生物物理学报　第二十三卷　第六期　二○○七年十二月
ACTA BIOPHYSICA SINICA　Vol.23　No.6　Dec. 2007

用支持向量机识别 β- 发夹模体

胡秀珍[1,2]，　李前忠[1]

（1. 内蒙古大学理工学院物理系，呼和浩特 010021; 2. 内蒙古工业大学理学院物理系，呼和浩特 010059）

摘要：基于蛋白质序列，提出了一种新的超二级结构模体 β- 发夹的预测方法。利用离散增量构成的向量来表示序列信息，并将 6 个离散增量输入支持向量机，在六维向量空间中寻找最优超平面，将 β- 发夹和非 β- 发夹进行分类。计算结果表明，利用所设计的算法预测 β- 发夹，有较高的预测能力。对于训练集，5- 交叉检验的预测总精度为 81.24%，相关系数为 0.57，β- 发夹敏感性为 83.06%；对于独立的检验集，预测总精度为 78.34%，相关系数 0.56，β- 发夹敏感性为 77.24%。将此预测模型应用于 CASP6 的 63 个蛋白质进行检验，得到较好结果。

关键词：超二级结构；β- 发夹模体；离散增量；支持向量机

中图分类号：Q61

0　引　　言

由于蛋白质的空间结构决定其功能，因此蛋白质的结构预测一直是一个重要的研究课题。蛋白质超二级结构是两个或几个二级结构单元被连接多肽连接起来，进一步组成有特殊的几何排列的局域空间结构[1]。常见的简单超二级结构有：α-α、α-β 和 β-β 模体。对于 β-β 模体，如果两个 β-strand 被连接多肽（loop）连接，而且两个 strand 之间存在一个或多个氢键，则称这种超二级结构为 β- 发夹（hairpin）[2]；否则将 β-β 模体称为非 β- 发夹。

超二级结构预测是一个新兴研究领域。1997 年，Sun 等[3]利用人工神经网络（ANN）对 11 种不同的常见超二级结构模体进行了预测，并且得到了 70%~80%的预测精度，相关系数为 0.4~0.5，这种超二级结构预测属于对不同模体数的预测。2002 年，Cruz 等[4]利用片段的 14 个得分改善了蛋白质中 β- 发夹的预测方法，预测精度是 47.7%；Kuhn 等[2]在 2004 年以氨基酸序列做输入，将 strand-loop-strand模体分为局域 hairpin 和非局域 diverging turn，预测精度是 75.9%；2005 年，Kumar 等[5]使用两种机器记忆技术：支持向量机（SVM）和 ANN模型，对 EVA 的序列相似性小于 33%的 2880 非冗余蛋白质中 β- 发夹进行了研究，预测精度是 79.2%。

本文基于蛋白质的氨基酸序列，利用离散增量构成的向量来表示序列信息，并将离散增量输入支持向量机，提出了预测 β-β 模体中的 β- 发夹的理论方法。通过对序列相似性小于 40%（SCOP 的 ASTRAL 1.65 版）、分辨率高于 3Å 的 3088 个非冗余蛋白质的研究，对 8671 个模体（其中 β- 发夹模体数为 6028 个，非 β- 发夹模体数为 2 643 个）进行预测，得到较好结果。为了检验本方法的预测能力，我们将预测方法应用于最近的数据库 CASP6（http://predictioncenter.llnl.gov/casp6/）中的 63 个蛋白质的 163 个 β-β 模体，预测精度为 71.17%，β- 发夹敏感性为 79.17%，与 Kumar[5]对 63 个蛋白质的 β- 发夹预测结果相比，不但 β- 发夹敏感性提高了，还找出了一些 Kumar 识别不出来的 β- 发夹模体。

1　材料及方法

1.1　数据集

本文使用的数据集是从 ArchDB 数据库直接下载的（网站 http://sbi.imim.es/cgi-bin/archdb/loops.pl）、由 Oliva 和 Bates 等[6,7]整理好的 β- 发夹和非 β- 发夹模体。ArchDB 数据库是依据已知的蛋白质结构对 loop 构象分类的超二级结构数据库，该数据库的基本数据来自 DSSP 库，因此超二级结构模

收稿日期：2007-02-10

基金项目：国家自然科学基金资助项目（30560039）；内蒙自然科学基金资助项目（200508010509, 200607010101）

通讯作者：李前忠，电话：(0471)4992958，

E-mail：qzli@imu.edu.cn

体选取准确、可靠，本文使用的 β- 发夹和非 β- 发夹模体就是来自此数据库的 ArchDB40 数据集。

ArchDB40 是序列相似性小于 40%（SCOP 的 ASTRAL 1.65 版）、分辨率高于 3Å 的蛋白质结构域 loop 分类数据集。ArchDB40 中包含 3088 个非冗余蛋白质链，按照 loop 连接的规则二级结构，分为 alpha-alpha、beta-beta link、beta-beta hairpin、alpha-beta 和 beta-alpha 五种类型超二级结构。我们只选取 beta-beta hairpin 作为 β- 发夹数据集，选取 beta-beta link 作为非 β- 发夹数据集，两种数据集分别包括 6 216 和 2 964 个模体。两序列模式平均长分别为 14.9 和 15.2 个氨基酸，依据 Cruz[4] 的观点，固定模式长应为 15 个氨基酸。

1.2 计算方法

1.2.1 最佳固定序列模式长的选取及序列片段的截取方式

分两个步骤进行。第一步：确定研究对象的 loop 长，并给出最佳固定序列模式长。β- 发夹和非 β- 发夹模体对应的序列按 loop 长度进行统计，具体统计情况见图 1。统计结果发现：loop 包含 2~8 个氨基酸残基的 β- 发夹模体数为 6 028 个，占此类模体总数的 97%；非 β- 发夹模体数为 2 643 个，占其总数的 89%。因此，可以选取 loop 长 2-8 个氨基酸的序列作为研究对象。还发现：loop 包含 2~5 个氨基酸残基的模体数最多，结合计算结果，最佳固定模式长选取 12 个氨基酸。

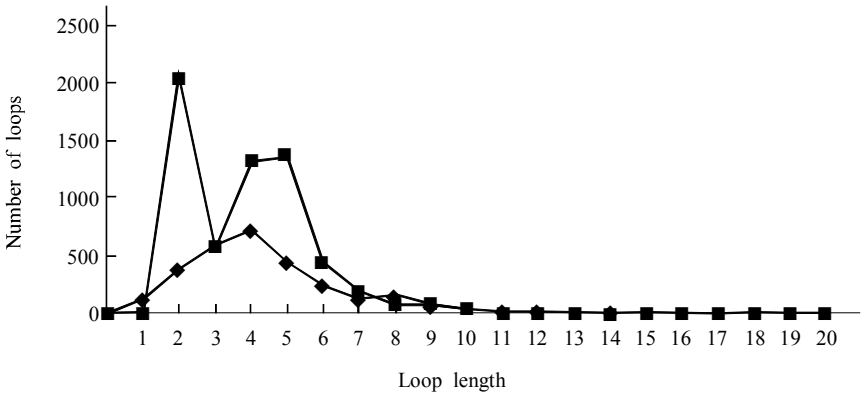

Fig.1 The distribution of the sequence number of loops with different loop lengths in the β-hairpin and non-β-hairpin. —◆—: Non-hairpin; —■—: Hairpin

第二步：在第一步确定 loop 长 2~8 个氨基酸的序列为研究对象和固定模式长选取 12 个氨基酸的基础上，对每一个序列片段采取三种截取方式：

1) 以 loop 为中心，当 loop 长为奇数时，所取序列模式左侧比右侧多取一个氨基酸残基；当 loop 长为偶数时，所取序列模式两侧取相同的氨基酸残基数。长度不足的，两端以空位补齐。如图 2(A)所示。

2) 以 loop 的起始点作为固定模式的第 5 位点，长度不足的，两端以空位补齐。如图 2(B)所示。

3) 以 loop 的终止位点作为固定模式的第 8 位点，长度不足的，两端以空位补齐。如图 2(C)所示。

以上的截取方式，综合了 Cruz[4]、Kumar[5]和 Kuhn[2]等对固定模式片段的截取方法。根据不同的截取方式得到的序列片段，构成不同的固定长序列模式集合，共得到三个不同的序列模式集合。

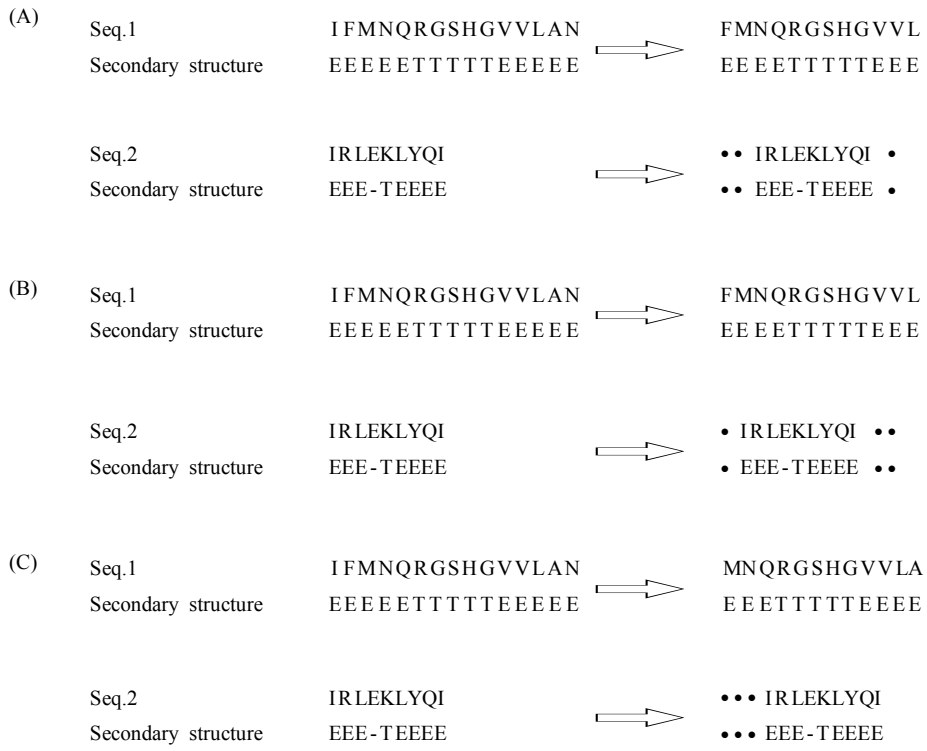

Fig.2 The diagram of three patterns fixed-length. (A) Shows the pattern that the loop locates the center of pattern fixed-length; (B) Shows that first amino acid of loop locates the fifth position of pattern fixed-length; (C) Shows that the end of loop locates the eighth position of pattern fixed-length. Note: first row is amino acid sequences, second row is secondary structures corresponding sequences. Boldfaces are amino acids and secondary structures of corresponding loops; • is a gap

1.2.2　离散量和离散增量

早在 1978 年 Laxton 就给出了离散量的定义，本文具体定义如下[8-10]，对于 t 个符号的状态空间 $\boldsymbol{S}$，m_i 表示第 i 个状态出现的个数，其离散源 $\{m_1, m_2, \cdots, m_t\}$ 的离散量为：

$$D(S)=M\log M-\sum_i m_i\log m_i \tag{1}$$

其中 $M=\sum_i m_i$，离散量的单位为哈特。

一般地，对于两个具有相同信息符号空间的离散源 $X(n_1, n_2, \cdots, n_t)$和 $S(m_1, m_2, \cdots, m_t)$，定义其离散增量为：

$$ID(X,S)=D(X+S)-D(X)-D(S) \tag{2}$$

其中 $D(X+S)$是混合离散源$(n_1+m_1, n_2+m_2, \cdots, n_t+m_t)$的离散量。$ID(X,S)$也可写成：$ID(X,S)=D(M,N)-\sum_i D(m_i,n_i)$形式，具体计算公式为：

$$D(M,N)=(M+N)\log(M+N)-M\log M-N\log N \tag{3}$$

$$D(m_i,n_i)=(m_i+n_i)\log(m_i+n_i)-m_i\log m_i-n_i\log n_i \tag{4}$$

其中 $N=\sum_i n_i$，$M=\sum_i m_i$。若 m_i 与 n_i 其中之一为零，则 $D(m_i,n_i)=0$

可以证明，离散增量满足：$0\leqslant ID(X,S)\leqslant D(N,M)$，$D(N,M)$是离散增量 $ID(X,S)$的极大值，根据 $ID(X,S)$值可以比较两离散源的相似程度。$ID(X,S)$越小，说明两个离散源越相似；$ID(X,S)$越大，说明两个离散源相似性越差。

例如，对含有 12 个氨基酸的固定序列模式片段，以每一位点的氨基酸（20 种氨基酸和一个空位）出现的频数作为参数，则构成的离散源为 $\boldsymbol{S}$: $\{m_1, m_2, \cdots, m_t\}$，$i=1, 2, \cdots, t$; $t=12\times21$。分别以 β- 发夹和非 β- 发夹训练集构成离散源，利用公式(1)计算离散源 $\boldsymbol{S}$ 的离散量，对任一序列片段使用公式(2)都可以得到 2 个离散增量；在 **1.2.1** 的描述中，每一个片段都有三种截取方式，因此每一序列片段都可得到 6 个离散增量。这样就可以把氨基酸序列片段所包含的特征信息全部转化到 6 个离散增量中。

1.2.3 支持向量机（SVM）方法

SVM 是 Vapnik[11]等人提出来的一类新型机器学习方法，SVM 的基本思想是基于统计学习理论，用非线性映射把输入数据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，使得超平面与不同类样本集之间的距离最大，从而达到最大的泛化能力，即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。另外，由于 SVM 算法是一个凸优化问题，因此局部最优解一定是全局最优解。这些特点是其它学习算法，如 ANN 学习算法所不及的。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-2.83 程序包[12]，输入参数为 **1.2.2** 中计算的 6 个离散增量。

1.2.4 基于离散量的 SVM 识别算法

SVM 使用的核函数为径向机函数（RBF）。首先使用训练集，我们用上面得到的 6 个离散增量构成一个六维向量，输入给 SVM 进行训练，SVM 先对训练集的数据进行归一化（scale）；然后进行网格化搜索最佳参数 c 和 gamma，使 5- 交叉检验得到的平均预测率最高，本文的最佳 c 和 gamma 值分别为 32768 和 0.00048828125；由于 β- 发夹和非 β- 发夹数据集所包含的序列数不平衡，我们采用加权重的方法来消除这种不平衡带来的大类淹没小类的影响[13]，本文的 β- 发夹权重因子（$w1$）为 2/7，非 β- 发夹权重因子（w-1）为 5/7，最后产生一个分类器，就用这个分类器对独立的检验集进行分类。

1.2.5 精确度评价指标

使用标准检验估算的方法：Acc、MCC、$Q_{o(H)}$、$Q_{o(NH)}$、$Q_{p(H)}$、$Q_{p(NH)}$分别表示预测精度、相关系数、β- 发夹的敏感性、非 β- 发夹的敏感性、β- 发夹的特异性、非 β- 发夹的特异性。

$$Acc=\frac{(p+r)}{(p+r+o+u)}\times 100$$

$$MCC=\frac{(p\times r)-(o\times u)}{\sqrt{(p+u)(p+o)(r+u)(r+o)}}$$

$$Q_{o(H)}=\frac{p}{p+u}\times 100,\quad Q_{o(NH)}=\frac{r}{r+o}\times 100$$

$$Q_{p(H)}=\frac{p}{p+o}\times 100,\quad Q_{o(NH)}=\frac{r}{r+u}\times 100$$

上式中 p 为真阳性样本序列数，r 为真阴性样本序列数，u 假阴性样本序列数，o 为假阳性样本序列数。

2 计算结果及讨论

2.1 β- 发夹和非 β- 发夹的预测

随机选取 β- 发夹模体中的 5 000 个模体作为训练集，其余的 1 028 个模体作为独立的检验集；非 β- 发夹模体中的 2 000 个模体作为训练集，剩下的 643 个模体作为独立的检验集。对 β- 发夹模体的预测结果见表 1。

Table 1 The predictive results, respectively for 5-fold cross validation of training set and independent testing set

	Acc	Mcc	$Q_{o(H)}$	$Q_{o(NH)}$	$Q_{p(H)}$	$Q_{p(NH)}$
Training set	81.24%	0.57	83.06%	76.70%	89.91%	64.43%
Testing set	78.34%	0.56	77.24%	80.09%	86.12%	68.76%

从表 1 的预测结果看到：对训练集 5- 交叉检验的预测精度达到了 81.24%，相关系数为 0.57，β- 发夹的敏感性为 83.06%，特异性达到 89.91%，预测结果比较理想；对独立的检验集预测精度达到了 78.34%，相关系数为 0.56，β- 发夹的敏感性为 77.24% ，特异性达到 86.12%，相对来讲是一个比较好的结果。无论是预测精度还是 β- 发夹的敏感性都高于 Cruz[4]、Kumar[5]和 Kuhn[2]等的预测结果。

Kumar[5]也使用了 SVM 模型，预测精度为 79.2%，但利用了单序列信息、进化 profile、surface accessibility 和二级结构信息，相比之下，我们选用的参数少，输入 SVM 的参数维数低，预测精度也高。与 Cruz[4]、Kumar[5]选取固定序列模式长为 17 个氨基酸残基相比，我们的固定序列模式长选取较短。Kuhn[2]等选取的固定序列模式长也是 12 个氨基酸，但使用的 loop 长为 2~8 个氨基酸的 β- 发夹和非 β- 发夹，只覆盖了发夹和非发夹总数的 75%，而在本文中，loop 长 2~8 个氨基酸的 β-

发夹和非 β- 发夹覆盖了发夹和非发夹总数的 95%，相比之下我们的研究对象覆盖面大。

另外，我们给出了训练集 5- 交叉检验的 ROC 曲线及独立的检验集的 ROC 曲线，以证明使用的 SVM 分类器有较强的识别能力。图 3(A)为训练集的使用 5- 交叉检验的 ROC 曲线，曲线下的面积为 0.87；图 3(B)为独立的检验集的 ROC 曲线，曲线下的面积为 0.86。

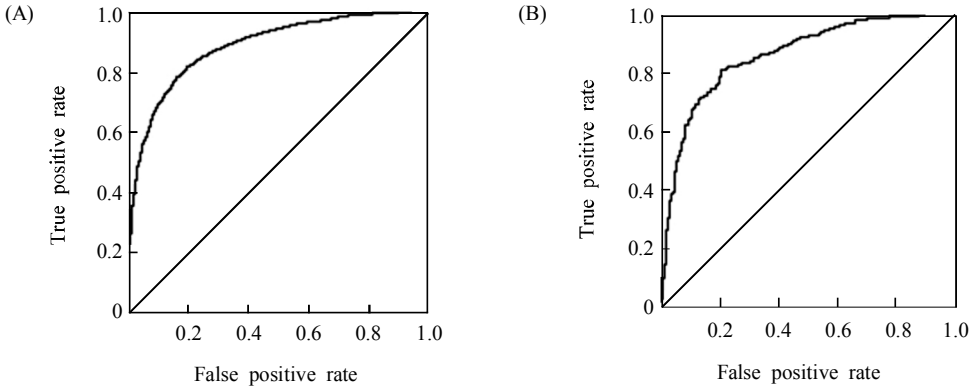

Fig.3 ROC curves in the training set for 5-fold cross validation and in the independent testing set. (A) Shows ROC curve of training set for 5-fold cross validation; (B) shows ROC curve of independent testing set

2.2　对 CASP6 的 63 个蛋白质的预测检验

用 CASP6 的 63 个蛋白质来检验我们的计算方法，首先对每一个蛋白质给出如下信息：(i)用 PSIPRED[14]对蛋白质序列进行二级结构预测，给出 β-β 模式；(ii)用 PROMOTIF[15]给出 β- 发夹模体，其中与 β-β 模式（loop 长 2~8 个氨基酸）精确匹配的有 26 个，46 个是非精确匹配的（预测和观察区域交叠，但几个残基的二级结构不匹配）；(iii)共给出 72 个 β- 发夹和 91 个非 β- 发夹模体（loop 长 2~8 个氨基酸）。

对 72 个 β- 发夹的预测结果是：精确匹配的 26 个模体中有 21 个被预测出来（β- 发夹的敏感性为 80.77%）；46 个非精确匹配的模体中有 36 个被预测出来（β- 发夹的敏感性为 78.26%）。具体情况见表 2。

Table 2-1 Prediction results for β-hairpins of 63 proteins in the CASP6 dataset

Exact matches the β-hairpins		Exact matches the β-hairpins		Exact matches the β -hairpins	
RWVYKLNGVTLEVNRV	Yes	CLIVEIGGVYFVRR	Yes	SIHVDGEGTCLVT	No
EKFVLENGVL	Yes	TAIVQIRNREMPVKVT	No	VKTLIVLDNAGGVYAVVI	Yes
LGIVSGGRLI	Yes	RYELRNGEIRAT	Yes	YGTYGMVSESGEHFNAI	No
ISGEFSLFAKGYWVENGEIA	Yes	FRIHAIAGGYRFLT	Yes	GQVKVKFDVTPDGRVDNVQILS	Yes
NIVIKLLEVNGNHAIKIS	Yes	KMIDVALRVDGVEVDRIR (*)	Yes	(*)	
TREFSLRLANGDLDQYTD	Yes	MIQMGTKFYQI	No	MMFHVRTDSNHDVLMD	Yes
LTIQVNGVP	Yes	VIDESSHFVSVA	No	KFYQIDSTGKLSE	Yes
LNGVALHFEGGKATAAERFI	Yes	KAIVVADGQKSV (*)	Yes	RLFVAESESGEVVGFAAF	Yes
TDLAVLSDGELQLTTI (*)	Yes	KVELRQVECINGNIFLHLGAV	Yes	HYQAFIRDELDENKWYKF	Yes

Note: *mark shows correct prediction β-hairpin by our method, but incorrect prediction by Kumar's method. Boldfaces show amino acids in loops; “Yes” and “No” show correct and incorrect prediction β-hairpins by our method. Kumar's results are from reference 5

Table 2-2 Prediction results for β-hairpins of 63 proteins in the CASP6 dataset

Non-exact matches the β-hairpins		Non-exact matches the β-hairpins		Non-exact matches the β-hairpins	
IKVTVTNSFFEV	No	VGFKVKGPSGIGGIVRI	No	AIIMKVDKDRQMVVL	Yes
RWIGNMMFHVR	No	LGLVYDIQIDDQNNVKVLMTM	Yes	ILRVMLIPVSEDYLIFSIL	Yes
KVERMSKTVYTVDWV	Yes	PVRFTDAQGNQHEGIIT	Yes	GGIVRIERNREKVEFAI	No
REIELGGAKLWEVAYGF	Yes	VICKPIGPSKVYVS	Yes	YVNFYIANGGII (*)	Yes
IVYYFTEDFFRLVV	Yes	RLLEGERGPWVQI	Yes	RLYTHPDGRIVVVP	Yes
FNVEIKVLKDGKTGTFRT	Yes	FVSVAPFAATYPFEIWI	Yes	LAFDREGYRL	Yes
VLRVGRFEDDGYFCTIEV-TATSTVT	Yes	RMVDFHGWMMPLHY (*)	Yes	LEVNRVEGIGDFVDIEV	No
VISFEGGKLKVRVKAI	No	DLFIATTGYTGEAGYEIAL (*)	Yes	LIFSILTEFGVSKVTPIK	Yes
YSYKYVHDDGRVSYPLCFIF	Yes	TFSPTLGYSIALARV (*)	Yes	FVHFRNVKYLGEHRFE	Yes
FVIDDAKNIYIYVSG	Yes	VNVLFVDDEAKTNQIFAR-RRLSFDCTATLK	Yes	HWIITDANGKTSEVQ	No
WYKFNDDKVSVV	No	KDFRIEIEYERTEEHPRIFTKVH-LKYIFKF	No	RILKGGTAYQT	Yes
GWRTFDVNGEKLTVVNLL	Yes	YNLIGVITHQGANSES-GHYQAFIR	Yes	MCCFARPGVVLLSW	Yes
VYFEVPRPKLLRIR	Yes	VVLVRKVGAPGNPEFALGAV(*)	Yes	QVEYFNSKLKQKFTLTLG (*)	Yes
FAVFSGKYFKGESPIGSVYLF	Yes	FLFAMAIARDANPRSGSWYE-LAR	Yes	MEVTTDHGVIKL	Yes
KIKYTGGELCI	Yes	HLHFITEDKTSGGHVLNLQFD	Yes	KALYSGMLNASGGVID	Yes
TVRGVFIVDARGVIRTMLYY	No				

Note: *mark shows correct prediction β-hairpin by our method, but incorrect prediction by Kumar's method. Boldfaces show amino acids in loops; "Yes" and "No" show correct and incorrect prediction β-hairpins by our method. Kumar's results are from Reference 5

对 72 个 β- 发夹和 91 个非 β- 发夹模体的预测结果是：β- 发夹被预测对的有 57 个，$Q_{o(H)}$= 79.17%，非 β- 发夹预测对的有 59 个，$Q_{o(NH)}$= 64.84%；Acc=71.17%。用我们设计的计算方法预测 63 个蛋白质中的 β- 发夹，基础是基于 PSIPRED 给出的 β-β 模式，如果 β-β 模式的精确度很高，那么就能得到较高的 β- 发夹的预测精度。

将 63 个蛋白质的 β- 发夹预测情况与 Kumar 等[5]对 β- 发夹的预测对比：1）Kumar 对 β- 发夹预测的敏感性只有 65.28%，我们的结果是 79.17%；尤其是非精确匹配的模体，被我们预测出来的数目比 Kumar 多（Kumar 预测对 25 个）。2）我们识别出了一些 Kumar 没识别出来的 β- 发夹模体，如，"KAIVVADGQKSV"、"TDLAVLSDGELQLTTI"、"YVNFYIANGGII"、"RMVDFHGWMMPLHY" 等。

参考文献：

[1] 阎隆飞, 孙之荣. 蛋白质分子结构. 北京：清华大学出版社, 1999. 43~44

[2] Kuhn M, Meiler J, Baker D. Strand-loop-strand motifs: prediction of hairpins and diverging turns in proteins. *Proteins*: *Struct Funct Bioinform*, 2004,54:282~288

[3] Sun ZR, Rao X, Peng L, Xu D. Prediction of protein super secondary structures based on artificial neural network method. *Protein Eng*, 1997,10:763~769

[4] Cruz X, Hutchinson EG, Shepherd A, Thornton JM. Toward predicting protein topology: an approach to identifying B hairpins. *Proc Natl Acad Sci USA*, 2002,99:11157~11162

[5] Kumar M, Bhasin M, Natt NK, Raghava GPS. BhairPred: prediction of b-hairpins in a protein from multiple alignment information using ANN and SVM techniques. *Nucleic Acids Res*, 2005,33:154~159

[6] Oliva B, Bates PA, Querol E, Aviles FX, Sternberg MJE. An automated classification of the structure of protein loops. *J Mol Biol*, 1997,266:814~830

[7] Espadaler J, Fuentes NF, Hermoso A, Querol E, Aviles FX, Sternberg MJE, Oliva B. ArchDB: automated protein loop classification as a tool for structural genomics. *Nucleic Acids Res*, 2004,32:185~188

[8] Laxton RR. The measure of diversity. *J Theor Biol*, 1978,71: 51~67

[9] Li QZ, Lu ZQ. The prediction of the structural class of

protein: application of the measure of diversity. *J Theor Boil*, 2001,213(3):493~502

[10] 徐克学. 生物数学. 北京: 科学出版社, 2001.277~286

[11] Cortes C, Vapnik VM. Support vector networks. *Machine Learning*, 1995,20:273~297

[12] Chang CC, Lin CJ. LIBSVM: a library for support vector machines. 2001, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[13] 刘 爽, 贾传莹, 陈 鹏. 一种自动选择参数的加权支持向量机算法. 计算机工程与应用, 2006,2:64~66

[14] Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*, 1999,292: 195~202

[15] Hutchinson EG, Thornton JM. PROMOTIF——a program to identify and analyze structural motifs in proteins. *Protein Sic*, 1996,5:212~220

THE β-HAIRPIN MOTIFS PREDICTION USING SUPPORT VECTOR MACHINE

HU Xiu-zhen[1,2], LI Qian-zhong[1]

(*1. Department of Physics, College of Sciences and Technology, Inner Mongolia University, Hohhot 010021, China;*
2. Department of Physics, College of Sciences, Inner Mongolia University of Technology, Hohhot 010059, China)

Abstract: Based on the protein sequence, a new method for predicting supersecondary structure motif, β-hairpins, is proposed. By using of the composite vector with increment of diversity to express the information of sequence, and input the increment of diversity to support vector machine(SVM)，SVM can find the optimization hyper plane in six dimension space to classify the β-hairpins and the non-β-hairpins. The result indicates that the higher predicting accuracy of β-hairpin motifs is obtained by using of our method. For training set 5-fold cross validation, the overall accuracy of prediction, Matthew's correlation coefficient (MCC) and sensitivity for β-hairpins are 81.24%, 0.57 and 83.06%, respectively. For independent testing set, the overall accuracy of prediction, MCC and sensitivity for β-hairpins are 78.34%, 0.56 and 77.24%, respectively. In addition, the performance of the method was also evaluated by predicting the 63 proteins in the CAPS6 dataset. And the better results are obtained by using our method.

Key Words: Super secondary structure；β-hairpin motif；Increment of diversity；
Support vector machine

This work was supported by grants from The National Natural Sciences Foundation of China (30560039) and the Natural Science Foundation of the Inner Mongolia of China (200508010509, 200607010101)

Received: Feb 10, 2007

Corresponding author: LI Qian-zhong, Tel：+86(471)4992958，E-mail：qzli@imu.edu.cn