第 4 卷◆第 3 期◆版本 1.0◆2022 年
文章类型：论文丨刊号（ISSN）：2705-1102(P) / 2705-1110(O)

基于序列信息的血红素配体结合残基识别

李彩艳　马勇　张慧敏　姚可　丁海麦*
包头医学院
DOI:10.12238/bmtr.v4i3.5204

[摘　要] 血红素在生命活动中发挥重要作用,识别血红素结合位点的残基,有助于更好地理解血红蛋白的生物学功能,揭示血红蛋白作用机制,并为蛋白质设计提供有价值信息。本文通过整理Biolip数据库中,血红素配体与蛋白质结合残基及其近邻残基信息,以血红素结合残基和非结合残基的氨基酸信息、亲疏水性、带电性、氨基酸分类等性质作为指标,并利用矩阵打分方法判别血红素结合位点,取得较好结果。
[关键词] 蛋白质配体；位点识别；血红素；序列信息
中图分类号：Q61　文献标识码：A

Recognition of Heme Ligand Binding Residues Based on Sequence Information

Caiyan Li　Yong Ma　Huimin Zhang　Ke Yao　Haimai Ding*
Baotou Medical College

[Abstract] Heme plays an important role in life activities. The identification of residues of heme binding sites is helpful to better understand the biological function of hemoglobin, reveal the mechanism of hemoglobin action, and provide valuable information for protein design. In this paper, the information of heme ligand and protein binding residues and their neighboring residues in Biolip database were sorted out, and the amino acid information, hydrophilicity and hydrophobicity, electric property and amino acid classification of heme binding residues and non binding residues were as indicators, and the matrix scoring method was used to identify the heme binding sites and good results were obtained.
[Key words] protein ligand; site recognition; heme; sequence information

引言

蛋白质经常通过与其他小分子配体相互作用参与各种生物过程,所以准确预测蛋白质配体结合位点,对理解蛋白质功能具有非常重要意义[1]。但基于实验的方法识别蛋白质质配体结合位点,不仅昂贵且耗时。利用计算机结合数学统计等方法可以快速预测结合位点,可以作为实验室测定一种补充,为实验室测定蛋白质配体结合位点提供理论指导。血红素是铁卟啉化合物,存在于所有生物中,其功能具有多样性,在很多生命活动中起着不可替代作用[2-4]。例如,骨质疏松症的发病原因可能与绝经后体内血红素中铁蓄积状态有关[5],肝癌与血红素加氧酶1表达有关[6],故很有必要开发预测血红素方法。

1 材料及方法

1.1数据集的建立

本文选取Biolip数据库中含血红素蛋白链同源性小于40%,分辨率高于3Å,氨基酸长度超过50的蛋白链254条。考虑到血红素和残基的结合,不仅和结合残基相关,也和结合残基附近氨基酸有关,所以采用滑动窗口截取序列片段。设定长度为K的固定滑窗窗口,K一般选取奇数,可以把不同长度的氨基酸序列分割成固定长度。设定滑窗的中心位置作为靶点,从序列的第一个氨基酸其作为靶点,以此类推,直到序列的最后一个氨基酸。长度不够K时,左右末端空位补齐。若靶点位置为血红素结合位点,则将该样本定义为正样本,若靶点位置为非血红素结合位点,则将该样本定义为负样本,本文使用的数据集有4589个血红素结合位点,66137个非血红素结合位点。由此,得到4589个长度为K的正样本,66137个长度K的负样本。由于负集大约是正集的14倍,为平衡正负集,从负集中随机选取与正集数量相等的14个集合进行预测,结果取14个的平均。

1.2氨基酸的物化性质

前人在蛋白质配体结合残基的研究中[4-6],发现氨基酸信息、电荷信息、亲疏水信息、氨基酸分类等信息作为参数时,可以较好地识别一些蛋白质配体结合残基。参照前人做法,考虑血红素结合残基及其附近的将氨基酸信息、亲疏水信息(分为6类见表1)、电荷信息(分为3类见表2)、氨基酸分类信息(分为7类见表3)。

表1 氨基酸亲疏水性

Classification	Amino Acids	Classification	Amino Acids
Strongly hydrophilic	R,D,E,N,Q,K,H	Proline	P
Strongly hydrophobic	L,I,V,A,M,F	Glycine	G
Weakly hydrophilic	S,T,Y,W	Cysteine	C

表2 氨基酸带电性

Classification	Amino Acids
positive charged	K,R,P
negative charged	D,E
uncharged	N, Q, H, L, I, V, A, M, F, S, T, Y, W, C, G

表3 氨基酸分类

Classification	Amino Acids	Classification	Amino Acids
Aliphatic	G A V L I	Basic	K R H
Aromatic	F W Y	Acidic	D E N Q Z
Sulphur	C M	Aliphatic hydroxyl	S T
proline	P		

1.3方法

位置权重打分方法是一种分类方法,在蛋白质二级结构研究中曾被应用。具体算法如下：

函数为：$s = \dfrac{\sum_{i=1}^{L} C_i f_{i,j}}{\sum_{i=1}^{L} C_i f_{i,max}}$

其中 $C_i = \dfrac{100}{\ln n}(\sum_{i=1}^{n} p_{i,j} \ln p_{i,j} + \ln n)$

其中 n 为指标个数，$p_{i,j}$ 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各位置出现的氨基酸频次，N_i 表示在第i个位点上出现的氨基酸频次的总和，$f_{i,max}$ 表示位置频数矩阵的第i列的最大值。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,序列片段属于 s 高的类别。

参考文献[1],预测结果采用 $Recal$、$Prcision$、ACC、$F1-score$、MCC 指标,分别定义为：

$$Recal = \frac{TP}{TP+FN} \qquad Prcision = \frac{TP}{TP+FP}$$

$$F1-score = \frac{2 \times Recal \times Prcision}{Recal + Prcision}$$

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

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

其中,TP 表示正确预测正集的数量；FN 表示将正集识别为负集的数量；TN 表示正确识别负集的数量；FP 表示将负集识别为正集的数量。

2 结果

采用打分矩阵方法,分别以20种氨基酸、带电性、亲疏水性、氨基酸分类为指标,及其某些指标的2-mer和3-mer指标,对正负集进行分类。移动窗口长度从9到19的奇数,都进行了判别,发现亲疏水性和带电性移动窗口选取13时结果较好,氨基酸分类和20氨基酸指标选取移动窗口长度为11时结果较好,其他窗口长度也就未给出,结果见表4。

3 讨论

由以上结果发现,对于亲疏水性和氨基酸分类,3-mer判别效果较好,而对于带电性,1-mer效果好,氨基酸指标的2-mer和1-mer差异不大。总体而言,氨基酸分类指标和20种氨基酸指标在区别血红素结合氨基时,使用PSSM方法,结果相对较好,氨基酸分类3-mer精度指标MCC达到0.23,总精度ACC也达到了61.17%,20种氨基酸精度1-mer指标MCC达到0.32,而使用20种氨基酸r指标时,总精度ACC达到了65.59%,Precision和F1-score分别为74.81%和53.49%,。文献综合使用PSSM、DWT、DCT、PSA特征预测时,时总精度ACC为76.49%且MCC为0.407,好于本文结果,但Precision和F1-score分别为34.07%和47.56%,远低于本文结果。

本文通过研究分析血红素结合残基和非结合残基的一些特性,并在此基础上通过PSSM方法进行预测,能在一定程度上识别血红素结合残基,但本文预测结果还有待提高,以后将综合20种氨基酸信息以及氨基酸分类3-mer等相对预测结果好的指标结合其他方法进行预测。

[基金项目]

内蒙古大学生创新创业训练计划项目(S202119127006,S202119127007,S202119127013)；内蒙自然科学基金项目(2020MS08015)；内蒙古自治区高等学校科学研究项目(ZSZX21303)；包头医学院科学基金项目(BYJJ-SZZX202011,BYJJ-ZROM202209)。

表 4 PSSM 采用不同参数判别结果

方法	参数	窗口长度	Recal(%)	Precision(%)	ACC(%)	F1-score(%)	MCC
PSSM	亲疏水性 1-mer	13	29.68	61.47	55.56	39.98	0.13
	亲疏水性 2-mer	13	27.78	64.52	56.29	38.63	0.15
	亲疏水性 3-mer	13	24.43	67.82	56.43	35.88	0.17
	带电性 1-mer	13	32.53	59.80	55.30	42.11	0.12
	带电性 2-mer	13	44.80	56.86	55.36	50.10	0.11
	带电性 3-mer	13	42.84	54.03	53.18	47.78	0.07
	氨基酸分类 1-mer	11	42.23	65.57	60.00	51.35	0.21
	氨基酸分类 2-mer	11	28.46	87.56	58.01	40.34	0.20
	氨基酸分类 3-mer	11	47.55	65.35	61.17	55.01	0.23
	氨基酸 1-mer	11	53.98	70.45	65.59	60.95	0.32
	氨基酸 2-mer	11	42.01	74.81	63.85	53.49	0.31

[参考文献]

[1]Hubbard,R.E.(ed.)(2006)Structure-Based Drug Discovery: An Overview [J].RSC Publishing, Cambridge, UK.

[2] 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].Front Genet 2020 ,Mar 19;11:214.

[3] HU, X Z, GE, R, FENG, Z X. Recognizing five molecular ligand-binding sites with similar chemical structure[J].J Comput Chem,2020 Jan 15;41(2):110-118.

[4] LIU, L, HU, XZ ,FENG, ZX,etal. Prediction of acid radical ion binding residues by K-nearest neighbors classifier[J]. BMC Mol Cell Biol,2019 Dec 11;20(Suppl3):52.

[5]徐又佳,XIHuang.降低“铁过载”:一个防治绝经后骨质疏松的新方案[J].中华骨质疏松和骨矿盐疾病杂志,2012,(5):1-6.

[6]王波,于景翠,周虹.血红素加氧酶-Ⅰ与实体肿瘤[J].医学分子生物学杂志,2008,(4):11.

作者简介：

李彩艳(1975--),女,汉族,山西汾阳人,研究生,包头医院,副教授,研究方向：生物信息。

***通信作者：**

丁海麦(1975--),男,汉族,山西永济人,研究生,包头医学院,副教授,研究方向：生物化学。