

# 结合位点进化距离与支持向量机的 蛋白质分类方法

李玉岗<sup>1)</sup> 张 法<sup>2)</sup> 刘志勇<sup>2)</sup>

<sup>1)</sup>(北京理工大学计算机科学技术学院智能信息技术北京市重点实验室 北京 100081)

<sup>2)</sup>(中国科学院计算技术研究所 北京 100080)

**摘 要** 生物信息学的一个关键的研究课题是理解细胞的分子机制,这依赖于对基因所决定的每一条蛋白质的含义或者功能的理解.一般通过与一条或多条功能已知的蛋白质的相似性比较来推测未知蛋白质的功能,其中,基于支持向量机的一些算法取得了很好的成果.SVM-pairwise 算法是当前最好的基于支持向量机的算法中的一个,该方法利用两条序列的相似性来将蛋白质序列转化为固定长度的向量.文中提出了一种新的利用支持向量机算法对蛋白质序列进行分类的方法,这种方法使用位点进化距离代替两条序列的比对得分,该方法比 SVM-pairwise 有着显著的改善,在蛋白质结构分类数据库(SCOP)上进行的实验表明,该方法具有比 SVM-pairwise 更好的分类性能.

**关键词** 生物信息学;内核;位点进化距离;支持向量机;蛋白质结构分类数据库  
中图法分类号 TP18

## Combining Position-Specific-Value Method and SVM for Remote Protein Classification

LI Yu-Gang<sup>1)</sup> ZHANG Fa<sup>2)</sup> LIU Zhi-Yong<sup>2)</sup>

<sup>1)</sup>(Beijing Key Laboratory for Intelligent Information Technology,  
School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081)

<sup>2)</sup>(Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080)

**Abstract** An important research topic in bioinformatics is to understand the meaning and function of each protein encoded in the genome. One of the most successful approaches to this problem is via sequence similarity with one or more proteins whose functions are known. The SVM based methods are among the most successful ones. Currently, one of the most accurate homology detection method is the SVM-pairwise method. This method combines the pairwise sequence similarity with Support Vector Machine. This paper presents an alternative for SVM-based protein classification. The method, SVM-PSV, uses a new sequence similarity kernel, the Position Specific Values (PSV) kernel, for use with Support Vector Machines (SVMs) to solve the protein classification problem. The resulting algorithm gives better recognizing accuracy in the comparison with state-of-art methods, including SVM-pairwise, in the experiments of the detection of the homology based on the SCOP database. In the respect of computational efficiency, this method is significantly better than the SVM-pairwise one.

**Keywords** bioinformatics; kernel; PSV; SVM; SCOP

收稿日期:2005-08-25;最终修改稿收到日期:2006-11-26. 本课题得到国家自然科学基金(60503060,90612019,60752001)资助. 李玉岗,男,1971年生,博士,研究方向为计算机算法、并行处理与生物信息学. E-mail: lyg@ncic.ac.cn. 张 法,男,1974年生,博士,助理研究员,研究方向为计算机算法、并行处理与生物信息学. E-mail: zf@ncic.ac.cn. 刘志勇,男,1946年生,博士,研究员,研究领域为计算机算法和体系结构、并行处理、网络与图像处理. E-mail: zyliu@ict.ac.cn.

## 1 引 言

理解分子机制的一个关键问题是理解由染色体中的基因所决定的蛋白质的含义或者功能. 通过比较一条蛋白质序列与一条或者几条功能已知的蛋白质序列的相似性, 对蛋白质序列进行分类, 从而达到推断该蛋白质的功能的目的, 是解决这个问题的有效途径之一. 用于蛋白质的分类问题的算法可以分为两类: 产生式方法和判别式方法. 迄今为止, 每一类算法都包含很多不同的实现方法. 产生式方法包括基于两条序列比对算法<sup>[1-6]</sup>、基于蛋白质家族的序列谱(profiles)<sup>[7]</sup>和基于隐马尔可夫模型的算法<sup>[8-10]</sup>. 同时, 基于判别式的算法也很多<sup>[11-17]</sup>, 其中, 基于支持向量机的算法<sup>[12-16]</sup>占了相当的比重, 也是现在所采用的精确度最高的算法之一. 文献[12-13]采用了基于隐马尔可夫的 Fisher 方法; 文献[15-16]采用了频谱内核(spectrum kernel); 而文献[14]则是在两条序列相似性算法(如 Smith-Waterman 算法, BLAST 算法等)的基础上, 将不定长的序列转化为固定长度的向量, 向量的每个分量是该序列与相应的蛋白质家族中的成员序列的相似度.

产生式方法的原理是: 针对某个蛋白质家族建立模型, 然后评价一个要进行分类的序列与该模型相符的程度. 如果相符程度超出了某个事先设定的阈值, 则认为这条序列属于模型所对应的家族, 否则, 认为它不属于这个家族. 判别式方法的原理与产生式方法的原理有着显著的区别, 它首先将已知的蛋白质序列分为正样本和负样本, 属于所考虑的蛋白质家族的序列为正样本, 否则为负样本, 在此基础上进行有监督的学习, 并利用学习的结果来判断未知序列是否属于相应的蛋白质家族. 由于产生式方法只利用了正样本序列的信息, 而判别式方法既利用了正样本序列的信息, 又利用了负样本序列的信息. 因此, 一般来说, 判别式方法的精确度要高于产生式方法.

按照对蛋白质序列分类的精确程度进行划分, 现有的蛋白质分类算法可分为 4 类<sup>[14]</sup>. 第 1 类是基于两条序列比对的算法, 是精确度最低的, 出现的时间也最早. 这一类算法中, 精确度最高的算法是基于动态规划的算法<sup>[1-3]</sup>, FASTA<sup>[5]</sup>和 BLAST<sup>[4]</sup>在动态规划算法的基础上, 结合启示性方法, 比单纯的动态规划算法快, 但这是以牺牲精确性作为代价的. 第 2 类算法以隐马尔可夫模型<sup>[9-10]</sup>和序列谱<sup>[7]</sup>为代表,

将一组相似的序列进行整体统计, 然后利用统计结果与未知的单个序列相比较, 来对未知序列进行归类. 第 3 类算法不但考虑了相似的序列, 还将数据库中的大量的没有经过标识的序列考虑进去, 作为对比信息用于对未知序列的分类, SAM-T98<sup>[8]</sup>和 PSI-BLAST<sup>[6]</sup>是其代表. 第 4 类方法使用有监督的机器学习算法, 如贝叶斯神经网络(BNN)<sup>[11,17]</sup>等. 这一类算法显式地将样本序列划分为正样本和负样本, 然后进行学习, 最后利用学习的结果对未知的样本进行归类. 采用这类算法, 长度不固定的序列空间的序列首先要转换到长度固定的向量空间的向量. 迄今为止, 有很多不同的方法用于这一空间变换<sup>[12-17]</sup>. 文献[12-13]首先建立一个隐马尔可夫模型, 然后利用一个蛋白质家族中的成员序列来优化该模型, 使得这个蛋白质家族中的序列在这个模型上得到的结果最好, 最后利用该模型来将蛋白质序列转化为向量空间中的向量. 文献[15]利用了字符串频谱, 将固定长度(比如 3)的所有可能的子序列数作为转化后的向量的长度. 每一维的值为相应子序列在序列中出现的次数. 文献[16]的思想与文献[15]基本相同, 但是在计算子序列的出现次数时, 采用不精确匹配的字符串匹配. 文献[14]采用基于两条序列比对的方法, 序列对应的向量的维数等于训练样本(包括正样本和负样本)的个数, 每一维对应一条训练样本, 给定一条序列, 它所对应的向量的分量的值为该序列与相应分量对应的样本序列的比对的分值, 采用 Smith-Waterman 算法来求两条序列的比对的分值. 而文献[17]结合了 2-gram 和 6 字母等价组, 具体地说, 将 20 个氨基酸分为 6 组, 每一组中的所有字母由一个代表代替. 一条序列对应的向量分为两部分, 第一部分与文献[16]类似, 不同之处是采用得子序列的长度为 2; 第二部分将序列转化为由 6 个代表字母组成的序列, 再采用与第一部分相同的方法. 这样, 一条序列就转化成维数为  $20^2 + 6^2 = 436$  的向量. 其中文献[15-16]是任何基于字符串的分类问题的通用算法. 文献[12-13]训练模型时只采用正样本, 而文献[14]在映射、训练的整个过程中都是既利用了正样本, 又利用了负样本, 因此精确度要高于其他方法.

本文提出了一种新的将序列从序列空间映射到向量空间的方法, 称为位点进化距离法 Position Specific Value(PSV), 用来将蛋白质序列转化为向量, 在此基础上, 结合使用支持向量机算法用于蛋白质的分类及同源检测. 采用位点进化距离法从序列

空间到向量空间进行映射,映射后向量的每个分量的值为序列中相应位置开始的长度为  $k$  的子序列到该蛋白质家族的最近共同祖先序列的从相应位置开始的长度为  $k$  的子序列的编辑距离(进化距离)<sup>[18]</sup>. 该方法基于如下假设:对于一组相似性越强的序列(如一个蛋白质家族),从其最近的共同祖先进化到其中任意一条序列所产生的插入、删除,比进化到不属于该组的一条序列的过程中所发生的插入、删除要少;因此,从对应位置开始的长度为  $k$  的子序列之间的编辑距离,前者小于后者的要占大多数.

作为用于蛋白质同源检测(或者分类)分类方法,其价值在于快速地在训练集上进行训练,在此基础上,能够迅速、准确对未知样本进行分类. 本文所提出的用于蛋白质同源检测的方法 SVM-PSV,其时间复杂度要明显得优于 SVM-pairwise. 在训练过程中的向量化阶段, SVM-pairwise 的时间复杂度是  $O(N^2mp)$ , 而 SVM-PSV 的时间复杂度是  $O(Nmk^2)$ , 这里  $m$  和  $p$  两条蛋白质序列的长度,  $N$  是训练集中的序列条数, 一般有  $k^2 \ll p$  ( $k$  取值 3~5 之间). 训练过程的优化阶段, 两种方法所用时间大致相同, 都是支持向量机所需要的  $O(N^2)$ . 当一个未知序列需要对其进行归类时, SVM-pairwise 需要  $O(Nmp)$  的计算时间将其向量化, 而 SVM-PSV 只需要  $O(mk^2)$ . 进一步, 由于  $k$  取值较小, 所以可以将所有的子序列对的编辑距离事先算好, 这样, 从序列到向量的转化时间可以提高  $k^2$  倍, 即训练时需要  $O(Nm)$  的时间来将序列转化为向量, 而一条未知序列在归类前, 需要  $O(m)$  的时间来转化为向量. 同时, 在蛋白质结构分类数据库(SCOP)上进行的实验表明, 该方法的精确度也比其他方法好.

本文第 2 节详细介绍 PSV 方法和使用该方法将序列从序列空间映射到向量空间; 第 3 节介绍用于比较几种同源检测方法的实验; 第 4 节给出实验结果并进行分析; 第 5 节给出结论.

## 2 基于位点进化距离的内核 (PSV KERNEL)

支持向量机是一种基于内核的机器学习算法, 给定一组固定维数的向量, 通过区分正样本与负本来解决学习问题. 这类算法首先将训练样本映射到一个特征空间(这个特征空间的维数可能比原来高很多), 然后在特征空间中, 通过学习来寻找一个能够将正样本和负样本区分开的超平面. 得到了这

样的超平面后, 对于未知的样本, 支持向量机首先将它映射到特征空间, 然后根据该样本的特征向量位于超平面的哪一端来对该样本做出判断.

支持向量机所找到的超平面具有如下特点: 如果存在很多能够将正、负训练样本区分开的超平面, 那么支持向量机所给出的超平面具有如下特征: 它不但能够将两个集合分开, 还使得每个集合中离分界面最近的点到分界面的距离最大. 统计学习理论认为, 这一特点使得在区分以前未见过的样本时, 具有最大的一般性<sup>[19]</sup>.

采用支持向量机算法的一个很重要的必要条件是: 输入的样本集必须是固定长度的向量的集合, 而蛋白质是长度不定的由氨基酸组成的序列, 无法直接采用支持向量机算法. SVM-Fisher 算法借助于隐马尔可夫模型来完成从蛋白质序列空间到长度固定的向量空间的转化. 首先, 利用训练集中的正样本来训练隐马尔可夫模型, 然后, 对于任意的序列(正样本、负样本以及待区分的样本), 根据训练好的模型来求其梯度向量, 梯度向量的每一个分量对应隐马尔可夫模型的一个参数. 在此基础上, 利用支持向量机算法来学习并区分未知序列. SVM-pairwise 利用两条序列的相似性来将蛋白质序列转化为固定长度的向量, 转化后向量的维数是训练样本的个数, 序列所对应的向量的每个分量为这条序列与相应的训练样本序列联配的得分值. SVM-pairwise 的分类性能明显优于 SVM-Fisher, 但是所花的时间大约是后者的  $O(N)$  倍<sup>[14]</sup>.

这里, 我们利用位点进化距离特征映射(Position-Specific-Value)来完成从蛋白质序列空间到固定维数的特征向量空间的映射, 然后在映射后的向量空间中采用支持向量机算法来解决蛋白质分类问题.

位点进化距离内核背后的基本思想是: 每一个蛋白质家族(或者超家族), 其成员序列都是由某个共同的最近祖先序列进化而来, 不同的蛋白质家族的最近共同祖先各不相同. 对于某个给定的蛋白质家族, 从蛋白质家族的最近共同祖先进化到任意一个成员序列的过程中, 发生的“插入”, “删除”以及“替换”要比进化到非成员序列少. 这样, 对于一个家族的成员序列来说, 其大多数长度为  $k$  的子序列到该家族共同祖先的对应位置的长度为  $k$  的子序列的进化距离要比一个不属于这个家族的序列的相同位置的子序列的进化距离小.

PSV 本质上就是一个从有限长度的序列空间

到  $M$  维的向量空间的映射. 这个映射定义如下: 给定一个蛋白质家族, 一条序列映射到如下向量: 向量的下标对应于子序列在序列中的位置, 分量的值为相应子序列到这个蛋白质家族的共同祖先序列的相应位置的子序列的进化距离. 由于序列是长短不一的(同一个蛋白质家族中的成员序列的长度也不尽相同), 我们当前采用的方法是以最长的序列为准, 在比较短的序列右边添加 ‘\_’ 来对齐. 设  $n$  是最长序列的长度, 对于给定的  $k > 1$ , 定义  $M = n - k + 1$ . 假设一个家族的蛋白质序列为  $S_1, S_2, \dots, S_N$ . 其最近共同祖先为  $X$ ,  $S_{i,j}$  表示序列  $S_i$  的第  $j$  条长度为  $k$  的子序列(从位置  $j$  开始),  $X_j$  表示序列  $X$  的第  $j$  条长度为  $k$  的子序列, 而  $D_i$  表示序列  $S_i$  的在向量空间中的映射,  $d_{i,j}$  表示  $D_i$  的第  $j$  个分量, 也就是由  $S_{i,j}$  到  $X_j$  的编辑距离. 所谓的编辑距离, 是用来表示序列之间的“距离”, 侧重于将一个序列通过一系列的对单个字符的编辑将其转化为另一个, 合法的操作包括对单个字符的“插入”、“删除”和“替换”. 对“插入”、“删除”和“替换”设定罚分或者权值, 编辑距离就是将一个序列转化为另一个所需要的最小的权值(罚分).

例如, 如果对单个字符的“插入”、“删除”和“替换”的罚分都是 1, 那么, “vinter”和“writers”之间的编辑距离为 5.

```
v - i n t e r -
w r i t e r s
```

现在, 我们定义从一条序列  $S_i$  到  $M$  维向量空间中向量  $D_i$  的映射. 首先定义  $S_{i,j}$  到  $D_i$  的映射, 也就是它们之间的编辑距离  $d_{i,j}$  为

$$d_{i,j} = \varphi_k(S_{i,j}) = \text{edit\_distance}(S_{i,j}, X_{i,j}) \quad (1)$$

那么,  $S_i$  所对应的向量为

$$\begin{aligned} D_i &= \Phi_k(S_i) = (d_{i,1}, d_{i,1}, \dots, d_{i,M}) \\ &= (\phi_k(S_{i,1}), \phi_k(S_{i,2}), \dots, \phi_k(S_{i,M})) \end{aligned} \quad (2)$$

而基于位点进化距离的内核函数就是特征空间中特征向量的内积函数:

$$K_k(S_i, S_j) = \langle \Phi_k(S_i), \Phi_k(S_j) \rangle \quad (3)$$

但是, 蛋白质家族的最近共同祖先是未知的, 因此必须用某种方法来“找到”或者代替它. 我们采用的方法是在计算某个序列的第  $j$  个子序列到最近共同祖先的编辑距离时, 用其到蛋白质家族中所有成员序列的第  $j$  个子序列的编辑距离的平均值来代替. 这样, 相当于以蛋白质家族的成员序列所映射到的向量的重心向量所对应的序列来代替该家族的最近共同祖先. 因此式(1)~(3)变为

$$\begin{cases} d_{i,j} = \frac{1}{N} \sum_{k=1}^N \text{edit\_distance}(S_{i,j}, S_{k,j}) \\ \Phi(S_i) = (d_{i,1}, d_{i,2}, \dots, d_{i,M}) \\ K(S_i, S_j) = \langle \Phi(S_i), \Phi(S_j) \rangle \end{cases} \quad (4)$$

### 3 蛋白质分类实验描述

我们设计了两个实验, 第一个实验比较用于蛋白质分类(同源检测)的几种方法: SVM-pairwise, FPS, SVM-Fisher, SAM, PSI-BLAST, SVM-pairwise \_ BLAST, SVM-pairwise +, PSI-BLAST 和 SVM-PSV 的精确度; 第二个实验则是用于比较 SVM-pairwise 和 SVM-PSV 的计算效率.

在实验中, 我们通过测试算法将蛋白质结构分类数据库(SCOP)<sup>[20]</sup>中的蛋白质结构域划分为不同的超家族的能力来评价算法的性能. 所用的蛋白质结构域序列来自 Astral 数据库(astral.stanford.edu<sup>[21]</sup>), 按照 E-value 的阈值  $10^{-25}$  来剔除相似序列, 得到 4352 条不同的序列, 然后归类为家族、超家族. 对于任意一个蛋白质家族, 该家族的成员序列作为正测试样本, 不属于该家族, 但属于同一个超家族的序列作为正训练样本. 这样得到 54 个家族(见附表 1), 每个家族包含 10 条以上的序列, 5 个超家族. 负样本来自不包含该蛋白质家族的折叠类型(Fold), 按照正测试样本和正训练样本的比例随机分成测试样本和训练样本<sup>[14]</sup>. 这也是文献[14]所采取的方法.

在实验中, 我们采用免费软件(the Gist Support vector machine free software)作为我们的支持向量机算法的实现<sup>[22]</sup>(<http://microarray.cpmc.columbia.edu/gist/>). 支持向量机的关键是其内核函数(kernel function), 我们采用的内核函数是所输入的向量对的内积.

在计算子序列对之间的进化距离以及在对 SVM-pairwise 算法进行从序列到向量的转化时, 我们都采用了 Smith-Waterman 算法, 采用了缺省参数: 对于空位开放的罚分为 11, 空位延伸的罚分为 1, 采用的得分矩阵是 BLOSUM62. 隐马尔可夫模型采用 Hmmer 软件<sup>[22]</sup>([predict.sanger.ac.uk/mirrors/hmm/hmm.html](http://predict.sanger.ac.uk/mirrors/hmm/hmm.html))和 Clustalw 软件<sup>[23]</sup>([bimas.dcert.nih.gov/clustalw/clustalw.html](http://bimas.dcert.nih.gov/clustalw/clustalw.html)). 在 SVM-fisher 算法中, 在将序列映射到向量的过程中, 也采用这两种软件. 至于支持向量机的内核函数, SVM-PSV 采用点积函数, 其余基于向量机的算法则采取

了与文献[14]相同的内核. 其余方法的实现细节与文献[14]相同.

为了评价各种方法的识别能力, 我们采用了两个指标: ROC (Receiver Operating Characteristic) 分值和假阳性比例的均值. ROC 分值是这样得到的: 根据不同的阈值, 将真阳性作为假阳性的函数画一条曲线, 曲线下的面积规范化以后即为 ROC 的分值<sup>[24]</sup>. ROC 分值在一定程度上代表了正确区分正、负样本的概率, 一个能够把所有的正负样本完美区分的分离器的 ROC 得分值是 1, 而一个随机的分离器的 ROC 的得分几乎为零. 平均的假阳性比是指负样本中那些得分超过正样本的平均得分的样本的比例.

为了评价算法的计算效率, 我们设计实验比较 SVM-pairwise 和 SVM-PSV 的计算效率. 两种方法的运行时间都包括两部分: (1) 将训练样本和测试样本转化为固定维数的向量; (2) 利用支持向量机算法进行训练, 并利用训练结果对测试集中的样本进行分类 (这里的训练、测试样本都是在转化后的向量空间上). 由于采用相同的训练、测试集, 并且支持向量机算法在优化阶段的计算量与向量空间的维数基本无关, 所以第一部分的时间大致相等 (训练支持向量机所需的时间为  $O(N^2)$ ). 同样, 将一条序列转化为向量以后, 再进行分类所用的时间为  $O(1)$ . 将训练集中的序列转化为向量阶段, SVM-pairwise 和 SVM-PSV 的所需要的时间分别为  $O(N^2mp)$  和  $O(KNm k^2)$ . 将一条未知序列转化为向量, SVM-pairwise 和 SVM-PSV 的所需要的时间分别为  $O(mp)$  和  $O(Km k^2)$ . 因此我们可以通过比较将蛋白质序列映射到向量空间的向量的时间来评价 SVM-pairwise 和 SVM-PSV 的计算效率, 具体地说, 我们通过计算将每一个家族的正训练样本、负训练样本、正测试样本、负测试样本映射到向量空间的间来评价其计算效率, 而不考虑在映射后的向量空间中训练向量机和利用训练好的向量机进行分类. 在评价计算时间时, 我们采用曙光 3000 超级服务器作为计算平台.

## 4 精度与效率的对比与分析

实验结果如图 1、图 2 和图 3 所示. 在图 1、图 2 所示的两个图中, 分别表示了 9 个不同的分类算法的两个用于衡量精确度的指标: ROC 分值和 RFP 的均值. 每个图中, 越高的曲线表示相应的算法的精

确度越高. 从 ROC 值来看, 如图 1 所示, SVM-PSV 的 ROC 值比 FPS, SVM-Fisher, SAM, PSI-BLAST, SVM-pairwise\_BLAST, SVM-pairwise+, PSI-BLAST 都要高很多, 也比 SVM-pairwise 稍微高一些. 而按照 RFP 均值, 如图 2 所示, SVM-PSV 的比包括 SVM-pairwise 在内的算法都要精确. 具体地说, SVM-PSV 算法的精确度在对蛋白质家族 3.1.8.3, 3.2.1.3, 3.2.1.6, 3.2.1.7, 3.32.1.11, 3.32.1.8, 3.42.1.5 和 3.42.1.8 的实验中明显地高于其它算法.

实际上, SVM-PSV 还可以提高其识别性能, 目前所采用的模型还比较初级. 首先, 对于不同长度的序列, 我们采取了在其后面加入 ‘\_’ 的方法将其对齐; 其次, 在计算一条序列的子序列到家族的最近共同祖先的进化距离时, 采用了子序列到所有该家族的成员序列的相应位置上的子序列的进化距离的平均值来代替; 最后, 当前一个家族中的成员序列太少, 也影响了识别的性能.

在计算复杂性方面, SVM-PSV 算法明显好于 SVM-pairwise. 首先分析训练的时间复杂度, 二者均包含 SVM 优化过程, 在这个过程中, 计算的复杂度与向量的维数基本无关, 只与训练集中的样本的数量有关. 因此, 时间复杂度为  $O(N^2)$ , 二者是相同的. 在将序列转化为向量的过程中, SVM-pairwise 需要进行  $N^2$  次两条序列的相似度, 采用 Smith-Waterman 算法, 每次相似度计算的时间复杂度为  $O(mp)$ , 总的时间复杂度为  $O(N^2mp)$ . 而 SVM-PSV 的向量化过程只需要计算  $M$  次长度为  $k$  的子序列对的编辑距离, 计算复杂度为  $O(k^2)$ , 总的计算时间复杂度为  $O(KNm k^2)$ ,  $K$  是正样本的个数. 考虑到实际应用中,  $k$  的值要小于 6 (一般 3~5), 而  $p$  则可能是数百, 数千, 甚至更大, 因此有  $k^2 \ll p$ ,  $\frac{N}{K} \approx 100$  和  $m \approx p > M$ . 为了讨论的方便, 我们假定  $m = p$ , 这样, SVM-PSV 就比 SVM-pairwise 要快  $O\left(\frac{Nm}{Kk^2}\right)$  或者  $O(m)$ .

由图 3 可以看出, 从整体来看, SVM-PSV 对比 SVM-pairwise 的平均加速比是 10 左右. 具体地说, 除了蛋白质家族 2.44.1.2 和 7.3.10.1 外, SVM-PSV 对比 SVM-pairwise 的加速比从 4.2 到 17.6 不等. 蛋白质家族 2.44.1.2 获得的加速比为 0.7, 而家族 7.3.10.1 的加速比是 1.8. 原因有两个, 第一, 这两个蛋白质家族的很多序列长度太小; 第二,

训练样本数量太小.

当然,有很多方法可以用于提高 SVM-pairwise 的向量化速度,如 SVM-pairwise\_BLAST, SVM-

pairwise+, 采用启发式算法代替动态规划, 可以将总的时间复杂度降为  $O(N^2m)$ . 从图 1 和图 2 可以看出, 这是以大幅度地降低识别性能为代价的.

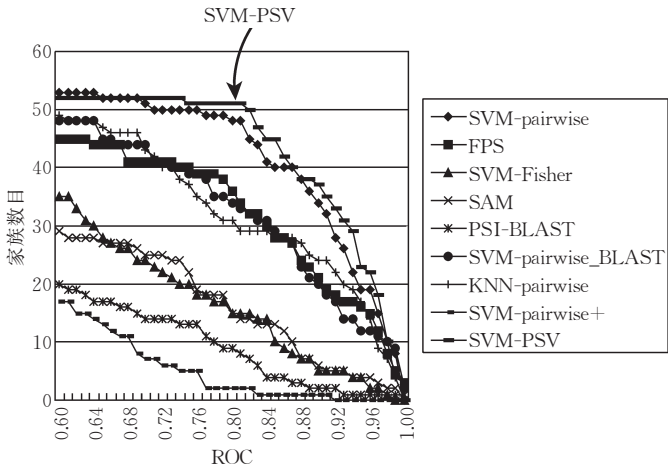


图 1 9 种蛋白质序列的同源检测方法的 ROC 值与每种算法所能达到该指标的家族数目的关系

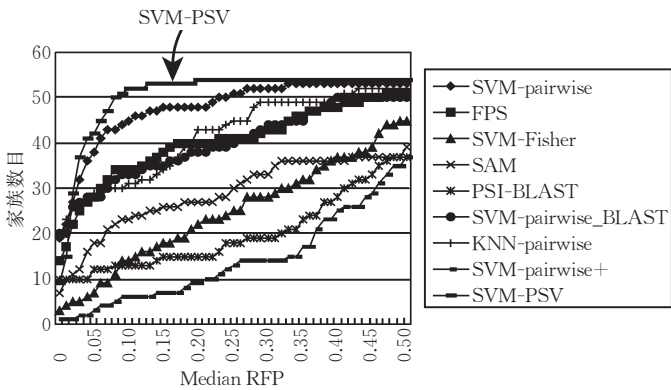


图 2 9 种蛋白质序列的同源检测方法的 median RFP 值与每种算法所能达到该指标的家族数目的关系

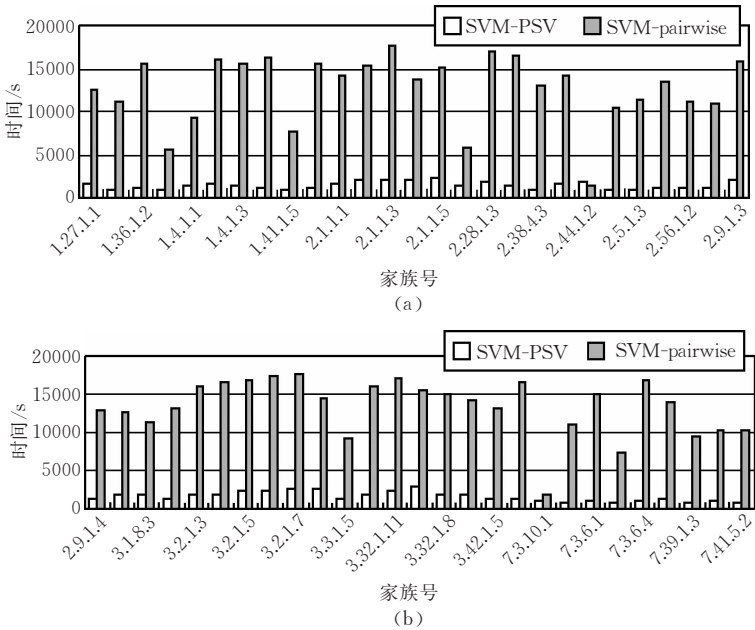


图 3 SVM-PSV 和 SVM-Pairwise 的从序列映射到向量所需时间的比较

## 5 结 论

本文提出了一种新的用于蛋白质分类的算法——SVM-PSV 算法. 这种算法的思想是一个蛋白质家族(或者超家族)的成员序列到该家族(超家族)的最近祖先的进化距离要比非成员序列近. 在 9 种算法中, 该算法即是精确最高的, 在时间复杂度上, 比精确度相近的 SVM-pairwise 算法有着明显地改进, 可以获得  $O(m)$  的加速比.

我们在蛋白质结构分类数据库 SCOP<sup>[20]</sup> 上设计实验测试该方法的性能, 结果表明,

SVM-PSV 比 SVM-pairwise+, FPS, SVM-Fisher, SAM, PSI-BLAST, SVM-pairwise\_BLAST 和 PSI-BLAST 都要精确很多, 也比 SVM-pairwise 精确度高. 而在计算效率方面, SVM-PSV 比 SVM-pairwise 平均快 10 倍左右.

如何对进一步提高 SVM-PSV 的精确度, 是值得进一步研究的课题.

**致 谢** 我们特别感谢 William Standford Noble 教授允许我们使用他们在以前的工作中的数据和结果!

## 参 考 文 献

- [1] Needleman S B, Wunsch C D. A general method applicable to the search for similarities in the amino acid sequences of two proteins. *Journal of Molecular Biology*, 1970, 48(3): 443-453
- [2] Smith T F, Waterman M S. Identification of common molecular subsequences. *Journal of Molecular Biology*, 1981, 147(1): 195-197
- [3] Smith T F, Waterman M S. Comparison of biosequences. *Advances in Applied Mathematics*, 1981, 2: 482-489
- [4] Altschul S F, Gish W, Miller W, Myers E W, Lipman D J. A basic local alignment search tool. *Journal of Molecular Biology*, 1990, 215(3): 403-410
- [5] Pearson W R. Rapid and sensitive sequence comparisons with FASTP and FASTA. *Methods in Enzymology*, 1985, 183: 63-98
- [6] Altschul S F, Madden T L, Schaffer A A, Zhang J, Zhang Z, Miller W, Lipman D J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, 1997, 25(17): 3389-3402
- [7] Gribskov M, Lüthy R, Eisenberg D. Profile analysis. *Methods in Enzymology*, 1990, 183: 146-159
- [8] Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, 1998, 14(10): 846-856
- [9] Krogh A, Brown M, Mian I, Sjolander K, Haussler D. Hidden Markov models in computational biology: Applications to protein modeling. *Journal of Molecular Biology*, 1994, 235(5): 1501-1531
- [10] Baldi P, Chauvin Y, Hunkapiller T, McClure M A. Hidden Markov models of biological primary sequence information. *Proceedings of the National Academy of Sciences of the United States of America*, 1994, 91(3): 1059-1063
- [11] Pasquier C, Promponas V J, Hamodrakas S J. PRED-CLASS: Cascading neural networks for generalized protein classification and genome-wide applications. *PROTEINS: Structure, Function, and Genetics*, 2001, 44(3): 361-369
- [12] Jaakkola T, Diekhans M, Haussler D. A discriminative framework for detecting remote protein homologies. *Journal of Computational Biology*, 2000, 7(1-2): 95-114
- [13] Jaakkola T, Diekhans M, Haussler D. Using the Fisher kernel method to detect remote protein homologies//*Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology*. Heidelberg, Germany, 1999: 149-158
- [14] Liao L, Noble W S. Combining pairwise sequence similarity and support vector machines for remote protein homology detection//*Proceedings of the 6th International Conference Computational Molecular Biology*. Washington, DC, USA, 2002: 225-232
- [15] Leslie C, Eskin E, Noble W S. The spectrum kernel: A string kernel for SVM protein classification//*Proceedings of the Pacific Symposium on Biocomputing (PSB-2002)*. Hawaii, USA, 2002: 564-575
- [16] Leslie C, Eskin E, Weston J, Noble W. Mismatch string kernels for SVM protein classification//*Proceedings of the Advances in Neural Information Processing Systems*. Vancouver, Canadian, 2002: 1441-1448
- [17] Wang J T L, Ma Q, Shasha D, Wu C H. New techniques for extracting features from protein sequences. *IBM Systems Journal*, 2001, 40(2): 426-441
- [18] Sankoff D, Kruskal J. *Times Wraps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Reading, MA: Addison-Wesley, 1983
- [19] Vapnik V N. *Statistical Learning Theory. Adaptive and Learning Systems for Signal Processing, Communications, and Control*. New York: Wiley, 1998
- [20] Murzin A G, Brenner S E, Hubbard T, Chothia C. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 1995, 247(4): 536-540
- [21] Brenner S E, Koehl P, Levitt M. The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Research*, 2000, 28(1): 254-256
- [22] Eddy S. Multiple alignment using hidden Markov models//*Rawlings C ed. Proceedings of the 3rd International Conference on Intelligent Systems for Molecular Biology*. Cambridge, United Kingdom, 1995: 114-120
- [23] Thompson J D, Higgins D G, Gibson T J. CLUSTALW: Improving the sensitivity of progressive multiple sequence

alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research, 1994, 22(22): 4673-4680

[24] Gribskov M, Robinson N L. Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. Computers and Chemistry, 1996, 20(1): 25-33

附表 1.

Long-chain cytokines	Short-chain cytokines	Phage repressors	Bacterial repressors
Homeodomain	Recombinase DNA-binding domain	Myb	S100 proteins
Calmodulin-like	Bacterial repressors	V set domains (antibody variable domain-like)	C1 set domains (antibody constant domain-like)
C2 set domains	I set domains	E set domains	Legume lectins
Galectin (animal S-lectin)	Anticodon-binding domain	Single strand DNA-binding domain, SSB	Cold shock DNA-binding domain-like
Eukaryotic proteases	Plastocyanin/azurin-like	Multidomain cupredoxins	Phosphotyrosine-binding domain (PTB)
Fatty acid binding protein-like	Plant virus proteins	Insect virus proteins	Animal virus proteins
alpha-Amylases, N-terminal domain	beta-glycanases	Tyrosine-dependent oxidoreductases	Glyceraldehyde 3-phosphate dehydrogenase-like, N-terminal domain
Formate/glycerate dehydrogenases, NAD-domain	Lactate & malate dehydrogenases, N-terminal domain	6-phosphogluconate dehydrogenase-like, N-terminal domain	Amino-acid dehydrogenase-like, C-terminal domain
FAD-linked reductases, N-terminal domain	FAD/NAD-linked reductases, N-terminal and central domains	Nucleotide and nucleoside kinases	RecA protein-like (ATPase-domain)
Extended AAA-ATPase domain	G proteins	Thioltransferase	Glutathione S-transferases, N-terminal domain
Glutathione peroxidase-like	EGF-type module	Spider toxins	Long-chain scorpion toxins
Short-chain scorpion toxins	Plant defensins	Nuclear receptor	LIM domain
Rubredoxin	Desulforedoxin		



**LI Yu-Gang**, born in 1971, Ph. D. . His current research interests include computer algorithms, parallel processing and bioinformatics.

**ZHANG Fa**, born in 1974, Ph. D. , assistant professor. His current research interests include computer algorithms, parallel processing and bioinformatics.

**LIU Zhi-Yong**, born in 1946, Ph. D. , professor. His current research interests include computer algorithms and architectures, parallel processing, networks, and image processing.

Background

The major aim of this paper is to investigate the accurate and the computing efficiencies of existing methods and lay down a sound base for further implementation. The authors find a new method of protein classification, whose computing efficiency is higher and shows a better accuracy in the experi-

ments.

The project is supported in part by the National Natural Science Foundation of China, which aims at finding a more accurate protein structure predicting method.