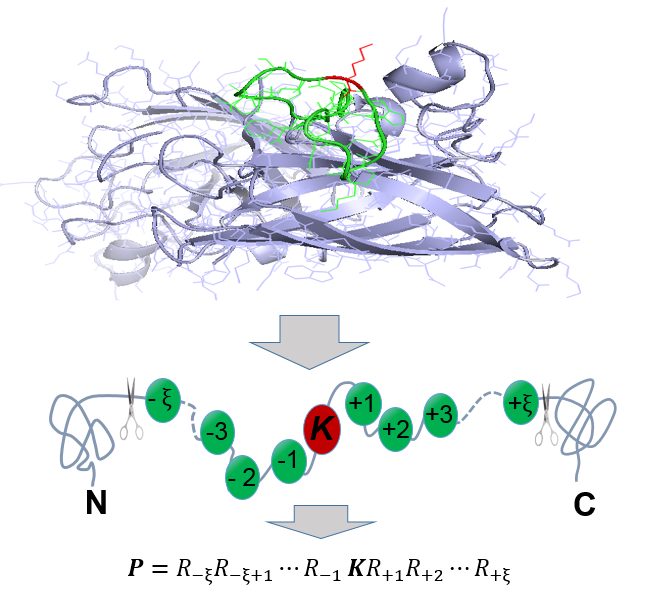
# **Peptide represent**



According to the general PseAAC, the peptide sequence ***P*** can be formulated as

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

# **Pseudo Amino Acid Composition of Chou (PAAC) RSTPAA\_1**

**(可计算整个序列的，也可用来计算片断的，但主要是计算蛋白质序列，计算片断的结果不好，参数调节是为了保证在同一数量级，常取0.005)**

In developing a statistical method for predicting the attribute of peptides in proteins, the most important procedures was to formulate the peptide samples with an effective mathematical expression that could truly reflect the intrinsic correlation with the desired target. To realize this, various feature vectors were proposed to express peptides by extracting their different features into the pseudo amino acid composition (PseAAC).

Described as {Chou, 2001 #2479;Chou, 2005 #2432}, the typical PseAA composition of the protein of Eq. (1) can be formulated as

(2)

where the 20+ components are given by .

(3)

where is the weight factor and is the tier correlation factor, which reflects the sequence order correlation between all of the most contiguous residues as formulated by

(4)

with

(5)

where , , and are the hydrophobicity value, hydrophilicity value and side chain mass for the amino acid respectively. Note that before substituting the values of hydrophobicity, hydrophilicity, and side chain mass, they all are subjected to a standard conversion, as described by the following equation:

(6)

where the symbols and are the original hydrophobicity and hydrophilicity values for , and is the side chain mass for . In other words, , , . And the symbol < > means taking the average of the quantity therein over 20 native amino acids, and SD means the corresponding standard deviation. The also is 5 in this study.

# **Grey information (greyPsePssm\_seq)**

**在通过blast得到PSSM矩阵的基础上计算灰成份，n=1时得60维，n=2时得80，blastpssm\_seq.m是我编的通过matlab调用blast程序，在调用时要注意安装目录，路径和环境变量的设置。**

According to {Schaffer, 2001 #1353}, the sequence evolution information of protein with amino acid residues can be expressed by a matrix, as given by

|  |  |  |
| --- | --- | --- |
|  |  | (7) |

where represents the original score of amino acid residue in the *i*-th ( sequential position of the protein that is being changed to amino acid type *j* during the evolution process. Here, the numerical codes 1, 2, …, 20 are used to denote the 20 native amino acid types according to the alphabetical order of their single character codes . The scores in **Eq.1** were generated by using PSI-BLAST to search the UniProtKB/Swiss-Prot database through three iterations with 0.001 as the -value cutoff for multiple sequence alignment against the sequence of the protein . To make every element in **Eq.1** within the range of , a conversion was performed through the standard sigmoid function to make it become

|  |  |  |
| --- | --- | --- |
|  |  | (8) |

where

|  |  |  |
| --- | --- | --- |
|  |  | (9) |

Now, we can use the grey system theory {Deng, 1989 #2686} to extract useful information from **Eq.9** to define the componentsof **Eq.1**. Why do we choose such theory in this study? As is known, if the information of a system investigated is fully known, it is called a “white system”; if completely unknown, a “black system”; if partially known, a “grey system”. Although we have some reliable experimental information for the current system, it is extremely complicated, lacking sufficient information and needing to process some uncertain information. Accordingly, it is a typical grey system, for which the model based on the grey system theory is particularly effective.

According to {Lin, 2011 #2539}, we can extract the following information from the *j*-th column of **Eq.8**

|  |  |  |
| --- | --- | --- |
|  | = | (10) |

where

|  |  |  |
| --- | --- | --- |
|  |  | (11) |

and

|  |  |  |
| --- | --- | --- |
|  |  | (12) |

Therefore, when using the grey model approach to extract the protein sequence evolution information via the PSSM**,** we can extract a total of quantities. Thus, **Eq.1** canbe quantitatively converted to

|  |  |  |
| --- | --- | --- |
|  |  | (13) |

where

|  |  |  |
| --- | --- | --- |
|  | (*j*=1,2, | (14) |

where is the occurrence frequency of the *j*-th amino acid in the protein concerned, and , , and are the weight factors, which were all set to 1 in the current study.

# **Fragment Disorder（文件夹VSL2DIR中的getdisorder.m）**

**先下载VSL2（文件夹中有），然后配置电脑，注意安装目录，路径和环境变量的设置，参数只能选奇数，且不能超过序列的长度，否则不出结果。**

Although, under physiological conditions, disordered regions in proteins do not have fixed three-dimensional structures, its functional importance has been increasingly recognized and was used to predict protein structures and functions for that they play various roles in signaling and regulation by multiple binding of proteins and high-specificity low affinity interactions.

On the basis of disorder score calculated by VSL2, this study encoded the disorder status of each amino acid in the protein sequence. The disorder scores of lysine site and its surrounding amino acids formed the part features of which can be listed as:

(15)

Where is the disorder score of the residue of protein sequence fragment , . The disorder score value for each amino acid was calculated by VSL2.

# **4)Fragment Grey info（和（2）中一样，只是选取PSSM矩阵中的子阵）**

The Grey information of peptide. Feature 2 is the information of the whole sequence.

# **5)AA index（GetSeqAAindex.m）直接调用它，可以更改AAindex矩阵，例如可以用林老师提到的那三个，每个氨基酸可得对应列数那么维特征，所以总特征维数是20\*index的个数**

Since that the structure and function of proteins are largely dependent on the composition of various properties of each of the 20 amino acids, amino acid physicochemical properties have been successfully used in predicting AA site. With the assistance of studies in multivariate statistical analyses, AAIndex has become a well-known database of amino acids’ biochemical and physicochemical properties. In this database, there are five multidimensional patterns of attribute covariation comprise of polarity (AAFactor 1), secondary structure (AAFactor 2), molecular volume (AAFactor 3), codon diversity (AAFactor 4), and electrostatic charge (AAFactor 5). Here, these five amino acid factors will be called amino acid factors and used to encode each amino acid in the lysine fragment. For the protein sequence fragment, the components, may served as a part of the feature vector in Eq.3, are given by:

(16)

Where are the five amino acid factor scores of the residue of protein sequence fragment , . The score values for each amino acid on the five factors can be found in reference.

# **6)EBAG（直接调用EBAG）**

In the method, the 20 amino acid residues were divided into four different classes according to their physicochemical property: the hydrophobic group = {A, F,G, I, L, M, P, V, W}, the polar group = {C, N, Q, S, T, Y}, the acidic group = {D, E}, and the basic group ={H, K, R}.Given a protein sequence p fragment with 2 + 1 amino acid residues, we used the above classification to transform it into four binary sequences as follows:

(17)

Then, the peptide listed in **Eq.1** can be formed as:

(18)

# **7)PWAA（直接调用PWAA）**

Position weight amino acid composition can reveals the sequence-order information around some PTM sites, and it had been used in identifying viral AA sites as well as methylation sites. To reflect this kind of information, the PseAAC of Eq.1 was defined by

|  |  |  |
| --- | --- | --- |
|  |  | (19) |

where

|  |  |  |
| --- | --- | --- |
|  |  | (20) |

where is the same as in Eq.1, and

|  |  |  |
| --- | --- | --- |
|  |  | (21) |

**8) MT Probability（调用extr\_prjft2\_Qiu.m，此特征有点过拟合，只有样本足够多才稳定，提广性不强，不建议使用）**

According to the general PseAAC {Chou, 2011 #2402}, the peptide sequence of **Eq.1** can be formulated as

|  |  |  |
| --- | --- | --- |
|  |  | (22) |

where

|  |  |  |
| --- | --- | --- |
|  |  | (23) |

and

|  |  |  |
| --- | --- | --- |
|  |  | (24) |

In **Eq.23** is the conditional probability of amino acid occurring at the left 1st position (see **Eq.1**) given that its closest right neighbor is , is the conditional probability of amino acid occurring at the left 2nd position given that its closest right neighbor is , and so forth. Note that in **Eq.23**, only and are of non-conditional probability since the right neighbor of and the left neighbor of are always .

**9)knn score(ExtrKNN及KNNScore相关命名的函数（如**KNNScore\_Blo

# **是为了计算距离的），有些是子函数，注意变量的输入格式。我编此程序时是先提取所用样本的KNNScore，所以输出是一个矩阵，在预测中如果要用到的话，需要改编它。**

Local sequence clusters often exist around phosphorylation sites because the same post-translational modification family usually share similar patterns in local sequences. As a better choice for depicting the character, K nearest neighbor score is the ratio between the numbers of positive and negative sites among the pre-defined number of neighbors of a given peptide. It had also used for phosphorylation prediction. The neighborhood of a peptide is defined as a certain percentage of the training dataset ranked by the pair-wise similarity between the target peptide and the peptides in the dataset. In this study, the sequence similarity calculation was based on the BLOSUM62 matrix; the neighborhood was set to 0.25, 0.5,1,2, and 4% of the training dataset, respectively; and the flanking sequence size was set to 13 amino acids. Then, the peptide listed in Eq.1 can be formed as:

(25)

where are the ratio between the numbers of positive and negative sites when the neighborhood was set to 0.25, 0.5, 1, 2, and 4% of the training dataset, respectively.