生物化学与生物物理进展 *Progress in Biochemistry and Biophysics*, 2006, 33(2): 155~162
www.pibb.ac.cn

基于支持向量机融合网络的蛋白质折叠子识别研究 *

施建宇[1)**] 潘 泉[1)] 张绍武[1,2)] 梁 彦[1)]
([1)]西北工业大学自动化学院，西安 710072；[2)]西北工业大学生命科学院，西安 710072)

摘要 在不依赖于序列相似性的条件下，蛋白质折叠子识别是一种分析蛋白质结构的重要方法. 提出了一种三层支持向量机融合网络，从蛋白质的氨基酸序列出发，对 27 类折叠子进行识别. 融合网络使用支持向量机作为成员分类器，采用“多对多”的多类分类策略，将折叠子的 6 种特征分为主要特征和次要特征，构建了多个差异的融合方案，然后对这些融合方案进行动态选择得到最终决策. 当分类之前难以确定哪些参与组合的特征种类能够使分类结果最好时，提供了一种可靠的解决方案来自动选择特征信息互补最大的组合，保证了最佳分类结果. 最后，识别系统对独立测试样本的总分类精度达到 61.04%. 结果和对比表明，此方法是一种有效的折叠子识别方法.

关键词 折叠子识别，支持向量机，分类器融合，动态选择
学科分类号 Q617

通常认为蛋白质的三维空间结构是由它的一级结构——氨基酸序列决定的，了解氨基酸序列如何形成蛋白质的三维空间结构一直是分子生物学中最重要的目标之一. 为了能够实现这样一个目标，人们针对蛋白质结构的不同层次和方面展开了很多研究，比如二级结构预测[1]，结构类预测[2]，折叠子预测[3~5]，同源寡聚体预测[6]等. 折叠子包含一个或多个蛋白质超家族，这些蛋白质超家族中的蛋白质核心结构相同，根据蛋白质肽链的拓扑相似性，可以定义一系列的折叠子[3]. 每一个折叠子的结构内核有确定的结构特征，具有相同折叠子的不同蛋白质具有相同的内核结构特征. 具有相同折叠子的蛋白质，在结构上的相似性可能与蛋白质肽链拓扑学，以及肽链空间排布与侧链堆积的物理化学因素有关. 在不依赖于序列相似性的条件下，蛋白质折叠子识别是一种分析蛋白质结构的重要方法.

目前国外一些研究组已经开展了蛋白质折叠子识别的研究并取得了一定的进展[3~5,7]，而国内对此研究较少. 在他们的工作中，折叠子识别的特征提取方法是基于氨基酸组成成分[7]，或者是氨基酸组成成分和若干种物化特征的简单复合[3,4]. 单一的氨基酸组成成分不能充分表达折叠子信息，多种特征经过简单复合后虽然能表达更多的信息，但是那种串联方式的复合，随着特征种类增多，特征向量维数增大，噪声也逐渐增大，且事先难以确定哪些参与组合的特征种类能够使分类结果最好. Chinnasamy 等[5]进一步考虑多种特征之间的互补信息，提出了一种 BAYESPROT 系统，使用多个贝叶斯分类器以并联方式对特征进行复合(融合)，取得了更好的分类结果.

支持向量机[8]是近年来国际上新兴的一种机器学习方法，由于出色的学习性能，该技术已成为当前的研究热点，且已被成功地应用于生物信息学的很多方面：蛋白质家族[9]、转录起始点[10]、蛋白质亚细胞定位[11,12]、蛋白质折叠子识别[4]等方面. 本文提出一种三层支持向量机融合网络，对 27 类蛋白质折叠子进行预测研究.

1 材料与方法

1.1 数据库

本文使用的数据库来自文献[4]. 数据库最早由 Dubchak 等[3]从 PDB 中挑选样本构建而成，折叠子

*国家自然科学基金资助项目(60404011).
** 通讯联系人. Tel: 029-88494352, E-mail: snake5947@hotmail.com
收稿日期：2005-08-23，接受日期：2005-09-30

被分为 128 类，后来 Ding 等[4]仅挑选出其中数目较多的 27 类进行研究，之后，Chinnasamy 等[5]使用该数据库对折叠子进行了进一步的研究，他们的研究均表明该数据库是合理的. 折叠子数据库包含了一个训练集和一个独立测试集，训练集和独立测试集样本数分别为 313 和 385，任意两条样本序列的同源性分别小于 35%和 40%，均可分为 27 类. 从各类样本数目分布来看，数据库是不均衡的，这对折叠子识别系统来说是一个不小的挑战. 整个数据库可以从 http://www.nersc.gov/~cding/protein/ 下载获得.

所有样本从氨基酸序列经特征提取形成氨基酸组成成分(*C*)、极性(*P*)、极化性(*Z*)、范德瓦尔斯量(*V*)、疏水性(*H*)和预测的二级结构(*S*)等 6 种特征数据，特征信息见表 1. 其中二级结构是基于神经网络和序列比对的方法并结合蛋白质进化信息进行预测的[13].

Table 1 Six groups of fold features extracted from protein sequence

Symbol	Feature	Dim
C	Amino acid composition	20
P	Polarity	21
Z	Polarizability	21
V	Normalized Van Der Waals volume	21
H	Hydrophobicity	21
S	Predicted secondary structure	21

Symbol: Short name of feature group; Feature: Full name of feature group; Dim: Dimension of feature vector.

表 1 所列几种特征信息中 *P*、*Z*、*V*、*H* 和 *S* 都包含一定的氨基酸关联信息. 文献[4]采用了全局蛋白质序列描述特征提取方法[3]：首先将 20 个基本氨基酸分别按照极性、极化性、范德瓦尔斯量、疏水性和预测的二级结构等物化特性分为若干组，比如预测的二级结构可分为螺旋(Helix)、叠片(Strand)、卷曲(Coil)和未知(Unknown) 4 组. 然后使用 3 种描述子(Descriptors)来描述蛋白质特征，第一种是组成成分(Composition)，描述了给定氨基酸特征各组的全局组成成分；第二种是交替频率(Transition)，描述了沿着序列顺序特征各组两两变更的频率；第三种是分布(Distribution)，描述了沿着序列方向各组特征的分布模式.

1.2 串联复合特征

文献[3, 4]的研究表明，将多种特征串联起来形成一个更高维的特征是一种有效的方法，能提高分类精度. 例如，对于氨基酸组成成分和预测的二级结构这两种特征，一个蛋白质序列可表示为如下特征向量：$x=[c_1, c_2, \cdots, c_{20}, s_1, s_2, \cdots, s_{21}]$，多种特征复合的特征向量以此类推. 多种特征经过串联方式的复合后虽然能表达更多的信息，但是随着特征种类增多，特征向量维数增大，噪声也逐渐增大，导致最终分类精度降低. 由于无法明确表达或量化各种特征之间的冲突 / 互补关系，因此在分类之前我们难以确定哪些参与组合的特征种类能够使分类结果最好. 文献[4]使用支持向量机和神经网络对折叠子进行识别，其最好分类结果对应特征为氨基酸组成成分、疏水性和预测的二级结构. 文献[4]的研究表明支持向量机比神经网络更有效.

1.3 BAYESPROT 系统

文献[5]提出了一种 BAYESPROT 系统，首先采用文献[4]的特征提取方法，将一个蛋白质序列分别表示为 3 个特征向量：氨基酸组成成分、预测的二级结构和串联所有 6 种特征而形成的复合特征，然后将 3 个特征向量分别输入到 3 个贝叶斯分类器进行判别，最后使用平均概率投票算法 (mean probability voting) 对 3 个分类器的判别结果进行最终的决策. 文献[5]的研究表明：利用不同的特征对于分类器模式的互补信息，可以提高分类性能，它能整合来自多信息源的信息，降低单信息源中存在的不确定性，从而提高系统的整体性能.

1.4 支持向量机

支持向量机 (support vector machines，SVM)是一种基于统计学习理论的分类方法[8]，它对“维数灾难”不敏感并且能够足够有效地处理数据量大和特征输入空间大的分类问题，并且存在高效和高质量的算法实现[14]，这方便了 SVM 在各个领域中的应用，有关 SVM 更详细的信息可以从文献[14]获得. SVM 存在多种多类分类的算法，比如“一对多”(One-Versus-Rest[15]或 One-Versus-All[4])，“一对一”(One-Versus-One) [15] 或称“多对多”(All-Versus-All)[4]，有向非循环图法(DAGSVM)[16]以及其他正在验证和证明之中的多类分类算法[17,18]. 研究结果表明[18]，“多对多”和 DAGSVM 更适合在实践中使用，且 SVM 的核函数选用径向基函数通常能取得较好的分类效果. 本文采用 SVM 作为成员分类器，核函数选用径向基函数，采用“多对多”的多类分类算法.

1.5 支持向量机融合网络

支持向量机融合网络 (support vector machines

fusion network，SFN) 分为三层：输入层 (input layer)、融合层 (fusion layer) 和决策层 (decision layer)，其构架图如图 1 所示.

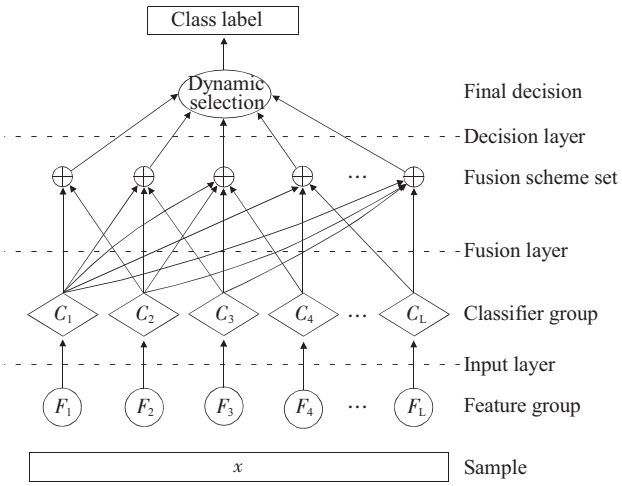

Fig. 1　The architecture of support vector machines fusion network

设折叠子样本空间 X 中包含 c 类样本，F_i 表示样本第 i 种特征数据，共有 L 种特征. 每一个样本分别经过输入层、融合层和决策层之后，得出分类判别结果，详细算法如下.

输入层：首先将所有样本映射到每一个特征空间，然后根据各种特征自检测分类精度的大小或者先验知识来确定主要特征和次要特征，每一种特征对应一个成员分类器. 设经过特征选取有 m 种主要特征，L-m 种次要特征，相应地有 m 个主要分类器，L-m 个次要分类器.

融合层：根据输入层确定的主次分类器使用训练集 Z 可构建差异融合方案集 $R=\{R_s \mid s=1,\cdots,2^{L-m}\}$，其中主分类器参与每一个融合方案 R_s.

设 $D_i(x)=\{d_{i,1}(x)\cdots d_{i,j}(x)\cdots d_{i,c}(x)\}$表示成员分类器 C_i 对样本 x 的分类结果，$d_{i,j}(x)$表示成员分类器 C_i 将样本 x 分为第 j 类的概率(后验概率)，其中 $i=1, \cdots, L_s, j=1, \cdots, c$, L_s 表示融合方案 R_s 所使用成员分类器的数目，$s=1, \cdots, 2^{L-m}$. 于是对于每一个样本 x 经过 L_s 个分类器都可以得到一个决策轮廓 (*DP*) 矩阵[19]：

$$DP_s(x)=\begin{bmatrix} d_{1,1}(x) \cdots d_{1,j}(x) \cdots d_{1,c}(x) \\ \cdots \\ d_{i,1}(x) \cdots d_{i,j}(x) \cdots d_{i,c}(x) \\ \cdots \\ d_{L_s,1}(x) \cdots d_{L_s,j}(x) \cdots d_{L_s,c}(x) \end{bmatrix} \quad (1)$$

文献[20]表明概率乘积规则[21]往往适用于各个分类器的输入来自不同特征的场合，于是本文采用概率乘积作为融合方案 R_s 的融合规则来计算样本 x 属于第 j 类的概率 $P_{s,j}(x)$：

$$P_{s,j}(x)=\frac{\prod_{i=1}^{L_s} d_{i,j}(x)}{P(j)^{L_s-1}},\ j=1, \cdots, c \quad (2)$$

令 $j_s^*=\arg\max_j(P_{s,j}(x)), j=1, \cdots, c$，则 j_s^*就是融合方案 R_s 对样本 x 的融合判别结果. (2)式中 $P(j)$表示第 j 类样本的先验概率，可由对训练样本集的估计得出：

$$P(j)=\frac{N_j}{N},\ j=1, \cdots, c \quad (3)$$

式中 N_j 为训练样本集中属于第 j 类的样本数目，N 为训练样本的总数目.

决策层：借鉴局部精度动态选择算法[19]的思想，对多个融合方案进行动态决策. 其计算步骤如下：

步骤 1. 给定一个样本 x，使用所有的融合方案进行类别判定，如果判定结果均相同，则 x 的类别为判定结果，结束判别. 否则转入第 2 步.

步骤 2. 根据区域阈值 K (推荐 K=10)为每一个融合方案 R_s 评估局部精度. 在训练集 Z 中寻找 K 个经过 R_s 判定 (局部自检验) 属于第 j_s^*类的且离 x 最近的样本，这些被选出的 K 个训练样本中被正确判定的比例就是融合方案 R_s 对第 j_s^*类的局部精度. 局部精度最高的融合方案构成了集合 M_1，如果 M_1 只包含一个元素，则 x 的类别为该融合方案的判定结果，结束判别. 否则转入第 3 步.

步骤 3. 检查 M_1 中所有元素对样本 x 是否有相同的判定结果，相同则采用该判定结果，结束判别. 否则选择出现最多的判定结果，结束判别. 若还是不能选出结果 (所有 M_1 元素的判别结果均不相同)，转入第 4 步.

步骤 4. 类似上两步，具有第二高局部精度的融合方案形成了集合 M_2，如果 M_2 只包含一个元素，则 x 的类别为该融合方案的判定结果，结束判别. 如果 M_2 中包含多个元素，则检查他们对样本 x 是否有相同的判定结果，相同则采用该判定结果，结束判别. 否则选择出现最多的判定结果，结束判别. 若 M_2 所有元素的判别结果均不相同，转入第 5 步.

步骤 5. 在融合方案集合 M_1 或者 M_2 中随机选择一个元素的判别结果，结束判别.

1.6 评估参数

为了评估本文方法的有效性，文中采用了多类分类系统评估参数：总分类精度 Q，每类样本的分类精度 Q_j 和 Matthews 相关系数[6] $MCC(j)$. 它们分别定义为：

$$Q=\frac{\sum_{j=1}^{c} p(k)}{N} \tag{4}$$

$$Q_j=\frac{p(j)}{N_j} \tag{5}$$

$$MCC(j)=\frac{p(j)n(j)-u(j)o(j)}{\sqrt{p(j)+u(j))(p(j)+o(j))(n(j)+u(j))(n(j)+o(j))}} \tag{6}$$

式中 N 为样本总数，c 为样本类别总数，N_j 为类别 j 的样本数，$p(j)$为第 j 类样本的正确分类数，$n(j)$为非 j 类样本的正确分类数，$u(j)$为第 j 类样本中被错分为其他类别的样本数，$o(j)$为其他类别的样本被错分为第 j 类样本的数目.

此外，为了进行理论分析和对比，文中还采用了“绝对可靠”(Oracle) [19]的分类结果来评估融合网络的整体性能. “绝对可靠”是指对于样本 x 只要有一个分类器能够正确识别就算整体分类正确，这表现了融合网络的理论上界.

2 结果与讨论

实验中所有特征数据都按照如下方式进行归一化：首先计算训练集特征的最小值和最大值，并进行归一化，于是所有训练样本特征向量均被线形变换到[0,1]，然后使用同样的变换系数，对测试集进行归一化. 经过各种特征自检测分析，可知氨基酸组成成分和预测的二级结构是两种主要特征，其他 4 种的物化性质为次要特征，因此可构建 $2^4 = 16$ 种融合方案，每一种融合方案均采用概率乘积规则进行融合；在动态决策的时候，由于个别类别样本数较少，这里采用自动调整区域阈值的方法，当特征空间中某个区域样本数小于阈值 K 时，则将 K 的值为该区域样本数，否则 K 选取推荐值 10.

2.1 结果

为了验证本文方法的有效性，我们与文献[5]的结果 (以下简称 BAYESPROT) 和文献[4]的最好结果 (以下简称 Ding) 进行了详细的对比. 因为 Ding 没有给出混淆矩阵，所以我们无法计算出它的 *MCC* 值，只能给出了总分类精度和各类分类精度的对比. 且又因为 Ding 的结果小数点后只有一位有效数字，所以本文结果与其对比时，采用四舍五入的方式. 表 2 给出了独立测试集的实验结果.

2.2 测试集分类结果对比分析

从表 2 可看出，与 Ding 相比，SFN 结果的总精度提高了 5.0%，有 15 类分类精度提高了，其中 13 类精度提高了 10%以上，最大差值为 28.6%，有 4 类分类精度降低了，其中 3 类精度降低了 10%以上，最大差值为 24.1%，有 8 类分类精度相当. 图 2 显示了 SFN 与 Ding 的各类分类精度差值.

与 BAYESPROT 的分类精度相比，SFN 结果的总精度提高了 2.86%，有 13 类分类精度提高了，其中 6 类精度提高了 20%以上，最大差值为 57.14%，有 5 类分类精度降低了，其中 3 类精度降低了 10%以上，最大差值为 20.00%，有 9 类分类精度相当. 图 3 显示了 SFN 与 BAYESPROT 的各类分类精度差值. 与 BAYESPROT 的 *MCC* 值相比，SFN 有 17 类提高了，其中 11 类的 *MCC* 提高了 0.1 以上，最大差值为 0.302，有 8 类降低了，其中 3 类的 MCC 降低了 0.1 以上，最大差值为 0.206，有 2 类 *MCC* 相等.

对于 27 类折叠子数据库，且每类的样本数目又较少，尽管 SFN 取得的精度不算太高(61.04%)，但应注意对于 27 类问题其随机分类精度仅为 1/27=3.7%. 上述结果表明，本文方法是一种有效的折叠子识别算法.

2.3 SFN 分类性能分析

从识别的速度来看，整个识别过程中融合层与决策层只花费非常少的时间，最耗时的任务是训练分类器，对应 6 组特征需要训练 6 个分类器. 但是实验中我们发现 SFN 并不像看起来那么耗时，使用复合高维特征 (所有特征) 所花费的训练时间只比 6 组低维特征所花费的训练时间的总和少一些，这表明 SFN 完成一次识别的时间并非太多，能够满足在线识别的要求. 此外，SFN 还有一个优点就是可以通过并行计算来完成识别任务.

从表 2 中可以看出，与 Oracle 的结果相比，SFN 的总分类精度相差 8.31%，其中有 12 类分类精度已达到理论上界，说明我们采用的分类器和融合方法是合理的. 但是有 15 类折叠子在理论上还存在很大改进空间，精度最大差值为 37.5%，且 Oracle 的结果只有 69.35%，表明选用的特征并没有很好地覆盖特征空间，需要进一步研究具有更好互补关系的特征，才能提高 SFN 的分类精度. 图 4 显示了 Oracle 与 SFN 的各类分类精度差值.

Table 2 The comparison of SFN, BAYESPROT and Ding in the independent test

Fold		Oracle	SFN		BAYESPROT		Ding
Index	Class	Q_j / %	Q_j / %	*MCC*	Q_j / %	*MCC*	Q_j / %
α							
1	1	83.33	83.33	0.829	83.33	0.829	83.3
3	2	100.00	100.00	0.901	88.89	0.837	77.8
4	3	65.00	45.00	0.554	65.00	0.618	35.0
7	4	100.00	62.50	0.612	62.50	0.713	50.0
9	5	100.00	100.00	0.825	100.00	0.654	100.0
11	6	66.67	66.67	0.513	22.22	0.278	66.7
β							
20	7	77.27	68.18	0.608	77.27	0.644	71.6
23	8	41.67	33.33	0.384	8.33	0.146	16.7
26	9	92.31	69.23	0.705	69.23	0.543	50.0
30	10	33.33	33.33	0.462	33.33	0.462	33.3
31	11	75.00	37.50	0.417	50.00	0.623	50.0
32	12	42.11	42.11	0.350	31.58	0.379	26.3
33	13	75.00	50.00	0.571	50.00	0.396	50.0
35	14	50.00	50.00	0.364	25.00	0.278	25.0
39	15	85.71	71.43	0.618	57.14	0.606	57.1
α/β							
46	16	83.33	77.08	0.698	87.50	0.589	77.1
47	17	75.00	66.67	0.809	58.33	0.659	58.3
48	18	46.15	38.46	0.389	30.77	0.396	48.7
51	19	44.44	37.04	0.339	37.04	0.454	61.1
54	20	58.33	58.33	0.430	33.33	0.298	36.1
57	21	62.50	62.50	0.714	37.50	0.519	50.0
59	22	71.43	64.29	0.576	50.00	0.513	35.7
62	23	85.71	85.71	0.591	28.57	0.289	71.4
69	24	25.00	25.00	0.278	25.00	0.237	25.0
α+β							
72	25	37.50	37.50	0.417	25.00	0.267	12.5
87	26	48.15	37.04	0.351	40.74	0.338	37.0
110	27	100.00	96.30	0.920	96.30	0.938	83.3
Q / %		69.35	61.04	—	58.18	—	56.0

Index: Fold index in original dataset, Class: Class ID of all folds, Oracle: Oracle result of our approach, SFN: Result of our approach, BAYESPROT: Result of BAYESPROT [5], Ding: Best result of Ding[4], and CHS features are used, Q /%: Overall accuracy (in percent) for 27-class folds, Q_j /%: Accuracy (in percent) for each class, *MCC*: Matthews correlation coefficient for each class.

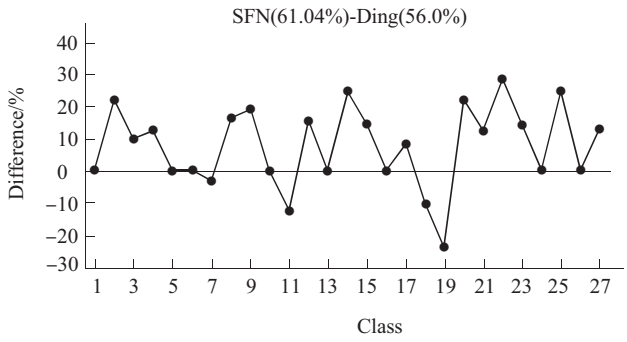

Fig. 2 The difference of accuracy for each class between SFN and Ding

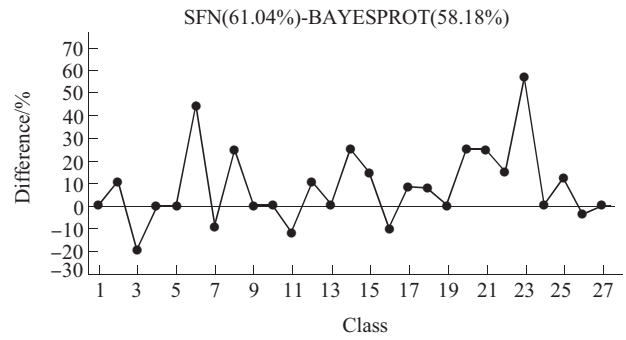

Fig. 3 The difference of accuracy for each class between SFN and BAYESPROT

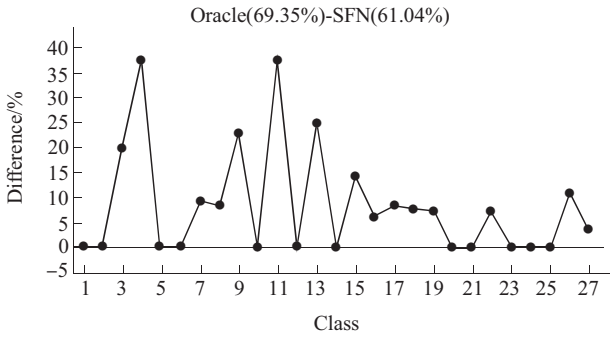

Fig. 4 The difference of accuracy for each class between Oracle and SFN

2.4 多种特征之间的冲突与互补分析

为了分析多种特征的冲突与互补，我们统计了SFN 融合层中的 16 种方案对独立测试集的总分类精度，如表 3 所示.

从表 3 中可看出，Scheme 1 (只有主要特征参与融合的结果) 与 Scheme 14 (主要特征与次要特征融合的最好结果) 的总分类精度只相差 1.82%，这也再次验证了文献[4, 5]的研究——氨基酸组成成分和二级结构是折叠子的主要特征，其他的物理化学特征是次要特征. 主要特征和次要特征融合之后，通常能改善分类性能，体现了特征之间的互补，但应注意到 Scheme 3、4 和 9 的总分类精度并不比 Scheme 1 效果好，且发现 Scheme 16 (所有特征均参与融合) 的结果反而比 Scheme 14 降低了 1.78%，可知参与融合的特征种类并非越多越好，所选用的各种特征包含的信息之间并非完全互补，而是存在着某种冲突.

Table 3 The overall accuracy of 16 fusion schemes in independent test

Scheme	*C*	*P*	*Z*	*V*	*H*	*S*	*Q* / %
1	√					√	57.40
2	√	√				√	58.18
3	√		√			√	57.14
4	√			√		√	57.40
5	√				√	√	58.18
6	√	√	√			√	58.70
7	√	√		√		√	57.92
8	√	√			√	√	57.66
9	√		√	√		√	56.62
10	√		√		√	√	58.44
11	√			√	√	√	58.96
12	√	√	√	√		√	58.44
13	√	√	√		√	√	58.44
14	√	√		√	√	√	59.22
15	√		√	√	√	√	58.70
16	√	√	√	√	√	√	58.18

Scheme: ID of fusion schemes, *C*, *P*, *Z*, *V*, *H*, *S*: Six groups of feature, see also table 1, *Q* / %: Overall accuracy (in percent) of each fusion scheme for 27-class folds, √ : The feature group used in fusion scheme.

通常蛋白质都具有多种特征且每种特征均包含了一定的信息，但是目前我们无法明确表达或量化不同种特征之间的冲突 / 互补关系，因此在进行最终决策之前，难以确定哪种组合方案能够使全局分类精度最好，且全局分类精度最好的融合方案不一定保证所有局部区域的分类精度最好. 本文方法为解决这种困难提供了一种可靠的解决方案，在无法获知特征之间冲突 / 互补关系时，SFN 能够自动选出对应于所有局部区域的局部分类精度最好的相应融合方案，从而达到最佳分类结果.

2.5 结论

本文提出的三层支持向量机融合网络(SFN)融合了多种特征信息，降低了单信息源中存在的不确定性，提高了识别系统的整体性能，是一种有效的折叠子识别方法. 同时也发现各种特征包含的信息之间并非完全互补，而是存在着某种冲突，且无法明确表达或量化不同种特征之间的冲突 / 互补关系，导致在分类之前难以确定哪些参与组合的特征

种类能够使分类结果最好. 对于这类问题，SFN 提供了一种可靠的解决方案来自动选择特征信息互补最大的组合，保证了最佳分类结果. 如果我们能够获得更多的样本，采用更多种的携带互补信息的折叠子特征，并进一步考虑更多氨基酸之间的关联信息，就能更全面地描述特征空间，SFN 的分类能力必将提高. 为了进一步验证 SFN 的有效性，我们将在下一步研究中应用 SFN 于蛋白质的其他分类问题中.

参 考 文 献

1 Kneller D G, Cohen F E, Langridge R. Improvements in protein secondary-structure prediction by enhanced neural networks. J Mol Biol, 1990, **214** (1): 171～182

2 Zhang C T, Chou K C. An optimization approach to predicting protein structural class from amino acid composition. Protein Sci, 1992, **1** (3): 401～408

3 Dubchak I, Muchnik I, Mayor C, *et al*. Recognition of a protein fold in the context of the SCOP classification. Proteins, 1999, **35** (4): 401～407

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

5 Chinnasamy A, Sung W K, Mittal A. Protein structure and fold prediction using tree-augmented naive bayesian classifier. J Bioinform Comput Biol, 2005, **3** (4): 803～820.

6 张绍武, 潘 泉, 陈润生, 等. 基于支持向量机的蛋白质同源寡聚体分类研究. 生物化学与生物物理进展, 2003, **30** (6): 879～883
Zhang S W, Pan Q, Chen R S, *et al*. Prog Biochem Biophys, 2003, **30** (6): 879～883

7 Nakashima H, Nishikawa K, Ooi T. The folding type of a protein is relevant to the amino acid composition. J Biochem, 1986, **99** (1): 153～162

8 Vapnik V. The Nature of Statistical Learning Theory. New York: Spinger-Verlag, 1995. 1～188

9 Jaakkola T, Diekhans M, Haussler D. Using the fisher kernel method to detect remote protein homologies. In: Lengauer T, eds. Proceedings of The Seventh International Conference on Intelligent Systems for Molecular Biology. Menlo Park: AAAI Press, 1999. 149～158

10 Zien A, Ratsch G, Mika S, *et al*. Engineering support vector machine kernels that recognize translation initiation sites. Bioinformatics, 2000, **16** (9): 799～807

11 Hua S J, Sun Z R. Support vector machine approach for protein subcellular localization prediction. Bioinformatics, 2001, **17** (8): 721～728

12 Cai Y D, Liu X J, Xu X B, *et al*. Support vector machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect. J Cell Biochem, 2002, **84** (2): 343～348

13 Rost B, Sander C. Combining evolutionary information and neural networks to predict protein secondary structure. Proteins 1994, **19** (1): 55～72

14 Joachims T. Making large-scale SVM learning practical. In: Schölkopf B, eds. Advances in Kernel Methods:Support Vector Learning. Cambridge: MIT Press, 1999. 169～184

15 Kre(el U. Pairwise classification and support vector machines. In: Schölkopf B, eds. Advances in Kernel Methods:Support Vector Learning. Cambridge: MIT Press, 1999. 255～268

16 Platt J, Cristianini N, Shawe-Taylor J. Large margin dags for multiclass classification. In: Jordan M I, eds. Proceedings of Neural Information Processing Systems. Cambridge: MIT Press, 2000. 547～553

17 Crammer K, Singer Y. On the learnability and design of output codes for multiclass problems. In: Cesa-Bianchi N, eds. Proceedings of the Thirteen Annual Conference on Computational Learning Theory. San Francisco: Morgan Kaufmann Publishers, 2000.35～46

18 Hsu C W, Lin C J. A comparison of methods for multi-class support vector machines. IEEE Transactions in Neural Networks, 2002, **13** (2): 415～425

19 Kuncheva L I. Switching between selection and fusion in combining classifiers: an experiment. IEEE Transactions on SMC (Part B), 2002, **32** (2):146～156.

20 Webb A R. 王 萍等译. 统计模式识别. 第 2 版. 北京: 电子工业出版社, 2004. 218～226
Webb A R. Translated by Wang P, *et al*. Statistical pattern recognition. 2nd. Beijing: Publishing House of Electronics Industry, 2004. 218～226

21 Kittler J, Hatef M, Duin R, *et al*. On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, **20** (3): 226～239

Protein Fold Recognition With Support Vector Machines Fusion Network*

SHI Jian-Yu[1)**], PAN Quan[1)], ZHANG Shao-Wu[1,2)], LIANG Yan[1)]

([1)] *College of Automation, Northwestern Polytechnical University, Xi'An* 710072, China;

[2)] *College of Life Science, Northwestern Polytechnical University, Xi'An* 710072, China)

Abstract One of the important approaches to structure analysis is protein fold recognition, which is often applied when there is no significant sequence similarity between structurally similar proteins. A framework with a three-layer support vector machines fusion network (SFN) is presented. The framework is applied to 27-class protein fold recognition from primary structure of proteins. SFN uses support vector machines as member classifiers, and adopts All-Versus-All as multi-class categorization. Six groups of features are divided into major and minor ones by SFN, and several diversity fusion schemes are correspondingly built. The final decision is made by dynamic selection of the results of all fusion schemes. When it is still difficult to know what kind of fusion of feature groups can achieve good prediction, SFN is a dependable solution by selecting the optimal fusion of feature groups automatically, which can ensure the best recognition. Overall recognition system achieves 61.04% fold prediction accuracy on the independent test dataset. The results and the comparison with other approaches demonstrate the effectiveness of SFN, and thus encourage its further exploration.

Key words protein fold recognition, support vector machines (SVM), classifier fusion, dynamic selection

---

*This work was supported by a grant from The National Natural Science Foundation of China (60404011).

**Corresponding author . Tel: 86-29-88494352, E-mail: snake5947@msn.com

Received: August 23, 2005 Accepted: September 30, 2005