DOI:10.12113/202103014

# 基于氨基酸组分和位点保守信息识别蛋白质-HEME结合残基

李彩艳[1],马 勇[1],邢俊凤[1],郭国栋[1],武一凡[1],闻昊坤[1],丁海麦[2*],张改梅[3*]
(1.包头医学院 计算机科学与技术学院,内蒙古 包头 014000; 2.包头医学院 基础与法医学院,
内蒙古 包头 014000; 3.呼和浩特第一医院,呼和浩特 010051)

**摘 要**:血红素是一种重要的、常用的配体,在电子传递、催化、信号转导和基因表达等方面发挥着重要作用,准确预测蛋白质与血红素相互作用的结合残基是结构生物信息学的主要挑战之一。本文下载整理了Biolip数据库中HEME配体与蛋白质结合的信息,统计分析了结合残基和非结合残基的氨基酸组分和位点保守性信息并将其作为预测特征参数,用Fisher-PSSM判别法识别HEME结合残基,计算结果表明优化特征参数的Fisher-PSSM判别法得到了较好的预测结果。

**关键词**:蛋白质配体; 结合位点; 统计分析; 血红素HEME
**中图分类号**:Q61 **文献标志码**:A **文章编号**:1672-5565(2022)03-189-06

# Identification of protein binding residues HEME based on amino acid component and conservative information

LI Caiyan[1], MA Yong[1], XING Junfeng[1], GUO Guodong[1], WU Yifan[1], WEN Haokun[1], DING Haimai[2*], ZHANG Gaimei[3*]
(1.*School of Computer Science and Technology, Baotou Medical College, Baotou* 014000, *Inner Mongolia, China*;
2.*School of Medical School of Foundation, Baotou Medical College, Baotou* 014000, *Inner Mongolia, China*;
3.*Hohhot First Hospital, Hohhot* 010051, *China*)

**Abstract**: HEME is an important and commonly used ligand that plays an important role in electron transfer, catalysis, signaling transduction, and gene expression. Accurate prediction of the binding residues of protein-HEME interactions is one of the main challenges in structural bioinformatics. In this study, the information of HEME ligand and protein was downloaded from Biolip database, and amino acid components and site conservative information of binding residues were and nonbinding residues were satistically analyzed and used as prediction characteristic parameters. HEME binding residues were identified by Fisher-PSSM criterion, and calculation results showed that the Fisher-PSSM criterion of optimizing characteristic parameters had good prediction results.

**Keywords**: Protein ligand; Binding site; Statistical analysis; HEME

蛋白质通过与配体的相互作用来执行生物学功能[1-3],所以准确地识别蛋白质结合残基与配体结合位点是理解蛋白质生物学功能,药物设计和疾病治疗的关键。血红素HEME辅因子是一种极其通用的辅基,对几乎所有的生物执行功能都至关重要[4-7]。例如亚铁血红素辅因子通常与血红素蛋白结合,在多种生物过程中发挥重要作用,包括电子转移、氧气转运、金属离子储存、化学催化、基因表达、细胞信号转导等[8-10]。所以对血红素结合位点残基的识别有助于更好地理解血红素结合蛋白的生物学功能,揭示血红素-蛋白相互作用的机制,为生物启发蛋白设计提供有价值的线索[11]。然而,实验测定血红素结合残基既费时又费力还耗材,因此,很有必要开发能够预测血红素结合残基的计算方法。

国内外很多学者对蛋白质与配体结合位点进行了研究。如2008年,Jessica等人[12]对$Zn^{2+}$配体的

**收稿日期**:2021-03-31;**修回日期**:2021-06-27.
**基金项目**:内蒙自然科学基金项目(No.2020MS08015);内蒙古大学生创新创业训练计划项目(No.S202119127006,S202119127007).
**作者简介**:李彩艳,女,副教授,研究方向:生物信息学. E-mali: licaiyanding@126.com.
*通信作者:丁海麦,男,副教授,研究方向:生物化学 E-mail: 102007210@btmc.edu.cn.
张改梅,女,副教授,研究方向:医学 E-mail: 249756044@qq.com.

结合位点进行预测;Babor 等人基于 3D 结构开发了 CHED 算法,预测了 $Zn^{2+}$, $Co^{2+}$, $Ni^{2+}$, $Fe^{2+}$, $Cu^{2+}$, $Mn^{2+}$金属的结合位点[13-14],2019 年 Zhu 等人利用机器学习方法预测了多种蛋白质配体等[15]。2013 年,Zhang 等人[16]收集了配体和蛋白质之间相互作用,整理得到 Biolip 数据库,这是一个半手工蛋白质离子配体数据库,比较全面地注释了蛋白质配体及其结合残基信息。Biolip 数据库中每个条目都包含了对以下内容的注释:配体结合残基、配体结合亲和力、催化位点、委员会编号、基因本体术语和其他数据库的交叉链接等。数据库中包含了极其广泛和精准的配体蛋白数据,之后很多学者使用 Biolip 数据库中的配体信息来预测蛋白质配体结合位点。如 2016 年,Hu 等人[17]使用 SVM 方法较好地识别了 Biolip 数据库中 $Cu^{2+}$, $Fe^{2+}$, $Fe^{3+}$等金属离子配体的结合位点;2017 年 Gao 等人[18]统计分析了金属配体结合残基序列片段的信息,使用 SVM 算法对 Biolip 数据库中 $Zn^{2+}$, $Co^{2+}$, $Ni^{2+}$, $Fe^{2+}$, $Cu^{2+}$, $Mn^{2+}$等金属离子配体的结合位点进行预测等等。

2011 年,Liu 等人[19]等人使用支持向量机的方法,考虑血红素配体结合残基及其附近残基的溶剂可及性面积、进化保守性、深度和突出性等特征,对含有血红素配体的 141 条无冗的蛋白结合位点进行了预测,得到总精度 76.49% 和 MCC 为 0.407。Liu 等人[20]也利用支持向量机方法,对同样的蛋白序列,通过结合序列的拓扑特征和结构特征来识别血红素结合残基,得到总精度 85.99% 和 MCC 为 0.489。2019 年,Zhao 等人[21]使用 SXGBsite 方法对 Biolip 数据库中 27 条含血红素蛋白质进行预测,得到总精度 96.2%和 MCC 为 0.618。

本文从 Biolip 数据库中下载了蛋白质和血红素结合的相关数据,并且进行了整理,然后从中提取了有益信息,使用 Fisher 判别法和矩阵打分方法进行了计算,得到了较好的预测结果,并与前人进行了比较,为 HEME 与蛋白质结合提供有益信息。

## 1 数据及方法

### 1.1 数据集

从 Biolip 数据库下载整理了已知血红素与蛋白质结合信息,得到蛋白链 2 952 条,筛选分辨率好于 3Å、序列长度大于 50 个残基,序列相似性低于 40% 的蛋白质链 254 条。由于蛋白质序列中,残基和血红素配体结合不仅仅由残基本身决定,也受周围残基的影响。因此,采用移动窗口的方法截取蛋白质序列片段。将移动窗口中心位置含有配体结合残基的序列片段定义为正集,否则定义为负集。得到正集片段数为 4 589,负集片段数为 66 137。由于负集片段数大于正集片段数,约是正集的 14 倍,本文采用随机抽样的方法,把负集随机分为 14 组,使每组负集的片段数与正集相等,最后取 14 次结果的平均值作为最终结果。由于周围环境对结合残基的影响未知,所以滑动窗口长分别选取 7,9,11,13,15,17,19,并通过计算得出最优窗口长。

### 1.2 数据集的统计分析

#### 1.2.1 氨基酸组分信息

由文献[17-18]可知,氨基酸组份信息在区分正负集序列片段方面是一个较好的参数,所以我们对 HEME 正负集序列片段中 20 种氨基酸组份出现概率进行统计分析(见图 1),空心柱体表示正集,实心柱体表示负集,*X* 轴表示 20 种氨基酸,*Y* 轴表示相应氨基酸组份在正负集合中出现概率。发现正集片段中 G、H、C 明显高于负集片段,而 E、D 则在负集片段中含量明显高于正集片段,同时我们注意到 V、K、W、Y 也在正负集中差异较大。所以氨基酸组份信息可以作为区分序列片段是参数。

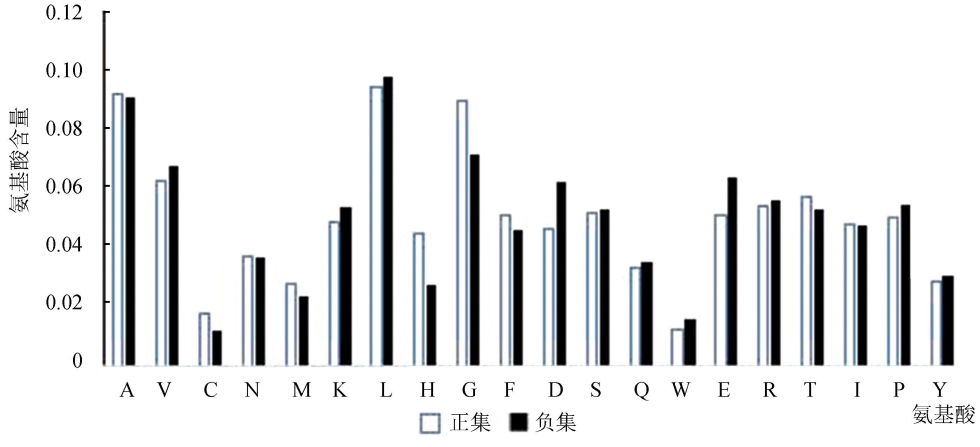


**图 1 血红素片段正负集氨基酸组份含量**

**Fig.1 Amino acid composition content of positive and negative sets in hemoglobin fragments**

### 1.2.2　位点氨基酸保守性信息

利用 WEBLOGO 软件[22],对血红素片段的正负集氨基酸位点保守性信息进行了统计分析,我们以动窗口长度 19 为例,统计结果(见图 2),横坐标为位点,纵坐标为各位点的氨基酸保守性,氨基酸字母高度代表了在此位点上氨基酸出现的相对频率。正集片段中心即位置 10 表示血红素配体结合残基,血红素配体结合残基偏好使用 L,F,H,R,I 等氨基酸,在结合残基附近氨基酸位点保守性都较强。在相同位点处正负集片段保守性有着显著差异,比如对于位置 11 处正集的偏好残基为 G、L、A、T、F 等,而负集在这个位点处偏好残基为 A、K、V 等,再如对于位置 12 处正集的偏好残基为 G、A、L 等,而负集在此位点处偏好残基仍为 A、K、V 等。因此,位点氨基酸保守性信息有助于识别正负集序列片段。

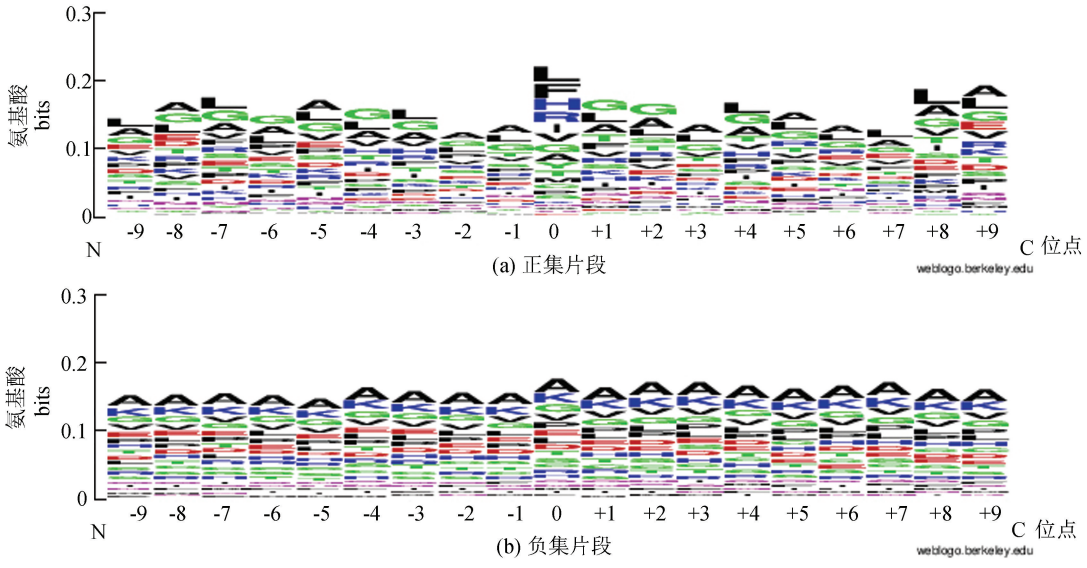


**图 2　血红素片段正负集氨基酸位点保守性**

**Fig.2　Position conservation of positive and negative amino acid in hemoglobin fragments**

## 1.3　方法

### 1.3.1　Fisher 判别法

Fisher 判别法在两类识别方面,具有较好的性能[23],该方法已成功应用于蛋白质超二级结构预测[24]。在本文的应用中,以氨基酸组份信息为特征指标为例,考虑正负集每个序列片段上 21(20 种氨基酸和一个伪氨基酸)维特征指标出现频次;对正负集各 4 589 个序列片段,统计 21 个特征指标出现频数。

$X_i^{(1)}(x_{i1}^{(1)},x_{i2}^{(1)},\cdots\cdots x_{in}^{(1)})$　　$X_i^{(2)}(x_{i1}^{(2)},x_{i2}^{(2)},\cdots\cdots x_{in}^{(2)})$　　($i$ 为片段数,$n$ 为特征数 )

分别计算各指标在正负集中的总体均值、距离、协方差:

$$\bar{x}_i^{(1)}=\frac{1}{4589}\sum_{k=1}^{4589}x_{ki}^{(1)},\quad \bar{x}_i^{(2)}=\frac{1}{4589}\sum_{k=1}^{4589}x_{ki}^{(2)}$$

$$d_i=\bar{x}_i^{(1)}-\bar{x}_i^{(2)}\qquad (i=1,2\cdots n)$$

$$S_{ij}=\sum_{k=1}^{p}(x_{ki}^{(1)}-\bar{x}_i^{(1)})(x_{kj}^{(1)}-\bar{x}_j^{(1)})+\sum_{k=1}^{q}(x_{ki}^{(2)}-\bar{x}_i^{2})(x_{kj}^{(2)}-\bar{x}_j^{(2)})$$

建立判别函数

$$F(x_1,x_2,\cdots\cdots x_n)=C_1x_1+C_2x_2+\cdots+C_nx_n$$

将平均值代入判别函数,计算判别值

$$\bar{y}^{(1)}=C_1\bar{x}_1^{(1)}+C_2\bar{x}_2^{(1)}+\cdots\cdots+C_n\bar{x}_n^{(1)}$$

$$\bar{y}^{(2)}=C_1\bar{x}_1^{(2)}+C_2\bar{x}_2^{(2)}+\cdots\cdots+C_n\bar{x}_n^{(2)}$$

分界点为 $C=\dfrac{\bar{y}_1+\bar{y}_2}{2}$,最后将判别组数据代入判别函数,进行鉴别 $y=C_1x_1+C_2x_2+\cdots C_nx_n$。

若 $\bar{y}_1>\bar{y}_2$:当 $y>C$,该片段属于正集片段;当 $y<C$,该片段属于负集片段;

若 $\bar{y}_1<\bar{y}_2$:当 $y>C$,该片段属于负集片段;当 $y<C$,该片段属于正集片段。

### 1.3.2　PSSM 算法

PSSM 算法是一种较好的分类方法,被应用于超二级结构预测等研究中并取得了不错的效果[25-26],具体算法如下:

打分函数为:$s=\dfrac{\sum_{i=1}^{L}C_if_{i,j}}{\sum_{i=1}^{L}C_if_{i,\max}}$

其中 $C_i$ 为位点保守性参量:$C_i=\dfrac{100}{\ln 21}(\sum_{\iota=1}^{21}p_{\iota,j}\ln p_{\iota,j}+\ln 21)$

其中 $p_{i,j}$ 为位置概率矩阵的矩阵元:

$$P_{i,j} = \frac{(f_{i,j} + \frac{\sqrt{N_i}}{21})}{(N_i + \sqrt{N_i})}$$

$f_{i,j}$ 表示位置频数矩阵的第 $i$ 列、第 $j$ 各氨基酸出现的频次 $N_i$ 表示在第 $i$ 个位点上 出现的氨基酸的总和，$f_{i,\max}$ 表示位置频数矩阵的第 $i$ 列的最大值。

以位点氨基酸信息为基础参数，通过训练集构造标准打分矩阵，对于检验集的每条片段，得到两个打分 $s$ 值，哪个分数高，片段就被判为那个集。同时，打分值也是一个比较好的预测参数，因此本文也把打分值作为预测特征参数用于 Fisher 计算。

1.3.3 预测结果的评价方法

采用 5 交叉检验，即把两类把样本随机分为 5 份，每次都选取 4/5 作训练集，1/5 作检验集，交叉进行 5 次，最后取 5 次平均结果。

通常结合残疾的识别都采用五交叉检验，并且使用下列评价指标：敏感性（*Sn*）、特异性（*Sp*）、总精度（*ACC*）和相关系数（*MCC*）表示，分别定义为：

$$S_n = \frac{TP}{TP + FN} S_p = \frac{TN}{TN + FP}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

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

此外，文献[16]中也采用 Re*cal* 、Pr*cision* 、*ACC* 、*F*1 − *score* 、*MCC* 来评价预测结果，其中 Re*cal* 即上文提到敏感性（*Sn*），Pr*cision* 与 *F*1 − *score* 计算方法如下：

$$\text{Pr}cision = \frac{TP}{TP + FP}$$

$$F1 - score = \frac{2 \times \text{Re}cal \times \text{Pr}cision}{\text{Re}cal + \text{Pr}cision}$$

其中，*TP* 表示正确识别金属离子配体结合残基的数量；*FN* 表示将金属离子配体结合残基识别为非金属离子配体结合残基的数量；*TN* 表示正确识别金属离子配体非结合残基的数量；*FP* 表示将金属离子配体非结合残基识别为金属离子配体结合残基的数量。

## 2 结果及讨论

### 2.1 位置权重矩阵打分算法的预测结果及讨论

以氨基酸位点保守性为特征指标，使用位置权重矩阵打分算法进行预测，选取移动窗口为 7、9、11、13、15、17、19 长度，判别结果（见表 1）。从 *Sn* 来看移动窗口长度为 7、9、11、13 时结果较好，都超过到 53%；从 *Sp* 来看，移动窗口长度为 15、17、19 时结果较好，都超过 81%；移动窗口长度为 9、11、13 时，预测 *ACC* 和 *MCC* 结果较好，分别超过了 65% 和 0.32。相对来讲，窗口长度为 11 时，相关系数为 0.32，总精度为 65.59%，预测结果较好。另外我们发现使用位置权重矩阵打分算法预测结果不高，但相关系数都大于 0.30。

**表 1 以氨基酸参数位置权重矩阵打分算法判别法判别结果**

**Table 1 Performance of position weight matrix discrimination algorithm using amino acid as parameter**

| 方法 | 参数 | 窗口长度 | *Sn*/% | *Sp*/% | *ACC*/% | *MCC* |
|---|---|---|---|---|---|---|
| PSSM | 氨基酸 | 7 | 56.03 | 73.72 | 64.87 | 0.30 |
| | | 9 | 54.41 | 76.88 | 65.65 | 0.32 |
| | | 11 | 53.98 | 77.21 | 65.59 | 0.32 |
| | | 13 | 53.17 | 77.51 | 65.34 | 0.32 |
| | | 15 | 48.22 | 81.63 | 64.93 | 0.32 |
| | | 17 | 40.88 | 86.56 | 63.72 | 0.31 |
| | | 19 | 41.12 | 86.34 | 63.73 | 0.31 |

### 2.2 Fisher 判别法的预测结果及讨论

以组分氨基酸为特征指标，使用 Fisher 判别法进行预测。对于每条训练集的序列，都可以得到 21（20 种氨基酸和一个伪氨基酸）维特征参数，选取移动窗口为 7、9、11、13、15、17、19 长度，判别结果（见表 2）。发现以氨基酸组份为参数时，从 *Sn* 来看 7 个窗口结果相差不大，相对来讲 7、9、11 稍好，大约都在 61%；从 *Sp* 来看，移动窗口选取 15、17、19 时较好，约在 65% 以上；*ACC* 和 *MCC* 结果也相差不大。相对来讲，窗口长度为 9 时结果稍好，总精度为 63.17%，相关系数为 0.32。

以位点氨基酸保守性信息为特征指标，使用 Fisher 判别法进行预测，对于每条训练集长度为 *L* 的序列，都可以得到 21×*L* 维特征参数，选取移动窗口为 7、9、11、13、15、17、19 长度，判别结果见表 2。以位点氨基酸保守性信息为参数时，发现窗口的改变时 *Sn* 、*Sp* 、*ACC* 、*MCC* 这四个指标几乎无太大改变，除了窗口长度为 7 时 *MCC* 略差；相对来讲，

移动窗口长度为 13、15 时预测结果略占优势,总精度67.79%,相关系数 0.36。基于同样的特征参数,该预测结果比用位置权重矩阵打分算法要好一些。

**表 2　以氨基酸参数 Fisher 判别法判别结果**

**Table 2　Performance of Fisher discrimination algorithm by the amino acid parameter**

| 方法 | 参数 | 窗口长度 | *Sn*/% | *Sp*/% | *ACC*/% | *MCC* |
|---|---|---|---|---|---|---|
| FisherFisher | 氨基酸组份 | 7 | 61.21 | 62.80 | 62.01 | 0.24 |
| | | 9 | 61.89 | 64.46 | 63.17 | 0.26 |
| | | 11 | 61.28 | 64.20 | 62.74 | 0.25 |
| | | 13 | 60.97 | 64.85 | 62.91 | 0.26 |
| | | 15 | 60.93 | 65.24 | 63.09 | 0.26 |
| | | 17 | 60.69 | 65.37 | 63.03 | 0.26 |
| | | 19 | 60.67 | 64.74 | 62.70 | 0.25 |
| | 位点氨基酸保守信息 | 7 | 66.05 | 65.81 | 65.93 | 0.32 |
| | | 9 | 67.44 | 67.23 | 67.33 | 0.35 |
| | | 11 | 67.29 | 67.94 | 67.62 | 0.35 |
| | | 13 | 67.33 | 67.64 | 67.49 | 0.35 |
| | | 15 | 67.97 | 67.62 | 67.79 | 0.36 |
| | | 17 | 67.42 | 67.40 | 67.41 | 0.35 |
| | | 19 | 67.73 | 67.47 | 67.60 | 0.35 |

### 2.3　优化特征参数 Fisher 判别法的预测探索

由以上计算结果可知,利用 Fisher 判别法,以氨基酸组份信息为参数时移动窗口长度为 9 结果较好,以位点氨基酸保守信息为参数时移动窗口长度为 15 结果较好,综合这两个指标进行探索,特征参数为 15×21+21 共 336 个指标,利用 Fisher 判别法,判别结果(见表 3)。发现综合位点氨基酸保守信息和氨基酸组份信息时,判别结果和只考虑位点氨基酸保守信息结果无太大差异,可能由于指标参数太多,影响判别结果,所以,考虑先作主成分分析,再进行判别。

由前面结果可知,以氨基酸为特征指标,位置权重矩阵打分算法进行预测,移动窗口选取 11 个氨基酸长度相关系数较高,于是选取氨基酸位置权重矩阵打分与 Fisher 判别法中氨基酸组份信息的最优窗口组合,进行预测。先用位置权重矩阵打分算法对 11 个窗口氨基酸进行打分,将两个分数作为特征指标加入到以组份氨基酸为特征指标的 9 个窗口长度的 Fisher 判别法中,这样特征指标共 23 个,判别结果(见表 3)。发现以组分氨基酸为特征指标时,Fisher 判别法加入位置权重矩阵打分值以后预测结果有很大提高,相关系数从原来的 0.26 提高到 0.35,说明这两种方法结合有利于预测。如果进一步加入亲疏水性,极性等指标,结果可能会更好。另外以位点氨基酸信息的 Fisher 判别法用 15 窗口长度加入 11 窗口长度的位置权重矩阵打分值,但预测精度没有提高,结果也就没有给出,可能由于特征指标太多影响预测结果,下一步将结合主成分分析进行预测。

### 2.4　与文献[19]结果比较

文献[19]中使用 SVM 方法,综合了 PSSM、RASA、DPX、CX 或这些特征的组合构造了 15 个基于结构的分类器,部分预测结果(见表 4)。比较而言,该方法更为简单。文献[19]中使用的单一指标最好的是 PSSM,利用 Fisher 判别法,使用单一指标最好是氨基酸位点保守性,结果较好一些,综合使用多种指标时,文献[19]结果更好一些,所以下一步工作也考虑综合多个指标进行判别。

**表 3　不同窗口优化组合判别法判别结果**

**Table 3　Performance of different window optimization combination discrimination methods**

| 方法 | 参数 | 窗口长度 | *Sn*/% | *Sp*/% | *ACC*/% | *MCC* |
|---|---|---|---|---|---|---|
| Fishersher | 位点氨基酸保守信息+氨基酸组份信息 | 位点氨基酸保守信息选取 15 氨基酸组份选取 9 | 67.94 | 67.62 | 67.78 | 0.36 |
| Fisher | 氨基酸组份+PSSM 值 | 位置权重矩阵打分选取 11 氨基酸组份选取 9 | 66.27 | 68.75 | 67.51 | 0.35 |

**表 4　与文献[19]结果比较**

**Table 4　Comparison with of reference[19] results**

| 方法 | 参数 | Re*cal*/% | Pr*ecisiom*/% | *ACC*/% | *F*1-*score*/% | *MCC* |
|---|---|---|---|---|---|---|
| SVM | PSSM | 74.55 | 29.88 | 72.98 | 42.64 | 0.34 |
| SVM | PSSM+RASA+DPX+CX | 79.08 | 34.04 | 76.49 | 47.10 | 0.41 |
| Fisher | 位点氨基酸保守性 | 66.27 | 68.75 | 67.51 | 67.83 | 0.35 |

## 3 结 论

从 Biolip 数据库中,整理出与血红素结合的蛋白质链,并利用 Fisher 判别法和位置权重打分矩阵进行识别血红素结合残基。利用 Fisher 判别法时,考虑 20 种氨基酸组份信息、位点氨基酸保守信息以及两种信息的优化组合,取得较好预测结果。但与前人工作相比,预测结果稍差,在以后工作中将考虑氨基酸的二级结构信息,亲疏水性、极性等指标进行判别,以进一步提高预测结果。

## 参考文献(References)


[1]HU X Z, FENG Z X, ZHANG X J, et al. The identification of metal ion ligand-binding residues by adding the reclassified relative solvent accessibility[J]. Frontiers in Genetics, 2020,11:214.DOI: 10.3389.fgene.2020.00214.

[2]HU X Z, GE R, FENG Z X.Recognizing five molecular ligand-binding sites with similar chemical structure[J].Journal of Computational Chemistry, 2020, 41(2):110-118. DOI: 10.1002/jcc.26077.

[3]LIU L,HU X Z ,FENG Z X,etal. Prediction of acid radical ion binding residues by K-nearest neighbors classifier[J]. BMC Molecular and Cell Biology,2019,20(S3):52.DOI:10.1186/S12860-019-0238-8.

[4]SCHNEIDER S, MARLES-WRIGHT J, SHARP K H,et al. Diversity and conservation of interactions for binding heme in b-type heme proteins[J]. Natural Product Reports, 2007, 24:621-630.

[5]KENNETH L. An overview on GPCRs and drug discovery: Structure-based drug design and structural biology on GPCRs [J].Methods in Mollecular Biology,2009,552:51-66.

[6]BIKIE D E, BOECHI L, CAPAPECE L, et al. Modeling heme proteins using atomistic simulations[J]. Phys Chem Chem Phys 2006,8:5611-5628.

[7]AJAY A P G,MAURICIO L,BENJAMIN F S.HeMoQuest: A webserver for qualitative prediction of transient heme binding to protein motifs[J].BMC Bioinformatics,2020,21(1): 124.

[8]TERWILLIGER N B. Functional adaptations of oxygen-transport proteins[J]. Experimental Biology and Medicine 1998, 201:1085-1098.

[9]REEDY C J,GIBNERY B R.Heme protein assemblies[J]. Chem Rev 2004, 104:617-649.

[10]MENSE S M, ZHANG L. Heme: A versatile signaling molecule controlling the activities of diverse regulators ranging from transcription factors to MAP kinases[J]. Cell Research, 2006,16:681-692.

[11]SCHNEIDER S,MARLES-WRIGHT J, SHARP K H, et al. Diversity and conservation of interactions for binding heme in b-type heme proteins [J]. Nature Product Reports, 2007, 24:621-630.

[12]EBERT J, ALTMAN R. Robust recognition of zinc binding sites in proteins[J].Protein Science, 2008,17(1):54-65.

[13]SOBOLEV V, EDELMAN M. Web tools for predicting metal binding sites in proteins[J]. Israel Journal of Chemistry, 2013, 53(3/4): 166-172.

[14]BABOR M, GERZON S, RAVEH B, et al. Prediction of transition metal—binding sites from apo protein structures [J]. Proteins: Structure, Function, and Bioinformatics, 2008, 70(1): 208-217.

[15]ZHU Y H, HU J, QI Y. Boosting granular support vector machines for the accurate prediction of protein-nucleotide binding sites[J].Combinatorial Chemistry & High Throughput Screening,2019,22(7):455-469.

[16]YANG J, ROY A, ZHANG Y. BioLiP: A semi-manually curated database for biologically relevant ligand-protein interactions[J]. Nucleic Acids Research, 2013, 41(D1): D1096-D1103.

[17]HU X Z, DONG Q W.Recognizing metal and acid radical ion-binding sites by integrating ab initio modeling with template-based transferals[J].Bioinformatics,2016,32(23): btw396

[18]CAO X Y, HU X Z, ZHANG X J, etal. Identification of metal ion binding sites based on amino acid sequences[J]. Plos One,2017,12(8):13.

[19]LIU R, HU J. HemeBIND: A novel method for heme binding residue prediction by combining structural and sequence information[J]. BMC Bioinformatics, 2011, 12: 207.

[20]LIU R, HU J. HemeBIND: Computational prediction of Heme-binding residues by exploiting residue interaction network[J]. PLoS One,2011;6(10):e25560.

[21]ZHAO Z Q,XU Y H, ZHAO Y, SXGBsite: prediction of protein-ligand binding sites using sequence information and extreme gradient boosting[J]. Genes (Basel), 2019, 10(12):965.

[22]LIU T,LIN Y ,WEN X,et al.BindingDB: A web-accessible database of experimentally determined protein-ligand binding affinities[J]. Nucleic Acids Research,2007,35:198-201.

[23]徐克学.生物数学[M].北京:科学出版社,2002:40-50.
XU Kexue. Biological mathematics[M]. Beijing: Science Press,2002:40-50.

[24]李彩艳.蛋白质超二级结构库的建立及其序列统计分析[D].呼和浩特:内蒙古工业大学,2007:5.
LI Caiyan. Construction of a protein supersecondary structure library and sequence statistical analysis[D]. Hohhot: Inner Mongolia University of Technology,2007:5.

[25]姜雪,胡秀珍.打分矩阵方法在β一发夹模体识别中的应用[J]. 生物信息学, 2007(4):156-158.
JIANG Xue, HU Xiuzhen. Application of grade matrix method in the mode identification of β[J]. China Journal of Bioinformatics, 2007(4):156-158.

[26]王春莲,张晓东.基于打分矩阵的多类蛋白质折叠子的预测[J]. 生物信息学, 2011(9):42-45.
WANG Chunlian, ZHANG Xiaodong. prediction of multi-protein folding based on scoring matrix[J]. Chinese Journal of Bioinformatics, 2011(9):42-45.