20 kinds of amino acids [62] |
| Dipeptide Composition(DC) | | transforming the variable length of proteins to fixed length feature vectors [62] |
| Composition of K-Spaced Amino Acid Pairs (CKSAAP) | | Extracting important intrinsic correlation information of protein sequences in multidimensional space [63-65] |
| Grouped Dipeptide Composition (GDC) | | A variation of the DPC descriptor which generates 25 descriptors [66] |
| Grouped Tripeptide Composition (GTC) | | Another variation of TPC descriptor which generates 125 descriptors [66] |
| Conjoint Triad (CT) | | Calculating the frequency of occurrence of each triad [67] |
| K-Spaced Conjoint Triad (KSCTriad) | | Combining CT and considers the continuous amino acid units that are separated by any *k* residues [68] |
| Composition (C)  Transition (T)  Distribution (D) | | Calculating composition descriptors  Calculating transition descriptors  Calculating distribution descriptors [69-71] |
| Encoding Based on Grouped Weight (EBGW) | | Capturing the continuity and discontinuity features based on grouped weight coding [72] |
| Auto Covariance (AC) | | Measuring the correlation of the same property between two residues separated by distance of *l* [73] |
| Moreau-Broto autocorrelation (Morean-Broto) | | Measuring the physiochemical and position information between two amino acid [74] |
| Moran Autocorrelation (Moran) | | Measuring the physiochemical information of adjacent amino acid [75] |
| Geary Autocorrelation (Geary) | | Measuring the physiochemical information and generate positive values [76, 77] |
| Quasi-Sequence-Order (QSO) | | Obtaining the sequence distribution patters for a specific physicochemical property [78] |
| Pseudo-Amino Acid Composition (PseAAC) | | Extracting the physicochemical information and sequence order information [79, 80] |
| Amphiphilic Pseudo-Amino Acid Composition (APAAC) | | Extracting the type-2 pseudo amino acid composition [79, 80] |
| Amino Acid Composition PSSM (ACC-PSSM) | | Calculating process of amino acid composition PSSM [81, 82] |
| Dipeptide Composition PSSM (DPC-PSSM) | | Extracting the sequence-order information in the PSSM [82] |
| Bi-gram PSSM (Bi-PSSM) | | Calculating the frequency of the transition between amino acids [83] |
| Auto Covariance PSSM (AC-PSSM) | | Measuring the correlation of the same property between two residues separated by lag [84] |
| Pseudo PSSM (PsePSSM) | | Calculating the PsePSSM feature vector according to the pseudo amino acid composition [85] |
| AB-PSSM | | Calculating feature vector based on averaged PSSM over blocks [86] |
| Secondary Structure Composition (SSC) | | Calculating feature based normalized count of frequency of the structural motifs present at the amino-acid residue positions [87] |
| Accessible Surface Area composition (ASA) | | Calculating feature based on normalized sum of accessible surface area [87] |
| Torsional Angles Composition (TAC) | | Calculating features based four different types of torsional angles [87] |
| Torsional Angles bigram (TA-bigram) | | Calculating feature based on the bigram of the torsional angles [87] |
| Structural Probabilities bigram (SP-bigram) | | Calculating feature based on structural probabilities for each position of amino acid residue [87] |
| Torsional Angles Auto-Covariance (TAAC) | | Calculating feature from the torsional auto-covariance [87] |
| Structural Probabilities Auto-Covariance (SPAC) | | Calculating feature from the structural probabilities [87] |

Type 1: Sequence information

1. Amino acid composition (AAC)

The amino acid composition is the fraction of each amino acid type within a protein. The amino acid composition gives 20 features and the fractions of all 20 natural amino acids are calculated as

 (24)

where  is the number of the amino acid type  and  is the length of the sequence.

1. Dipeptide composition (DC)

The dipeptide composition is used to transform the variable length of proteins to fixed length feature vectors.A dipeptide composition has been used earlier by Grassmann et al. and Reczko and Bohr for the development of fold recognition methods.We adopt the same dipeptide composition-based approach in developing a deep neural networks-based method for predicting protein-protein inter-action.The dipeptide composition gives a fixed pattern length of 400.Dipeptide composition encapsulates information about the fraction of amino acids as well as their local order.The dipeptide composition is defined as

 (25)

wherethe number of dipeptide represented by amino acid typeand.

1. -gap dipeptide composition introduces (g-GapDC)

The -gap dipeptide composition introduces (-Gap DC) extracts important intrinsic correlation information of protein sequences in multidimensional space. For a protein ,  of the sequence length C, the calculation process of the -Gap DC can be expressed as

 (26)

 (27)

where  represents the number of residues in the primary structure of the two amino acids, represents the number of occurrences of the -th feature in the -gap dipeptide, and  represents the frequency at which the -th feature in the -gap dipeptide appears in the sequences. When , there is no gap between two adjacent amino acid residues, and when , it means that one amino acid residue is between two adjacent amino acid residues. From equation (9), a protein sequence can be represented as a 400-dimensional feature vector.

1. Grouped di-peptide composition (GDC)

The Grouped di-peptide composition encoding is another variation of the DPC descriptor. It is composed of a total of 25 descriptors that are defined as:

 (28)

where  is the number of the di-peptide type , and  is the length of the sequence.

1. Grouped tri-peptide composition (GTC)

The Grouped tri-peptide composition encoding is also a variation of TPC descriptor, which generates 125 descriptors, defined as:

 (29)

where  is the number of the tri-peptide type , and  is the length of the sequence.

1. Conjoint triad (CT)

First, 20 amino acids are clustered into seven classes based on dipoles and volumes of side chains. Considering the interaction between the amino acid and its vicinal amino acids, the three continuous amino acids are regarded as a unit, so that we can obtain  triad types. We calculate the frequency of occurrence of each triad called . Then the 343-dimensional feature vector is obtained according to equation (6).

 (30)

1. *k*-Spaced Conjoint Triad (KSCTriad)

The *k*-spaced conjoint triad descriptor is based on the conjoint ctriad descriptor,

which not only calculates the numbers of three continuous amino acid units, but also considers the continuous amino acid units that are separated by any *k* residues (The default maximum value of *k* is set to 5). For example, AxRxT is a 1-spaced triad. Thus, the dimensionality of the KSCTriad encoded feature vector is 343 (*k*+1).

1. Composition, transition and distribution (CTD)

8.1 Composition

For CTD, taking hydrophobicity as an example, all amino acids are classified into three categories: polar, neutral, and hydrophobic. The replacement sequence consists of three types, and the composition descriptors of the polar, neutral, and hydrophobic residues of the protein can be calculated as follows:

 (31)

where  is the number of amino acid type  in the coding sequence and  is sequence length.

8.2 Transition

The transition descriptor first converts the original sequence into a replacement sequence, and T includes three characteristics, the dipeptide composition frequency from the polar group to the neutral group and the composition frequency from the neutral group to the polar group. Transitions between the neutral group and the hydrophobicity and these between hydrophobic group and the polar group are defined in the same way. The T descriptor is defined as follows:

 (32)

Where  and  are the  and  dipeptide frequency, respectively.  is the sequence length.

8.3 Distribution

For each group (polar, neutral and hydrophobic), the D descriptor can generate five values. We obtain the position of the first, 25%, 50%, 75% and 100% of the specific encoded group sequence and then divided the position by the whole sequence. Given sequence MTTTVPKVFAFHEF. It can be represented as '32223213323213' according to Hydrophobicity\_PRAM900101. '1' represents polar, '2' represents neutral, '3' represents hydrophobicity. Take '3' for example, there are 6 residues encoded '3'. The first '3' is 1. The second '3' is . The third '3' is . The fourth '3' is . The fifth '3 is . The position in the first, the second, the third, the fourth, the fifth '3' of whole sequence are 1, 1, 8, 9, 14, respectively. So the distribution descriptor for '3' are ,,,,.

The C descriptor generates a 39-dimensional feature vector, the T descriptor generates a 39-dimensional feature vector, and the D descriptor generates a 195-dimensional feature vector. For each protein sequence, the CTD generates a 273-dimensional feature vector.

1. Encoding based on grouped weight coding (EBGW)

Studies have shown that different amino acids have different physical and chemical properties. The EBGW can capture the sequence and physicochemical information based on grouped situation. These amino acids are classified into four categories:

neutral and hydrophobic amino acids 

neutral and polarity amino acids 

acidic amino acids 

basic amino acids .

Thus, we can get three combinations, each of which can partition the 20 amino acid residues into two disjoint group:  vs , or vs, and  vs . Let be a protein sequence, we can transform it into three binary sequences by three homomorphic mapswhich are defined as follows:

 (33)

 (34)

 (35)

where , and we call as characteristic sequence of the protein sequence.

Then , , are three binary sequences of length *L*. These sequences are divided into a number of sub-sequences of increasing length successively. A fixed parameter *N* is set and the sub-sequence can be expressed as, where represents the integer operator. Calculate the frequency of 1 in each sub-sequence, each  can be converted into an *N*-dimensional feature vector. To sum up, for a protein sequence *P* with length *L*, a 3*N*-dimension vector can be obtained.

Type 2: Physicochemical information

1. Auto covariance (AC)

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

 (36)

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

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

 (37)

where is a physicochemical index, is the length of the protein sequence, means the numerical value of the physicochemical index for the amino acid  at position ,  is the average value for physicochemical index along the whole sequence. In such a way, the length of AC feature vector is , where *N* is the number of physicochemical indices extracted from AAindex; ** is the maximum of sequence.

1. Moreau-Broto autocorrelation (Morean-Broto)

Moreau-Broto autocorrelation descriptor is defined as:

 (38)

where , and  indicate the  and the  amino acids of the protein sequence, respectively.  and indicate the normalized physicochemical values of  and . The  is the parameter that needs to be adjusted.

1. Moran autocorrelation (Moran)

Moran autocorrelation descriptor is defined as:

 (39)

where  represents the mean value of whole protein sequence for specific physicochemical property.

1. Geary autocorrelation (Geary)

Geary autocorrelation descriptor is defined as:

 (40)

1. Quasi-sequence-order descriptors (QSO)

The sequence order features can also be used for representing amino acid distribution patterns of a specific physicochemical property along protein or peptide sequence. These descriptors are derived from both the Schneider-Wrede physicochemical distance matrix and the Grantham chemical distance matrix between each pair of the 20 amino acids. Theth rank sequence-order-coupling number is defined as

 (41)

whereis the distance between the two amino acids at positionand. Maxlag is the maximum lag and the length of the protein must be not less than maxlag. The maxlag is equal to 30 in the experiment. For each amino acid type, the type-1 quasi-sequence-order descriptor can be defined as

 (42)

where is the normalized occurrence of amino acid type-1 and is a weighting factor. The type-2 quasi-sequence-order is defined as

 (43)

In addition to the Schneider-Wrede physicochemical distance matrix used by Chou et al., another chemical distance matrix by Grantham is also used here. The sequence-order features produce a total ofdescriptors.

1. Pseudo-amino acid composition (PseAAC)

Pseudo-amino acid composition are utilized to extract the physicochemical information. Pseudo-amino acid composition (PseAAC) represents the composition information of the protein and sequence order information. The feature vector of PseAAC can be represented:

 (44)

where the first 20-dimentional feature vector represents the amino acid composition information, and the latter  dimensional vector represents the order information of protein sequence.  is the length of amino acid sequence.

 (45)

where  is the normalized frequency of 20 amino acids in protein.  is the layer sequence correlation factor calculated according to the equation (2). Sequence correlation factor can be obtained from the hydrophobicity, hydrophilicity, and side-chain mass of the amino acid.

1. Amphiphilic pseudo amino acid composition (APAAC)

APAAC (http://www. csbio. sjtu. edu. cn/bioinf/PseAAC/type2. htm) are also called type-2 pseudoamino acid composition. The definitions of these qualities are similar to PAAC descriptors. First, two variables are derived from the original hydrophobicity valuesand hydrophilicity valueof 20 amino acids

 (46)

 (47)

whereand,the hydrophobicity and hydrophilicity correlation functions, are defined respectively as

 (48)

where sequence order factors can be defined as

 (49)

 (50)

 (51)

Then a set of descriptors called "Amphiphilic Pseudo Amino Acid Composition” are defined as

 (52)

 (53)

whereis the weight factor and is taken as, andis equal to 30 in the experiment. So, we produce a total ofdescriptors.

Type 3: Evolutionary information

In order to obtain the evolutionary information of the amino acid sequence, all the protein sequences in the dataset are compared with the non-redundant database SwissProt using the PSI-BLAST program. The program can search sequences based on the iterative BLAST search method. Evolutionary information in the position specific scoring matrix (PSSM) plays an important role in biological system analysis.

During the running process, the parameter E-value threshold of PSI-BLAST is set to 0.001, the maximum number of iterations is set to 3, and the remaining parameters are set by default. Then PSSM of each protein sequence is obtained. For a protein sequence whose length is , PSSM is shown in equation (49).

 (54)

where each row of the PSSM represents a log likelihood score for amino acid substitutions occurring at corresponding positions in the query sequence. Where  represents the  position of query sequence being mutated to type  during evolution process. The scores are positive integers or negative integers. A positive integer indicates that more mutations have occurred in the alignment and a negative integer indicates that fewer substitutions have occurred in the alignment.

1. Amino acid composition PSSM (AAC-PSSM)

In this calculating process of amino acid composition PSSM (AAC-PSSM), the PSSM is standardized by the logic function. PSSM elements are map to the interval [0,1].

 (55)

PSSM are converted to feature vector by AAC-PSSM via equation (51)

 (56)

where ,  represents the composition information of the  amino acid residue, which is the average score of  amino acid in PSSM.

1. Dipeptide composition PSSM (DPC-PSSM)

The PSSM contains evolutionary information. ACC-PSSM only represents the composition information from PSSM, and loses the order information, which is insufficient to fully represent the evolutionary information. Dipeptide composition PSSM (DPC-PSSM) can reflect the sequence-order information in the PSSM, which converts the character signal into the numerical signal, and the extracted feature vector can be expressed as

 (57)

where , the dimension of DPC-PSSM is 400.

1. Bi-gram PSSM (Bi-PSSM)

For Bi-gram PSSM, the frequency of the transition from the  amino acids to the  amino acids is calculated:

 (58)

Therefore, there are 400 possible cases for , then the Bi-gram PSSM eigenvector for each protein sequence is:

 (59)

1. Auto covariance PSSM (AC-PSSM)

AC-PSSM 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:

 (60)

whereis one of the residues,is the length of the protein sequence,is the PSSM score of amino acidat position,is the average score for amino acid  along the whole sequence:

 (61)

In such a way, the number of AC variables can be calculated as, where  is the maximum of sequence.

1. Cross covariance PSSM (CC-PSSM)

CC-PSSM 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:

 (62)

whereare two different amino acids andis the average score for amino acidalong the sequence. Since the CC variables are not symmetric, the total number of CC variables is.

1. Auto-cross covariance PSSM (ACC-PSSM)

ACC-PSSM 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.

1. Pseudo PSSM (PsePSSM)

According to the pseudo amino acid composition, we obtain the PsePSSM feature vector:

 (63)

where, each protein sequence can generate the dimensional feature vector. The first 20-dimensional vector represents the composition information of the PSSM matrix, and the remaining  dimensional feature vector represents the order evolutionary information. PsePSSM can transform an inconsistent protein sequence into a consistent numerical vector by feature extraction.

1. AB-PSSM

AB-PSSM is based on the averaged PSSM profiles over blocks, each with 5 percent of a sequence. Thus, a protein sequence, regardless of its length, is divided into 20 blocks and each block consists of 20 features (derived from the 20 columns in PSSMs). Mathematically, for the  block, the feature  is a  dimensional feature vector, which is generated by using the following equation:

 (64)

where  is the size of the  block, which is 5 percent of the length of a sequence and Pej i is a  vector extracted from the PSSM profile at the  position in the  block. For each sequence, there are a total of 20 blocks; therefore, the final feature is a 400-dimensional vector.

Type 4: Structural information

1. Secondary structure composition (SSC)

This feature is the normalized count or frequency of the structural motifs present at the amino-acid residue positions. There are three types of motifs: *α*-helix (H), *β*-sheet (E) and random coil (C). SPIDER2 returns a vector *SS* of dimension *L* × 1 containing this information. Thus, we can define this feature as following:

 (65)

where, *L* is the length of the protein and

 (66)

where, *SSj* is the structural motif at position *j* of the protein sequence and *fi* is one of the 3 different motif symbols.

1. Accessible surface area composition (ASA)

The accessible surface area composition is the normalized sum of accessible surface area defined by:

 (67)

where ASA is the vector of accessible surface area of dimension *L* × 1 containing the values of accessible surface area for all the amino acid residues.

1. Torsional angles composition (TAC)

Four different types of torsional angles: *ϕ*, *ψ*, *τ* and *θ* are returned by SPIDER2 for each residue. First, we convert each of them into radians from degree angles and then take sign and cosine of the angles at each residue position. Thus, we get a matrix of dimension. We denotethis matrix by *T*. Torsional angles composition is defined as

 (68)

1. Torsional angles bigram (TA-bigram)

The Bigram for the torsional angles is similar to that of the PSSM matrix and is defined as:

 (69)

1. Structural probabilities bigram (SP-bigram)

Structural probabilities for each position of the amino-acid residue are given in the SPD2 file as a matrix of dimension *L* × 3, which we denote by *P*. The Bigram of the structural probabilities is similar to that of PSSM matrix and is defined as:

 (70)

1. Torsional angles auto-covariance (TAAC)

This feature is also derived from the torsional angles and is defined as:

 (71)

1. Structural probabilities auto-covariance (SPAC)

This feature is also derived from the structural probabilities and is defined as:

 (72)

**5 Feature Selection and Dimensionality Reduction Methods**

ALLFEATURE integrated a step of feature selection and dimensionality reduction methods in **Table 3**.

Table 3. Feature Selection and Dimensionality Reduction Methods

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Feature Selection Method** | | **Description** | **Dimensionality Reduction Method** | **Description** |
| Lasso | Using Lasso liner model to recursively eliminate features [31, 88] | | K-means | Clustering data by separating samples in n groups of equal variances [34] |
| ElasticNet | Using ElasticNet model to recursively eliminate features [89] | | T-SNE | Visualizing high-dimensional data [36] |
| L1-SVM | Using SVM with L1 penalty model to recursively eliminate features [90] | | Principal Component Analysis (PCA) | Linear dimensionality reduction using singular value decomposition [35] |
| CHI2 | Retrieving best features based on test [29, 91] | | Kernel PCA (KPCA) | Non-linear dimensionality reduction through use of kernels [35] |
| Pearson Correlation (PC) | Retrieving best features based on Pearson correlation [32] | | Locally linear embedding (LLE) | Reducing projection of data which preserves distances within local neighborhoods [105] |
| ExtraTree | Using ExtraTree model to recursively eliminate features [92] | | Truncated Singular Value Decomposition (TSVD) | Linear dimensionality reduction by means of truncated singular value decomposition [106] |
| xgBosst | Using xgBoost model to recursively eliminate features [40. 93] | | Non-negative matrix factorization (NMF) | Reducing dimension by finding two non-negative metrices [107] |
| SVM-RFE | Using linear SVM model to recursively eliminate features [30, 100] | | Multi-dimensional Scaling (MDS) | Reducing dimension by modeling data as distances in a geometric space [108] |
| LOG-RFE | Using Logistic Regression model to recursively eliminate features [94] | | Independent Component Analysis (ICA) | Reducing dimension by finding components with some sparsity [109] |
| Mutual Information (MI) | Retrieving best features based mutual information [95] | | Factor Analysis (FA) | Reducing dimension by performing a maximum likelihood estimate [110] |
| Minimum Redundancy Maximum Relevance (MRMR) | Selecting features that still having high correlation to the classification variable [96] | | Agglomerate Feature (AF) | Recursively merges feature instead of samples [111] |
| Joint Mutual Information (JMI) | Retrieving best features based joint mutual information [97] | | Gaussian Random Projection (GRP) | Reducing the dimension by projecting the original input space using the Gaussian distribution [112] |
| Maximum Relevance Maximum Distance (MRMD) | Retrieving best features by measuring relevance and redundancy between features [98] | | Sparse Random Projection (SRP) | Reducing dimension by projecting the original input space using a sparse random matrix [113] |
| ReliefF | Retrieving best features by calculating and ranking a feature score for each feature [33] | | Autoencoder | Reducing the dimension using encode and decode neural network [114] |
| Trace Ratio | Retrieving best features by calculating the corresponding score in trace ratio form [99] | | Gaussian Noise Autoencoder (GNA) | Corrupting input before being passed to autoencoder neural network [115] |
| Gini Index | Retrieving best features by constructing the measure function based on Gini-Index [100] | | Variational Autoencoder (VA) | Neural network can be trained with stochastic gradient descent [116] |
| Spectral Feature Selection (SPEC) | Retrieving best features based on structure induced [101] | | - | - |
| Fisher Score | Retrieving best features based on scores of features under the Fisher criterion [102] | | - | - |
| T Score | Retrieving best features based on their t-score [103] | | - | - |
| Information Gain (IG) | Retrieving best features based on their information gain [104] | | - | - |

This section provides user a step-by-step instruction illustrating the workflow of our tool. There are five main Python scripts in our tool: “DNA\_Feature\_Extraction.py”, “RNA\_Feature\_Extraction.py”, “PROTEIN\_Feature\_Extraction.py”, “Feature\_selection.py”, “Feature\_Reduction.py”, “Feature\_Evaluation.py”, “Feature\_Evaluation\_NN.py”.

1. “DNA\_Feature\_Extraction.py” is the main program used to extract 16 different types of feature descriptors from provided DNA sequence.
2. “RNA\_Feature\_Extraction.py” is the main program used to extract 12 different types of feature descriptors from provided RNA sequence.
3. “PROTEIN\_Feature\_Extraction.py” is the main program used to extract 32 different types of feature descriptors from provided Protein sequence.
4. “Feature\_selection.py” is the main program used to select features from generated feature descriptors. The program includes 23 feature selection methods.
5. “Feature\_Reduction.py” is the main program used to reduce dimension of features from generated feature descriptors. The program includes 13 feature dimension reduction methods.
6. “Feature\_Evaluation.py” is the program used to evaluate the performance of one feature selection method based on classification accuracies from 10 different classifies.
7. “Feature\_Evaluation\_NN.py” is the program used to evaluate the performance of one feature selection method based on classification accuracies from 3 different neural network classification methods.

The flowchart of tool:



**4 Data Preparation**

Implementing “DNA\_Feature\_Extraction.py”, “RNA\_Feature\_Extraction.py”, “PROTEIN\_Feature\_Extraction.py” require user to provide DNA, RNA or Protein sequence data (FASTA format). These three programs will generate feature descriptor file (csv format).

Implementing “Feature\_Selection.py” and “Feautre\_Evaluation.py” require user to provide generated feature descriptors from the step of feature extraction and labels of feature descriptors.

Implementing “Feature\_Reuction.py” only require generated feature descriptors from the step of feature extraction.

**5 DNA Feature Extraction**

“DNA\_Feature\_Extraction.py” contains 16 feature extraction methods for DNA sequencing data.

|  |  |
| --- | --- |
| DNA Feature Extraction Method | DNA Extraction Number |
| Kmer | 1 |
| Reverse Compliment Kmer | 2 |
| Pseudo dinucleotide composition | 3 |
| Pseudo k-tuple nucleotide composition | 4 |
| Dinucleotide-based auto covariance | 5 |
| Dinucleotide-based cross covariance | 6 |
| Dinucleotide-based auto-cross covariance | 7 |
| Trinucleotide-based auto covariance | 8 |
| Trinucleotide-based cross covariance | 9 |
| Trinucleotide-based auto-cross covariance | 10 |
| Nucleic acid composition | 11 |
| Di-nucleotide composition | 12 |
| Tri-nucleotide composition | 13 |
| zcurve | 14 |
| monoMonoKGap | 15 |
| monoDiKGap | 16 |

1. **Kmer**

For kmer descriptor, the DNA sequences are represented as the occurrence frequencies of  neighboring nucleic acids. The Kmer (k=3) descriptor can be calculated as:

 (1)

where  is the number of kmer type , while  is the length of a nucleotide sequence.

1. **Reverse compliment kmer (RCKmer)**

The reverse compliment kmer is a variant of kmer descriptor, in which the kmers are not expected to be strand specific. For instance, for a DNA sequence, there are 16 types of 2-mers (i.e. 'AA', 'AC', 'AG', 'AT', 'CA', 'CC', 'CG', 'CT', 'GA', 'GC', 'GG', 'GT', 'TA', 'TC', 'TG', 'TT'), ‘TT’ is reverse compliment with ‘AA’. After removing the reverse compliment kmers, there are only 10 distinct kmers in the reverse compliment kmer approach ('AA', 'AC', 'AG', 'AT', 'CA', 'CC', 'CG', 'GA', 'GC', 'TA').

1. **Pseudo dinucleotide composition (PseDNC)**

PseDNC 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, the PseDNC feature vector of D is defined:



 (2)

where  is the normalized occurrence frequency of dinucleotides in the RNA sequence, the parameter  is an integer, representing the highest counted rank (or tier) of the correlation along a RNA sequence, w is the weight factor ranged from 0 to 1,  is called the correlation factor that reflects the sequence-order correlation between all the most contiguous dinucleotides along a RNA sequence, which is defined:

 (3)

where the correlation function is given by

 (4)

where  is the number of physicochemical indices, in this approach, 6 indices reflecting the local RNA structural properties are employed to generate the PseDNC feature vector,  represents the numerical value of the  physicochemical index of the dinucleotide .

1. **Pseudo k-tuple nucleotide composition (PseKNC)**

PseKNC extends the PseDNC approach by incorporating *k*-tuple nucleotide composition. Given a DNA sequence D, the feature vector of D is defined:



 (5)

where  is the number of the total counted ranks (or tiers) of the correlations along a DNA sequence;  is the frequency of oligonucleotide that is normalized to ; is a weight factor; is given by

 (6)

which represents the  structural correlation factor between all the  most contiguous dinucleotides. The correlation function  is defined by

 (7)

where  is the number of physicochemical indices, in this study, 6 indices reflecting the local DNA structural properties are employed to generate the PseKNC feature vector;  represents the numerical value of the  physicochemical index for the dinucleotide .

1. **Dinucleotide-based auto covariance (DAC)**

The DAC 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:

 (8)

where  is a physicochemical index,  is the length of the RNA sequence  means the numerical value of the physicochemical index  for the dinucleotide  at position ,  is the average value for physicochemical index  along the whole sequence:

In such a way, the length of DAC feature vector is , where  is the number of physicochemical indices, and  is the maximum of sequence.

1. **Dinucleotide-based cross covariance (DCC)**

Given a DNA sequence, the DCC 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:

 (9)

Where  are two different physicochemical indices,  is the length of the DNA sequence,  is the numerical value of the physicochemical index  for the dinucleotide .  is the average value for physicochemical index value  along the whole sequence:

 (10)

In such a way, the length of the DCC feature vector is ,  is the maximum of sequence; N is the number of physicochemical indices.

1. **Dinucleotide-based auto-cross covariance (DACC)**

DACC is a combination of DAC and DCC. Therefore, the length of the DACC feature vector is , where N is the number of physicochemical indices and  is the maximum of sequence.

1. **Trinucleotide-based Auto Covariance (TAC)**

The Trinucleotide-based auto covariance (TAC) encoding measures the correlation of the same

physicochemical index between trinucleotides separated by lag nucleic acids along the sequence, and can be calculated as:

 (8)

where  is a physicochemical index,  is the length of the DNA sequence  means the numerical value of the physicochemical index  for the dinucleotide  at position , is the average value for physicochemical index  along the whole sequence. The dimension of TAC feature vector is , where  is the number of physicochemical indices, and  is the maximum of sequence.

1. **Trinucleotide-based Cross Covariance (TCC)**

The trinucleotide-based cross covariance (TCC) encoding measures the correlation of twodifferent physicochemical indices between two trinucleotides separated by nucleic acids along the sequence. The TCC encoding can be calculated as:

 (9)

Where  are two different physicochemical indices,  is the length of the DNA sequence,  is the numerical value of the physicochemical index  for the dinucleotide .  is the average value for physicochemical index value  along the whole sequence:

 (10)

In such a way, the length of the TCC feature vector is ,  is the maximum of sequence; N is the number of physicochemical indices.

1. **Trinucleotide-based auto-cross covariance (TACC)**

The trinucleotide-based auto-cross covariance (TACC) encoding is a combination of TAC and TACC encoding. Thus, the dimension of the TACC encoding is , where N is the number of physicochemical indices and  is the maximum of DNA sequence.

1. **Nucleic Acid Composition (NAC)**

The Nucleic Acid Composition (NAC) encoding calculates the frequency of each nucleic acid type in a nucleotide sequence. The frequencies of all 4 natural nucleic acids (i.e. “ACGT or U”) can be calculated as:

 (11)

where *N*(*t*) is the number of nucleic acid type *t*, while *N* is the length of a nucleotide sequence.

1. **Di-Nucleotide Composition (DNC)**

The Di-Nucleotide Composition gives 16 descriptors. It is defined as:

 (12)

where is the number of di-nucleotide represented by nucleic acid types *r* and *s*.

1. **Tri-Nucleotide Composition (TNC)**

The Tri-Nucleotide Composition gives 64 descriptors. It is defined as:

 (13)

where is the number of tri-nucleotide represented by nucleic acid types *r, s* and *t*.

1. **zCurve Mathematical Formula (zCurve)**

Z-curve theory is often used in genomic sequence analysis. It has got three components in three axis. They are de\_ned as following.

 (14)

Three features will generate using the zCurve method.

1. **MonoKGap Theoretical Description (MonoKGap)**

When -kgap=n then the  features will exist for DNA and RNA but 

features will exist for protein.

When -kgap=1, feature structure will be X\_X.

When -kgap=2, feature structure will be X\_X, and X\_\_X.

When -kgap=3, feature structure will be X\_X. X\_\_X, and X\_\_\_X.

Described with appropriate examples:

When -kgap=1 then only sixteen features will exist for DNA and RNA but four hundred

(400) features will exist for protein. Features will be numbers of A\_A, A\_C, A\_G, A\_T, C\_A,

C\_C, C\_G, C\_T, G\_A, G\_C, G\_G, G\_T, T\_A, T\_C, T\_G, and T\_T of the whole sequence of

DNA respectively.

When -kgap=2 then only thirty two (32) features will exist for DNA and RNA but eight

hundred (800) features will exist for protein. Features will be numbers of A\_A, A\_C, A\_G, A\_T, C\_A,C\_C, C\_G, C\_T, G\_A, G\_C, G\_G, G\_T, T\_A, T\_C, T\_G,T\_T, A\_\_A, A\_\_C, A\_\_G,

A\_\_T, C\_\_A, C\_\_C, C\_\_G, C\_\_T, G\_\_A, G\_\_C, G\_\_G, G\_\_T, T\_\_A, T\_\_C, T\_\_G, and T\_\_T of

the whole sequence of DNA respectively.

When -kgap=3 then only forty eight (48) features will exist for DNA and RNA, but one

thousand and two hundred (1,200) features will exist for protein. Features will be numbers

of A\_A, A\_C, A\_G, A\_T, C\_A,C\_C, C\_G, C\_T, G\_A, G\_C, G\_G, G\_T, T\_A, T\_C, T\_G,T\_T, A\_\_A, A\_\_C, A\_\_G,A\_\_T, C\_\_A, C\_\_C, C\_\_G, C\_\_T, G\_\_A, G\_\_C, G\_\_G, G\_\_T, T\_\_A, T\_\_C, T\_\_G, T\_\_T, A\_\_\_A, A\_\_\_C, A\_\_\_G, A\_\_\_T, C\_\_\_A, C\_\_\_C, C\_\_\_G, C\_\_\_T, G\_\_\_A, G\_\_\_C, G\_\_\_G,G\_\_\_T, T\_\_\_A, T\_\_\_C, T\_\_\_G, and T\_\_\_T of the whole sequence of DNA respectively.

1. **MonoDiKGap Theoretical Description (MonoDiKGap)**

When kgap=n then the  features will exist for DNA and RNA, but features will exist for protein.

When -kgap=1, feature structure will be X\_XX.

When -kgap=2, feature structure will be X\_XX, and X\_\_XX.

When -kgap=3, feature structure will be X\_XX, X\_\_XX, and X\_\_\_XX.

Described with appropriate examples:

When -kgap=1 then only sixty four (64) features will exist for DNA and RNA, but eight thousand (8,000) features will exist for protein. Features will be numbers of A\_AA, A\_AC,A\_AG, A\_AT, A\_CA, A\_CC, A\_CG, A\_CT, A\_GA, A\_GC, A\_GG, A\_GT, A\_TA, A\_TC, A\_TG,A\_TT, C\_AA, C\_AC, C\_AG, C\_AT, C\_CA, C\_CC, C\_CG, C\_CT, C\_GA, C\_GC, C\_GG, C\_GT,C\_TA, C\_TC, C\_TG, C\_TT, G\_AA, G\_AC, G\_AG, G\_AT, G\_CA, G\_CC, G\_CG, G\_CT, G\_GA,G\_GC, G\_GG, G\_GT, G\_TA, G\_TC, G\_TG, G\_TT, T\_AA, T\_AC, T\_AG, T\_AT, T\_CA, T\_CC,T\_CG, T\_CT, T\_GA, T\_GC, T\_GG, T\_GT, T\_TA, T\_TC, T\_TG, and T\_TT of the whole sequence of DNA respectively.

When -kgap=2 then only hundred and twenty eight (128) features will exist for DNA and RNA, but sixteen thousand (16,000) features will exist for protein. Features will be numbers of A\_AA, A\_AC,A\_AG, A\_AT, A\_CA, A\_CC, A\_CG, A\_CT, A\_GA, A\_GC, A\_GG, A\_GT, A\_TA, A\_TC, A\_TG,A\_TT, C\_AA, C\_AC, C\_AG, C\_AT, C\_CA, C\_CC, C\_CG, C\_CT, C\_GA, C\_GC, C\_GG, C\_GT,C\_TA, C\_TC, C\_TG, C\_TT, G\_AA, G\_AC, G\_AG, G\_AT, G\_CA, G\_CC, G\_CG, G\_CT, G\_GA,G\_GC, G\_GG, G\_GT, G\_TA, G\_TC, G\_TG, G\_TT, T\_AA, T\_AC, T\_AG, T\_AT, T\_CA, T\_CC,T\_CG, T\_CT, T\_GA, T\_GC, T\_GG, T\_GT, T\_TA, T\_TC, T\_TG,T\_TT, A\_\_AA, A\_\_AC, A\_\_AG, A\_\_AT, A\_\_CA, A\_\_CC, A\_\_CG, A\_\_CT, A\_\_GA, A\_\_GC,A\_\_GG, A\_\_GT, A\_\_TA, A\_\_TC, A\_\_TG, A\_\_TT, C\_\_AA, C\_\_AC, C\_\_AG, C\_\_AT, C\_\_CA,C\_\_CC, C\_\_CG, C\_\_CT, C\_\_GA, C\_\_GC, C\_\_GG, C\_\_GT, C\_\_TA, C\_\_TC, C\_\_TG, C\_\_TT,G\_\_AA, G\_\_AC, G\_\_AG, G\_\_AT, G\_\_CA, G\_\_CC, G\_\_CG, G\_\_CT, G\_\_GA, G\_\_GC, G\_\_GG,G\_\_GT, G\_\_TA, G\_\_TC, G\_\_TG, G\_\_TT, T\_\_AA, T\_\_AC, T\_\_AG, T\_\_AT, T\_\_CA, T\_\_CC,T\_\_CG, T\_\_CT, T\_\_GA, T\_\_GC, T\_\_GG, T\_\_GT, T\_\_TA, T\_\_TC, T\_\_TG, and T\_\_TT of the whole sequence of DNA respectively.

**6 RNA Feature Extraction**

“RNA\_Feature\_Extraction.py” contains 12 feature extraction methods for RNA sequencing data.

|  |  |
| --- | --- |
| RNA Feature Extraction Method | RNA Extraction Number |
| Kmer | 1 |
| Reverse Compliment Kmer | 2 |
| Pseudo dinucleotide composition | 3 |
| Dinucleotide-based auto covariance | 4 |
| Dinucleotide-based cross covariance | 5 |
| Dinucleotide-based auto-cross covariance | 6 |
| Nucleic acid composition | 7 |
| Di-nucleotide composition | 8 |
| Tri-nucleotide composition | 9 |
| zcurve | 10 |
| monoMonoKGap | 11 |
| monoDiKGap | 12 |

**1. Kmer**

For kmer descriptor, the RNA sequences are represented as the occurrence frequencies of  neighboring nucleic acids. The Kmer (k=3) descriptor can be calculated as:

 (1)

where  is the number of kmer type , while  is the length of a nucleotide sequence.

1. **Reverse compliment kmer (RCKmer)**

The reverse compliment kmer (2,4) is a variant of kmer descriptor, in which the kmers are not

expected to be strand-specific. For instance, for a DNA sequence, there are 16 types of 2-mers (i.e. 'AA', 'AC', 'AG', 'AT', 'CA', 'CC', 'CG', 'CT', 'GA', 'GC', 'GG', 'GT', 'TA', 'TC', 'TG', 'TT'), ‘TT’ is reverse compliment with ‘AA’. After removing the reverse compliment kmers, there are only 10 distinct kmers in the reverse compliment kmer approach ('AA', 'AC', 'AG', 'AT', 'CA', 'CC', 'CG', 'GA', 'GC', 'TA').

1. **Pseudo dinucleotide composition (PseDNC)**

PseDNC is an approach incorporating the contiguous local sequence-order information and the global sequence-order information into the feature vector of the RNA sequence. Given a RNA sequence D, the PseDNC feature vector of D is defined:



 (2)

where  is the normalized occurrence frequency of dinucleotides in the RNA sequence, the parameter  is an integer, representing the highest counted rank (or tier) of the correlation along a RNA sequence, w is the weight factor ranged from 0 to 1,  is called the correlation factor that reflects the sequence-order correlation between all the most contiguous dinucleotides along a RNA sequence, which is defined:

 (3)

where the correlation function is given by

 (4)

where  is the number of physicochemical indices, in this approach, 6 indices reflecting the local RNA structural properties are employed to generate the PseDNC feature vector,  represents the numerical value of the  physicochemical index of the dinucleotide .

1. **Dinucleotide-based auto covariance (DAC)**

The DAC 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:

 (5)

where  is a physicochemical index,  is the length of the RNA sequence  means the numerical value of the physicochemical index  for the dinucleotide  at position ,  is the average value for physicochemical index  along the whole sequence:

In such a way, the length of DAC feature vector is , where  is the number of physicochemical indices, and  is the maximum of sequence.

1. **Dinucleotide-based cross covariance (DCC)**

Given a RNA sequence, the DCC 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:

 (6)

Where  are two different physicochemical indices,  is the length of the RNA sequence,  is the numerical value of the physicochemical index  for the dinucleotide .  is the average value for physicochemical index value  along the whole sequence:

 (7)

In such a way, the length of the DCC feature vector is ,  is the maximum of sequence; N is the number of physicochemical indices.

1. **Dinucleotide-based auto-cross covariance (DACC)**

DACC is a combination of DAC and DCC. Therefore, the length of the DACC feature vector is , where N is the number of physicochemical indices and  is the maximum of sequence.

1. **Nucleic Acid Composition (NAC)**

The Nucleic Acid Composition (NAC) encoding calculates the frequency of each nucleic acid type in a nucleotide sequence. The frequencies of all 4 natural nucleic acids (i.e. “ACGT or U”) can be calculated as:

 (8)

where *N*(*t*) is the number of nucleic acid type *t*, while *N* is the length of a nucleotide sequence.

1. **Di-Nucleotide Composition (DNC)**

The Di-Nucleotide Composition gives 16 descriptors. It is defined as:

 (9)

where is the number of di-nucleotide represented by nucleic acid types *r* and *s*.

1. **Tri-Nucleotide Composition (TNC)**

The Tri-Nucleotide Composition gives 64 descriptors. It is defined as:

 (10)

where is the number of tri-nucleotide represented by nucleic acid types *r, s* and *t*.

1. **zCurve Mathematical Formula (zCurve)**

Z-curve theory is often used in genomic sequence analysis. It has got three components in three axis. They are de\_ned as following.

 (11)

Three features will generate using the zCurve method.

1. **MonoKGap Theoretical Description (MonoKGap)**

When kgap=n then the  features will exist for DNA and RNA.

When -kgap=1, feature structure will be X\_X.

When -kgap=2, feature structure will be X\_X, and X\_\_X.

When -kgap=3, feature structure will be X\_X. X\_\_X, and X\_\_\_X.

Described with appropriate examples:

When -kgap=1 then only sixteen (16) features will exist for DNA and RNA but four hundred

(400) features will exist for protein. Features will be numbers of A\_A, A\_C, A\_G, A\_T, C\_A,

C\_C, C\_G, C\_T, G\_A, G\_C, G\_G, G\_T, T\_A, T\_C, T\_G, and T\_T of the whole sequence of

DNA respectively.

When -kgap=2 then only thirty two (32) features will exist for DNA and RNA but eight

hundred (800) features will exist for protein. Features will be numbers of A\_A, A\_C, A\_G, A\_T, C\_A,C\_C, C\_G, C\_T, G\_A, G\_C, G\_G, G\_T, T\_A, T\_C, T\_G,T\_T, A\_\_A, A\_\_C, A\_\_G,

A\_\_T, C\_\_A, C\_\_C, C\_\_G, C\_\_T, G\_\_A, G\_\_C, G\_\_G, G\_\_T, T\_\_A, T\_\_C, T\_\_G, and T\_\_T of

the whole sequence of DNA respectively.

When -kgap=3 then only forty eight (48) features will exist for DNA and RNA, but one

thousand and two hundred (1,200) features will exist for protein. Features will be numbers

of A\_A, A\_C, A\_G, A\_T, C\_A,C\_C, C\_G, C\_T, G\_A, G\_C, G\_G, G\_T, T\_A, T\_C, T\_G,T\_T, A\_\_A, A\_\_C, A\_\_G,A\_\_T, C\_\_A, C\_\_C, C\_\_G, C\_\_T, G\_\_A, G\_\_C, G\_\_G, G\_\_T, T\_\_A, T\_\_C, T\_\_G, T\_\_T, A\_\_\_A, A\_\_\_C, A\_\_\_G, A\_\_\_T, C\_\_\_A, C\_\_\_C, C\_\_\_G, C\_\_\_T, G\_\_\_A, G\_\_\_C, G\_\_\_G,G\_\_\_T, T\_\_\_A, T\_\_\_C, T\_\_\_G, and T\_\_\_T of the whole sequence of DNA respectively.

1. **MonoDiKGap Theoretical Description (MonoDiKGap)**

When -kgap=n then the  features will exist for DNA and RNA, but

 features will exist for protein.

When -kgap=1, feature structure will be X\_XX.

When -kgap=2, feature structure will be X\_XX, and X\_\_XX.

When -kgap=3, feature structure will be X\_XX, X\_\_XX, and X\_\_\_XX.

Described with appropriate examples:

When -kgap=1 then only sixty four (64) features will exist for DNA and RNA, but eight thousand (8,000) features will exist for protein. Features will be numbers of A\_AA, A\_AC,A\_AG, A\_AT, A\_CA, A\_CC, A\_CG, A\_CT, A\_GA, A\_GC, A\_GG, A\_GT, A\_TA, A\_TC, A\_TG,A\_TT, C\_AA, C\_AC, C\_AG, C\_AT, C\_CA, C\_CC, C\_CG, C\_CT, C\_GA, C\_GC, C\_GG, C\_GT,C\_TA, C\_TC, C\_TG, C\_TT, G\_AA, G\_AC, G\_AG, G\_AT, G\_CA, G\_CC, G\_CG, G\_CT, G\_GA,G\_GC, G\_GG, G\_GT, G\_TA, G\_TC, G\_TG, G\_TT, T\_AA, T\_AC, T\_AG, T\_AT, T\_CA, T\_CC,T\_CG, T\_CT, T\_GA, T\_GC, T\_GG, T\_GT, T\_TA, T\_TC, T\_TG, and T\_TT of the whole sequence of DNA respectively.

When -kgap=2 then only hundred and twenty eight (128) features will exist for DNA and RNA, but sixteen thousand (16,000) features will exist for protein. Features will be numbers of A\_AA, A\_AC,A\_AG, A\_AT, A\_CA, A\_CC, A\_CG, A\_CT, A\_GA, A\_GC, A\_GG, A\_GT, A\_TA, A\_TC, A\_TG,A\_TT, C\_AA, C\_AC, C\_AG, C\_AT, C\_CA, C\_CC, C\_CG, C\_CT, C\_GA, C\_GC, C\_GG, C\_GT,C\_TA, C\_TC, C\_TG, C\_TT, G\_AA, G\_AC, G\_AG, G\_AT, G\_CA, G\_CC, G\_CG, G\_CT, G\_GA,G\_GC, G\_GG, G\_GT, G\_TA, G\_TC, G\_TG, G\_TT, T\_AA, T\_AC, T\_AG, T\_AT, T\_CA, T\_CC,T\_CG, T\_CT, T\_GA, T\_GC, T\_GG, T\_GT, T\_TA, T\_TC, T\_TG,T\_TT, A\_\_AA, A\_\_AC, A\_\_AG, A\_\_AT, A\_\_CA, A\_\_CC, A\_\_CG, A\_\_CT, A\_\_GA, A\_\_GC,A\_\_GG, A\_\_GT, A\_\_TA, A\_\_TC, A\_\_TG, A\_\_TT, C\_\_AA, C\_\_AC, C\_\_AG, C\_\_AT, C\_\_CA,C\_\_CC, C\_\_CG, C\_\_CT, C\_\_GA, C\_\_GC, C\_\_GG, C\_\_GT, C\_\_TA, C\_\_TC, C\_\_TG, C\_\_TT,G\_\_AA, G\_\_AC, G\_\_AG, G\_\_AT, G\_\_CA, G\_\_CC, G\_\_CG, G\_\_CT, G\_\_GA, G\_\_GC, G\_\_GG,G\_\_GT, G\_\_TA, G\_\_TC, G\_\_TG, G\_\_TT, T\_\_AA, T\_\_AC, T\_\_AG, T\_\_AT, T\_\_CA, T\_\_CC,T\_\_CG, T\_\_CT, T\_\_GA, T\_\_GC, T\_\_GG, T\_\_GT, T\_\_TA, T\_\_TC, T\_\_TG, and T\_\_TT of the whole sequence of DNA respectively.

**7 PROTEIN Feature Extraction**

“Protein\_Feature\_Extraction.py” contains 32 feature extraction methods for Protein sequencing data.

|  |  |
| --- | --- |
| Protein Feature Extraction Method | Protein Extraction Number |
| Amino acid composition | 1 |
| Composition of k-spaced amino acid pairs | 2 |
| Dipeptide composition | 3 |
| Grouped dipeptide composition | 4 |
| Grouped tripeptide composition | 5 |
| Cojoint triad | 6 |
| k-spaced cojoint triad | 7 |
| Composition | 8 |
| Transition | 9 |
| Distribution | 10 |
| Encoding based on grouped weight | 11 |
| Auto covariance | 12 |
| Moran autocorrelation | 13 |
| Geary autocorrelation | 14 |
| Quasi-sequence-order | 15 |
| Pseudo-amino acid composition | 16 |
| Amphiphilic pseudo-amino acid composition | 17 |
| Amino Acid Composition PSSM | 18 |
| Dipeptide composition PSSM | 19 |
| Pseudo PSSM | 20 |
| Auto covariance PSSM | 21 |
| Cross covariance PSSM | 22 |
| Auto Cross covariance PSSM | 23 |
| Bigram-PSSM | 24 |
| AB-PSSM | 25 |
| Secondary structure composition | 26 |
| Accessible surface area composition | 27 |
| Torsional angles composition | 28 |
| Torsional angles bigram | 29 |
| Structural probabilities Bigram | 30 |
| Torsional angles auto-covariance | 31 |
| Structural probabilities auto-covariance | 32 |

**Sequence information**

1. **Amino acid composition (AAC)**

The amino acid composition is the fraction of each amino acid type within a protein.The amino acid composition gives 20 features and the fractions of all 20 natural amino acids are calculated as

 (1)

where  is the number of the amino acid type  and  is the length of the sequence.

1. **Dipeptide composition (DC)**

The dipeptide composition is used to transform the variable length of proteins to fixed length feature vectors.A dipeptide composition has been used earlier by Grassmann et al. and Reczko and Bohr for the development of fold recognition methods.We adopt the same dipeptide composition-based approach in developing a deep neural networks-based method for predicting protein-protein inter-action.The dipeptide composition gives a fixed pattern length of 400.Dipeptide composition encapsulates information about the fraction of amino acids as well as their local order.The dipeptide composition is defined as

 (2)

wherethe number of dipeptide represented by amino acid typeand.

1. **-gap dipeptide composition introduces (g-GapDC)**

The -gap dipeptide composition introduces (-Gap DC) extracts important intrinsic correlation information of protein sequences in multidimensional space. For a protein ,  of the sequence length C, the calculation process of the -Gap DC can be expressed as

 (3)

 (4)

where  represents the number of residues in the primary structure of the two amino acids, represents the number of occurrences of the -th feature in the -gap dipeptide, and  represents the frequency at which the -th feature in the -gap dipeptide appears in the sequences. When , there is no gap between two adjacent amino acid residues, and when , it means that one amino acid residue is between two adjacent amino acid residues. From equation (9), a protein sequence can be represented as a 400-dimensional feature vector.

1. **Grouped di-peptide composition (GDC)**

The Grouped di-peptide composition encoding is another variation of the DPC descriptor. It is composed of a total of 25 descriptors that are defined as:

 (5)

where  is the number of the di-peptide type , and  is the length of the sequence.

1. **Grouped tri-peptide composition (GTC)**

The Grouped tri-peptide composition encoding is also a variation of TPC descriptor, which generates 125 descriptors, defined as:

 (6)

where  is the number of the tri-peptide type , and  is the length of the sequence.

1. **Conjoint triad (CT)**

First, 20 amino acids are clustered into seven classes based on dipoles and volumes of side chains. Considering the interaction between the amino acid and its vicinal amino acids, the three continuous amino acids are regarded as a unit, so that we can obtain  triad types. We calculate the frequency of occurrence of each triad called . Then the 343-dimensional feature vector is obtained according to equation (6).

 (7)

1. ***k*-Spaced Conjoint Triad (KSCTriad)**

The *k*-spaced conjoint triad descriptor is based on the conjoint ctriad descriptor,

which not only calculates the numbers of three continuous amino acid units, but also considers the

continuous amino acid units that are separated by any *k* residues (The default maximum value of *k*

is set to 5). For example, AxRxT is a 1-spaced triad. Thus, the dimensionality of the KSCTriad encoded feature vector is 343 (*k*+1).

1. **Composition, transition and distribution (CTD)**

**8.1 Composition**

For CTD, taking hydrophobicity as an example, all amino acids are classified into three categories: polar, neutral, and hydrophobic. The replacement sequence consists of three types, and the composition descriptors of the polar, neutral, and hydrophobic residues of the protein can be calculated as follows:

 (8)

where  is the number of amino acid type  in the coding sequence and  is sequence length.

**8.2 Transition**

The transision descriptor first converts the original sequence into a replacement sequence, and T includes three characteristics, the dipeptide composition frequency from the polar group to the neutral group and the composition frequency from the neutral group to the polar group. Transitions between the neutral group and the hydrophobicity and these between hydrophobic group and the polar group are defined in the same way. The T descriptor is defined as follows:

 (9)

Where  and  are the  and  dipeptide frequency, respectively.  is the sequence length.

**8.3 Distribution**

For each group (polar, neutral and hydrophobic), the D descriptor can generate five values. We obtain the position of the first, 25%, 50%, 75% and 100% of the specific encoded group sequence and then divided the position by the whole sequence. Given sequence MTTTVPKVFAFHEF. It can be represented as '32223213323213' according to Hydrophobicity\_PRAM900101. '1' represents polar, '2' represents neutral, '3' represents hydrophobicity. Take '3' for example, there are 6 residues encoded '3'. The first '3' is 1. The second '3' is . The third '3' is . The fourth '3' is . The fifth '3 is . The position in the first, the second, the third, the fourth, the fifth '3' of whole sequence are 1, 1, 8, 9, 14, respectively. So the distribution descriptor for '3' are ,,,,.

The C descriptor generates a 39-dimensional feature vector, the T descriptor generates a 39-dimensional feature vector, and the D descriptor generates a 195-dimensional feature vector. For each protein sequence, the CTD generates a 273-dimensional feature vector.

1. **Encoding based on grouped weight coding (EBGW)**

Studies have shown that different amino acids have different physical and chemical properties. The EBGW can capture the sequence and physicochemical information based on grouped situation. These amino acids are classified into four categories：

neutral and hydrophobic amino acids 

neutral and polarity amino acids 

acidic amino acids 

basic amino acids .

Thus, we can get three combinations, each of which can partition the 20 amino acid residues into two disjoint group:  vs , or vs, and  vs . Let be a protein sequence, we can transform it into three binary sequences by three homomorphic mapswhich are defined as follows:

 (10)

 (11)

 (12)

where , and we call as characteristic sequence of the protein sequence.

Then , , are three binary sequences of length *L*. These sequences are divided into a number of sub-sequences of increasing length successively. A fixed parameter *N* is set and the sub-sequence can be expressed as, where represents the integer operator. Calculate the frequency of 1 in each sub-sequence, each  can be converted into an *N*-dimensional feature vector. To sum up, for a protein sequence *P* with length *L*, a 3*N*-dimension vector can be obtained.

**Physicochemical information**

1. **Auto covariance (AC)**

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

 (13)

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

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

 (14)

where is a physicochemical index, is the length of the protein sequence, means the numerical value of the physicochemical index for the amino acid  at position ,  is the average value for physicochemical index along the whole sequence. In such a way, the length of AC feature vector is , where *N* is the number of physicochemical indices extracted from AAindex; ** is the maximum of sequence.

1. **Moreau-Broto autocorrelation (Morean-Broto)**

Moreau-Broto autocorrelation descriptor is defined as:

 (15)

where , and  indicate the  and the  amino acids of the protein sequence, respectively.  and indicate the normalized physicochemical values of  and . The  is the parameter that needs to be adjusted.

1. **Moran autocorrelation (Moran)**

Moran autocorrelation descriptor is defined as:

 (16)

where  represents the mean value of whole protein sequence for specific physicochemical property.

1. **Geary autocorrelation (Geary)**

Geary autocorrelation descriptor is defined as:

 (17)

1. **Quasi-sequence-order descriptors (QSO)**

The sequenceorder features can also be used for representing amino acid distribution patterns of a specific physicochemical property along protein or peptide sequence. These descriptors are derived from both the Schneider-Wrede physicochemical distance matrix and the Grantham chemical distance matrix between each pair of the 20 amino acids. Theth rank sequence-order-coupling number is defined as

 (18)

whereis the distance between the two amino acids at positionand. Maxlag is the maximum lag and the length of the protein must be not less than maxlag. The maxlag is equal to 30 in the experiment. For each amino acid type, the type-1 quasi-sequence-order descriptor can be defined as

 (19)

where is the normalized occurrence of amino acid type-1 and is a weighting factor. The type-2 quasi-sequence-order is defined as

 (20)

In addition to the Schneider-Wrede physicochemical distance matrix used by Chou et al., another chemical distance matrix by Grantham is also used here. The sequence-order features produce a total ofdescriptors.

1. **Pseudo-amino acid composition (PseAAC)**

Pseudo-amino acid composition are utilized to extract the physicochemical information. Pseudo-amino acid composition (PseAAC) represents the composition information of the protein and sequence order information. The feature vector of PseAAC can be represented:

 (21)

where the first 20-dimentional feature vector represents the amino acid composition information, and the latter  dimensional vector represents the order information of protein sequence.  is the length of amino acid sequence.

  (22)

where  is the normalized frequency of 20 amino acids in protein.  is the layer sequence correlation factor calculated according to the equation (2). Sequence correlation factor can be obtained from the hydrophobicity, hydrophilicity, and side-chain mass of the amino acid.

1. **Amphiphilic pseudo amino acid composition (APAAC)**

APAAC (http://www. csbio. sjtu. edu. cn/bioinf/PseAAC/type2. htm) are also called type-2 pseudoamino acid composition. The definitions of these qualities are similar to PAAC descriptors. First, two variables are derived from the original hydrophobicity valuesand hydrophilicity valueof 20 amino acids

 (23)

 (24)

whereand,the hydrophobicity and hydrophilicity correlation functions,are defined respectively as

 (25)

where sequence order factors can be defined as

 (26)

 (27)

 (28)

Then a set of descriptors called "Amphiphilic Pseudo Amino Acid Composition"(APAAC)are defined as

 (29)

 (30)

whereis the weight factor and is taken as, andis equal to 30 in the experiment. So, we produce a total ofdescriptors.

**Evolutionary information**

In order to obtain the evolutionary information of the amino acid sequence, all the protein sequences in the dataset are compared with the non-redundant database SwissProt using the PSI-BLAST program. The program can search sequences based on the iterative BLAST search method. Evolutionary information in the position specific scoring matrix (PSSM) plays an important role in biological system analysis.

During the running process, the parameter E-value threshold of PSI-BLAST is set to 0.001, the maximum number of iterations is set to 3, and the remaining parameters are set by default. Then PSSM of each protein sequence is obtained. For a protein sequence whose length is , PSSM is shown in equation (32).

  (31)

where each row of the PSSM represents a log likelihood score for amino acid substitutions occurring at corresponding positions in the query sequence. Where  represents the  position of query sequence being mutated to type  during evolution process. The scores are positive integers or negative integers. A positive integer indicates that more mutations have occurred in the alignment and a negative integer indicates that fewer substitutions have occurred in the alignment.

1. **Amino acid composition PSSM (AAC-PSSM)**

In this calculating process of amino acid composition PSSM (AAC-PSSM), the PSSM is standardized by the logic function. PSSM elements are map to the interval [0,1].

  (32)

PSSM are converted to feature vector by AAC-PSSM via equation (14)

 (33)

where ,  represents the composition information of the  amino acid residue, which is the average score of  amino acid in PSSM.

1. **Dipeptide composition PSSM (DPC-PSSM)**

The PSSM contains evolutionary information. ACC-PSSM only represents the composition information from PSSM, and loses the order information, which is insufficient to fully represent the evolutionary information. Dipeptide composition PSSM (DPC-PSSM) can reflect the sequence-order information in the PSSM, which converts the character signal into the numerical signal, and the extracted feature vector can be expressed as

 (34)

where , the dimension of DPC-PSSM is 400.

1. **Bi-gram PSSM (Bi-PSSM)**

For Bi-gram PSSM, the frequency of the transition from the  amino acids to the  amino acids is calculated:

 (35)

Therefore, there are 400 possible cases for , then the Bi-gram PSSM eigenvector for each protein sequence is:

 (36)

1. **Auto covariance PSSM (AC-PSSM)**

AC-PSSM 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:

 (37)

whereis one of the residues,is the length of the protein sequence,is the PSSM score of amino acidat position,is the average score for amino acid  along the whole sequence:

 (38)

In such a way, the number of AC variables can be calculated as, where  is the maximum of sequence.

1. **Cross covariance PSSM (CC-PSSM)**

CC-PSSM 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:

 (39)

whereare two different amino acids andis the average score for amino acidalong the sequence. Since the CC variables are not symmetric, the total number of CC variables is.

1. **Auto-cross covariance PSSM (ACC-PSSM)**

ACC-PSSM 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.

1. **Pseudo PSSM (PsePSSM)**

According to the pseudo amino acid composition, we obtain the PsePSSM feature vector:

  (40)

where, each protein sequence can generate the dimensional feature vector. The first 20-dimensional vector represents the composition information of the PSSM matrix, and the remaining  dimensional feature vector represents the order evolutionary information. PsePSSM can transform an inconsistent protein sequence into a consistent numerical vector by feature extraction.

1. **AB-PSSM**

AB-PSSM is based on the averaged PSSM profiles over blocks, each with 5 percent of a sequence. Thus, a protein sequence, regardless of its length, is divided into 20 blocks and each block consists of 20 features (derived from the 20 columns in PSSMs). Mathematically, for the  block, the feature  is a  dimensional feature vector, which is generated by using the following equation:

 (41)

where  is the size of the  block, which is 5 percent of the length of a sequence and Pej i is a  vector extracted from the PSSM profile at the  position in the  block. For each sequence, there are a total of 20 blocks; therefore, the final feature is a 400-dimensional vector.

**Structural information**

1. **Secondary structure composition (SSC)**

This feature is the normalized count or frequency of the structural motifs present at the amino-acid residue positions. There are three types of motifs: *α*-helix (H), *β*-sheet (E) and random coil (C). SPIDER2 returns a vector *SS* of dimension *L* × 1 containing this information. Thus we can define this feature as following:

 (42)

where, *L* is the length of the protein and

 (43)

where, *SSj* is the structural motif at position *j* of the protein sequence and *fi* is one of the 3 different motif symbols.

1. **Accessible surface area composition (ASA)**

The accessible surface area composition is the normalized sum of accessible surface area defined by:

 (44)

where ASA is the vector of accessible surface area of dimension *L* × 1 containing the values of accessible surface area for all the amino acid residues.

1. **Torsional angles composition (TAC)**

Four different types of torsional angles: *ϕ*, *ψ*, *τ* and *θ* are returned by SPIDER2 for each residue. First, we convert each of them into radians from degree angles and then take sign and cosine of the angles at each residue position. Thus we get a matrix of dimension. We denotethis matrix by *T*. Torsional angles composition is defined as

 (45)

1. **Torsional angles bigram (TA-bigram)**

The Bigram for the torsional angles is similar to that of the PSSM matrix and is defined as:

 (46)

1. **Structural probabilities bigram (SP-bigram)**

Structural probabilities for each position of the amino-acid residue are given in the SPD2 file as a matrix of dimension *L* × 3, which we denote by *P*. The Bigram of the structural probabilities is similar to that of PSSM matrix and is defined as:

 (47)

1. **Torsional angles auto-covariance (TAAC)**

This feature is also derived from the torsional angles and is defined as:

 (48)

1. **Structural probablities auto-covariance (SPAC)**

This feature is also derived from the structural probabilities and is defined as:

 (49)

**8 Feature Selection**

“Feature\_Selection.py” includes 20 supervised features selection methods. User need provide method name (See below), feature descriptors and label of feature descriptors.

1. **Lasso**

Lasso linear model is a regression method that performs both variable selection and regularization in order to enhance the prediction accuracy. Therefore, Lasso regression can be used in feature selection. The best model is selected by cross-validation. The optimization objective for Lasso is:

(1)

1. **Elastic Net**

Elastic Net is a regularized regression method that linearly combines the and penalties of the lasso and ridge methods.

1. **Chi-squared test (chi2)**

Chi-squared test ( test) is one of statistical method for feature selection. The chi-squared test is applied to determine whether there is significant difference between two events. is a measure of how much expected counts E and observed counts N deviate from each other (Chen, et al., 2009). A high value of indicates that the hypothesis of independence, which implies that expected and observed counts are similar, is incorrect. This score can be sued to select the n features with the highest values for test chi-square statistic from x, which must contain only non-negative features such as Booleans or frequencies, relative to the classes.

where is the observed value and is the expected value.

1. Pearson Correlation (pc)

The Pearson correlation is also known as the “product moment correlation coefficient”. Pearson Correlation is one of the simplest methods to explore features’ relation to the response variable. It is a measure of the linear correlation between two variables X and Y. The resulting value lies in [-1:1], with -1 meaning perfect negative correlation that means one variable increase whereas the other decrease, with +1 meaning perfect positive correlation and 0 meaning no linear correlation between two variables.

The Pearson’s correlation is commonly represented by

where n is sample size, are the individual sample points indexed with , (the sample mean), and analogously for .

**Use the following command to perform the Chi2 feature selection:**

tcsh% python Feature\_selection.py pc num\_to\_select feature\_descriptors.csv label.txt

1. ReliefF (reliefF)

Relief was designed for application to binary classification problems with discrete or numerical features. It calculates a feature score for each feature which can then be applied to rank and select top scoring features for feature selection. Relief feature scoring is based on the identification of feature value differences between nearest neighbor instance pairs. If a feature value difference is observed in a neighboring instance pair with the same class, the feature score decreases. Alternatively, if a feature value difference is observed in a neighboring instance pair with different class values, the feature score increases.

ReliefF algorithm is updated version of Relief algorithm. ReliefF algorithm is one of the most successful filtering feature selection methods. Relief algorithms are commonly applied to genetic analyses, where epistasis is common the algorithms implemented in this package can be applied to almost any supervised classification data set and supports.

**Use the following command to perform the Chi2 feature selection:**

tcsh% python Feature\_selection.py relieff num\_to\_select feature\_descriptors.csv label.txt

1. Linear Discriminant Analysis (lda)

Linear Discriminant Analysis (Blei, et al., 2003) is a generative probabilistic model for collections of discrete datasets such as text corpora.

**Use the following command to perform the Chi2 feature selection:**

tcsh% python Feature\_selection.py lda num\_to\_select feature\_descriptors.csv label.txt

Unsupervised Methods

1. Principal Component Analysis (pca)

PCA (Pearson, 1901) is used to decompose a multivariate dataset in a set of successive orthogonal components that explain a maximum amount of the variance.

**Use the following command to perform the Chi2 feature selection:**

tcsh% python Feature\_selection.py pca num\_to\_select feature\_descriptors.csv

1. Kernal Principal Component Analysis (kpca)
2. Locally-linear embedding (lle)
3. Singular Value Decomposition (svd)
4. Multi-dimensional Scaling (mds)
5. Independent Component Analysis (ica)
6. Factor Analysis (fa)
7. Non-negative matrix factorization (nmf)

Information Theory Methods

1. Mutual Information (mi)