**巴豆酰化预测算法设计**

**数据预处理：**

数据去冗余：CD-HIT

简介：CD-HIT stands for Cluster Database at High Identity with Tolerance. The program (cd-hit) takes a fasta format sequence database as input and produces a set of 'non-redundant' (nr) representative sequences as output. In addition cd-hit outputs a cluster file, documenting the sequence 'groupies' for each nr sequence representative.

输入的文件是fasta格式的序列文件，通过序列比对聚类（Cluster）的方法去除冗除、相似的序列，最后输出一个非冗除（non-redundant，nr）的序列文件。 另外，还有一个序列比对的结果。

数据的筛选，清理，不空，去噪，归一化，标准化，降维，转换

加入滑动窗口：①穷举法得到长度②采用已有问题的经验，从信息熵的概念出发确定滑动窗口的大小

**特征提取和选择：**

考虑PSSM矩阵信息：对每个氨基酸用PSSM矩阵中对应的行向量作为编码

氨基酸序列正交编码：用20位二进制数表示，\_用00000000000000000000表示

基于氨基酸组成的编码（AAC） ：它将每个固定长度的肽段特征用２０种氨基酸的组成比例表示 ，构成 １ 个 ２０ 维度的向量 。 AAC 编码只关注每种氨基酸在肽段中的比例 ，不关注其所处的位置信息 。

N‐AAC 编码是在 AAC 编码的基础上 ，以赖氨酸为中心 ，通过选取不同长度的肽段（例如 11、21 、31 、41 个残基） ，分别统计其氨基酸组成比例 ，最后将不同长度下的氨基酸组成整合成 １ 个向量 。 相对于AAC ，N‐AAC 编码的优势在于 ，它不仅统计了整个序列窗口的氨基酸组成情况 ，同时也关注了泛素化位点周围多个局部的氨基酸组成信息

正交编码（BINARY） ：正交编码是 １ 种二进制的数字编码形式 ，它可以特异性地表征序列上各氨基酸的位置信息 。 每个氨基酸由 20 个二进制的编码（0 或 1）组成 ，根据序列长度 N 的不同 ，最终的编码维度也不一样 ，共有 20 × N 维 。 虽然正交编码所组成的向量很直观 ，但由于其变量只有 ０ 和 １ ，单调的变量往往不利于机器学习方法进行特征学习 。

基于 k 空格氨基酸对的编码（CKSAAP） ：CK‐SAAP 反映的是氨基酸序列周围的 motif 信息 。 近些年来 ，CKSAAP 被广泛应用于构建各翻译后修饰类型位点的预测中 ，获得了较好的评价 。对于 １ 个肽段 ，固定 k 值 ，将会有２０ 种类型的氨基酸构成的 400 个氨基酸对的组成形式 。 在实际应用中 ，k 往往取多组数值 。 本文 k 取 0 、

1 、2 、3 、4 、5 共 6 个数值 ，所以最终每个小鼠蛋白质序列泛素化位点周围特征窗口将由 ６ × ４００ ＝ ２ ４００ 个维度的向量表示 。

N‐CKSAAP 编码是在 CKSAAP 编码的基础上

根据氨基酸理化性质和结构特性 ，将 ２０ 种氨基酸分成 1０ 大类（表 １）之后 ，再考虑氨基酸对的组成比例的编码 。 一般来说 ，N‐CKSAAP 不仅达到了降维处理的目的 ，还从编码上考虑了不同氨基酸之间的理化性质相似性 。 同理 ，BINARY 编码也可演化出N‐BINARY 编码 。

分类名称 氨基酸类型

极性 N 、Q 、S 、D 、E 、C 、T 、K 、R 、H 、Y 、W

带正电荷 K 、H 、R

带负电荷 D 、E

带电荷 K 、H 、R 、D 、E

疏水性 A 、G 、C 、T 、I 、V 、L 、K 、H 、F 、Y 、W 、M

脂肪族 I 、V 、L

芳香族 F 、Y 、W 、H

较小 P 、N 、D 、T 、C 、G 、S 、V

极小 A 、S 、G 、C

脯氨酸 P