2 蛋白质的物理化学属性

福建师范大学福清分校

游文杰

蛋白质的物理化学特征，简称物化特征(Physicochemical Feature, PF)。它对识别不同结构、功能和相互作用的蛋白质或二肽来说有很大作用。2006年新加坡国立大学计算机科学学院生物信息学与药物设计团队开发了PROFEAT[1]网站，该网站可以对蛋白质的各种物化特征的详细计算。2011年他们联合四川大学化学院的老师对PROFEAT网站进行了更新[2]。目前PROFEAT网站可以计算11种描述蛋白质和肽类的物化特征，以下将对PROFEAT网站中的物化特征计算方式和原理进行介绍。

### 2.1 基于氨基酸组成的特征

氨基酸组成(Amino acid Composition, AAC)通过计算每种氨基酸数量在序列总氨基酸数量中的占比来获得，氨基酸组成的计算公式为：



其中表示类型为的氨基酸数量，表示蛋白质序列的总长度，这样20种氨基酸会计算出20个描述值。

### 2.2 基于二肽组成的特征

二肽是通过-羧基和-氨基脱水缩合而形成的蛋白质片段，-羧基、-氨基均连接在除羧基那个碳原子外，其分子上的第一个碳原子上。它的分子中一般含有一个肽键。可以通过判断氨基酸种类、结构来判断其是否属于同一种二肽。氨基酸二肽组成(Dipeptide Composition, DC)的计算公式如下所示：



其中表示与类型的氨基酸中所含二肽的数量。这样总共会得到氨基酸的二肽组成20\*20共400个描述值。

### 2.3 基于组成、转换和分布的特征

蛋白质序列的参数向量有很多，其中包含“组成(C)”、“转换(T)”和“分布(D)”3种描述符。它们代表着沿蛋白质或二肽序列的特定结构或物理化学特征的氨基酸分布模式[3,4]。从七个方面计算这三个物化特征：疏水性、规范化范德华体积、极性、极化性、电荷、二级结构和溶剂可溶性。每一种特征均有3种特性：极性、中性和疏水性。即20种氨基酸依据这个维度被分成3组。

我们以疏水性为例来计算这些特征，所有的氨基酸有3部分组成：极性、中性和疏水性。参数向量“组成”包含3个值：蛋白质中极性残基全局组成、中性残基全局组成和疏水性残基全局组成。参数向量“转换”包含3个值：一个极性残基紧跟一个中性残基或一个中性残基紧跟一个极性残基的频率、一个极性残基紧跟一个疏水性残基或一个疏水性残基紧跟一个极性残基的频率、一个中性残基紧跟一个疏水性残基或一个疏水性残基紧跟一个中性残基的频率。参数向量“分布”的三组残基(极性、中性、疏水性)分别包含5个值：整个序列中给定的第一个残基所在的片段、序列25%所处的残基、序列50%所处的残基、序列75%所处的残基和序列100%所处的残基分布。因此，对于疏水性的性质来说，共有3种描述符，21个描述值。进而得到7种物化性质的147个描述值，这就是氨基酸的组成、转换和分布(Composition, transition and distribution, CTD)特征的维数。

由上文分析，现将CTD特征7种物化属性的详情总结如表2.2中。

**表2.2 蛋白质CTD特征7种物化属性及其参数向量维数**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 性质 | 参数向量维数 | | | 总维度 |
| 组成 | 转换 | 分布 |
| 疏水性 | 3 | 3 | 15 | 21 |
| 规范化范德华体积 | 3 | 3 | 15 | 21 |
| 极性 | 3 | 3 | 15 | 21 |
| 极化性 | 3 | 3 | 15 | 21 |
| 电荷 | 3 | 3 | 15 | 21 |
| 二级结构 | 3 | 3 | 15 | 21 |
| 溶剂可溶性 | 3 | 3 | 15 | 21 |

### 2.4 基于Moreau-Broto自相关指数的特征

蛋白质物化特征的自相关指数中有Moreau-Broto自相关指数、Geary自相关指数和Moran自相关指数3种。这里仅介绍效果较好的Moreau-Broto自相关指数(Moreau-Broto Autocorrelation, MB)。自相关指数描述了基于氨基酸分布的特定结构或物化特征上，两个物质之间的相关程度。每一种自相关指数都含有8个描述符，240个描述值。在计算指数前，它们均需要进行归一化处理，归一化公式为：



其中，为20种氨基酸特性的平均值，为方差。

Moreau-Broto自相关指数的计算公式如下，在识别跨膜蛋白类型和蛋白质二级结构[1]中经常使用，其识别准确率分别在82-94%和91-94%中间。



其中，是自相关指数的间隔，和表示在和位置的氨基酸特性。是序列的长度。

归一化的Moreau-Broto自相关指数定义为：



### 2.2.5 基于伪氨基酸组成的特征

根据维基百科中的介绍，伪氨基酸组成(Pseudo-amino acid Composition, PseAAC)是由Kuo-Chen Chou提出，它代表蛋白质样本，用于改善蛋白质亚细胞定位和膜蛋白类型识别。自从伪氨基酸组成的概念出现后，其数字表达模型也被广泛用于各种与蛋白质有关的识别中。2011年，PROFEAT网站的开发团队在新增了这个计算伪氨基酸特性的功能。在增加这个功能之前，这些功能的算法已经在识别蛋白质、蛋白质与蛋白质之间的关系、定量小分子的结构活性关系研究中，进行了广泛的测试，并在很多已出版的论文中得到使用[2]。 PseAAC弥补了AAC未考虑氨基酸顺序的缺陷，它不仅考虑每个蛋白质序列，同时考虑序列中氨基酸排列顺序的影响。PseAAC的描述符和描述值取决于所选氨基酸的特性和参数的取值，其定义如下所示，结果通过维数组表示：





其中，是矩阵的转置，PROFEAT网站默认为30。为权重因子，一般建议将其设置为0.05，表示20种氨基酸中第种氨基酸出现的次数。表示第等级的相关因子，它可以反映第等级中连续性最大的残基之间的序列顺序相关性，其计算公式为：



其中，表示氨基酸第个的氨基酸特性(标度值)，其计算值均经过公式进行归一化处理。是氨基酸所含特性的总数，比如氨基酸的特性包含亲水性值，疏水性值和侧边链值3个元素，则取3。

### 2.2.6 基于序列顺序的特征

所介绍氨基酸物化特性中最后一个特征是基于序列顺序的特征(Sequence-order, SO)，它代表着氨基酸的分布模式。其描述符来自Schneider-Wrede物化距离矩阵和Grantham化学距离矩阵[1]。序列顺序特征中有2种描述符，共160个描述值。第等级序列顺序耦合数的定义如下，耦合数字反映了所有连续残基第等级的耦合模式。



其中表示在和处两种氨基酸的特性(标度值)之差。

对于序列所有氨基酸而言，第一种类型的Quasi-sequence-order的数学描述为：



其中，表示20种氨基酸中第种氨基酸出现的次数，是个权重因子，一般设置。

第二种氨基酸类型的Quasi-sequence-order的数学描述为：



现将本文所使用的蛋白质物化特征中的向量特征、描述符以及描述值汇总如下表2.3所示：

**表2.3 所介绍的物化属性的相关信息**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 特征 | 向量特征 | 特征顺序 | 描述符 | 描述值 |
| 氨基酸组成 | 氨基酸组成 | 1 | 1 | 20 |
| 二肽组成 | 二肽组成 | 2 | 1 | 400 |
| 组成、转换、分布 | 组成 | 5 | 7 | 21 |
| 转换 | 5 | 7 | 21 |
| 分布 | 5 | 7 | 21 |
| Moreau-Broto自相关指数 | 归一化Moreau-Broto | 6 | 8 | 240 |
| 伪氨基酸组成 | 伪氨基酸组成 | 7 | 1 | 50 |
| 序列顺序 | 序列顺序 | 8 | 2 | 160 |

参考文献

[1] Li, Z. R., Lin, H. H., Han, L. Y., Jiang, L., Chen, X., & Chen, Y. Z. (2006). PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucleic Acids Research, 34(suppl\_2), W32-W37.

[2] Rao, H. B., Zhu, F., Yang, G. B., Li, Z. R., & Chen, Y. Z. (2011). Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucleic acids research, 39(suppl\_2), W385-W390.

[3] Semenov, Y. S., Popova, S. K., & Lebedev, M. P. (2009). Phase composition and distribution of alloying elements in the transition layer. Journal of applied mechanics and technical physics, 50(6), 1089-1094.

[4] McDermott, J. E., Bruillard, P., Overall, C. C., Gosink, L., & Lindemann, S. R. (2015). Prediction of multi-drug resistance transporters using a novel sequence analysis method. F1000Research, 4..