# BBSdb, an open resource for bacterial biofilm-associated proteins

Zhiyuan Zhang1, 2\*, Yuanyuan Pan2, Wajid Hussain3, Guozhong Chen2, Erguang Li1,\*

1Medical School, Nanjing University, Nanjing, Jiangsu, 210093, China.

2Department of Medical Information Engineering, School of Medical Information, Wannan Medical College, Wuhu 241000, China.

3Advanced Biomaterials and Tissue Engineering Center, College of Life Sciences and Technology, Huazhong University of Science and Technology, Wuhan 430074, China.

\*Corresponding authors: Email: [erguang@nju.edu.cn](mailto:erguang@nju.edu.cn;), zhangzhiyuan@wnmc.edu.cn

**Supplementary Material**

**Content**

### **S1.** Summary

### **S2.** Used Feature Descriptors for Protein sequences

### S1. Summary

## 1.1The dimension of sequence-based d**escriptor**

Table S1 dimension of 46 sequence-based descriptors

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Description | Descriptor | Dimension | Description | Descriptor | Dimension |
| the frequencies of 20 different amino acids | AAC(k=1) | 20 | Enhanced Amino Acid Composition | EAAC(k=7) | 20 |
| the occurrences of 20 different amino acids | AAC\_count(k=2) | 20 | Enhanced Amino Acid Composition | EAAC(k=8) | 20 |
| AAindex | AAindex | 5000 | Enhanced Amino Acid Composition | EAAC(k=9) | 20 |
| Composition of k-spaced Amino Acid Pairs | CKSAAP(k=2) | 1200 | Enhanced Amino Acid Composition | EAAC(k=10) | 20 |
| Composition of k-spaced Amino Acid Pairs | CKSAAP(k=3) | 1600 | Enhanced Amino Acid Composition (occurrences) | p\_EAAC(k=5) | 20 |
| Composition of k-spaced Amino Acid Pairs | CKSAAP(k=4) | 2000 | Enhanced Amino Acid Composition (occurrences) | p\_EAAC(k=6) | 20 |
| Composition of k-spaced Amino Acid Pairs | CKSAAP(k=5) | 2400 | Enhanced Amino Acid Composition (occurrences) | p\_EAAC(k=7) | 20 |
| Composition of k-spaced Amino Acid Pairs | CKSAAP(k=6) | 2800 | Enhanced Amino Acid Composition (occurrences) | p\_EAAC(k=8) | 20 |
| The Composition descriptor | CTDC | 39 | Enhanced Amino Acid Composition (occurrences) | p\_EAAC(k=9) | 20 |
| The Distribution descriptor | CTDD | 195 | Enhanced Amino Acid Composition (occurrences) | p\_EAAC(k=10) | 20 |
| The Transition descriptor T | CTDT | 39 | Grouped Amino Acid Compositio | GAAC | 5 |
| Conjoint Triad | CTriad(k=0) | 343 | Enhanced GAAC | EGAAC(k=5) | 5 |
| Conjoint Triad | CTriad(k=1) | 343 | Enhanced GAAC | EGAAC(k=6) | 5 |
| Conjoint Triad | CTriad(k=2) | 343 | Enhanced GAAC | EGAAC(k=7) | 5 |
| Conjoint Triad | CTriad(k=3) | 343 | Enhanced GAAC | EGAAC(k=8) | 5 |
| Conjoint Triad | CTriad(k=4) | 343 | Enhanced GAAC | EGAAC(k=9) | 5 |
| Conjoint Triad | CTriad(k=5) | 343 | Enhanced GAAC | EGAAC(k=10) | 5 |
| Pseudo-Amino Acid Composition | PAAC | 23 | Grouped Di-Peptide Composition | GDPC | 25 |
| Dipeptide Deviation from Expected Mean | DDE | 400 | Geary correlation | Geary | 240 |
| The Dipeptide Composition | DPC | 400 | Grouped Tri-Peptide Composition | GTPC | 125 |
| Enhanced Amino Acid Composition | EAAC(k=5) | 20 | The Tripeptide Composition | TPC | 8000 |
| Enhanced Amino Acid Composition | EAAC(k=6) | 20 | A combination of AAC and the extension of DPC | aac\_dpc | 420 |
| Composition (C), transition (T) and distribution (D) | ctd | 273 | A combination of AAC and the extension of TPC | aac\_tpc | 8020 |

## 1.2 Performance evaluation

Table S2. Performance evaluation indicators for 46 sequence-based feature descriptors using RF classifier

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **feature descriptors** | **Sensitivity** | **Specificity** | **Accuracy** | **MCC** | **Recall** | **Precision** | **F1-score** |
| AAC | 0.73 | 0.67 | 0.70 | 0.41 | 0.73 | 0.70 | 0.71 |
| AAC\_count | 0.67 | 0.69 | 0.68 | 0.37 | 0.67 | 0.70 | 0.68 |
| aac\_dpc | 0.63 | 0.59 | 0.61 | 0.22 | 0.63 | 0.61 | 0.62 |
| aac\_tpc | 0.67 | 0.58 | 0.62 | 0.25 | 0.67 | 0.62 | 0.64 |
| AAindex | 0.67 | 0.58 | 0.62 | 0.25 | 0.67 | 0.61 | 0.63 |
| CKSAAP\_2 | 0.67 | 0.68 | 0.68 | 0.36 | 0.67 | 0.69 | 0.68 |
| CKSAAP\_3 | 0.66 | 0.68 | 0.67 | 0.35 | 0.66 | 0.69 | 0.67 |
| CKSAAP\_4 | 0.66 | 0.70 | 0.68 | 0.36 | 0.66 | 0.70 | 0.67 |
| CKSAAP\_5 | 0.68 | 0.69 | 0.68 | 0.37 | 0.68 | 0.70 | 0.68 |
| CKSAAP\_6 | 0.66 | 0.70 | 0.68 | 0.37 | 0.66 | 0.70 | 0.67 |
| ctd | 0.66 | 0.58 | 0.62 | 0.24 | 0.66 | 0.62 | 0.63 |
| CTDC | 0.60 | 0.59 | 0.60 | 0.20 | 0.60 | 0.60 | 0.60 |
| CTDD | 0.60 | 0.58 | 0.59 | 0.17 | 0.60 | 0.59 | 0.59 |
| CTDT | 0.68 | 0.63 | 0.65 | 0.31 | 0.68 | 0.66 | 0.66 |
| CTriad\_0 | 0.61 | 0.62 | 0.61 | 0.23 | 0.61 | 0.62 | 0.61 |
| CTriad\_1 | 0.61 | 0.62 | 0.61 | 0.23 | 0.61 | 0.62 | 0.61 |
| CTriad\_2 | 0.61 | 0.62 | 0.61 | 0.23 | 0.61 | 0.62 | 0.61 |
| CTriad\_3 | 0.61 | 0.62 | 0.61 | 0.23 | 0.61 | 0.62 | 0.61 |
| CTriad\_4 | 0.61 | 0.62 | 0.61 | 0.23 | 0.61 | 0.62 | 0.61 |
| CTriad\_5 | 0.61 | 0.62 | 0.61 | 0.23 | 0.61 | 0.62 | 0.61 |
| DDE | 0.67 | 0.66 | 0.66 | 0.33 | 0.67 | 0.68 | 0.67 |
| DPC | 0.65 | 0.66 | 0.66 | 0.32 | 0.65 | 0.68 | 0.66 |
| EAAC\_10 | 0.68 | 0.70 | 0.69 | 0.39 | 0.68 | 0.70 | 0.69 |
| EAAC\_5 | 0.67 | 0.67 | 0.67 | 0.35 | 0.67 | 0.68 | 0.67 |
| EAAC\_6 | 0.68 | 0.64 | 0.66 | 0.32 | 0.68 | 0.66 | 0.67 |
| EAAC\_7 | 0.70 | 0.69 | 0.69 | 0.39 | 0.70 | 0.69 | 0.70 |
| EAAC\_8 | 0.68 | 0.70 | 0.69 | 0.38 | 0.68 | 0.70 | 0.69 |
| EAAC\_9 | 0.71 | 0.68 | 0.70 | 0.39 | 0.71 | 0.70 | 0.70 |
| EGAAC\_10 | 0.62 | 0.58 | 0.60 | 0.20 | 0.62 | 0.60 | 0.61 |
| EGAAC\_5 | 0.63 | 0.58 | 0.61 | 0.21 | 0.63 | 0.61 | 0.61 |
| EGAAC\_6 | 0.60 | 0.66 | 0.63 | 0.27 | 0.60 | 0.65 | 0.62 |
| EGAAC\_7 | 0.51 | 0.56 | 0.54 | 0.07 | 0.51 | 0.54 | 0.52 |
| EGAAC\_8 | 0.62 | 0.62 | 0.62 | 0.23 | 0.62 | 0.62 | 0.62 |
| EGAAC\_9 | 0.65 | 0.67 | 0.66 | 0.33 | 0.65 | 0.67 | 0.66 |
| GAAC | 0.73 | 0.67 | 0.70 | 0.40 | 0.73 | 0.70 | 0.71 |
| GDPC | 0.69 | 0.68 | 0.69 | 0.38 | 0.69 | 0.70 | 0.69 |
| Geary | 0.70 | 0.69 | 0.69 | 0.39 | 0.70 | 0.70 | 0.70 |
| GTPC | 0.70 | 0.69 | 0.70 | 0.39 | 0.70 | 0.70 | 0.70 |
| p\_EAAC\_10 | 0.66 | 0.59 | 0.62 | 0.25 | 0.66 | 0.62 | 0.64 |
| p\_EAAC\_5 | 0.73 | 0.68 | 0.70 | 0.41 | 0.73 | 0.71 | 0.71 |
| p\_EAAC\_6 | 0.70 | 0.68 | 0.69 | 0.38 | 0.70 | 0.70 | 0.69 |
| p\_EAAC\_7 | 0.70 | 0.68 | 0.69 | 0.39 | 0.70 | 0.70 | 0.70 |
| p\_EAAC\_8 | 0.70 | 0.67 | 0.68 | 0.37 | 0.70 | 0.69 | 0.69 |
| p\_EAAC\_9 | 0.64 | 0.58 | 0.61 | 0.23 | 0.64 | 0.62 | 0.63 |
| PAAC | 0.70 | 0.68 | 0.69 | 0.37 | 0.70 | 0.69 | 0.69 |
| TPC | 0.71 | 0.67 | 0.69 | 0.39 | 0.71 | 0.70 | 0.70 |

Table S3 The prediction of 23 shared gene using developed predictor

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **gene\_name** | **protein\_name** | **uniprot** | **prediction** | **bacteria** |  |
| *b1551* | Uncharacterized protein YnfN | <https://www.uniprot.org/uniprotkb/P76157/entry> | yes | *E.coli* |  |
| *secG* | Protein-export membrane protein SecG | <https://www.uniprot.org/uniprotkb/P0AG99/entry> | yes | *E.coli* |  |
| *zapA* | Cell division protein ZapA | <https://www.uniprot.org/uniprotkb/P0ADS2/entry> | yes | *E.coli* |  |
| *wzzB* | Chain length determinant protein | <https://www.uniprot.org/uniprotkb/P76372/entry> | yes | *E.coli* |  |
| *ydfK* | Cold shock protein YdfK | <https://www.uniprot.org/uniprotkb/P76154/entry> | yes | *E.coli* |  |
| *yjcB* | Uncharacterized protein YjcB | <https://www.uniprot.org/uniprotkb/P32700/entry> | yes | *E.coli* |  |
| *csrA* | Carbon storage regulator | <https://www.uniprot.org/uniprotkb/P69913/entry> | yes | *E.coli* |  |
| *b4140* | UPF0716 protein FxsA | <https://www.uniprot.org/uniprotkb/P37147/entry> | yes | *E.coli* |  |
| *yifE* | UPF0438 protein YifE | <https://www.uniprot.org/uniprotkb/P0ADN2/entry> | yes | *E.coli* |  |
| *yheL* | Protein TusB | <https://www.uniprot.org/uniprotkb/P45530/entry> | yes | *E.coli* |  |
| *garP* | Probable galactarate/D-glucarate transporter GarP | <https://www.uniprot.org/uniprotkb/P0AA80/entry> | yes | *E.coli* |  |
| *infA* | Translation initiation factor IF-1 | <https://www.uniprot.org/uniprotkb/P69222/entry> | no | *E.coli* |  |
| *cspB* | Cold shock-like protein CspB | <https://www.uniprot.org/uniprotkb/P36995/entry> | no | *E.coli* |  |
| *cspG* | none | none | no | *E.coli* |  |
| *tyrU* | none | none | no | *E.coli* | tRNA |
| *cspI* | Cold shock-like protein CspI | <https://www.uniprot.org/uniprotkb/P0A986/entry> | no | *E.coli* |  |
| *groS* | Co-chaperonin GroES | <https://www.uniprot.org/uniprotkb/P0A6F9/entry> | no | *E.coli* |  |
| *exbD* | Biopolymer transport protein ExbD | <https://www.uniprot.org/uniprotkb/P0ABV2/entry> | no | *E.coli* |  |
| *argQ* | none | none | no | *E.coli* | tRNA |
| *sraF* | none | none | no | *E.coli* | tRNA |
| *metU* | none | none | no | *E.coli* | tRNA |
| *rplU* | Large ribosomal subunit protein bL21 | <https://www.uniprot.org/uniprotkb/P0AG48/entry> | no | *E.coli* |  |
| *greA* | Transcription elongation factor GreA | <https://www.uniprot.org/uniprotkb/P0A6W5/entry> | no | *E.coli* |  |

### S2. Used Feature Descriptors for Protein sequences

### 2.1 Amino Acid Composition (AAC)

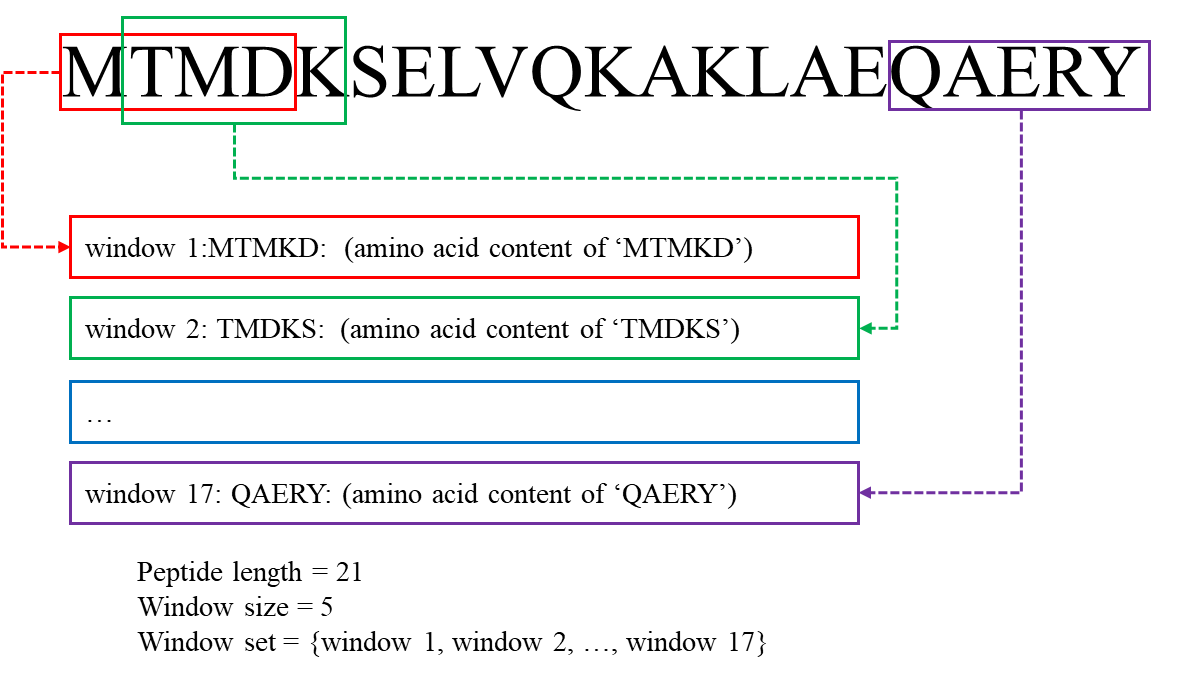
To include the compositional information, we computed the frequencies and the occurrences of 20 different amino acids in the peptide sequence, respectively. The feature vector for the AAC descriptor is represented below:

AAC(k)=(*V1,V2,⋯,Va,…,V20*),

where Va denotes the value of the amino acid type *a*, and k(k=1,2) denotes the method type; for example, 1 denotes the frequency while 2 denotes the occurrence number.

### 2.2 Enhanced Amino Acid Composition (EAAC)

The Enhanced Amino Acid Composition (EAAC) feature type calculates the AAC based on the sequence window of fixed length (the default value is 5) that continuously slides from the N- to C-terminus of each peptide and can be usually applied to encode the peptides with an equal length. An illustrated example of this encoding scheme is provided in the following **Figure S1**.



**Figure S1**. An illustrated example of the EAAC descriptor.

The EAAC can be calculated as:



where *N(t,win)* is the number of amino acid type *t* in the sliding window *win* and *N(win)* is the size of the sliding window *win*.

### 2.3 Enhanced Amino Acid Composition (occurrences)

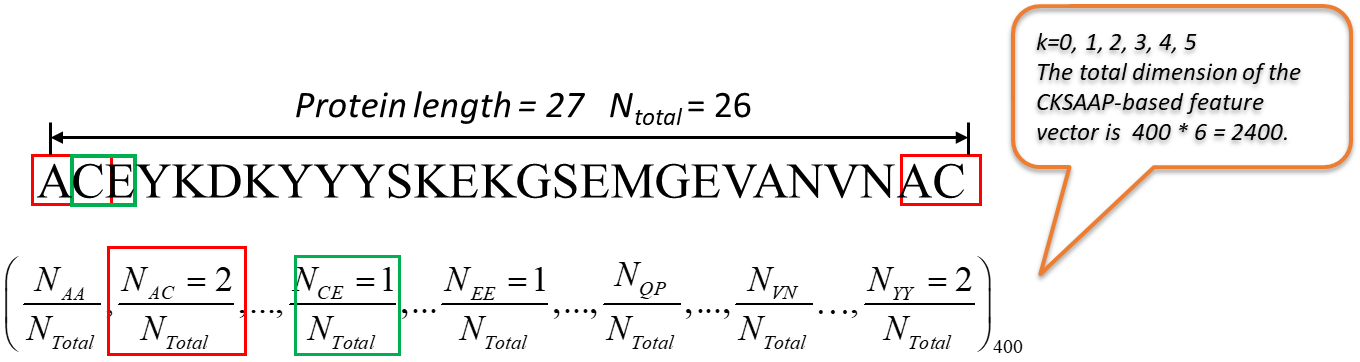
The Enhanced Amino Acid Composition (p\_EAAC) feature type is introduced here for the first time. It calculates the AAC (occurrences) based on the sequence window of fixed length (the default value is 5) that continuously slides from the N- to C-terminus of each peptide and can be usually applied to encode the peptides with an equal length.

## 2.4 Composition of *k*-spaced Amino Acid Pairs (CKSAAP)

The CKSAAP feature encoding calculates the frequency of amino acid pairs separated by any *k* residues (*k* = 0, 1, 2, … , 5. The default maximum value of *k* is 5) (Chen, et al., 2009; Chen, et al., 2007a; Chen, et al., 2007b; Chen, et al., 2008). Taking *k* = 0 as an example, there are 400 0-spaced residue pairs (i.e., AA, AC, AD,…, YY.). Then, a feature vector can be defined as:



The value of each descriptor denotes the composition of the corresponding residue pair in the protein or peptide sequence. For instance, if the residue pair AA appears *m* times in the protein, the composition of the residue pair AA is equal to *m* divided by the total number of *0*-spaced residue pairs (*Ntotal*) in the protein. For *k* = 0, 1, 2, 3, 4 and 5, the value of *Ntotal* is *P* – 1, *P* – 2, *P* – 3, *P* – 4, *P* – 5 and *P* – 6 for a protein of length *P*, respectively. An illustrated example of this encoding scheme (*k*=0) is provided in the following **Figure S2**.



## 2.5 Tri-Peptide Composition (TPC)

The Tripeptide Composition (TPC) (Bhasin and Raghava, 2004) gives 8000 descriptors, defined as:



where *Nrst* is the number of tripeptides represented by amino acid types *r*, *s* and *t.*

## 2.6 Di-Peptide Composition (DPC)

The Dipeptide Composition (Saravanan and Gautham, 2015) gives 400 descriptors. It is defined as:



## 2.7 Dipeptide Deviation from Expected Mean (DDE)

where *Nrs* is the number of dipeptides represented by amino acid types *r* and *s.*

The Dipeptide Deviation from Expected Mean feature vector (Saravanan and Gautham, 2015) is constructed by computing three parameters, i.e. dipeptide composition (*Dc*), theoretical mean (*Tm*), and theoretical variance (*Tv*). The above three parameters and the DDE are computed as follows. *Dc(r,s)*, the dipeptide composition measure for the dipeptide ‘*rs’*, is given as



where *Nrs* is the number of dipeptides represented by amino acid types *r* and *s* and *N* is the length of the protein or peptide. *Tm(r,s)*, the theoretical mean, is given by:



where *Cr* is the number of codons that code for the first amino acid and *Cs* is the number of codons that code for the second amino acid in the given dipeptide ‘*rs*’. *CN* is the total number of possible codons, excluding the three stop codons (i.e., 61). *Tv* (*r,s*), the theoretical variance of the dipeptide ‘*rs*’, is given by:



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



## 2.8 Grouped Amino Acid Composition (GAAC)

In the GAAC encoding, the 20 amino acid types are further categorized into five classes according to their physicochemical properties, e.g. hydrophobicity, charge and molecular size (Lee, et al., 2011b). The five classes include the aliphatic group (*g1*: GAVLMI), aromatic group (*g2*: FYW), positive charge group (*g3*: KRH), negative charged group (*g4*: DE) and uncharged group (*g5*: STCPNQ). GAAC descriptor is the frequency of each amino acid group, which is defined as:





where *N(g)* is the number of amino acid in group *g*, *N(t)* is the number of amino acid type *t,* and *N* is the length of the protein/peptide sequence.

## 2.9 Enhanced GAAC (EGAAC)

The Enhanced GAAC (EGAAC) is also for the first time proposed in this work. It calculates GAAC in windows of fixed length (default is 5) continuously sliding from the N- to C-terminal of each peptide and is usually applied to peptides with an equal length.



where *N(g, win)* is the number of amino acids in group *g* within the sliding window *win* and *N(win)* is the size of the sliding window *win*.

## 2.10 Grouped Di-Peptide Composition (GDPC)

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:



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

## 2.11 Grouped Tri-Peptide Composition (GTPC)

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



where *Nrst* is the number of tripeptides represented by amino acid type groups *r*, *s* and *t*. *N* is the length of a protein or peptide sequence.

## 1.12 Geary correlation (Geary)

The Geary autocorrelation descriptors (Sokal and Thomson, 2006) for a protein or peptide sequence are defined as:



where *d*, *P*, *Pi* and *Pi+d*, *nlag* have the same definitions as described above.

## 2.13 Composition/Transition/Distribution (CTD)

The Composition, Transition and Distribution (CTD) features represent the amino acid distribution patterns of a specific structural or physicochemical property in a protein or peptide sequence (Cai, et al., 2003; Cai, et al., 2004; Dubchak, et al., 1995; Dubchak, et al., 1999; Han, et al., 2004). 13 types of physicochemical properties have been previously used for computing these features. These include hydrophobicity, normalized Van der Waals Volume, polarity, polarizability, charge, secondary structures and solvent accessibility. These descriptors are calculated according to the following procedures: (i) The sequence of amino acids is transformed into a sequence of certain structural or physicochemical properties of residues; (ii) Twenty amino acids are divided into three groups for each of the seven different physicochemical attributes based on the main clusters of the amino acid indices of Tomii and Kanehisa (Tomii and Kanehisa, 1996). The groups of amino acids are listed in **Table S1**.

**Table S1.** Amino acid physicochemical attributes and the division of the amino acids into three groups according to each attribute.

|  |  |  |  |
| --- | --- | --- | --- |
| Attribute | Division | | |
| Hydrophobicity\_PRAM900101 | Polar: RKEDQN | Neutral: GASTPHY | Hydrophobicity: CLVIMFW |
| Hydrophobicity\_ARGP820101 | Polar: QSTNGDE | Neutral: RAHCKMV | Hydrophobicity: LYPFIW |
| Hydrophobicity\_ZIMJ680101 | Polar: QNGSWTDERA | Neutral: HMCKV | Hydrophobicity: LPFYI |
| Hydrophobicity\_PONP930101 | Polar: KPDESNQT | Neutral: GRHA | Hydrophobicity: YMFWLCVI |
| Hydrophobicity\_CASG920101 | Polar: KDEQPSRNTG | Neutral: AHYMLV | Hydrophobicity: FIWC |
| Hydrophobicity\_ENGD860101 | Polar: RDKENQHYP | Neutral :SGTAW | Hydrophobicity: CVLIMF |
| Hydrophobicity\_FASG890101 | Polar: KERSQD | Neutral: NTPG | Hydrophobicity: AYHWVMFLIC |
| Normalized van der Waals volume | Volume range: 0-2.78  GASTPD | Volume range: 2.95-94.0  NVEQIL | Volume range: 4.03-8.08  MHKFRYW |
| Polarity | Polarity value: 4.9-6.2  LIFWCMVY | Polarity value: 8.0-9.2  PATGS | Polarity value: 10.4-13.0  HQRKNED |
| Polarizability | Polarizability value: 0-1.08  GASDT | Polarizability value: 0.128-120.186  GPNVEQIL | Polarizability value: 0.219-0.409  KMHFRYW |
| Charge | Positive: KR | Neutral: ANCQGHILMFPSTWYV | Negative: DE |
| Secondary structure | Helix: EALMQKRH | Strand: VIYCWFT | Coil: GNPSD |
| Solvent accessibility | Buried: ALFCGIVW | Exposed: PKQEND | Intermediate: MPSTHY |

**2.14 CTDC**

Taking the hydrophobicity attribute as an example, all amino acids are divided into three groups: polar, neutral and hydrophobic (**Table S1**). The Composition descriptor consists of three values: the global compositions (percentage) of polar, neutral and hydrophobic residues of the protein. An illustrated example of this encoding scheme is provided in the following **Figure S4**. The Composition descriptor can be calculated as follows:



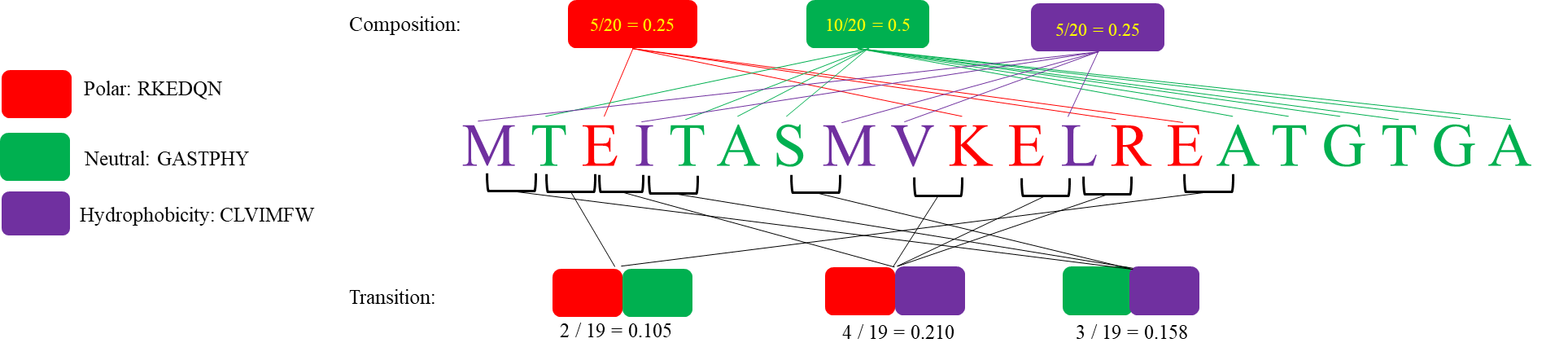
where *N(r)* is the number of amino acid type *r* in the encoded sequence and *N* is the length of the sequence.

## **2.15** CTDT

The Transition descriptor T also consists of three values (Dubchak, et al., 1995; Dubchak, et al., 1999): A transition from the polar group to the neutral group is the percentage frequency with which a polar residue is followed by a neutral residue or a neutral residue by a polar residue. Transitions between the neutral group and the hydrophobic group and those between the hydrophobic group and the polar group are defined in a similar way. The transition descriptor can then be calculated as:



where *N(r,s)* and *N(s,r)* are the numbers of dipeptides encoded as “*rs*” and “*sr*” respectively in the sequence, while *N* is the length of the sequence. An illustrated example of this encoding scheme is provided in the following **Figure S3**.



**Figure S3**. An illustrated example of the calculation of composition and transition descriptors. This example uses the hydrophobicity attribute.

## **2.16** CTDD

The Distribution descriptor consists of five values for each of the three groups (polar, neutral and hydrophobic) (Dubchak, et al., 1995; Dubchak, et al., 1999), namely the corresponding fraction of the entire sequence, where the first residue of a given group is located, and where 25, 50, 75 and 100% of occurrences are contained.

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

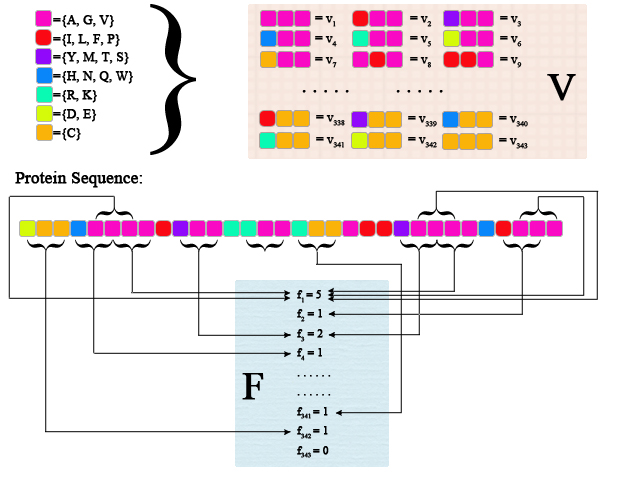
## 2.17 Conjoint Triad (CTriad)

The Conjoint Triad descriptor (CTriad) considers the properties of one amino acid and its vicinal amino acids by regarding any three continuous amino acids as a single unit (Shen, et al., 2007). First, the protein sequence is represented by a binary space (*V, F*), where *V* denotes the vector space of the sequence features, and each feature (*Vi*) represents a sort of triad type; *F* is the number vector corresponding to *V*, where *fi*, the value of the *i-*th dimension of *F*, is the number of type *Vi* appearing in the protein sequence.

For the amino acids that have been catalogued into seven classes, the size of *V* should be equal to 7ⅹ7ⅹ7=343. Accordingly, *i* = 1, 2, 3, …, 343. An illustrated example of this encoding scheme is provided in the following **Figure S4**.

In principle, the longer a protein sequence, the higher the probability to have larger values of *fi*, confounding the comparison of proteins with different lengths. Thus, we define a new parameter, *di*, by normalizing *fi* with the following equation:





**Figure S4**. Schematic diagram for constructing the vector space (*V, F*) of a given protein sequence. *V* is the vector space of the sequence features; each feature (*Vi*) represents a triad composed of three consecutive amino acids; *F* is the number vector corresponding to *V*, and the value of the *i-*th entry of *F*, denoted *fi*, is the number of occurrences that the triad associated with *Vi* appearing in the protein sequence. The figure was adapted from the Supplementary Figure in (Shen, et al., 2007).

## 2.18 Pseudo-Amino Acid Composition (PAAC)

This group of descriptors has been proposed in (Chou, 2001; Chou, 2005). Let , , for *i* = 1, 2, 3, … 20 be the original hydrophobicity values, the original hydrophilicity values and the original side chain masses of the 20 natural amino acids, respectively. They are converted to the following quantities by a standard conversion:



where and are normalized as and in the same manner.

Next, a correlation function can be defined as:



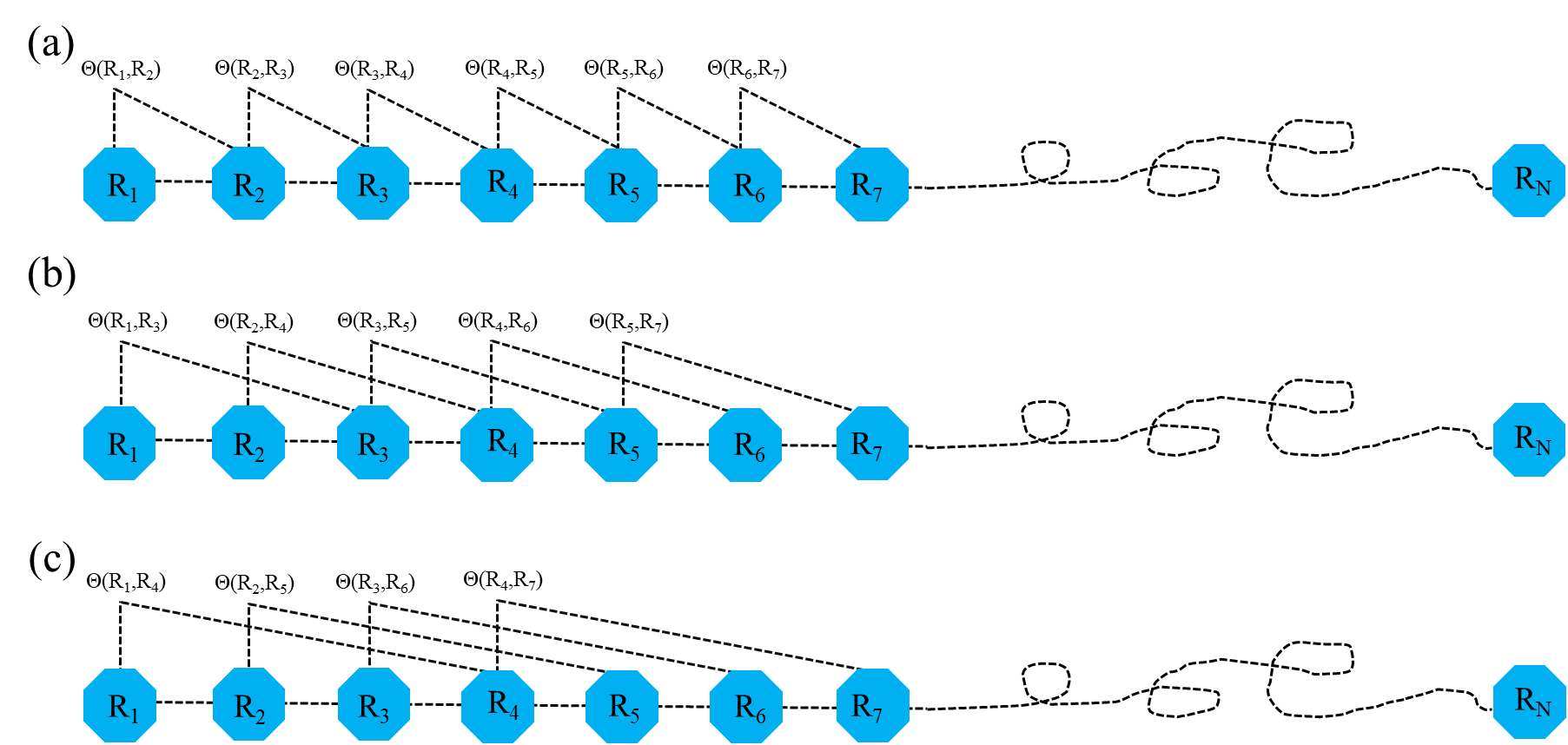
This correlation function is actually an averaged value for the three amino acid properties: hydrophobicity value, hydrophilicity value and side chain mass. Therefore, we can extend this definition of correlation function for one amino acid property or for a set of *n* amino acid properties.

For one amino acid property, the correlation can be defined as:



where *H(Ri)* is the amino acid property of amino acid *Ri* after standardization.

An illustrated example of the correlation function is provided in the following **Figure S5**.



**Figure S5**. A schematic drawing to show (a) the first-tier, (b) the second-tier, and (3) the third-tier sequence order correlation mode along a protein sequence. (a) reflects the coupling mode between all the most adjacent residues, (b) shows the coupling between the adjacent plus one residues, and (c) shows the coupling between the adjacent plus two residues. This figure is adapted from (Chou, 2001) for illustration purposes.

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



where *Hk(Ri)* is the *k*-th property in the amino acid property set for amino acid *Ri*.

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







…



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

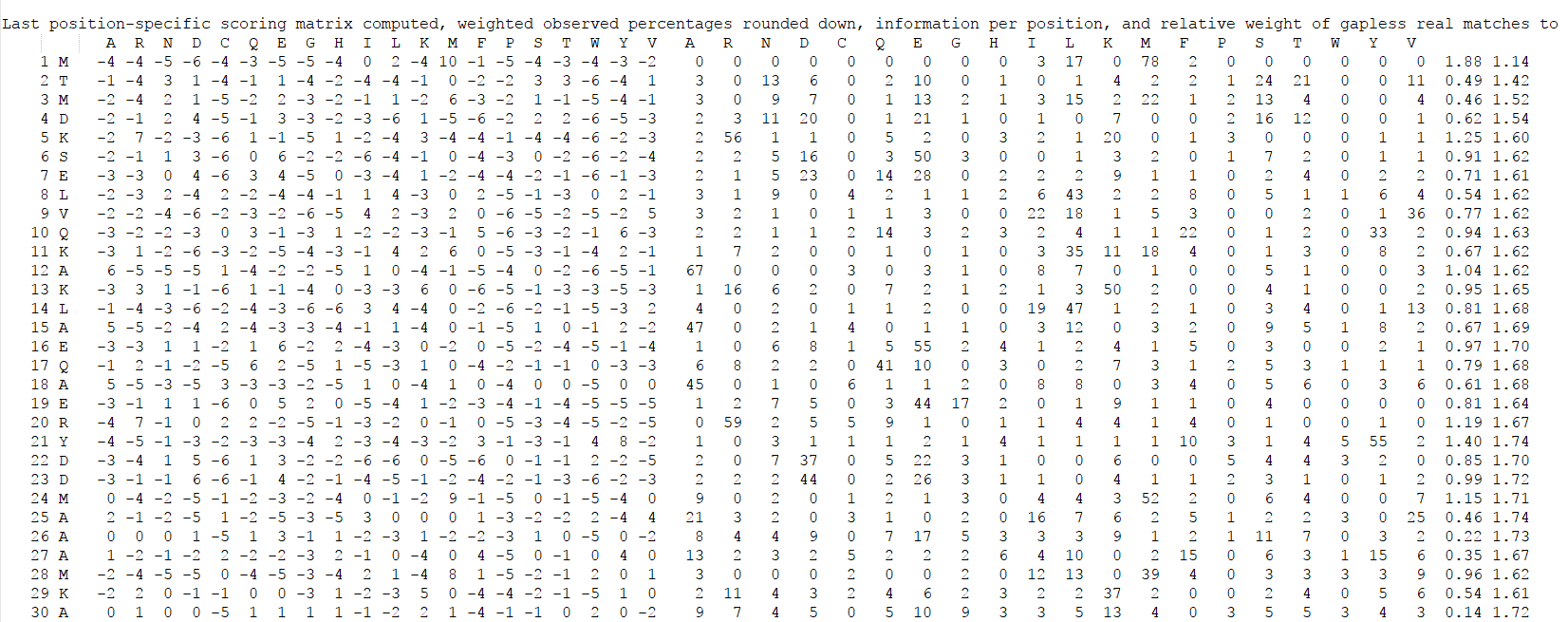




where *w* is the weighting factor for the sequence-order effect and is set to *w* = 0.05 in *BBSdb* as suggested by Chou *et al*. (Chou, 2001).

## 2.19 PSSM profile (PSSM)

This feature descriptor (Cai, et al., 2012; Radivojac, et al., 2010) is extracted from the Position-Specific Scoring Matrix (PSSM) profile. An illustrated example of PSSM profile is provided in **Figure S6**. The PSSM profile can be obtained by running PSI-BLAST (Altschul, et al., 1997) against the uniref 50 database. The PSSM descriptor is usually applied to encode the peptides with equal length. Each amino acid in the peptide is represented by a 20-dimensional vector.



**Figure S6**. Example of a PSSM profile.

## 2.20 AAindex (AAINDEX)

Physicochemical properties of amino acids are the most intuitive features for representing biochemical reactions and have been extensively applied in bioinformatics research. The amino acid indices (AAindex) database (Kawashima, et al., 2008) collects many published indices representing physicochemical properties of amino acids. For each physicochemical property, there is a set of 20 numerical values for all amino acids. Currently, 544 physicochemical properties can be retrieved from the AAindex database. After removing physicochemical properties with value 'NA' for any of the amino acids, 531 physicochemical properties were left. In contrast to the residue-based encoding methods of amino acid identity and evolutionary information, a vector of 531 mean values is used to represent a sample for various window sizes. The AAINDEX descriptor (Tung and Ho, 2008) can be applied to encode peptides of equal length.

## 2.21 aac\_dpc

This descriptor is a combination of AAC and the extension of dpc, measuring AAC and dpc information. In this descriptor, there are a total of 420 features. The first 20 features are the occurrence frequencies of 20 different amino acids extracted from AAC. The other 400 features are the dipeptide composition.

## 2.22 aac\_tpc

This descriptor is a combination of AAC and the extension of tpc, measuring AAC and tpc information.

**Acknowledgements**

The work was supported by grants from the NSFC (81871636 to EL).

**References**

Bhasin, M. and Raghava, G.P. (2004) Classification of nuclear receptors based on amino acid composition and dipeptide composition. *J Biol Chem*, 279, 23262-23266.

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003) Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993-1022.

Cai, C.Z.*, et al.* (2003) SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res*, 31, 3692-3697.

Cai, C.Z.*, et al.* (2004) Enzyme family classification by support vector machines. *Proteins*, 55, 66-76.

Cannata, N.*, et al.* (2002) Simplifying amino acid alphabets by means of a branch and bound algorithm and substitution matrices. *Bioinformatics*, 18, 1102-1108.

Chen, K.*, et al.* (2009) Prediction of integral membrane protein type by collocated hydrophobic amino acid pairs. *J Comput Chem*, 30, 163-172.

Chen, K., Kurgan, L. and Rahbari, M. (2007a) Prediction of protein crystallization using collocation of amino acid pairs. *Biochem Biophys Res Commun*, 355, 764-769.

Chen, K., Kurgan, L.A. and Ruan, J. (2007b) Prediction of flexible/rigid regions from protein sequences using k-spaced amino acid pairs. *BMC Struct Biol*, 7, 25.

Chen, K., Kurgan, L.A. and Ruan, J. (2008) Prediction of protein structural class using novel evolutionary collocation-based sequence representation. *J Comput Chem*, 29, 1596-1604.

Chen, X.*, et al.* (2013a) Incorporating key position and amino acid residue features to identify general and species-specific Ubiquitin conjugation sites. *Bioinformatics*, 29, 1614-1622.

Chen, Y.Z.*, et al.* (2012) SUMOhydro: a novel method for the prediction of sumoylation sites based on hydrophobic properties. *PLoS One*, 7, e39195.

Chen, Z.*, et al.* (2011) Prediction of ubiquitination sites by using the composition of k-spaced amino acid pairs. *PLoS One*, 6, e22930.

Chen, Z.*, et al.* (2013b) hCKSAAP\_UbSite: improved prediction of human ubiquitination sites by exploiting amino acid pattern and properties. *Biochim Biophys Acta*, 1834, 1461-1467.

Chen, Z.*, et al.* (2015) Towards more accurate prediction of ubiquitination sites: a comprehensive review of current methods, tools and features. *Brief Bioinform*, 16, 640-657.

Cheng, Y.Z. (1995) Mean Shift, Mode Seeking, and Clustering. *Ieee T Pattern Anal*, 17, 790-799.

Chou, K.C. (2000) Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. *Biochem Biophys Res Commun*, 278, 477-483.

Chou, K.C. (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins*, 43, 246-255.

Chou, K.C. (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics*, 21, 10-19.

Dubchak, I.*, et al.* (1995) Prediction of protein folding class using global description of amino acid sequence. *Proc Natl Acad Sci U S A*, 92, 8700-8704.

Dubchak, I.*, et al.* (1999) Recognition of a protein fold in the context of the Structural Classification of Proteins (SCOP) classification. *Proteins*, 35, 401-407.

Faraggi, E.*, et al.* (2009) Predicting continuous local structure and the effect of its substitution for secondary structure in fragment-free protein structure prediction. *Structure*, 17, 1515-1527.

Feng, Z.P. and Zhang, C.T. (2000) Prediction of membrane protein types based on the hydrophobic index of amino acids. *J Protein Chem*, 19, 269-275.

Frey, B.J. and Dueck, D. (2007) Clustering by passing messages between data points. *Science*, 315, 972-976.

Grantham, R. (1974) Amino acid difference formula to help explain protein evolution. *Science*, 185, 862-864.

Han, L.Y.*, et al.* (2004) Prediction of RNA-binding proteins from primary sequence by a support vector machine approach. *RNA*, 10, 355-368.

Heffernan, R.*, et al.* (2016) Highly accurate sequence-based prediction of half-sphere exposures of amino acid residues in proteins. *Bioinformatics*, 32, 843-849.

Heffernan, R.*, et al.* (2015) Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. *Sci Rep*, 5, 11476.

Horne, D.S. (1988) Prediction of protein helix content from an autocorrelation analysis of sequence hydrophobicities. *Biopolymers*, 27, 451-477.

Jain, A.K. (2010) Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31, 651-666.

Jain, A.K., Murty, M.N. and Flynn, P.J. (1999) Data clustering: A review. *Acm Comput Surv*, 31, 264-323.

Jones, D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*, 292, 195-202.

Kawashima, S.*, et al.* (2008) AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res*, 36, D202-205.

Kosiol, C., Goldman, N. and Buttimore, N.H. (2004) A new criterion and method for amino acid classification. *J Theor Biol*, 228, 97-106.

Lee, T.Y.*, et al.* (2011a) Incorporating distant sequence features and radial basis function networks to identify ubiquitin conjugation sites. *PLoS One*, 6, e17331.

Lee, T.Y.*, et al.* (2011b) Exploiting maximal dependence decomposition to identify conserved motifs from a group of aligned signal sequences. *Bioinformatics*, 27, 1780-1787.

Li, J. and Wang, W. (2007) Grouping of amino acids and recognition of protein structurally conserved regions by reduced alphabets of amino acids. *Sci China C Life Sci*, 50, 392-402.

Li, T.*, et al.* (2003) Reduction of protein sequence complexity by residue grouping. *Protein Eng*, 16, 323-330.

Liang, Y., Liu, S. and Zhang, S. (2015) Prediction of Protein Structural Classes for Low-Similarity Sequences Based on Consensus Sequence and Segmented PSSM. *Comput Math Methods Med*, 2015, 370756.

Lin, Z. and Pan, X.M. (2001) Accurate prediction of protein secondary structural content. *J Protein Chem*, 20, 217-220.

Liu, B.*, et al.* (2015a) Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy. *J Theor Biol*, 385, 153-159.

Liu, B.*, et al.* (2015b) 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*, 31, 1307-1309.

Liu, B.*, et al.* (2015c) Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res*, 43, W65-71.

Liu, X.*, et al.* (2002) Simplified amino acid alphabets based on deviation of conditional probability from random background. *Phys Rev E Stat Nonlin Soft Matter Phys*, 66, 021906.

M. Stigler, S. Francis Galton's Account of the Invention of Correlation. 1989.

Maaten, L.V.D. (2014) Accelerating t-SNE using tree-based algorithms. *J. Mach. Learn. Res.*, 15, 3221-3245.

Melo, F. and Marti-Renom, M.A. (2006) Accuracy of sequence alignment and fold assessment using reduced amino acid alphabets. *Proteins*, 63, 986-995.

Needleman, S.B. and Wunsch, C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48, 443-453.

Obradovic, Z.*, et al.* (2005) Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins*, 61 Suppl 7, 176-182.

Ogul, H. and Mumcuoglu, E.U. (2007) A discriminative method for remote homology detection based on n-peptide compositions with reduced amino acid alphabets. *Biosystems*, 87, 75-81.

Pearson, K. (1901) LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2, 559-572.

Peng, H.C., Long, F.H. and Ding, C. (2005) Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *Ieee T Pattern Anal*, 27, 1226-1238.

Peng, K.*, et al.* (2006) Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics*, 7, 208.

Peterson, E.L.*, et al.* (2009) Reduced amino acid alphabets exhibit an improved sensitivity and selectivity in fold assignment. *Bioinformatics*, 25, 1356-1362.

Radivojac, P.*, et al.* (2010) Identification, analysis, and prediction of protein ubiquitination sites. *Proteins*, 78, 365-380.

Rakshit, S. and Ananthasuresh, G.K. (2008) An amino acid map of inter-residue contact energies using metric multi-dimensional scaling. *J Theor Biol*, 250, 291-297.

Rokach, L. and Maimon, O. Clustering Methods. In: Maimon, O. and Rokach, L., editors, *Data Mining and Knowledge Discovery Handbook*. Boston, MA: Springer US; 2005. p. 321-352.

Sandberg, M.*, et al.* (1998) New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *J Med Chem*, 41, 2481-2491.

Saravanan, V. and Gautham, N. (2015) Harnessing Computational Biology for Exact Linear B-Cell Epitope Prediction: A Novel Amino Acid Composition-Based Feature Descriptor. *OMICS*, 19, 648-658.

Schneider, G. and Wrede, P. (1994) The rational design of amino acid sequences by artificial neural networks and simulated molecular evolution: de novo design of an idealized leader peptidase cleavage site. *Biophys J*, 66, 335-344.

Shen, J.*, et al.* (2007) Predicting protein-protein interactions based only on sequences information. *Proc Natl Acad Sci U S A*, 104, 4337-4341.

Sokal, R.R. and Thomson, B.A. (2006) Population structure inferred by local spatial autocorrelation: an example from an Amerindian tribal population. *Am J Phys Anthropol*, 129, 121-131.

Susko, E. and Roger, A.J. (2007) On reduced amino acid alphabets for phylogenetic inference. *Mol Biol Evol*, 24, 2139-2150.

Tomii, K. and Kanehisa, M. (1996) Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Eng*, 9, 27-36.

Tung, C.W. and Ho, S.Y. (2008) Computational identification of ubiquitylation sites from protein sequences. *BMC Bioinformatics*, 9, 310.

Wang, J. and Wang, W. (1999) A computational approach to simplifying the protein folding alphabet. *Nat Struct Biol*, 6, 1033-1038.

Wang, R., Xu, Y. and Liu, B. (2016) Recombination spot identification Based on gapped k-mers. *Sci Rep*, 6, 23934.

Xiao, N.*, et al.* (2015) protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics*, 31, 1857-1859.

Zuo, Y.*, et al.* (2017) PseKRAAC: a flexible web server for generating pseudo K-tuple reduced amino acids composition. *Bioinformatics*, 33, 122-124.

Zuo, Y.C. and Li, Q.Z. (2010) Using K-minimum increment of diversity to predict secretory proteins of malaria parasite based on groupings of amino acids. *Amino Acids*, 38, 859-867.