doi:10. 3969/j. issn. 1672 - 5565. 2013. 03. 04

用 ID-SVM 预测蛋白质的 ATP 结合位点

包文荣, 赵巨东*
(内蒙古工业大学理学院,内蒙古 呼和浩特 010051)

摘 要:从蛋白质序列出发,对经 Dr. G. P. S. Raghava 整理和使用过的 168 条非冗余的 ATP 与蛋白质结合氨基酸序列进行分段,对 ATP 与蛋白质结合位点进行了统计分析。在此基础上,利用 20 种氨基酸的亲疏水性将 20 种氨基酸约化为 6 类。以氨基酸组分和 6 类亲疏水紧邻为参数,用多样性增量(ID)方法将氨基酸组分和 6 类亲疏水紧邻降维并将降维后的特征参数输入支持向量机中运算,本文运算结果显示用氨基酸组分 ID 值和 6 类亲疏水紧邻 ID 值共同作为特征参数结果最优,在七交叉检验下的预测总精度达到了 99.67%,相关系数达到 0.993 4,好于前人的预测结果。

关键词: 多样性增量;支持向量机;6 类亲疏水紧邻;三磷酸腺苷(ATP)
中图分类号:Q61 **文献标识码**:A **文章编号**:1672 - 5565(2013) - 03 - 181 - 05

Use ID-SVM method to prediction ATP binding residues of a protein

BAO Wen-rong, ZHAO Ju-dong*
(*Faculty of Science, Inner Mongolia University of Technology, Hohhot* 010051, *China*)

Abstract: starting from the protein sequence, Dr. GPSRaghava are analyzing the 168 non-redundant ATP and protein segmentation which these are organized and used, statistical and analysis of ATP and protein-binding sites. On this basis, the use of 20 kinds of amino acids hydrophobicity of the 20 amino acids is reduced to 6. Close to the amino acid composition and 6 Hydrophobicity parameter increment of diversity (ID) close to the amino acid composition and 6 Hydrophobicity dimensionality reduction and dimensionality reduction characteristic parameter input support vector machine computing, this article the result of the operation is displayed next to the amino acid component ID value and 6 Hydrophobicity ID value as a common characteristic parameters of the best results, seven cross-examination under the forecast total accuracy of 99.67%, correlation coefficient of 0.9934, better than the previous forecast results.

Key words: Increment of diversity; Support Vector Machines; Diad; Seven Crosscheck; Adenosine-triphosphate

许多蛋白质的功能取决于其与小分子或配体的相互作用,ATP 是其中一种重要的配体,在蛋白质结构功能预测方面起着关键作用。ATP 在分子生物学领域是一种重要的能量分子和辅酶, ATP 与蛋白质结合为细胞内运输、肌肉收缩、细胞运动、调控各种代谢过程起重要作用。用传统实验的方法鉴别 ATP 与蛋白质结合位点是非常昂贵耗时和耗力的,因此就需要一种新的方法来预测,这种新方法(生物信息学)是借用计算机软件和计算机编程来实现预测的。用生物信息学方法研究 ATP 与蛋白质的相互作用是可行的,且目前预测准确率也在不断提高。ATP 与蛋白质结合位点的预测是重要的,它可以为医学相关领域提供很高价值的计算数据和理论依据,为相关领域在生物化学实验方面的研究提供更多的信息。Dr. G. P. S. Raghava 的科研小组 2009 年发表的一篇有关 ATP 与蛋白质结合位点预测的文章,该文章利用 ATPint 软件程序包预测总精度达 75.25%,相关系数 MCC 值为 0.5[1]。

收稿日期:2013 - 04 - 03;**修回日期**:2013 - 05 - 17.
基金项目:国家自然科学基金资助项目(51068020)。
作者简介:包文荣,女,内蒙古牙克石,硕士研究生,研究方向:生物信息学;E-mail:bwr2013@126. com.
* **通讯作者**:赵巨东,教授。E-mail:jdzhao@imut. edu. cn.

1 材料和方法

1.1 数据库

本文选取的数据集来自 SuperSite encyclopedia Dr. G. P. S. Raghava 小组从中提取 360 条与 ATP 结合的蛋白质序列利用 CD-HIT 程序去除不与蛋白质结合的无用序列[2],得到 267 条非冗余的 ATP 与氨基酸链结合的序列。其中,任意两条序列的相似度小于等于 40%;再用 LPC 软件检测这些序列[3],去掉对判别结合位点无意义的多余部分,最终得到 168 条非冗余的 ATP 与氨基酸结合序列(http://www. imtech. res. in/raghava/atpdataset)。这 168 条氨基酸序列中每个字母代表一个氨基酸,小写字母表示与 ATP 结合的氨基酸,大写字母表示不与 ATP 结合的氨基酸。如 168 条氨基酸序列中的一条:GLPTHLYKNFTVQELALKLKGKNQEFCLTAFMSGRS LVRACLSDAG HEHDTWF dtMLgfAIsAYAlKSRIALT VEDSPYPGTPGDLLELQiCPLNGYCE;在这条序列中与 ATP 结合的氨基酸共有 5 个,经过统计可知这 168 条序列中共有小写字母 3 058 个,小写字母表示与 ATP 结合的氨基酸,即有 3 058 个结合位点;大写字母 56 250 个,大写字母代表不与 ATP 结合的氨基酸。在这 168 条序列中,序列的长短不一,最长的一条序列由 959 个氨基酸组成,最短的一条序列只由 78 个氨基酸组成。

数据集按照下面的方法进行整理:为便于预测,将氨基酸序列切成片段,之前我们对各种长度的片段做过预测,发现片段的长度为 17 时预测结果最好:所以本文选定 17 片段长。用窗口滑动法截取片段,把每一条氨基酸序列的头和尾端各添加 8 个空位,这样就可确保每条氨基酸链上的每个氨基酸都能被窗口覆盖到。从序列头端的第一个空位开始滑动窗口,窗口宽度为 17,步长为 1。因为本文把 ATP 与氨基酸序列结合的位置为片段中心位置的片段做为正集,(除中心位点外其他位点作为结合位点的预测我们之前也做过,但发现预测结果都不如中心位点的预测结果好)即 17 片断中所有第 9 个位置是小写字母的片段作为正集,这样正集共有 3 058 个,其他 17 片段作为负集共有 56 250 个。由于正负集片段数量相差悬殊,造成数据不平衡,会对下面的预测带来很大问题,因此本文将负集随机分成了几部分,使负集片段数量与正集片段数量相当,本文对负集随机抽取 3 060 个片段进行研究和预测。

1.2 特征参数的提取和优化

1.2.1 氨基酸组分

构成蛋白质的氨基酸有 20 种,用 A、C、D、E、F、G、H、I、K、L、M、N、P、Q、R、S、T、V、W、Y 表示 20 种氨基酸,其中 n_1 表示氨基酸 A 的频数, n_2 表示氨基酸 C 的频数,…,n_{20}表示氨基酸 Y 出现的频数。

定义:片断中氨基酸出现的频数为 $n_i(i=1,2,3,\cdots 20)$,N 表示氨基酸的总数。

那么氨基酸组分

$$P_1 = \frac{n_i}{N} \text{ 其中}(i = 1,2,\cdots 20) \qquad (1)$$

所以可知氨基酸组分是一个 20 维的向量。

1.2.2 二联体组分

用氨基酸组分为特征可以比较简单地表示一条蛋白质,然而,却丢失了各氨基酸之间的相关信息。为此,我们引入氨基酸二联体的概念。所谓二联体即为一氨基酸序列中两个氨基酸的组合成为一个二联体。

紧邻二联体组分:为同一序列中相邻氨基酸二联体,因为氨基酸有 20 种,那么二联体就是两种氨基酸的组合,应该有 20x20 = 400 种。m_1 表示紧邻二联体 AA 出现的频数,m_2 表示紧邻二联体 AC 出现的频数,……,m_{400}表示紧邻二联体 YY 出现的频数。

定义:片段中氨基酸紧邻二联体出现的频数为 $m_j(j=1,2,3\cdots 400)$,M 表示紧邻二联体总数。

紧邻二联体组分

$$P_2 = \frac{mj}{M} \text{ 其中}(j = 1,2,\cdots,400) \qquad (2)$$

所以紧邻二联体组分是 400 维的向量。

约化后的氨基酸紧邻二联体组分:就是将 20 种氨基酸按照亲疏水性分为 6 类分别用字母 R、V、S、P、G、C 表示,那么 6 类亲疏水紧邻共有 6 × 6 = 36 种。b_1、b_2,…,b_{36} 分别表示 RR、RV、RS、RP、RG、RC、VR、VV、VS、VP、VG、VC、SR、SV、SS、SP、SG、SC、PR、PV、PS、PP、PG、PC、GR、GV、GS、GP、GG、GC、CR、CV、CS、CP、CG、CC 等 36 种 6 类亲疏水紧邻二联体出现的频数

定义:6 类亲疏水紧邻二联体出现的频数是 B 表示 6 类亲疏水紧邻二联体的总数。

6 类亲疏水紧邻二联体组分

$$P_3 = \frac{b_k}{B} \text{ 其中}(k = 1,2,\cdots 36) \qquad (3)$$

所以 6 类亲疏。紧邻二联体组分是 36 维的向量。

1.3 特征参数的降维

用 SVM 做预测时,输入到 SVM 中的参数维数过高会导致 SVM 过训练,因此需要对输入到 SVM 中的数据进行降维处理,本文利用生物数学中多样

性量和多样性增量的概念[4]，应用多样性增量原理对氨基酸组分（20 维）、6 类亲疏水紧邻（36 维）和紧邻（20×20 维）进行降维处理。这种降维的优势在于能把参数的高维数很有效的降低又能尽量减少在降维过程中丢失信息。

氨基酸组分（20 维）降维：先做出正集和负集的标准源，本文把 20 种氨基酸组分的每一维的加和作为标准源向量的一个元素，（做标准源时，应去除要与标准源做多样性增量计算的那个片段）这样构成两个标准源分别为：$s_1 = (a_1, a_2, \cdots a_{20})$；$s_2 = (b_1, b_2, \cdots b_{20})$ 多样性量公式为：

$$D(X) = D(n_1, n_2, \cdots, n_s) = N\log_b^N - \sum_{i=1}^{s} n_i \log_b^{n_i} \tag{4}$$

而这两个标准源的多样性量分别为 $D(S_1)$ 和 $D(S_2)$，用任意片段氨基酸组分 P_1 多样性量 $D(P_{1i})$ 与 $D(S_1)$ 和 $D(S_2)$ 分别作多样性增量运算，多样性增量公式为：

$$ID = D(X + Y) - D(X) - D(Y) \tag{5}$$

这样可以得到这个片段氨基酸组分的 2 个 *ID* 值 ID_1 和 ID_2。

紧邻二联体组分（400 维）降维：做出正集和负集的标准源，本文把紧邻二联体组分 P_2 的每一维的加和作为标准源向量的一个元素，（做标准源时，应去除要与标准源做多样性增量计算的那个片段）这样构成两个标准源为：$S_3 = (d_1, d_2 \cdots d_{400})$；$S_4 = (e_1, e_2 \cdots e_{400})$。用公式（4）计算 S_3 和的多样性量为 $D(S_3)$ 和 $D(S_4)$，用任意片段紧邻二联体组分 P_{2i} 的多样性量 $D(P_{2i})$ 与 $D(S_3)$ 和 $D(S_4)$ 分别作多样性增量运算，代入公式（5）可以得到任意片段紧邻二联体组分的 2 个 ID 值 ID_3 和 ID_4。

预测配体与氨基酸结合的位置通常都与氨基酸的物化性质有很大的关系，基于这点，将 20 种氨基酸按照亲疏水性约化为 6 类，并用约化为 6 类后的紧邻二联体组分作为一个参数进行预测。

6 类亲疏水紧邻二联体组分（36 维）降维：把 6 类亲疏水紧邻二联体组分 P_3 每一维的加和作为标准源向量的一个元素，（做标准源时，应去除要与标准源做多样性增量计算的那个片段）这样构成两个标准源为：$S_5 = (g_1, g_2 \cdots e_{36})$；$S_6 = (h_1, h_2 \cdots e_{36})$。用公式（4）计算 S_5 和 S_6 多样性量为 $D(S_5)$ 和 $D(S_6)$，用任意片段 6 类亲疏水紧邻二联体组分 P_{3i} 的多样性量 $D(P_{3i})$ 与 $D(S_5)$ 和 $D(S_6)$ 分别作多样性增量运算，代入公式（5）可以得到任意片段 6 类亲疏水紧邻二联体组分的 2 个 ID 值 ID_5 和 ID_6。

ID 表征了样品 *X* 和标准源信息参数分布的差异性，它提供了样品 *X* 特征的数学表示，由于信息参数的维数可能很大，这个方法自然地包含了把分类特征中冗余和不相关的特征剔除。

1.4　支持向量机（SVM）方法

支持向量机方法（Support Vector Machines，SVM）是由 V. Vapnik[5-6] 等人提出来的一类新型机器学习方法，是一种以统计学理论为基础，结构风险最小化的学习机学习方法，它在解决小样本 、非线性及高维模式识别问题中表现出许多特有的优势。目前已被成功地应用于蛋白质结构预测、蛋白质亚细胞定位及蛋白质折叠子的分类等多方面[7-10]。

输入支持向量机的参数为前面列举的氨基酸组分的 ID 值（2 维）、紧邻二联体组分的 ID 值（2 维）和 6 类亲疏水紧邻 ID 值（2 维），所以参数共 6 维即（ID_1、ID_2、ID_3、ID_4、ID_5、ID_6）将这三种参数分成两种组合 $Q_1 = (ID_1, ID_2, ID_3, ID_4)$ 和 $Q_2 = (ID_1, ID_2, ID_5, ID_6)$。把 Q_1 和 Q_2 分别输入 SVM 中进行预测。

1.5　精确度评价指标

本文做的是 7 折交叉检验：把正集和负集数据都随机分成 7 部分，分别从正、负集的这 7 份中拿出 6 份作为正集的训练集和负集的训练集，把正、负集中剩余的 1 份作为正集的检验集和负集的检验集。如此反复循环 7 次确保每一部分都被检验过，评价指标见表 1。

表 1　评价指标

Table 1　Evaluation

符号名称	描述
true positive（TP）	真阳性具有功能 F 的蛋白质预测为具有功能 F
false positive（FP）	假阳性不具有功能 F 的蛋白质预测为有功能 F
true negative（TN）	真阴性不具有功能 F 的蛋白质预测为不具有功能 F
false negative（FN）	假阴性具有功能 F 的蛋白质预测为不具有功能 F

敏感性：$sensitivity = \dfrac{TP}{TP + FN}$

特异性：$sensitivity = \dfrac{TN}{TN + FP}$

总精度：$accuracy = \dfrac{TP + TN}{TP + TN + FP + FN}$

相关系数：

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

2 结果与讨论

2.1 结果

用 ID 结合 SVM 方法预测 ATP 与蛋白质结合位点，将正集的 Q_1 向量的每一维前加上序列号和冒号即变成（$1:ID_1\ 2:ID_2\ 3:\ ID_3\ 4:ID_4$），负集 Q_1 同样每一维前面加上序列号（$1:ID_1^*\ 2:ID_2^*\ 3:ID_3^*\ 4:ID_4^*$）；再将正集数据和负集数据用数字（1）和（2）区分并排列好，格式是：（1）$1:ID_1\ 2:ID_2\ 3:\ ID_3\ 4:ID_4$；（2）$1:ID_1^*\ 2:ID_2^*\ 3:ID_3^*\ 4:ID_4^*$。将整理好的数据直接输入到 SVM 中即可；$Q_2$ 输入支持向量机的方式与 Q_1 相同，这是不再详述。此外，用 SVM 做预测很重要的就是确定参数 c 和 gamma 的最佳值，最后产生一个分类器，用这个分类器对结合位点进行预测，检验分类器的推广能力。注意用 SVM 做预测时，做交叉检验时需要把数据分成训练集和测试集，例如做 7-折交叉检验，就是把数据随机分成 7 份，每一份数据做一次独立检验，这样做 7 次独立检验后的平均值就是我们最后得到的 7-折交叉检验的结果（见表 2）。

表 2 选取不同特征的 ID 值组合作为参数用 SVM 预测 ATP 的结合位点

Table 2 select the ID values of the different characteristics of the composition as a parameter SVM prediction of ATP binding sites

组合方式		Sn	Sp	Ac	Mcc
二种特征 ID 组合(4 维)即	氨基酸组分 ID、紧邻二联体 ID	99.15%	99.50%	99.33%	0.986 6
	氨基酸组分 ID、6 类亲疏水紧邻 ID	99.51%	99.83%	99.67%	0.993 4

由表 2 可知，以氨基酸组分 ID、6 类亲疏水紧邻 ID 组合参数预测结果最佳。

2.2 推广应用

本文选用氨基酸组分的 ID 值和 6 类亲疏水紧邻 ID 值共同为参数用 SVM 进行预测，经过 7-折交叉检验得到了相当理想的预测结果，那么这个方法是否具有推广性和实用性，就需要寻找一个新的数据库，来验证我们的方法。新的数据库来源[11]，网址为 http://biomine. ece. ualberta. ca/ATPsite/（说明：这个数据库跟文章中的数据库有重复的氨基酸序列，但这个数据库要比文章所用的数据库包含的氨基酸序列更丰富）。运用同样的整理数据库的方法最终得到 3 156 个 17 片段长度的正集和 80 409 个 17 片段长度的负集，为了使数据平衡，对负集进行了随机选取，最终选取 3 072 个片段作为负集。经过 7-折交叉检验最终预测结果见表 3。

表 3 选取氨基酸组分 ID 和 6 类亲疏水紧邻 ID 作为特征参数用 SVM 预测的结果

Table 3 Select the amino acid component ID and (ADW) amino acid diad ID as the characteristic parameters of SVM prediction results

片段长度	正集片段总数	负集片段总数	TP	FP	TN	FN	Sn	Sp	MCC	Ac
17	3 156	3 072	3 131	25	3 046	26	99.18%	99.19%	0.983 6	99.18%

此方法预测的结果好于参考文献[11]中采用的 ATPsite 方法所进行的预测，该方法预测的最好结果是总精度达到 85.40%，相关系数为 0.433。而用本方法可看到总精度提高了将近 14 个百分点，相关系数也大大得到提高。

3 结论

根据 ATP 与蛋白质结合不同位置氨基酸序列的信息差异，选用不同参数，运用 ID 结合 SVM 算法取得了比较理想的预测效果。结果表明，对蛋白质上 ATP 结合位点的预测，参数选取很重要，本文研究中不仅选取了氨基酸组分信息，而且考虑了氨基酸之间位置关系即二联体组分信息，特别是约化后的二联体组分的选取对于预测的成功至关重要。

参考文献（References）

[1] Jagat S Chauhan, Nitish K Mishra, Gajendra PS Raghava. dentification of ATP binding residues of a protein from its primary se-

quence[J]. BMC Bioinformatics ,2009,10:434-438.

[2] Sobolev V, Sorokine A, Prilusky J. Automated analysis of interatomic contacts in proteins[J]. Bioinformatics, 1999, 15:327-332.

[3] Bauer RA, Günther S, Jansen D . dictionary of metabolite and drug binding sites in proteins[J]. Nucl Acids Res,2009, 37:95-200.

[4] McLachlan GJ. Discrimanant. Analysis and Statistical Pattern Recognition[J]. New YorkWiley, 1992,1:519-526.

[5] Vapnik V. The nature of statistical learning theory[M]. New York:Springer,1995.

[6] Vapnik V. Statistical learning theory[M]. Wiley-Interscience, 1998.

[7] Hu X Z, Li Q Z. Using support vector machine to predict β-turns and γ-turns in proteins[J]. Computational Chemistry,2008, 29(12): 1867-1875.

[8] Chou K C, Cai Y D. Using functional domain composition and support vector machines for prediction of protein subcellular location [J]. Journal of Biological Chemistry, 2002, 227: 45765-45769.

[9] Ding C H Q, Dubchak I. :Multi-class protein fold recognition using support vector machines and neural networks[J]. Bioinformatics,2001, 17(4): 349-358.

[10] Shi J Y, Pan Z, Zhang S W, Liang Y. Protein fold recognition with support vector machines fusion network[J]. Progress in Biochemistry Biophysics,2006, 3(2): 155-162.

[11] Ke Chen, Marcin J Mizianty, Lukasz K Kurgan. Accurate prediction of ATP-binding residues using sequence and sequence-derived structural descriptors[J]. BIBM,2010,9:43-48.

[12] Peter E. Wright, Lo Jolla, CA. Structure and identification of ADP-ribose recognition motifs of APLF and role in the DNA damage response[J]. PNAS ,2010, 22:409-433.

[13] Feng Z P. :Prediction of the subcellular location of prokaryotic protein based on a new representation of the amino acid composition[J]. Biopolymers,2001,58:491-499.

[114] Nakai,K, Kanehisa M. A knowledge base for predicting protein localization sites in eukaryotic cells[J]. Genomics,1992,14: 897-911.