Hans Journal of Computational Biology 计算生物学, 2014, 4, 1-11
Published Online March 2014 in Hans. http://www.hanspub.org/journal/hjcb
http://dx.doi.org/10.12677/hjcb.2014.41001

Hans汉斯

Prediction of Four Kinds of Supersecondary Structures in Enzymes by Using Ensemble Classifier Based on SVM

Sujuan Gao, Xiuzhen Hu*

College of Sciences, Inner Mongolia University of Technology, Huhhot
Email: *hxz@imut.edu.cn

Received: Feb. 28th, 2014; revised: Mar. 7th, 2014; accepted: Mar. 11th, 2014

Abstract

Enzymes are a kind of protein that has catalytic function. The study of supersecondary structures in enzymes plays an important role in the structure and function of enzymes. Based on enzyme sequence information, four kinds of supersecondary structures in enzymes were researched for the first time. Amino acids of sites and dipeptide components of sites were selected as parameters, for five selections of the best fixed-length pattern, the predictive results in 7-fold cross-validation were not ideal by using scoring function method; scores were selected as input parameters of support vector machine (SVM); the results were fused with weighted factors by using ensemble classifier; the better performance was obtained; the overall prediction accuracy was 72.64% and the Matthews correlation coefficient was above 0.57. Therefore, ensemble classifier based on SVM is an effective method to predict four kinds of supersecondary structures in enzymes.

Keywords

Enzyme; Supersecondary Structure; Scoring Function; Support Vector Machine; Ensemble Classifier

基于支持向量机的整体分类器算法预测酶蛋白质中四类简单超二级结构

高苏娟，胡秀珍

*通讯作者。

内蒙古工业大学理学院，呼和浩特
Email: *hxz@imut.edu.cn

收稿日期：2014年2月28日；修回日期：2014年3月7日；录用日期：2014年3月11日

摘　要

酶是一种具有催化功能的蛋白质，研究酶蛋白质中的超二级结构对研究酶的结构及功能有重要作用。本文从酶蛋白质序列出发，首次对酶蛋白质中的四类简单超二级结构进行研究。以位点氨基酸及其紧邻关联为参数，选取五种序列片段截取方式，采用7-交叉检验，使用矩阵打分方法预测的结果不理想；将矩阵打分值作为特征参数输入支持向量机，并用整体分类器进行加权融合，得到了较好的预测结果，预测总精度达到72.64%，Matthew's相关系数在0.57以上，因此，基于支持向量机的整体分类器方法是一种有效的预测酶蛋白质中超二级结构的方法。

关键词

酶蛋白质；超二级结构；矩阵打分；支持向量机；整体分类器

1. 引言

酶是活细胞内产生的具有高度专一性和催化效率的蛋白质，又称为生物催化剂，生命活动中引起新陈代谢的千千万万的化学变化几乎都是在酶的催化下进行的，酶与生命现象息息相关。因此，对于酶结构及功能的研究对生命科学的发展至关重要。近年来研究者们在酶蛋白质分子的功能研究上获得了较大的成果，比如关于酶与非酶[1] [2]、酶的亚类分类[3]-[7]方面的研究。但是对酶蛋白质结构的研究还相对较少，只有 2011 年 Liu 和 Hu[8]、2012 年 Long 和 Hu[9]对酶蛋白质中的β发夹模体进行了识别。

酶作为一种具有催化功能的蛋白质，它具有一般蛋白质分子所有的一级结构和高级结构，蛋白质的超二级结构(supersecondary structure)是指两个或几个二级结构单元被连接多肽(loop)连接起来，进一步组成有特殊几何排列的局域空间结构，简称 Motif[10]。简单超二级结构分为 β-loop-β、β-loop-α、α-loop-α 和 α-loop-β 四类。由于超二级结构是 α 螺旋、β 折叠简单排列形成的局域结构，有着比较强的序列信号，而且在三级结构中频繁出现，对蛋白质折叠及稳定性起重要作用，因此，学者们非常重视对超二级结构的研究，做了许多工作[11]-[18]。酶蛋白质中的超二级结构除了具有一般蛋白质中超二级结构的特点，还有其自身特点，常常参与形成一些结合位点和活性位点，执行复杂的生物学功能。例如，丝裂原活化蛋白激酶(mitogen-activated protein kinases, MAPKs)是信号从细胞表面传导到细胞核内部的重要传递者，其中就包含一个β-loop-α结构，氨基端的β折叠和羧基端的α螺旋之间形成一个裂隙，为ATP结合位点[19]。又如，SnRK3 是植物特有的一类蛋白激酶，又被称为类钙调磷酸酶 B 亚基互作蛋白激酶(calcineurin B-like calcium sensor-interacting protein kinases, CIPK)。CIPK 激酶在 C 端的酶结合区中含有一个抑制区域，与钙离子结合蛋白 CBL(calcineurin B-like calcium sensor, CBL)结合来激活这些激酶。而 CBL 蛋白有包含 4 个 α-loop-α 结构的保守核心区域[20]，每个 α-loop-α 结构的保守性与结合的激酶的差异有关。因此，酶蛋白质中的简单超二级结构对酶结构及功能研究有特殊意义。

对蛋白质中四类超二级结构的研究，2008 年 Hu 和 Li[16]、2010 年 Zou[15]等人取得了较好的预测结果。本文在前人研究各类蛋白质超二级结构的基础上，首次对酶蛋白质中的简单超二级结构进行研究，将 2261 个酶蛋白质的超二级结构，按照 loop 连接的二级结构类型，分为 β-loop-β、β-loop-α、α-loop-α

和 α-loop-β 四类。从超二级结构的一级序列出发，序列模式固定长度选取 24 个氨基酸残基，采用第 6 位点为 loop 的 N 端、第 19 位点为 loop 的 C 端、第 10 位点为 loop 的 N 端、第 15 位点为 loop 的 C 端和以 loop 序列为中心对齐五种片段截取方式，以位点氨基酸和位点氨基酸紧邻关联作为参数，分别采用矩阵打分算法和支持向量机方法的预测结果不理想，将支持向量机的预测结果通过整体分类器加权融合，进一步预测四类超二级结构，取得了较好的预测效果。

2. 材料和方法

2.1. 数据库的构建及统计分析

选取 SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/)数据库中 ASTRAL 1.75 版的序列相似性≤95%的 16712 个蛋白质，从中删除一些小蛋白质后，剩余 14977 个蛋白质。经过 Blastcluster 软件处理，得到序列相似性<25%的蛋白质有 8704 个，其中，序列片段长度大于 100 个氨基酸残基、分辨率<3.0 Å 的蛋白质有 4442 个。再从这 4442 个蛋白质中按照酶的 EC 编号[21]挑选出 2261 个酶蛋白质(其中包括氧化还原酶 393 个，转移酶 637 个，水解酶 776 个，裂解酶 199 个，异构酶 112 个、连接酶 115 个和同时属于两种以上的酶 29 个)。根据 dssp 数据库提供的二级结构，将 H、G、I 归为 α 螺旋，E、B 归为 β 折叠，其余为 loop。按照 loop 连接的二级结构类型，得到独立的超二级结构单元 53367 个，其中，β-loop-β(以下用“EE”表示)有 14037 个，β-loop-α(以下用“EH”表示)有 13391 个，α-loop-α(以下用“HH”表示)有 13539 个，α-loop-β(以下用“HE”表示)有 12400 个。对四类超二级结构序列片段进行统计分析，我们发现，loop 长度主要集中在 2-12 个氨基酸之间(见图 1)，包含四类超二级结构单元 45506 个，其中，EE 有 12956 个，EH 有 10646 个，HH 有 10682 个，HE 有 11222 个，分别占总数的 92.3%、79.5%、90.5%、78.9%。loop 长度在 2~12 个氨基酸之间的四类超二级结构中，序列片段长度主要分布在 6-30 个氨基酸之间(见图 2)，包含四类超二级结构 41793 个，具体有 EE12847 个、EH10090 个、HH8103 个、HE10753 个，分别占其中的 99.2%、94.8%、75.8%、95.8%。因此，我们以 loop 长度在 2~12 个氨基酸之间的序列片段长度在 6-30 个氨基酸之间的超二级结构为研究对象。

2.2. 计算方法

2.2.1. 四类超二级结构序列片段的截取

通过对四类超二级结构序列片段的统计分析，我们得到 EE、EH、HH、HE 的平均长度分别为 15 个氨基酸、19 个氨基酸、24 个氨基酸、19 个氨基酸。而且，α 螺旋的平均长度为 9 个氨基酸，loop 的平均长度为 4 个氨基酸，β 折叠的平均长度为 5 个氨基酸。因此，为了保证四类超二级结构的重要信息都不丢失，选取固定模式长度为 24 个氨基酸，保证 loop 两端连接的二级结构都能进入序列片段，同时，由于 loop 两端有较强的氨基酸保守性，比如，氨基酸 G 在 loop 两端出现较为频繁[10]，所以，我们采用以第 6 位点作为 loop 的 N 端、以第 19 位点作为 loop 的 C 端、以第 10 位点作为 loop 的 N 端、以第 15 位点作为 loop 的 C 端和以 loop 序列为中心对齐五种片段截取方式，见图 3。

2.2.2. 位点信息的统计分析及参数选取

对 2.2.1 的 5 种序列片段截取方式，分别使用 weblogo 软件进行位点保守性统计分析，由于篇幅限制，这里选取部分为例说明(见图 4)。以第 10 位点为 loop 的 N 端为例，(a) 图代表超二级结构 EE，(b) 图代表超二级结构 HH，比较(a)和(b)，(a) 图中第 10 位点到第 13 位点最保守的氨基酸都是 G，其中第 10 位点和第 11 位点氨基酸 D 的保守性次之，其它位点最保守的氨基酸多为 V、L；而(b)图中第 12、13 位点最保守的氨基酸是 P，第 10 位点和第 11 位点最保守的氨基酸虽然也是 G，但是次之保守的氨基酸分别是 L 和 P，其它位点最保守的氨基酸多为 L、A。说明同一种片段截取方式，不同超二级结构的保守性

不同。以超二级结构 HE 为例，(c) 图代表第 15 位点为 loop 的 C 端，(d) 图代表第 6 位点为 loop 的 N 端，

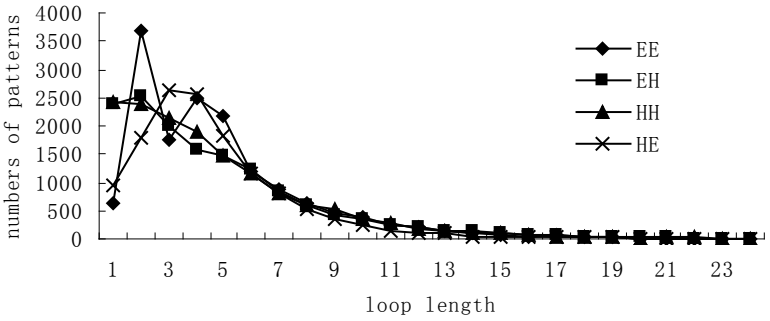


Figure 1. The distribution of sequence numbers with different loop length in the supersecondary structures
图 1. 不同 loop 长度对应的四类超二级结构数目

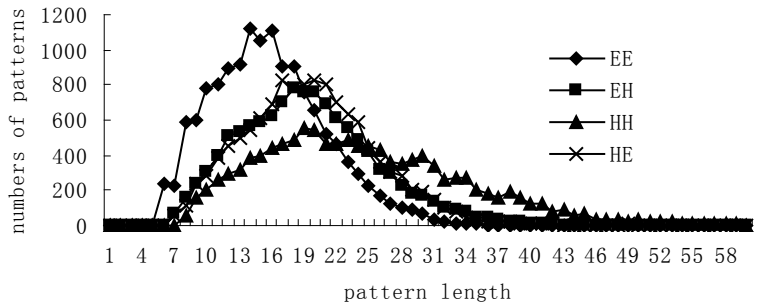
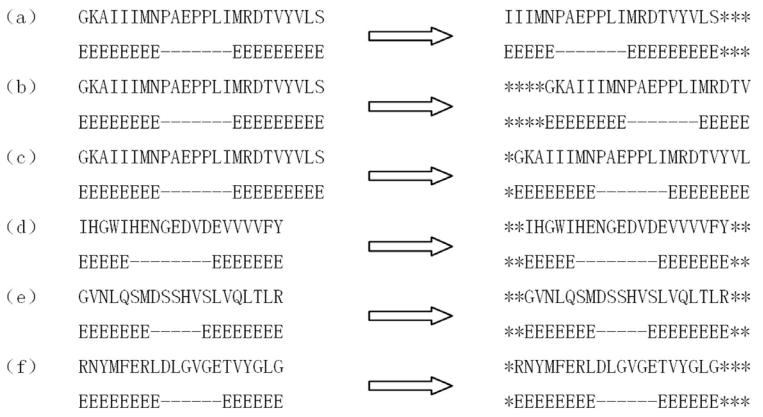
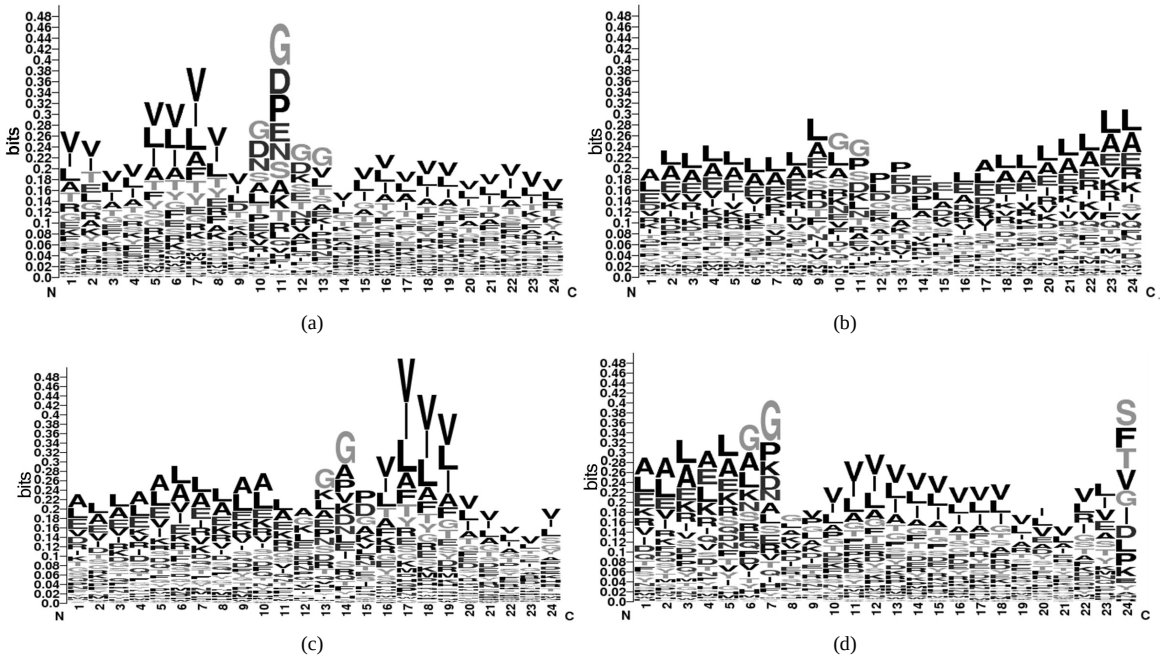


Figure 2. The distribution of pattern numbers with different pattern length
图 2. 不同序列片段长度对应的四类超二级结构数目



Note: first row is amino acid sequences, second row is secondary structures corresponding sequences, "*"is a terminal residue
注：第一行表示氨基酸序列，第二行表示序列对应的二级结构，“*”表示一个空位

Figure 3. The diagram of the best patterns fixed-length: (a) beginning of loop locates the sixth position (b) end of loop locates the nineteenth position (c) beginning of loop locates the tenth position (d) end of loop locates the fifteenth position (e) loop sequence locates the center (the length of loop is an odd number) (f) loop sequence locates the center (the length of loop is a even)
图 3. 最佳固定模式长度选取示意图：(a) 第 6 位点为 loop N 端；(b) 第 19 位点为 loop C 端；(c) 第 10 位点为 loop N 端；(d) 第 15 位点为 loop C 端；(e) 以 loop 序列为中心(loop 长为奇数)；(f) 以 loop 序列为中心(loop 长为偶数)



Note: the overall height of the stack indicates the position conservation, while the height of symbols within the stack indicates the relative frequency of each amino acid at that position

Figure 4. Sample of the position conservation: (a) beginning of loop locates the tenth position of EE (b) beginning of loop locates the tenth position of HH (c) end of loop locates the fifteenth position of HE (d) beginning of loop locates the sixth position of HE.
图 4. 位点氨基酸的保守性举例:(a)以第 10 位点为 loop 的 N 端(EE) (b)以第 10 位点为 loop 的 N 端(HH) (c)以第 15 位点为 loop 的 C 端(HE)(d)以第 6 位点为 loop 的 N 端(HE)

比较(c)和(d)，(c)图中第 13、14 位点最保守的氨基酸是 G，第 15 位点最保守的氨基酸是 P，其它位点最保守的氨基酸多为 V、A、L；而(d)图中，第 6、7、8 位点最保守的氨基酸是 G，其它位点最保守的氨基酸多为 A、L、V，另外，第 24 位点的氨基酸保守性有明显特点，最保守的氨基酸是 S、F、T、V。可见，不同的片段截取方式有着不同的位点氨基酸保守特性，所以 5 种截取方式的位点氨基酸保守信息可以作为预测参数。这里我们选取位点氨基酸(20 种氨基酸加一个空格)和其紧邻关联为参数。

2.2.3. 矩阵打分算法(PCSF)

矩阵打分方法在转录因子结合位点预测方面取得较好结果[22] [23]。本文以位点氨基酸及其紧邻关联的保守性作为参数，将酶蛋白质中的四类简单超二级结构用矩阵打分的方法分类。

矩阵相似性打分函数为： $mss = \dfrac{\sum_{i=1}^{L} c_i \left(w_{ij} - w_{i,\min} \right)}{\sum_{i=1}^{L} c_i \left(w_{i,\max} - w_{i,\min} \right)}$

其中，w_{ij} 是位置权重矩阵的矩阵元，$w_{ij} = \log \dfrac{p_{ij}}{p_{0j}}$，$L$ 为选取的蛋白质超二级结构序列模式的片断长度，以氨基酸为参数的矩阵是21行 L 列，以氨基酸紧邻关联为参数的矩阵是441行 $L-1$ 列。$w_{i,\min}$ 表示位置权重矩阵的第 i 列上出现的矩阵元最小值，$w_{i,\max}$ 表示第 i 列上出现的矩阵元最大值。p_{0j} 表示氨基酸 j 出现的背景概率。

p_{ij} 是位点位置概率，$p_{ij} = \dfrac{n_{ij} + \sqrt{N_i}/l}{N_i + \sqrt{N_i}}$。以氨基酸为参数时，$l = 21$，$j$ 表示 20 种氨基酸和空位，N_i

表示第 i 个位置上所有氨基酸出现的总数，n_{ij} 表示第 i 个位置上第 j 种氨基酸出现的频数；以氨基酸紧邻关联为参数时，$l=441$，N_i 表示第 i 个位置上所有氨基酸紧邻关联出现的总数，n_{ij} 表示第 i 个位置上第 j 种氨基酸紧邻关联出现的频数。

C_i 是位点保守性参量，反映被测序列在该位点与一致性序列在该位点的偏离性。

$$c_i=\frac{100}{\log l}\left(\sum_{j=1}^{l}p_{ij}\log p_{ij}+\log l\right)$$

可以证明：$0\le mss\le 1$，如果等于1，则说明未知序列每一位点上的氨基酸刚好与位置权重矩阵位点上的最大值所对应的氨基酸一致。对含有24个氨基酸的固定序列模式片段，以位点氨基酸(位点氨基酸紧邻关联)作为参数，四类超二级结构可以构建4个21行24列(441行23列)的标准打分矩阵，任一条待测序列片段分别以4个标准位置打分矩阵为标准，利用打分函数，确定相似程度，得到4个打分值，哪一个分值最大，该序列就被预测为哪一类超二级结构。

2.2.4. 支持向量机(SVM)方法

SVM是Vapnik[24] [25]等人提出来的一类新型机器学习方法，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[26]-[29]。SVM的基本思想是基于统计学习理论，用核函数映射的方法把输入矢量 x 映射到一个高维特征空间，在高维特征空间构造出最优超平面，使得该超平面在保证分类精度的同时，能够使超平面两侧的空白区域最大化，从而达到最大的泛化能力，其由有限的训练集样本得到的小的误差能够保证对独立的检验集仍保持小的误差。常用的核函数有多项式核函数$(Poly)$、径向基核函数(RBF)、Sig mod 核函数$(Sig\bmod)$，我们选择径向基核函数：$k(x,x_i)=\exp\left(-g\left\|x-x_i\right\|^2\right)$。另外，由于SVM算法是一个凸优化问题，局部最优解一定是全局最优解。因此SVM是一类很好的非线性模式识别分类器。SVM 算法已被很多学者编译成程序加以实现，常用的有libsvm、mysvm及svmlight等。这里我们使用的是libsvm-2.91程序包[30]，其中c和gamma值为缺省。

根据 2.2.3 的矩阵打分算法，以位点氨基酸(位点氨基酸紧邻关联)为参数，对于训练集，四类超二级结构可以构建 4 个标准打分矩阵，用于检验集，可以得到 4 个打分值。将这些矩阵打分值作为特征参数输入支持向量机进行预测。

2.2.5. 整体分类器

整体分类器已用于预测 27 类蛋白质折叠子[31]，取得了较好的预测效果。

定义整体分类器：

$$C=C_1\oplus C_2\oplus C_3\oplus\cdots\oplus C_\Omega \tag{1}$$

这里，$C_1,C_2,C_3,\cdots,C_\Omega$ 分别是基于支持向量机算法的 Ω 个单分类器，$\oplus$ 为分类器的融合。

将 Ω 个单分类器 $C_i\ (i=1,2,3,\cdots,\Omega)$ 整合为整体分类器 C 用于预测，可以由下式表示：

$$S_j=\sum_{i=1}^{\Omega}\tilde{\omega}_i SVM_i\left(X,Y_j\right)\ (j=1,2,3,4) \tag{2}$$

式中，$\tilde{\omega}_i$ 是单分类器 C_i 的权重系数，$0\le\tilde{\omega}_i\le 1$，$SVM_i\left(X,Y_j\right)$ 是任意一个待测序列 X 与单分类器 $C_i\ (i=1,2,3,\cdots,\Omega)$ 中第 j 类标准源 $Y_j\ (j=1,2,3,4)$ 的支持向量机，S_j 的最大值：

$$S\mu=\max\{S_1,S_2,S_3,S_4\}\ (\mu=1,2,3,4) \tag{3}$$

决定了待测序列 X 的预测类型。其中 max 表示取最大值，μ 就是待测序列 X 采用整体分类器预测所属的分类。

2.2.6. 系统检验

本文对分类结果的评价使用 7 交叉检验的方法，随机将数据库分为 7 个子集，依次取出 1 个子集作测试集，而其余 6 个子集作为训练集，此过程循环 7 次。

2.2.7. 精确度评价指标

定义第 i 类超二级结构(i = 1 代表 EE，i = 2 代表 EH，i = 3 代表 HH，i = 4 代表 HE)的敏感性指标 S_{ni}，特异性指标 S_{pi}， *Matthew's* 相关系数 M_i 及预测精度 S 分别为：

$$S_{ni} = \left[TP_i / \left(TP_i + FN_i\right)\right] \times 100\% \tag{4}$$

$$S_{pi} = \left[TP_i / \left(TP_i + FP_i\right)\right] \times 100\% \tag{5}$$

$$M_i = \frac{\left(TP_i \times TN_i\right) - \left(FN_i \times FP_i\right)}{\sqrt{\left(TP_i + TN_i\right) \times \left(TN_i + FP_i\right) \times \left(TP_i + FP_i\right) \times \left(TN_i + FN_i\right)}} \tag{6}$$

$$S = \frac{\sum_i TP_i}{N} \tag{7}$$

式中，TP_i 为该类中正确预测的样品数，FN_i 为该类中错误预测的样品数，TN_i 为其它类中正确预测的样品数，FP_i 为其他类被预测为此类的样品数，N 表示四类序列总数。

3. 结果与讨论

3.1. 矩阵打分方法的计算结果与讨论

以位点氨基酸(21AA)和位点氨基酸紧邻关联(441JL)分别为参数，对 5 种序列片段截取方式：以第 6 位点为 loop N 端(N6)；以第 19 位点为 loop C 端(C19)；以第 10 位点为 loop N 端(N10)；以第 15 位点为 loop C 端(C15)；以 loop 序列为中心对齐(Center)，用矩阵打分的方法预测酶蛋白质中四类超二级结构，7 交叉检验的结果见表 1。

从表中可以看出，21AA 的个别结果较好，比如超二级结构 EE 中 N6 和 C19 的敏感性分别达到 70.1%和 77.2%，而超二级结构 HH 的计算结果很不理想，C19、N10、C15、Center 的敏感性分别为 17.7%、22.5%、25.1%、25.1%，Matthew's 相关系数 0.1 左右，超二级结构 HE 中 N6 的敏感性也仅为 25%；441JL 结果的 HE 中 N6 敏感性较好，达到 79.2%，HH 中 C19 的特异性、HE 中 N10 的敏感性、EH 中 c15 和 center 的敏感性都达到 81.9%、81%、79.7%和 80%，结果较好，HH 的计算结果同样不理想，C19、N10、C15、Center 的敏感性分别为 25.5%、28.3%、24%、24.7%，Matthew's 相关系数 0.27 以上，C19 的预测总精度为 49.3%，结果较差。

3.2. SVM 的计算结果与讨论

将矩阵打分值作为特征参数输入 SVM，采用 7-交叉检验，将位点氨基酸为参数的打分值输入 SVM (21AA)，每种片段截取方式的预测结果见表 2。比较以位点氨基酸为参数的打分方法的预测结果(见表 1 中 21AA)，有些结果明显提高，比如，EE 中 N10、C15、Center 截取方式的敏感性分别由打分方法的 56%、56.9%、59.7%提高到现在的 78.9%、76.4%、81.5%，HE 中 N6 截取方式的敏感性由打分方法的 25%提高到现在的 67.4%，但是也有一些预测结果比打分方法的结果有所下降；将位点氨基酸紧邻关联为参数的打分值输入 SVM (441JL)，每种片段截取方式的预测结果(见表 2)比较以同样参数的打分方法的预测结果(见表 1 中 441JL)普遍有明显提高，比如，HH 中 N6、C19、N10、C15、Center 截取方式的敏感性分别由打分方法的 37.9%、25.5%、28.3%、24%、24.7%提高到现在的 47.2%、50.7%、46.5%、51%、43.9%，

Table 1. The predictive results of PCSF algorithm using 7_Fold cross-validation
表 1. 打分方法 7 交叉检验的预测结果

	21AA					441JL				
	N6	C19	N10	C15	Center	N6	C19	N10	C15	Center
S_{n1}(%)	70.1	77.2	56.0	56.9	59.7	38.3	22.3	39.3	45.1	48.0
S_{n2}(%)	49.1	39.6	60.1	52.8	56.4	58.9	82.3	66.6	79.4	80.0
S_{n3}(%)	41.5	17.7	22.5	25.1	25.1	37.9	25.5	28.3	24.0	24.7
S_{n4}(%)	25.0	61.0	60.2	62.8	61.7	79.2	62.4	81.0	71.2	73.2
S_{p1}(%)	43.2	44.8	52.4	52.1	53.4	57.3	53.7	73.1	71.8	73.8
S_{p2}(%)	60.9	58.6	49.9	49.3	52.4	68.2	37.2	54.8	49.0	50.9
S_{p3}(%)	32.9	47.3	43.0	44.1	39.2	67.0	81.9	60.0	66.8	62.0
S_{p4}(%)	69.7	55.0	54.1	53.0	56.3	43.5	67.2	48.3	56.1	57.8
S(%)	46.7	50.4	51.0	50.6	52.0	54.3	49.3	55.0	56.5	58.0
M_1	0.23	0.30	0.29	0.29	0.32	0.27	0.16	0.38	0.41	0.45
M_2	0.31	0.27	0.28	0.25	0.29	0.45	0.24	0.37	0.39	0.41
M_3	0.09	0.14	0.13	0.15	0.12	0.37	0.36	0.27	0.29	0.27
M_4	0.26	0.33	0.32	0.32	0.35	0.32	0.44	0.38	0.41	0.45

Table 2. The predicting results of SVM using 7_cross validation
表 2. 支持向量机方法 7 交叉检验的预测结果

	21AA					441JL				
	N6	C19	N10	C15	Center	N6	C19	N10	C15	Center
S_{n1}(%)	65.2	66.7	78.9	76.4	81.5	67.7	68.4	78.3	77.4	81.0
S_{n2}(%)	51.0	66.2	52.7	55.7	53.5	57.0	68.0	54.7	64.3	57.6
S_{n3}(%)	28.6	23.6	17.2	16.6	14.4	47.2	50.7	46.5	51.0	43.9
S_{n4}(%)	67.4	56.3	57.2	60.3	59.1	71.7	61.0	65.9	59.5	61.0
S_{p1}(%)	51.8	52.4	48.2	50.6	48.0	53.4	53.5	57.0	59.5	54.6
S_{p2}(%)	59.2	51.8	54.4	55.1	56.4	73.5	59.8	67.4	65.7	68.0
S_{p3}(%)	49.2	55.7	57.1	61.5	63.9	69.4	71.9	66.7	68.8	67.8
S_{p4}(%)	55.2	61.5	48.2	56.3	60.0	59.0	74.1	62.1	63.9	63.9
S(%)	54.5	54.8	53.3	54.2	54.1	61.8	62.8	62.2	63.8	61.9
M_1	0.34	0.35	0.36	0.37	0.37	0.39	0.40	0.48	0.50	0.47
M_2	0.34	0.34	0.30	0.32	0.33	0.50	0.45	0.45	0.49	0.47
M_3	0.21	0.22	0.19	0.21	0.20	0.45	0.49	0.43	0.47	0.42
M_4	0.38	0.38	0.37	0.35	0.38	0.46	0.53	0.46	0.45	0.45

预测总精度分别由打分方法的 54.3%、49.3%、55%、56.5%、58%提高到现在的 61.8%、62.8%、62.2%、63.8%、61.9%。为了进一步提高预测精度，我们将 SVM 的计算结果输入整体分类器。

3.3. 整体分类器的计算结果与讨论

将位点氨基酸紧邻关联为参数的五种片段截取方式作为 5 个单分类器，采用整体分类器进行加权融合，进一步预测酶蛋白质中四类超二级结构。5 个单分类器的支持向量机在不同权重系数下的预测结果见表 3(篇幅所限，只列出代表性的结果，其它略去)。平均预测总精度最高达到 72.64%，相关系数在 0.57 以上，相比表 2 中的预测结果，平均预测精度提高 8.8 个百分点以上，其它各项指标也显著提高。结果表明，整体分类器通过加权融合单分类器的计算结果，能有效的提取酶蛋白质中四类简单超二级结构中有益的预测信息，可以有效的提高预测精度。

目前，还没有预测酶蛋白质中四类简单超二级结构的相关文献，所以，我们只能参考前人对各类蛋白质中超二级结构的预测结果，见表 4。2008 年 Hu 和 Li[16]使用基于打分值和离散增量为组合向量的支持向量机识别算法对各类蛋白质中 4 类超二级结构进行分类，以氨基酸为参数，训练集的最高预测精度达到 71.7%，与本文相比，除了 EE 的特异性，本文中所有指标的预测结果都优于 Hu's。2010 年 Zou[15]等人用 SVM 和 IDQD 的方法也对各类蛋白质中的四类简单超二级结构进行预测，其中 IDQD 方法的预测结果最好，以氨基酸组成、二肽组分和氨基酸组成分布共同为参数，训练集的最高预测精度达到 77.7%，高于我们的预测总精度，但是，有些指标我们的算法高于 Zou's，比如，我们的算法中，EH、HH、HE 的特异性分别为 78%、83%和 74.3%，Matthew’s 相关系数分别为 0.63、0.65、0.64，而 Zou’s 算法中，EH、HH、HE 的特异性分别为 71%、73.3%和 69.5%，Matthew’s 相关系数分别为 0.56、0.58、0.55。

4. 结论

酶蛋白质中的简单超二级结构对酶的生物学功能有重要影响，本文首次对酶蛋白质中四类简单超二级结构进行了理论预测。首先建立研究的数据集，我们的数据集包含四类超二级结构单元 41793 个，这个大数量的数据集帮助我们更加有效地预测酶蛋白质中超二级结构，是我们研究工作的有利基础。从酶

Table 3. The overall results for ensemble classifier
表 3. 整体分类器的预测结果

$\tilde{\omega}_1$	$\tilde{\omega}_2$	$\tilde{\omega}_3$	$\tilde{\omega}_4$	$\tilde{\omega}_5$	S_{n1}(%)	S_{n2}(%)	S_{n3}(%)	S_{n4}(%)	S_{p1}(%)	S_{p2}(%)	S_{p3}(%)	S_{p4}(%)	M_1	M_2	M_3	M_4	S(%)
0.4017	0.3009	0.0999	0.0987	0.0988	80.2	69.6	63.6	75.4	62.6	78.0	83.0	74.3	0.57	0.63	0.65	0.64	72.64
0.3001	0.4013	0.0989	0.0998	0.0999	80.0	72.8	63.9	71.5	62.6	74.5	83.5	76.4	0.57	0.62	0.66	0.63	72.44
0.2010	0.3025	0.1982	0.0981	0.2002	81.2	69.3	60.4	70.2	60.8	73.7	84.7	73.4	0.55	0.59	0.64	0.60	70.78
0.1998	0.4056	0.2008	0.0945	0.0993	79.7	72.3	63.0	69.8	61.7	73.5	83.4	75.6	0.56	0.61	0.65	0.61	71.61
0.3103	0.3001	0.0943	0.2005	0.0948	80.6	71.0	63.0	71.8	62.0	75.2	84.1	74.8	0.57	0.62	0.65	0.62	72.03
0.3057	0.3033	0.1999	0.0974	0.0937	81.0	69.1	62.5	73.5	61.9	76.0	83.9	74.4	0.57	0.61	0.65	0.63	71.99
0.3111	0.1989	0.0919	0.2980	0.1001	81.0	69.3	62.0	71.5	61.5	76.0	84.0	72.7	0.56	0.61	0.64	0.60	71.37
0.1997	0.3033	0.1998	0.1981	0.0991	80.3	70.5	62.4	70.2	61.4	73.9	84.6	73.9	0.56	0.60	0.65	0.60	71.28

Table 4. Comparison among different predictive results
表 4. 不同方法预测结果的比较

	S_{n1}(%)	S_{n2}(%)	S_{n3}(%)	S_{n4}(%)	S_{p1}(%)	S_{p2}(%)	S_{p3}(%)	S_{p4}(%)	M_1	M_2	M_3	M_4	S(%)
Hu’s SVM	78.6	67.5	57.9	73.4	72.8	69.2	65.8	70.7	0.58	0.57	0.53	0.54	71.7
Zou’s IDQD	85.8	74.7	68.5	75.4	79.9	71.0	73.3	69.5	0.61	0.56	0.58	0.55	77.7
Our Classifier	80.2	69.6	63.6	75.4	62.6	78.0	83.0	74.3	0.57	0.63	0.65	0.64	72.64

序列出发，通过统计分析，选取最佳模式的长度为 24 个氨基酸，以位点氨基酸及其紧邻关联为参数，根据四类超二级结构的特点，采用五种片段截取方式，将计算出的矩阵打分值作为特征参数输入支持向量机，再通过整体分类器加权融合，得到较好的预测结果。

基于支持向量机算法的整体分类器能够有效的提高预测精度，首先，矩阵打分方法能降低参数的维数，而支持向量机算法能够融合有益的特征信息；其次，将各单分类器的计算结果进行加权融合可以提取更加有益的预测信息，从而提高预测的精度。本文只选择了位点氨基酸及其紧邻关联为参数，今后的工作中我们将考虑结合其它的有益参数，比如氨基酸的亲疏水性，柔性等等，也可以进一步调整分类器的个数和权重系数，进而获得更为理想的预测结果。

项目基金

国家自然科学基金资助项目(31260203, 30960090)。

参考文献 (References)

[1] Cai, Y.D. and Chou, K.C. (2005) Using Functional Domain Composition To Predict Enzyme Family Classes. *Journal of Proteome Research*, **4**, 109-111.

[2] Cai, Y.D., Guo, P.Z. and Chou, K.C. (2005) Predicting Enzyme Family Classes by Hybridizing Gene Product Composition and Pseudo-Amino Acid Composition. *Journal of Theoretical Biology*, **234**, 145-149.

[3] Chou, K.C. and Cai, Y.D. (2004) Using GO-PseAA Predictor to Predict Enzyme Sub-Class. *Biochemical and Biophysical Research Communications*, **325**, 506-507.

[4] Shen, H.B. and Chou, K.C. (2007) EzyPred: A Top-Down Approach for Predicting Enzyme Functional Classes and Subclasses. *Biochemical and Biophysical Research Communications*, **364**, 53-59.

[5] Shi, R.J. and Hu, X.Z. (2010) Predicting Enzyme Subclasses by Using Support Vector Machine with Composite Vectors. *Protein and Peptide Letters*, **17**, 599-604.

[6] Hu, X.Z. and Ting, W. (2011) Prediction of Enzyme Subclass by Using Support Vector Machine Based on Improved Parameters. 2011 7*th International Conference on Natural Computation*, Shanghai, 26-28 July 2011, 593-598.

[7] Wang, Y. and Hu, X.Z. (2011) Predicting of Oxidoreductase and Lyase Subclasses by Using Support Vector Machine. 2011 10*th IEEE/ACIS International Conference on Computer and Information Science*, Sanya, 16-18 May 2011, 27-31.

[8] Liu, X.X. and Hu, X.Z. (2011) Identifying the β-Hairpin Motifs in Enzymes by Using Support Vector Machine. 2011 10*th IEEE/ACIS International Conference on Computer and Information Science*, Sanya, 16-18 May 2011, 21-26.

[9] Long, H.X. and Hu, X.Z. (2012) Prediction β-Hairpin Motifs in Enzyme Protein Using Three Methods. 2012 8*th International Conference on Natural Computation* (*ICNC* 2012), Chongqing, 29-31 May 2012, 570-574.

[10] 阎隆飞, 孙之荣 (1999) 蛋白质分子结构.清华大学出版社, 北京, 43-56.

[11] Kuhn, M., Meiler, J. and Baker, D. (2004) Strand-Loop-Strand Motifs: Prediction of Hairpin and Diverging Turns in Proteins. *Protein*, **5**, 282-288.

[12] Cruz, X., Hutchinson, E.G., Shepherd, A., et al. (2002) Predicting Protein Topology: An Approach to Identifying Bhairpins. *Proceedings of the National Academy of Sciences*, **99**, ll157-1l162.

[13] Kumar, M., Bhasin, M., Natt, N.K., et al. (2005) BhairPred: Prediction of β-Hairpins in a Protein from Multiple Alignment Information Using ANN and SVM Techniques. *Nucleic Acids Research*, **33**, 154-159.

[14] 胡秀珍, 李前忠 (2006) 用离散量的方法识别蛋白质的超二级结构. *生物物理学报*, **6**, 424-428.

[15] Zou, D.S., He, Z.S., He, J.Y., et al. (2011) Supersecondary Structure Prediction Using Chou's Pseudo Amino Acid Composition. *Journal of Computational Chemistry*, **32**, 271-278.

[16] Hu, X.Z. and Li, Q.Z. (2008) Prediction of the β-Hairpins in Proteins Using Support Vector Machine. *The Protein Journal*, **27**, 115-122.

[17] Hu, X.Z., Li, Q.Z. and Wang, C.L. (2010) Recognition of β-Hairpin Motifs in Proteins by Using the Composite Vector. *Amino Acids*, **38**, 915-921.

[18] Sun, L.X., Hu, X.Z. and Li, S.B. (2012) Predicting βαβ Motifs Based on SVM by Using the ID and MS Values. 2012 5*th International Conference on BioMedical Engineering and Informatics* (*BMEI* 2012), Chongqing, 16-18 October

2012, 910-914.

[19] Wang, Z., Harkins, P.C., Ulevitch, R.J., Han, J.H., Cobb, M.H. and Goldsmith, E.J. (1997) The Structure of Mitogen-Activated Protein Kinase p38 at 2.1-Å Resolution. *Proceedings of the National Academy of Sciences*, **94**, 2327-2332.

[20] Batistic, O. and Kudla, J. (2004) Integration and Channeling of Calcium Signaling through the CBL Calcium Sensor/CIPK Protein Kinase Network. *Planta*, **219**, 915-924.

[21] Webb, E.C. (1992) Enzyme Nomenclature. Academic Press, SanDiego.

[22] Cartharius, K., Frech, K., Grote, K., et al. (2005) Mat Inspector and Beyond: Promoter Analysis Based on Transcription Factor Binding Sites. *Bioinformatics*, **21**, 2933-2942.

[23] Kel, A.E., GoBling, E., Reuter, I., et al. (2003) MATCHTM: A Tool for Searching Transcription Factor Binding Sites in DNA Sequences. *Nucleic Acids Research*, **31**, 3576-3579.

[24] Vapnik, V. (1995) The Nature of Statistical Learning Theory. Springer, New York.

[25] Vapnik, V. (1998) Statistical Learning Theory. Wiley-Interscience, Hoboken

[26] Hu, X.Z. and Li, Q.Z. (2008) Using Support Vector Machine to Predict β-Turns and γ-Turns in Proteins. *Computational Chemistry*, **29**, 1867-1875.

[27] Chou, K.C. and Cai, Y.D. (2002) Using Functional Domain Composition and Support Vector Machines for Prediction of Protein Subcellular Location. *Journal of Biological Chemistry*, **227**, 45765-45769.

[28] Ding, C.H.Q. and Dubchak, I. (2001) Multi-Class Protein Fold Recognition Using Support Vector Machines and Neural Networks. *Bioinformatics*, **17**, 349-358.

[29] Shi, J.Y., Pan, Z., Zhang, S.W. and Liang, Y. (2006) Protein Fold Recognition with Support Vector Machines Fusion Network. *Progress in Biochemistry Biophysics*, **3**, 155-162.

[30] Chang, C.C. and Lin, C.J. (2001) LIBSVM: A Library for Support Vector Machines. Software. http://www.Csie.ntu.edu.tw/cjlin/libsvm

[31] Shen, H.B. and Chou, K.C. (2006) Ensemble Classifier for Protein Fold Pattern Recognition. *Bioinformatics*, **22**, 1717-1722.