DOI:10.12113/202212005

基于支持向量机预测 C2H2 型锌指蛋白

刘　哲,李凤敏*
(内蒙古农业大学 理学院, 呼和浩特 010018)

摘　要:转录作为遗传信息传递的第一步,会受到多种转录因子的调控。转录因子(Transcription Factors, TF),是指能够在基因上游的特异核苷酸位点结合从而影响转录过程的蛋白质因子,锌指蛋白是数量最多的一类转录因子。由于锌指基序大多是不相同的,所以它们很可能结合不同的位点,行使多样化的调控功能。C2H2 型锌指蛋白是数量最多的一类锌指蛋白,本文构建 C2H2 型锌指蛋白数据集,提取了氨基酸单肽组分信息、平均化学位移、氨基酸二肽组分信息三类特征信息,利用支持向量机算法对锌指蛋白进行预测,在 Jackknife 检验下最高预测成功率为 87.86%。之后对氨基酸二肽组分信息特征参数进行不同方式的降维处理,降维后最高预测成功率为 90.21%。最后对三类特征信息进行融合,融合特征后最高预测成功率为 92.55%。对锌指蛋白进行预测,有助于更加深入地了解锌指蛋白的结构、功能和调控机理。

关键词:转录因子;锌指蛋白;特征信息;预测
中图分类号:Q61　　**文献标志码**:A　　**文章编号**:1672-5565(2024)02-140-08

Prediction of C2H2 zinc finger protein based on support vector machine

LIU Zhe, LI Fengmin*
(*College of Science*, *Inner Mongolia Agricultural University*, *Hohhot* 010018, *China*)

Abstract: The first step in the transmission of genetic information is transcription, transcription is regulated by a variety of transcription factors. Transcription factors can bind to specific nucleotide sites upstream of genes and then influence the transcription process. The category with the largest number of transcription factors is zinc finger protein. Because zinc finger motifs in zinc finger protein are different, so they can bind to different sites and perform different regulatory process. The category with the largest number of zinc finger protein is C2H2 zinc finger protein. In this paper, the data set of C2H2 zinc finger protein is established, and based on the three types of feature information including amino acid composition, auto-covariance average chemical shift and dipeptide composition. The zinc finger protein is predicted by using the algorithm of support vector machine, and the accuracy is 87.86% in Jackknife. After that, different methods are used to reduce the dimension of dipeptide composition, and the accuracy is 90.21% after dimension reduction. Finally, multi-feature information is used to predict, and the accuracy is 92.55%. Prediction of zinc finger protein in order to better understand the structure, function and regulation mechanism.

Keywords: Transcription factors; Zinc finger protein; Feature information; Prediction

中心法则由 Crick 于 1958 年提出,叙述了遗传信息的传递过程,是所有的细胞结构生物均需遵循的法则。遗传信息从 DNA 传递到 RNA 为转录过程,从 RNA 传递到蛋白质为翻译过程。转录过程是以双链 DNA 中的一条链作为模板,在 RNA 聚合酶的催化作用下合成 RNA[1]。转录是基因表达的调

收稿日期:2022-12-14;**修回日期**:2023-02-28;**网络首发日期**:2023-06-02.
网络首发地址:https://kns.cnki.net/kcms2/detail/23.1513.Q.20230601.1515.012.html.
基金项目:内蒙古自治区自然科学基金项目(No.2019MS03015).
*通信作者:李凤敏,女,教授,博导,研究方向:理论生物物理. E-mail: lfmbs@126.com.
引用格式:刘哲,李凤敏.基于支持向量机预测 C2H2 型锌指蛋白[J].生物信息学,2024,22(2):140-147.
LIU Zhe, LI Fengmin. Prediction of C2H2 zinc finger protein based on support vector machine[J]. Chinese Journal of Bioinformatics, 2024, 22(2): 140-147.

节控制和生物遗传信息传递中的一个重要环节,而且过程比较复杂,通常需要转录因子的协同帮助才可以完成。转录因子在分子生物学中具体是指一种能够结合在基因上游的特定位置,起正调节或负调节的蛋白质因子。正调节可以激活转录过程,提高转录效率,促进基因表达,在调节过程中起到正向推动作用;反之则为负调节,阻碍转录过程,降低转录效率,抑制基因表达,在调节过程中起到反向抑制作用,甚至会直接抑制转录过程的启动。在转录的起始过程,转录因子便可与 RNA 聚合酶形成一种复合体蛋白质,两者共同在转录的起始过程起作用。根据转录因子的作用特点不同,可以将其分为普遍转录因子和组织细胞特异性转录因子[2]。

1983 年,人类在非洲爪蟾卵母细胞的转录因子中第一次发现锌指蛋白,作为转录因子其中的一类,经常会在 DNA 结合蛋白中出现,能够对特定的碱基序列起识别作用。锌指蛋白是真核生物基因组中最广泛分布的蛋白质之一,在人类基因组里有大约 1%的序列中含有锌指蛋白。锌指蛋白的具体结构是由氨基酸环和锌离子组成,形状类似于人类的手指,故称这样的结构为锌指结构。锌指蛋白是一类含有锌指结构且必须与锌离子结合配位才能发挥作用的蛋白质[3-5]。

锌指蛋白的具体空间结构,是由半胱氨酸(Cys)残基和组氨酸(His)残基两者根据不同的数量和方式围绕锌离子所构成。根据具体空间结构的不同,Krishna 等[6]把锌指蛋白分为 C2H2 like, Gag knuckle, Treble clef, Zinc ribbon, Zn2/Cys6, TAZ2 domain like, Short zinc binding loops 和 Metallothionein,共计 8 种不同的折叠群(Fold group)。每一种折叠群都包含很多类型,但大部分类型的锌指蛋白都属于前三类折叠群[7]。具体来说,存在范围最广且生物功能最为重要的锌指蛋白有如下几种类型:C2H2 型锌指蛋白、RING 型锌指蛋白、PHD 型锌指蛋白以及 LIM 型锌指蛋白。C2H2 型锌指蛋白通常是指包含[C-x-C-x-H-x-H]结构域的锌指蛋白,作用机理是与 DNA 结合为结合蛋白,进而起到促进或抑制标靶基因表达的作用。C2H2 型锌指蛋白约占目前已知全部锌指蛋白的 45%,是锌指蛋白中数量最多的一类[8]。RING 型锌指蛋白通常是指包含[C-x-C-x-C-x-H-x-C-x-C-x-C-x-C]结构域的锌指蛋白,可以对转录过程中锌指蛋白的作用对象和作用活性产生影响[9]。PHD 型锌指蛋白通常是指包含[C-x-C-x-C-x-C-x-H-x-C-x-C-x-C]结构域的锌指蛋白,能够对染色质起到重塑作用,修饰表观遗传,调控识别核小体。LIM 型锌指蛋白通常是指包含[C-x-C-x-H-x-C-x-C-x-C-x-C-x-(C.H,D)]结构域的锌指蛋白,能够在肌动蛋白锚定过程中起到重要作用,也可对细胞骨架之间的相互作用产生影响。除上所述之外,锌指蛋白在基因翻译、mRNA 运输、细胞骨架组装、上皮细胞发育、细胞粘附、蛋白质折叠以及锌离子感应等方面均发挥重要的作用。

本文基于最新版本 UniProt 数据库建立了锌指蛋白家族的数据集,包含目前已知且通过实验验证的所有类型。最终选取 C2H2 型锌指蛋白和数量相近的非锌指蛋白共同构成预测数据集。通过输入单个特征信息和融合特征信息进行预测实验,最终结果是氨基酸二肽组分信息的预测成功率最高,达到 87.86%。在此基础上,对氨基酸二肽组分信息使用多种方式的降维,降维处理后利用支持向量机算法进行预测,最终得到最高预测成功率为 90.21%。预测结果表明:特征信息的降维处理对预测结果有较好的提升。在此基础上进行融合三类特征信息,利用支持向量机算法进行预测,融合特征信息最高预测成功率为 92.55%,预测结果表明: 特征信息的融合对预测结果有一定的提升作用。

1　材料与方法

1.1　数据集的构建

利用机器学习方法对蛋白质进行预测是如今生物信息学研究的重点方法,建立一个客观的、有代表性的数据集对于后续工作至关重要。UniProt (https://legacy.uniprot.org/)是一个信息丰富、资源广泛的蛋白质数据库,由 Swiss-Prot,TrEMBL 和 PIR-PSD 三大数据库整合而成。本文基于 UniProt 数据库,严格按照以下标准构建了锌指蛋白数据集:

1)在 UniProt 数据库,高级检索中输入关键词“zinc”及“finger”所得蛋白质序列,包含经过实验验证的序列 2 597 条,未经过实验验证的序列 469 483 条。

2)在高级检索选项“advanced”里添加限制条件“AND Reviewed”,选择经过实验验证的蛋白质序列,共得到序列 2 597 条。

3)去除“By similarity”,“Probably”等含糊不确定的关键字后,共得到序列2 597 条。

4)在“Sequence ”信息中选择蛋白质序列完整,去除片段序列后,共得到序列 2 479 条。

5)去除含有不确定氨基酸 Z,X,B,O,J,U 的蛋白质序列后,共得到序列 2 227 条。

6)根据蛋白质序列名称代码删除重复的蛋白质序列后,共得到序列 2 183 条。

由以上步骤最终得到各类锌指蛋白序列共2 183条,由于种类数量分布不平衡,C2H2 型锌指蛋白序列共有 1 601 条,占到了总数的 73.34%,其它类型只占有极少数,故将数据集中的 C2H2 型锌指蛋白序列筛选出来。为避免存在同源性误差,采用 CD-HIT 程序对数据集进行相似比对,序列相似性阈值设定为 25%。完成上述步骤后,共得到 363 条 C2H2 型锌指蛋白质序列。在 UniProt 数据库,高级检索中输入不含关键词“zinc”及“finger”的蛋白质序列,从中随机挑选 679 条非锌指蛋白序列。阈值设定为 25%,经过 CD-HIT 软件对数据集进行相似比对,最终获得 362 条非锌指蛋白序列。最终预测数据集由 363 条锌指蛋白序列和 362 条非锌指蛋白序列,共计 725 条蛋白质序列构成,详见表 1。

表 1 C2H2 型锌指蛋白及非锌指蛋白数据集中序列数目

Table 1 Number of sequences in the C2H2 zinc finger protein and non-zinc finger protein dataset

数据集	序列数量
C2H2 型锌指蛋白	363
非锌指蛋白	362
总计	725

1.2 特征参数的选取

1.2.1 氨基酸单肽组分信息

人体内有很多种类的蛋白质,它们性质有所不同,功能也各有差异,但组成成分都是 20 种氨基酸,分别为:A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W 和 Y。本文选取氨基酸单肽组分信息(Amino acid composition, AAC)作为特征参数之一,氨基酸单肽组分信息是蛋白质研究中最普遍应用的一类特征信息,具体是计算蛋白质序列中 20 种氨基酸出现的频率,也可以用 20 维特征向量表示[10]。定义为

$$P = [x_1, x_2, x_3, \cdots, x_i, \cdots, x_{20}] \tag{1}$$

$$x_i = \frac{m_i}{L} \tag{2}$$

上述公式中:L 指的是蛋白质序列长度,m_i指的是蛋白质序列中第 i 个氨基酸出现的次数。

1.2.2 氨基酸二肽组分信息

由蛋白质序列的组成可知,氨基酸对共有 20×20 等于 400 种组合。本文选取氨基酸二肽组分信息(Dipeptide composition,DC)作为特征参数之一,具体是计算两个相邻氨基酸残基的出现频率,也可以用 400 维特征向量表示[11]。氨基酸二肽组分信息相较于氨基酸单肽组分信息的优点是考虑了蛋白质序列中的氨基酸组合顺序。具体定义为

$$P = [f_1, f_2, f_3, \cdots, f_i, \cdots, f_{400}] \tag{3}$$

$$f_i = \frac{n_i}{L - 1} \tag{4}$$

上述公式中:L 指的是蛋白质序列长度,n_i指的是蛋白质序列中第 i 个二肽出现的次数。

1.2.3 平均化学位移

在对蛋白质的研究中,核磁共振技术发挥着重要作用,它可以对蛋白质在多个时间尺度上内部运动的相关信息进行跟踪记录。由于质子存在化学环境敏感性,所以在核磁共振技术手段中,质子会因为处于不同的化学环境而导致受到不同的磁场作用并产生不同的吸收频率。平均化学位移(Auto-covariance average chemical shift, acACS)即可以用各类不同质子相对于标准值的共振频率表示[12]。具体研究表明,蛋白质的平均化学位移与其二级结构有很大的相关性[13]。本文通过将蛋白质序列提交到 PSIPRED(PSIPRED workbench(ucl.ac.uk))网站获得数据集中蛋白质的二级结构,然后利用 python 程序将化学位移的结果提取出来。具体过程可表示为

$$ACS_i^k(j) = \frac{1}{N}\sum w_i^k(j) \tag{5}$$

上述公式中:i 指的是四种骨架原子{ ^{15}N, $^{13}C_\alpha$, $^{1}H_\alpha$, $^{1}H_N$ },k 指的是蛋白质二级结构的类别(H、E、C),j 指的是 20 种氨基酸,N 指的是蛋白质序列中氨基酸的个数。

对于蛋白质 P,序列中的每个氨基酸都被其平均化学位移取代,可以表示为

$$P = [A_1^i, A_2^i, A_3^i, \cdots, A_L^i]\ (i = {}^{15}N, {}^{13}C_\alpha, {}^{1}H_\alpha, {}^{1}H_N) \tag{6}$$

该化学位移的自相关协方差可表示为

$$\varphi_i^\lambda = \frac{1}{L-\lambda}\sum_{k=1}^{L-\lambda}[A_k^i - A_{k+\lambda}^i]^2$$
$$(i = {}^{15}N, {}^{13}C_\alpha, {}^{1}H_\alpha, {}^{1}H_N; \lambda < L) \tag{7}$$

最终 P_{acACS} 可表示为

$$P_{acACS} = [\varphi_i^0, \varphi_i^1, \varphi_i^2 \cdots \varphi_i^\lambda]$$
$$(i = {}^{15}N, {}^{13}C_\alpha, {}^{1}H_\alpha, {}^{1}H_N) \tag{8}$$

上述公式中:λ 指的是相关常数,L 指的是蛋白质序列的总长度。对于不同的蛋白质,为了预测可以取得更好的结果,应选择最优组合的骨架原子组合及 λ 值。

1.3 预测算法

选用支持向量机(Support vector machine, SVM)作为预测算法,1955 年 Vapnik 等[14]最早提出支持向量机算法,它是一种基于统计学理论的机器学习方法,后来也广泛用于蛋白质结构预测和功能预测。尽可能的利用最大间隔思想去降低分类器的

置信风险,这是支持向量机的核心思想。将数据从低维向量映射到高维向量,使得结构风险达到最小化,使正集和负集之间的距离最大化。具体原理见图 1。

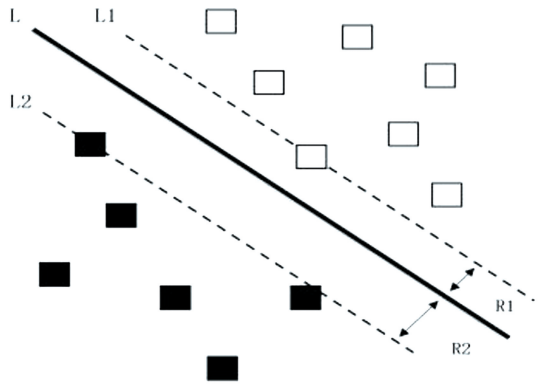

图 1　支持向量机原理图

Fig.1　Schematic diagram of SVM

图 1 中的黑色实心矩形和空心矩形各自代表着一类样本,黑色实线代表两类样本的分界线,黑色虚线代表距离分界线最近的样本,两条黑色虚线之间的距离为分类间隔。分类间隔的大小和误差成反比,通过调节分类间隔的大小来控制误差,得到的最优结果平面称为最优超平面。近年来,SVM 作为机器学习领域的热点工具,在处理小样本、高维度和非线性类的样本具有一定的优势,被广泛应用于生物学各个研究的领域中[15]。本文使用 LibSVM 支持向量机算法软件包进行预测。

1.4　降维去冗余方法

1.4.1　F-score 降维

在预测的过程中,不相关的特征向量和特征信息的冗余现象,都会导致增加不必要的计算过程,降低预测的准确率。为了消除特征信息的冗余,去除不相关特征向量,就需要用到特征选择技术。实现特征信息在空间维数上的压缩,获得最佳维数,在众多的特征中选择对分类识别最有效的特征,即为特征选择。目前具有代表性的特征选择技术有 F-score、最大相关-最小冗余(mRMR)、最大相关-最大距离(MRMD)、二项分布(BD)、递归特征消除法(RFE),主成分分析(PCA)和方差分析(ANOVA)等。本文选用 F-score 作为降维去冗余方法之一,F-score 最早是由 Chen Yi-Wei[16] 提出,它的执行思路是先按照设定逻辑对各个特征向量打分,所有的特征向量都按 F-score 值分数排序,分数越高表明该特征越具有区别性。将得分最高的作为第一个特征向量进行预测。然后根据分数高低,把其余的特征向量陆续加入到上一个特征向量后面,再进行预测,直到添加完所有的特征向量为止[17]。F-score 是一种衡量两类特征向量之间分辨能力的方法,可以实现对特征信息冗余的消除,获得最佳的维数,选择有效的特征向量[18]。具体可以表示为

$$F_i=\frac{(\bar{x}_i^{(+)}-\bar{x}_i)^2+(\bar{x}_i^{(-)}-\bar{x}_i)^2}{\frac{1}{n^+-1}\sum_{k=1}^{n^+}(x_{k,i}^{(+)}-\bar{x}_i^{(+)})^2+\frac{1}{n^--1}\sum_{k=1}^{n^-}(x_{k,i}^{(-)}-\bar{x}_i^{(-)})^2} \tag{9}$$

上述公式中:$\bar{x}_i$ 指的是全部样本中第 i 个特征样本的平均值,$\bar{x}_i^{(+)}$ 指的是正样本中第 i 个特征样本的平均值,$\bar{x}_i^{(-)}$ 指的是负样本中第 i 个特征样本的平均值;n^+ 指的是全部样本中正样本个数,n^- 指的是全部样本中负样本个数;$x_{k,i}^{(+)}$ 指的是正样本中第 k 个样本的第 i 个特征样本的值,$x_{k,i}^{(-)}$ 指的是负样本中第 k 个样本的第 i 个特征样本的值。

1.4.2 最大相关-最小冗余(mRMR)

选用最大相关-最小冗余(Maximal relevance and minimal redundancy, mRMR)作为第二个特征选择算法。为了消除特征信息冗余产生的不良影响,去除不相关特征向量,改善预测模型的稳定性和有效性,提高预测结果的可靠性和准确率,有很多特征选择方法已被提出[19]。在众多特征选择方法中,选用最大相关-最小冗余方法是因为它在基于不同的分类算法下,均可以显著改善特征选择结果可靠性和分类准确率,被广泛用于许多学科研究领域中。本文使用的 mRMR 程序由彭等[20] 开发(http://home.penglab.com/proj/mRMR/),mRMR 原理[21] 如下所示:

$$I(x;y)=\iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)}\mathrm{d}x\mathrm{d}y \tag{10}$$

上述公式中:x 和 y 是随机变量,$p(x)$ 和 $p(y)$ 指的是概率密度,$p(x,y)$ 指的是联合概率密度,$I(x;y)$ 指的是 x 和 y 之间的互信息。

最大相关和最小冗余的测度指标分别定义为:

$$maxD(S,c),D=\frac{1}{|S|}\sum_{x_i\in S}I(x_i;c) \tag{11}$$

$$minR(S),R=\frac{1}{|S|^2}\sum_{x_i,x_j\in S}I(x_i;x_j) \tag{12}$$

上述公式中:S 指的是特征集,$|S|$ 指的是特征集包含的特征数目;c 指的是目标类别;$I(x_i;c)$ 指的是特征 i 和目标类别 c 之间的互信息;$I(x_i;x_j)$ 指的是

特征 i 与特征 j 之间的互信息;D 指的是特征集 S 中各特征 x_i 与类别 c 之间互信息的均值,用于表示特征集与相应类别的相关性;R 指的是特征集 S 中各特征间互信息的大小,用于表示特征之间的冗余性[22]。

1.5 评价指标

为了对预测的结果进行检验和总结,很多检验方法被提出,例如:留一法(Jackknife 检验)、K 折交叉检验和自洽检验等。本文最终选用了 Jackknife 检验方法。Jackknife 检验方法中,数据集中的每个样本都将作为一个独立的测试样本,数据集中测试样本之外的其余样本作为训练集,依次将所有的样本都进行检验,使得到的结果可靠、客观和严谨。对预测结果的评估同样重要,本文选取以下指标作为对结果的评价:敏感性(Sensitivity, Sn),特异性(Specificity, Sp),马修斯相关系数(Matthew's correlation coefficient, MCC)和成功率(Accuracy, Acc)。Sn 可以表明预测结果的准确性;Sp 可以表明预测结果的可靠性;MCC 可以表明预测结果与实际数据的相关性;Acc 可以表明整个数据集的预测正确率[23]。具体表示为

$$Sn = \frac{TP}{TP + FN} \tag{13}$$

$$Sp = \frac{TN}{TN + FP} \tag{14}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TN + FN) \times (TP + FN) \times (TN + FP)}} \tag{15}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{16}$$

上述公式中:TP 指的是锌指蛋白序列中被正确预测的个数,TN 指的是非锌指蛋白序列中被正确预测的个数,FP 指的是非锌指蛋白序列中被错误预测为锌指蛋白的个数,FN 指的是锌指蛋白序列中被错误预测为非锌指蛋白的个数。

2 结果与分析

2.1 平均化学位移特征信息的最优组合选取

为更好预测锌指蛋白,需要对平均化学位移特征信息的两个主要参数进行最优选择,一个参数是四种骨架原子的最优组合,另一个参数是相关常数 λ 的最优选择。图 2 列出了平均化学位移的四种骨架原子不同组合方案预测成功率,包括单个骨架原子的预测成功率和多个骨架原子组合的预测成功率,预测结果表明,当骨架原子为 ^{15}N 时的成功率最高为 87.17%。图 3 表明在最优骨架原子 ^{15}N 的选择下,相关常数 λ 为 30 时的成功率最高。故本文平均化学位移的相关常数 λ 选择 30,四种骨架原子选择单骨架原子 ^{15}N。

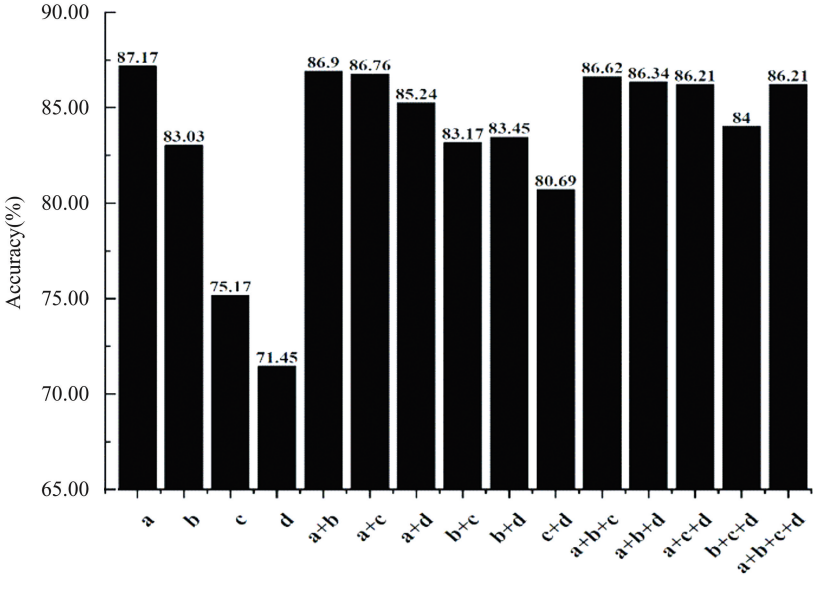

图 2 平均化学位移的不同骨架原子组合方案预测成功率

Fig.2 Accuracy of different skeleton atom combination schemes of auto-covariance Average Chemical Shift

注:字母 a 代表 ^{15}N,b 代表 $^{13}C_{\alpha}$,c 代表 $^{1}H_{\alpha}$,d 代表 $^{1}H_{N}$.

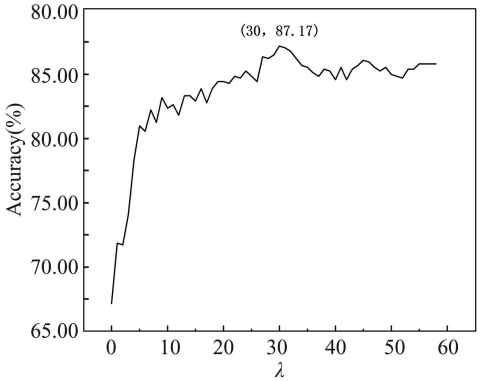

图 3　最优原子结合模式下不同 λ 对应的预测成功率

Fig.3　Accuracy of different λ in the optimal atoms combination mode

2.2　预测结果

2.2.1　单特征预测结果

本文基于支持向量机算法,Jackknife 检验方法,提取了氨基酸单肽组分信息、氨基酸二肽组分信息、平均化学位移三类特征信息,然后对数据集进行预测,详细结果见表 2。

表 2　不同特征参数的预测结果

Table 2　Prediction results of different feature parameters

特征(Feature)	Sn/%	Sp/%	MCC	Acc/%
AAC	90.08	85.64	0.76	87.86
DC	84.85	90.88	0.76	87.86
acACS	86.78	87.57	0.74	87.17

从表 2 总体来看,单个特征信息预测成功率都在 85%以上,马修斯相关系数都在 0.75 左右,其它各项评价指标也取得较好的结果,说明本文的预测具有较好的可靠性及参考价值。具体来看,氨基酸单肽组分信息和氨基酸二肽组分信息取得的成功率高且相近,预测结果都为 87.86%。氨基酸单肽组分信息的敏感性高于其它两个参数的敏感性,说明其更适合对锌指蛋白的预测。

2.2.2　氨基酸二肽组分信息特征降维结果

由于特征参数存在冗余现象,降低预测的准确率且增加不必要的计算过程,为了实现特征参数在空间维数上的压缩,获得最佳维数,本文选用 F-score 和最大相关-最小冗余(以下称 mRMR)两种方法对氨基酸二肽组分信息特征参数进行降维处理。使用 F-score 降维方法,如图 4 所示,颜色越趋于深红色,则说明相邻的氨基酸残基越具有更高的 F 值,并且更具有区别性,将作为第一个特征向量,依次再加入 F 值由高到低的其它特征向量进行预测。相反,颜色越趋于深蓝色,则不容易被区分识别。具体预测结果见表 3。

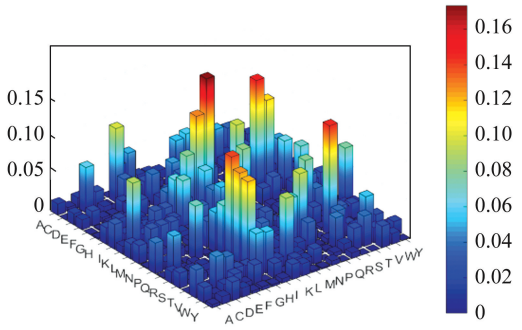

图 4　氨基酸二肽组分信息的 F 值三维热图

Fig.4　Three dimensional heat map of DC's F-score value

表 3　氨基酸二肽组分信息特征参数降维和不降维处理下的预测结果

Table 3　Prediction results of Dipeptide Composition after reduced and non-reduced dimension treatment

降维方法	Sn/%	Sp/%	MCC	Acc/%
未使用降维方法	84.85	90.88	0.76	87.86
F-score	88.15	90.88	0.79	89.52
mRMR	88.71	91.71	0.80	90.21

由表 3 可以看出,使用两种降维方法对氨基酸二肽组分信息特征处理后,预测成功率明显提升,各项评价指标也取得了更好的结果。其中 F-score 降维处理后,特异性变化不大,敏感性和马修斯相关系数均取得了提升,说明 F-score 降维后的特征参数对数据集中锌指蛋白的预测更加精准,结果更具可靠性。mRMR 降维后,各项评价指标的提升幅度明显,成功率达到了 90.21%,表明 mRMR 方法更适合对氨基酸二肽组分信息特征进行降维处理。

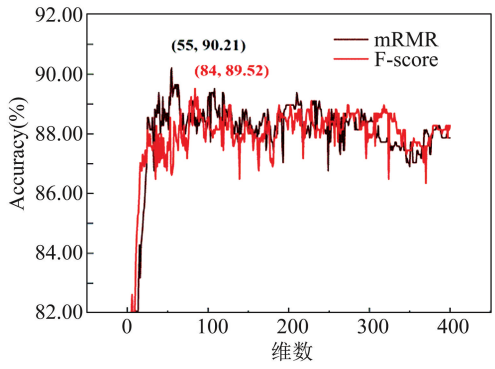

图 5　氨基酸二肽组分信息特征参数在不同的降维方法处理后的成功率

Fig.5　Accuracy of Dipeptide Composition after different dimension reduction methods

由图 5 可以看出,两种降维方法都明显提高了预测成功率,氨基酸二肽组分信息特征参数在 F-score 降维处理后,最终获得 84 维特征,成功率达到 89.52%;在 mRMR 降维后,最终获得 55 维特征,成功率达到 90.21%。总体而言,两种降维方法都起到

了良好效果,实现了特征参数在空间维数上的压缩,一定程度上消除了特征参数的冗余现象,增加了预测的准确率。

2.2.3 融合特征信息预测结果

在单特征信息预测的基础上,对平均化学位移、氨基酸单肽组分信息、氨基酸二肽组分信息三类特征信息进行融合,其中,氨基酸二肽组分信息采用 mRMR 降维后的结果。基于支持向量机算法,利用融合特征信息对数据集进行预测,结果见表 4。

表 4 融合特征参数的预测结果

Table 4 Prediction results of fusion feature parameters

特征(Feature)	Sn/%	Sp/%	MCC	Acc/%
AAC+DC	87.05	90.33	0.77	88.69
AAC+acACS	93.11	91.99	0.85	92.55
DC+acACS	92.01	92.82	0.85	92.41
AAC+DC+acACS	92.01	91.99	0.84	92.00

由表 4 可以看出,氨基酸单肽组分信息和平均化学位移两类特征信息融合、氨基酸二肽组分信息和平均化学位移两类特征信息融合和三类特征信息融合的预测成功率均高于三类单特征信息的预测成功率,说明特征信息融合方法对提升锌指蛋白的预测成功率具有一定的作用。具体来看,在两类特征信息融合后,氨基酸单肽组分信息和平均化学位移两类单特征信息融合后取得了最高的预测成功率,达到 92.55%,高于三类单特征信息的预测成功率。敏感性和马修斯相关系数结果也均高于三类单特征信息的结果和其它的两类特征信息融合的结果。说明氨基酸单肽组分信息和平均化学位移两类特征信息的融合对锌指蛋白的预测更具有优势,更加精确可靠。氨基酸二肽组分信息和平均化学位移两类特征信息融合后取得 92.41%的预测成功率,其它各项评价指标也都取得较好结果。氨基酸单肽组分信息和氨基酸二肽组分信息两类特征信息融合后的预测成功率未高于 mRMR 降维后的氨基酸二肽组分信息单特征信息预测成功率,具体的原因可能是氨基酸单肽组分信息包含的蛋白质信息不够全面和充分,而且本文中并未对氨基酸单肽组分信息进行降维处理,只对氨基酸二肽组分信息特征进行了两种方式的降维处理,氨基酸单肽组分信息和氨基酸二肽组分信息两类特征信息融合后加大了数据的冗余,影响最终预测结果。在三类特征信息融合后,预测结果取得 92%的成功率,各项评价指标结果较好,均高于三类单特征信息的预测评价指标结果,说明本文选取的三类特征信息对锌指蛋白的预测是有一定作用和意义的。

3 结 论

对 C2H2 型锌指蛋白的预测,可以深入了解锌指蛋白的结构和功能,为生物遗传、表观特征、医疗等方面的研究提供帮助。本文构建了 C2H2 型锌指蛋白和非锌指蛋白数据集,提取了三类特征信息,采用 F-score 和 mRMR 两种降维方法,利用支持向量机算法在 Jackknife 检验方法下对数据集进行预测。预测结果表明:

1)利用 F-score 和 mRMR 方法对氨基酸二肽组分特征信息进行降维,mRMR 降维方法好于 F-score 降维方法的预测结果。

2)对特征信息进行适当融合有助于提高预测成功率。在后期研究中,进一步选取蕴含 C2H2 型锌指蛋白结构特征的特征参数对锌指蛋白进行预测,力争得到更高的预测成功率。

参考文献(References)

[1] WEBSTER M W, WEIXLBAUMER A. The intricate relationship between transcription and translation[J]. Proceedings of the National Academy of Sciences of the United States of America, 2021, 118(21): e2106284118. DOI: 10.1073/PNAS.2106284118.

[2] 刘强,张贵友,陈受宜. 植物转录因子的结构与调控作用[J]. 科学通报,2000,45(14):1465-1474. DOI:10.3321/j.issn:0023-074X.2000.14.002.
LIU Qiang, ZHANG Guiyou, CHEN Shouyi. Structure and regulation of plant transcription factors[J]. Chinese Science Bulletin, 2000, 45(14): 1465-1474. DOI: 10.3321/j.issn:0023-074X.2000.14.002.

[3] 黄骥,王建飞,张红生. 植物 C2H2 型锌指蛋白的结构与功能[J]. 遗传,2004,26(3):414-418. DOI:10.16288/j.yczz.2004.03.030.
HUANG Ji, WANG Jianfei, ZHANG Hongsheng. Structure and function of plant C2H2 zinc finger protein[J]. Hereditas (Beijing), 2004, 26(3): 414-418. DOI: 10.16288/j.yczz.2004.03.030.

[4] MILLER J, MCLACHLAN A D, KLUG A. Repetitive zinc-binding domains in the protein transcription factor IMA from xenopus oocytes[J]. Journal of Trace Elements in Experimental Medicine, 2001, 14(2): 157-169. DOI: 10.1002/j.1460-2075.1985.tb03825.x.

[5] LEE M S, GIPPERT G P, SOMAN K V, et al. Three-di-

mensional solution structure of a single zinc finger DNA binding domain[J]. Science(New York, N.Y.), 1989, 245(4918): 635-637. DOI: 10.1126/science.2503871.

[6]KRISHNA S S, MAJUMDAR I, GRISHIN N V. Structural classification of zinc fingers: survey and summary [J]. Nucleic Acids Research, 2003, 31(2): 532-550. DOI: 10.1093/nar/gkg161.

[7]赵楠,赵飞,李玉花. 锌指蛋白结构及功能研究进展[J]. 生物技术通讯,2009,20(1):131-134. DOI:10.3969/j.issn.1009-0002.2009.01.037.
ZHAO Nan, ZHAO Fei, LI Yuhua. Advances in research on zinc finger protein [J]. Letters in Biotechnology, 2009, 20(1): 131-134. DOI: 10.3969/j.issn.1009-0002.2009.01.037.

[8]沈磐,杨冬,贺福初. C2H2 型锌指蛋白结合的 DNA 序列预测方法的研究进展[J]. 生物化学与生物物理进展,2017,44(7):573-579. DOI:10.16476/j.pibb.2017.0047.
SHEN Pan, YANG Dong, HE Fuchu. The advancement of the prediction methods for DNA-binding preferences of C2H2 zinc finger proteins [J]. Progress in Biochemistry and Biophysics, 2017, 44(7): 573-579. DOI: 10.16476/j.pibb.2017.0047.

[9]孙燕,苟德明,李文鑫. C2H2 型锌指蛋白研究进展[J]. 生命的化学,2001,21(6):473-475. DOI:10.3969/j.issn.1000-1336.2001.06.012.
SUN Yan, GOU Deming, LI Wenxin. Advances in research on C2H2 zinc finger protein [J]. Chemistry of Life, 2001, 21(6): 473-475. DOI: 10.3969/j.issn.1000-1336.2001.06.012.

[10]张振慧. 蛋白质分类问题的特征提取算法研究[D]. 长沙:国防科学技术大学,2006. DOI: 10.7666/d.y1101790.
ZHANG Zhenhui. Research on algorithm in feature extraction of protein classification [D]. Changsha: National University of Defense Technology, 2006. DOI: 10.7666/d.y1101790.

[11]AHMAD K, WARIS M, HAYAT M. Prediction of protein submitochondrial locations by incorporating dipeptide composition into chou's general pseudo amino acid composition [J]. The Journal of Membrane Biology, 2016, 249(3): 293-304. DOI: 10.1007/s00232-015-9868-8.

[12]FAN Guoliang, LI Qianzhong. Predict mycobacterial proteins subcellular locations by incorporating pseudo-average chemical shift into the general form of Chou's pseudo amino acid composition[J]. Journal of Theoretical Biology, 2012, 304: 88-95. DOI: 10.1016/j.jtbi.2012.03.017.

[13]姜燕.基于多信息融合预测单定位和多定位凋亡蛋白质亚细胞位置[D].呼和浩特:内蒙古大学,2015. DOI:10.7666/d.Y2840196.
JIANG Yan. Predicting the single and multiple subcellular location of apoptosis proteins based on multi-features fusion [D]. Hohhot: Inner Mongolia University, 2015. DOI: 10.7666/d.Y2840196.

[14]JING Xiaoyang, Li Fengmin. Predicting cell wall lytic enzymes using combined features [J]. Frontiers in Bioengineering and Biotechnology, 2021, 8(1): 1-8. DOI: 10.3389/fbioe.2020.627335.

[15]李明俊,李凤敏. 基于多信息融合识别核定位蛋白[J]. 内蒙古农业大学学报(自然科学版),2020,41(1):87-92. DOI:10.16853/j.cnki.1009-3575.2020.01.016.
LI Mingjun, LI Fengmin. Identification of nucleoprotein based on different features [J]. Journal of Inner Mongolia Agricultural University (Natural Science Edition), 2020, 41(1): 87-92. DOI: 10.16853/j.cnki.1009-3575.2020.01.016.

[16]CHEN Yiwei, LIN Zhiren. Combining SVMs with various feature selection strategies [M]. Berlin: Springe, 2006. 315-324. DOI:10.1007/978-3-540-35488-8_13.

[17]景晓洋. 热休克蛋白家族、细胞壁裂解酶的预测算法研究[D]. 呼和浩特:内蒙古农业大学,2021. DOI:10.27229/d.cnki.gnmnu.2021.000384.
JING Xiaoyang. The study on predictive algorithm for heat shock proteins and cell wall lytic enzymes [D]. Hohhot: Inner Mongolia Agricultural University, 2021. DOI: 10.27229/d.cnki.gnmnu.2021.000384.

[18]谢娟英,王春霞,蒋帅,等. 基于改进的 F-score 与支持向量机的特征选择方法[J]. 计算机应用,2010,30(4):993-996. DOI:1001-9081(2010)04-0993-04.
XIE Juanying, WANG Chunxia, JIANG Shuai, et al. Feature selection method combing improved F-score and support vector machine[J]. Journal of Computer Applications, 2010, 30(4): 993-996. DOI: 1001-9081(2010)04-0993-04.

[19]赖洪燕. 基于序列顺序与位置信息的启动子预测[D]. 成都:电子科技大学,2018.
LAI Hongyan. Based on sequence-order and position-correlation information recognizing promoters[D]. Chengdu: University of Electronic Science and Technology of China, 2018.

[20]DING C, PENG Hanchuan. Minimum redundancy feature selection from microarray gene expression data[J]. Journal of Bioinformatics and Computational Biology, 2005, 3(2): 185-205. DOI: 10.1142/s0219720005001004.

[21]PENG Hanchuan, LONG Fuhui, DING C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(8): 1226-1238. DOI: 10.1109/TPAMI.2005.159.

[22]李扬,顾雪平. 基于改进最大相关最小冗余判据的暂态稳定评估特征选择[J]. 中国电机工程学报,2013,33(34):179-186. DOI: 10.13334/j.0258-8013.pcsee.2013.34.024.
LI Yang, GU Xueping. Feature selection for transient stability assessment based on improved maximal relevance and minimal redundancy criterion [J]. Proceedings of the CSEE, 2013, 33(34): 179-186. DOI: 10.13334/j.0258-8013.pcsee.2013.34.024.

[23]张松,黄波,夏学峰,等. 蛋白质亚细胞定位的生物信息学研究[J].生物化学与生物物理进展,2007,34(6):573-579. DOI:10.3321/j.issn:1000-3282.2007.06.004.
ZHANG Song, HUANG Bo, XIA Xuefeng, et al. Bioinformatics research in subcellular localization of protein[J]. Progress in Biochemistry and Biophysics, 2007, 34(6): 573-579. DOI: 10.3321/j.issn:1000-3282.2007.06.004.