生物工程学报
journals.im.ac.cn
cjb@im.ac.cn

Chin J Biotech 2008, July 25; 24(7): 1140-1148
Chinese Journal of Biotechnology ISSN 1000-3061

研究报告

基于支持向量机分类的 RNA 共同二级结构预测

赵英杰, 王正志

国防科技大学机电工程与自动化学院, 长沙 410073

摘　要: 比较序列分析作为 RNA 二级结构预测的最可靠途径, 已经发展出许多算法。将基于此方法的结构预测视为一个二值分类问题: 根据序列比对给出的可用信息, 判断比对中任意两列能否构成碱基对。分类器采用支持向量机方法, 特征向量包括共变信息、热力学信息和碱基互补比例。考虑到共变信息对序列相似性的要求, 通过引入一个序列相似度影响因子, 来调整不同序列相似度情况下共变信息和热力学信息对预测过程的影响, 提高了预测精度。通过 49 组 Rfam-seed 比对的验证, 显示了该方法的有效性, 算法的预测精度优于多数同类算法, 并且可以预测简单的假节。

关键词: 比较序列分析, RNA 二级结构, 支持向量机, 相似性影响因子

RNA Secondary Structure Prediction Based on Support Vector Machine Classification

Yingjie Zhao, and Zhengzhi Wang

College of Mechatronics Engineering and Automation, National University of Defense Technology, Changsha 410073, *China*

Abstract: The comparative sequence analysis is the most reliable method for RNA secondary structure prediction, and many algorithms based on it have been developed in last several decades. This paper considers RNA structure prediction as a 2-classes classification problem: given a sequence alignment, to decide whether or not two columns of alignment form a base pair. We employed Support Vector Machine(SVM) to predict potential paired sites, and selected co-variation information, thermodynamic information and the fraction of complementary bases as feature vectors. Considering the effect of sequence similarity upon co-variation score, we introduced a similarity weight factor, which could adjust the contribution of co-variation and thermodynamic information toward prediction according to sequence similarity. The test on 49 Rfam-seed alignments showed the effectiveness of our method, and the accuracy was better than many similar algorithms. Furthermore, this method could predict simple pseudoknot.

Keywords: comparative sequences analysis, RNA secondary structure, support vector machine, similarity weight factor

RNA 作为生物大分子的一种, 其功能主要由其二级结构决定。随着人们对各种功能 RNA 研究的不断深入, 越来越凸现出 RNA 结构研究的重要性。对 RNA 二级结构的确定, 一般有两种方法: 生物学实验测定和计算预测。实验方法虽然准确, 但既耗时费力, 又代价高昂, 无法应对日益激增的序列数据, 而计算方法相对简单有效, 更适合处理海量数据, 已经成为批量结构确定的主要方法。

在过去的几十年里, 针对 RNA 二级结构预测问题, 已经提出了许多方法。然而, 作为计算分子生物学的一个经典问题, 从 RNA 初级序列中预测出正确的二级结构仍然是一个挑战性的问题。已有方法概括起来主要有三类: 基于热力学理论的自由能最小方法[1–3]、基于统计学习的方法[4–7]和基于概率的序

Received: November 23, 2007; **Accepted:** December 27, 2007
Corresponding author: Yingjie Zhao. Tel: +86-731-4574991; E-mail:matriz@163.com

列比较分析方法[8−11]。自由能最小方法是基于这样的一个假设: 具有最小自由能的分子结构应该是最稳定的, 它的参数来自于实验测定和推断, 是目前单序列预测的主要方法。其准确性主要依赖于能量参数的测量, 一般对短序列的预测精度较高, 而对长序列的预测结果不是很理想。由于测量方法及技术的限制, 有些参数在测量和估计时本身就存在误差, 而且实验证明, 有时分子的实际结构对应的自由能并不总是最小, 这都限制了该方法预测精度的提高。统计学习方法是对现有已知 RNA 二级结构数据库进行统计学分析, 采用机器学习的方法, 力图找出结构形成规律, 以用于结构预测。主要有随机上下文无关语法模型(Stochastic context-free grammars, SCFG)[4]、遗传算法(Genetic algorithm, GA)[5]、统计采样(Statistical sampling algorithm)[7]等方法, 这些方法的主要缺点是计算的复杂度较高。目前公认最有效的 RNA 二级预测方法是序列比较分析方法, 它主要是用于对同源 RNA 序列集进行共同二级结构推断。它采用的方法是共变探测, 即多序列比对中, 发生一定数量互补突变的两列, 则可能对应着碱基间的相互作用(形成碱基对)。这种方法需要大量的同源序列, 并且前提假设是它们有共同的二级结构, 另外对序列的保守性要求较高, 这在实际应用时不一定都能够满足, 也就限制了它的使用。

有鉴于此, 许多综合算法被提出: Hxmatch[12]、construct[13,14]、contrafold[15]、hxplot[16]和 ILM[17]都是将考虑了序列热力学信息的碱基配对概率矩阵和互信息矩阵(或是共变分值矩阵)结合在一起用于结构预测, 这些方法第一步先计算一个候选配对概率矩阵(矩阵中元素值为 1 的行列对应可能构成碱基对的列对), 然后采用各自结构组合方法得到最后的二级结构。但预测的主要依据仍是比较分析方法得到的互补突变信息。

比较序列分析方法又可以分成三类[18]: 一类是已知序列比对, 用互补突变探测来找出一致结构, 如 Pfold[5,19]、RNAalifold[20]、ILM[17]和 KnetFold[21,22]; 一类是采用 Sankoff 算法[23], 即同时进行序列比对和结构预测, 如 Foldalign[24,25]、Dynalign[26]和 PMcomp[27]; 最后一类先进行单条序列的结构预测, 然后通过结构比对来找出一致结构, 如 RNAforester[28]和 MARNA[29,30]。第一类方法中, 序列比对可以用“标准”的序列比对程序获得, 或是直接利用数据库中的参考比对, 这类方法对序列比对质量的要求较高, 多数比较分析算法属于这类。第二、三类方法的主要缺点在于它们高的计算代价和复杂度。

对于给定的已对齐同源 RNA 序列集, 共同结构预测可以看作一个分类问题: 考察比对中任意两列, 根据相关信息(主要指决定碱基配对的自由能信息和共变信息)来确定能否构成碱基对。本文基于第一类比较分析方法, 以热力学信息、共变信息以及互补碱基所占比例作为分类特征, 采用支持向量机分类器(Support vector machine, SVM)实现序列比对中潜在配对位置的预测, 最后采用茎组合规则, 得到最终的共同二级结构。

1 算法

本算法的思路是: 给定一组序列比对, 对所有比对列的两两组合, 首先计算出相应的特征矢量, 即共变信息(共变分值)、热力学信息(平均碱基配对概率矩阵)和互补碱基比例, 然后根据预先训练得到的模型, 进行分类预测, 得到一个潜在配对矩阵, 最后, 根据碱基配对规则和茎组合规则构造出最终的共同二级结构。

1.1 共变分值的计算

为定量度量比对中的互补突变, 一般采用传统的互信息分值[31,32]:

$$M(i,j)=\sum_{XY} f_{ij}(XY)\ln\frac{f_{ij}(XY)}{f_i(X)f_j(Y)} \tag{1}$$

其中, $f_i(X)$和 $f_j(Y)$分别是比对第 i 列和第 j 列中碱基 X 出现的频率, $X\in\{A,C,G,U\}$; $f_{ij}(XY)$表示比对中第 i 个位置出现碱基 X, 同时在第 j 个位置出现碱基 Y(同一序列中)的联合概率, $XY\in\{AU,UA,GC,CG,GU,UG\}$。通过互信息分值能够很好地探测到互补突变的发生, 但对进化过程中发生的单边互补突变情况(如图 1 所示)却无能为力。

这里采用 Hofacker[20]提出的共变分值来度量比对中的互补突变(包括单边和双边突变):

$$C(i,j)=\sum_{XY,X'Y'} f_{i,j}(XY)D(XY,X'Y')f_{i,j}(X'Y') \tag{2}$$

式中, $D(XY, X'Y')$是 XY、$X'Y'$间的 Hamming 距离。共变分值可以区分开保守的碱基对、单边互补突变

和双边互补突变的碱基对。此外, Hofacker 引入一个不一致序列罚分函数[20]:

$$q(i,j)=1-f_{i,j}^{comp}-f_{i,j}(-\bullet-) \quad (3)$$

式中, $f_{i,j}^{comp}$ 是比对第 i 和第 j 列中互补碱基对的比例。通常要从 $C(i, j)$中减去这个罚分, 已经证明, 这个函数对序列较少的比对效果较明显。

进化过程中的单边互补突变情况如图 1 所示。

i j

××××G×××U××××

××××G×××C××××

××××G×××U××××

××××G×××C××××

图 1 单边互补突变[20]示意图

Fig. 1 Schematic map of single conserved pair[20]

第 i 列对应一个保守的 G(或是 U), 第 j 列对应一个可变的 C 和 U(或是 A 和 G)。如果采用传统的互信息方法, 则 $M(i, j)=0$, 探测不到互补突变的发生, 但如果采用共变分值, 则有 $C(i, j)=4$, 说明存在碱基互补突变。

1.2 带权重的平均碱基配对概率矩阵

虽然共变分值能很好的探测进化中互补突变的发生, 但对于那些进化中保守的碱基对却无能为力(此时的共变分值为 0)。这也是所有比较分析法的一个共同缺点。在多数综合方法中, 都是通过引入热力学信息来增强预测的准确性。热力学信息的引入可以通过直接和间接两种方式: 直接引入是利用实验测定和推断的热力学参数, 但是热力学参数多达上千个, 而且计算时要考虑不同类型的结构组件, 计算较复杂, 且不易和其他类型参数组合; 间接引入通常采用由划分函数算法[33]得到的碱基配对概率矩阵, 它不仅受能量参数的不确定性影响相对较小[34], 而且计算方便, 能很好的与其他类型参数融合。

本文中采用带权重的平均碱基配对概率矩阵作为其中一个分类特征, 计算方法如下: 给定一个多序列比对, 首先剔除比对中引入的空位, 对其中每条序列(不含空位)使用 RNAfold[35,36]程序计算得到一个碱基配对概率矩阵, 然后根据序列比对结果, 在配对概率矩阵中引入空格(概率值设为 0), 最后将所有概率矩阵进行平均, 得到一个平均碱基配对概率矩阵 P 。

权重的引入主要是考虑序列相似性对预测的影响: 如果序列的相似性较高(保守程度较高), 则作为比较分析法中主要因素的共变分值的作用就会减少, 预测应当更倾向于依靠热力学信息, 此时要加大热力学参数的影响权重。为此, 定义一个相似度影响因子:

$$sp(i,j)=\frac{e^{0.1\cdot(s-s_0)}}{1+e^{0.1\cdot(s-s_0)}} \quad (4)$$

式中 s 表示比对中两列(i, j)的平均相似度:

$$s(i,j)=\frac{\sum_{k=1}^{N-1}\sum_{l=k+1}^{N}D(X_k(i,j),X_l(i,j))}{N(N-1)} \quad (5)$$

式中, $X_k(i, j)$表示比对中第 k 条序列的第 i、j 位置的碱基, D 为 Hamming 距离, N 是比对序列数。s_0 是一个平均相似度阈值, 因为采用共变分析方法的理想相似度为<70%, 所以这里 s_0=0.7。带权重的平均碱基配对概率矩阵 P_W 由平均配对概率矩阵 P 中元素和对应相似度影响因子得到:

$P_W(i,j)=P(i,j)\cdot sp(i,j)$。

1.3 支持向量机分类器

支持向量机(Support Vector Machine, SVM)[37]分类器作为一种广泛应用的机器学习工具, 在解决非线性分类、函数估计和密度估计方面有着很好的表现, 它能有效的避免过估计, 能处理大的特征空间以及进行信息压缩。随着生物学技术的飞速发展, 大量复杂生物数据不断累积, SVM 也日益成为生物数据分析处理的有力工具, 涉及序列分析领域的有: 蛋白质家族的聚类[38]、启动子分类[39]、蛋白质同源探测[40]、翻译起始位点预测[41]、剪切位点识别[42]、非编码 RNA 分类[43,44]、蛋白质结构预测[45]和 RNA 干扰活性(RNA interference activities)建模[46]等。

SVM 的基本思想基于模式分类的结构风险最小。对于二值问题 $y_i\in\{+1,-1\}$, 给定一系列训练矢量 $\mathbf{x}_i\in\mathbf{R}^d\ (i=1,2,\cdots,n)$, 目的是通过可用的训练样本, 构造一个最小错分二值分类器或决策函数。SVM 将训练矢量 $\mathbf{x}_i\ (i=1,2,\cdots,n)$ 用一个核函数 $K(\mathbf{x}_i,\mathbf{x}_j)$, 从训练空间 $\mathbf{R}^d$ 映射到更高维的特征空间 $\mathbf{H}$, 并找到一个最优分类超平面, 使这个超平面和每一类最临近数据点间的余量最大。SVM 的训练等价于解下面的凸二次最优问题:

$$Min:\ \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K\left(\mathbf{x}_i,\mathbf{x}_j\right)-\sum_{i=1}^{n}\alpha_i \quad (6)$$

上式满足下面的条件:

$\sum_{i=1}^{n} y_i\alpha_i = 0$, $0 \le \alpha_i \le C, i = 1,2,\cdots,n$。式中, n 是训练样本数, C 是调节参数(用于权衡训练误差和余量), α_i 是系数。决策函数为:

$$\mathrm{sgn}\left(\sum_{i=1}^{n} y_i\alpha_i K\left(\mathbf{x}_i,\mathbf{x}\right)+b\right) \quad (7)$$

普遍采用的两个核函数为:

多项式核函数:

$$K\left(\mathbf{x}_i,\mathbf{x}_j\right)=\left(\mathbf{x}_i\cdot\mathbf{x}_j+1\right)^d \quad (8)$$

径向基核函数(Radial basic function, RBF):

$$K\left(\mathbf{x}_i,\mathbf{x}_j\right)=\exp\left(-\gamma\left\|\mathbf{x}_i-\mathbf{x}_j\right\|^2\right) \quad (9)$$

式中, $\gamma = 1/\sigma$, σ 称为核的宽度。

对于特定的数据集, 只要选择合适的核函数并调节参数 C, 经过有效的学习训练即可得到一个理想的 SVM 分类模型。

RNA 二级结构预测是一个二值分类问题: 根据给定的信息, 判断比对中两列能否构成碱基对。分类特征为共变分值、平均碱基配对概率矩阵和碱基互补比例。共变分值探测比对中的互补突变, 平均碱基配对概率矩阵作为补充, 能很好的探测到保守的碱基配对, 碱基互补比例则给出比对中碱基组成的偏好。这里采用 Libsvm[47]作为训练和预测工具。

1.4 考虑假节的茎组合规则

经过 SVM 分类预测, 得到一个潜在配对矩阵以表示可能构成碱基对的位点, 但这其中存在着假阳性数据, 为剔除这些数据的影响, 采用最大茎分值组合方法来得到最后的二级结构。对每个茎采用共变分值、配对概率和互补碱基比例进行打分(茎的最小长度限定为 2), 选择最大分值的茎作为最后的结构组件。茎分值 H_{score} 定义为:

$$H_{score}=\sum_{(i,j)\in H} w_1 C(i,j)+w_2 P(i,j)+w_3 F(i,j) \quad (10)$$

式中, $C(i, j)$是共变分值, $P(i, j)$是平均权重配对概率, $F(i, j)$是互补碱基比例。w_i 是相应的权重, 且满足 $\sum w_i = 1$。

为了考虑假节结构, 采用了类似于迭代环匹配算法(Iterated loop matching algorithm, ILM)[17]的策略: 每次选择有最大分值的预测茎, 然后将包含在这个茎中的碱基对从原始序列中“剔除”(下次选择时不再考虑这些碱基对), 重复上面的步骤直到没有茎被选出。最后将选出的茎按碱基原始排列顺序组合成最终的二级结构。

2 训练及测试数据

2.1 训练数据的构造

为了得到较好的预测结果, 训练集应该尽量包含各类型 RNA 比对及其结构。欧洲核糖体 RNA 数据库(European ribosomal RNA database)[48]、CRW(Comparative RNA website and project)[49]数据库和 Rfam(RNA families database of alignments and CMs)[50,51]数据库都提供了比较详细的 RNA 序列、比对及结构信息。欧洲核糖体 RNA 数据库主要包括 SSU(Small subunit)和 LSU(Large subunit) rRNA 的信息, CRW 数据库以三种核糖体 RNA(5 S、16 S 和 23 S rRNA)、转移 RNA(tRNA)和两类催化内含子 RNA (group Ⅰ 和 group Ⅱ)为主, 而 Rfam 数据库则包含了几乎所有非编码 RNA(ncRNA)家族的信息。这里采用 Rfam[50,51]数据库(版本 8.0)中的 43 组“published” seed 比对及相应结构作为训练集, 比对长度从 26 到 577, 序列数量从 6 到 821, 序列相似度从 40.8%到 97.8%, 包括了不同类型、长度、序列数和相似度的比对。对序列数>40 的家族, 只选出 40 条作为代表序列。代表序列的选择如下: 对于每一个 RNA 家族, 首先计算其两两相似性矩阵 S (两条序列的相似性由它们间的 Hamming 距离和序列长度的比值定义), 对相似性矩阵 S 的每一行求和并按降序排列, 则第一行 $S(1,:)$就意味着对应序列和其他序列的相似度最高; 然后, 找出 $S(1,:)$中的最大值 $S(1, j)$=max $S(1,:)$, 这个值说明, 对应第 j 列的序列和第一行的序列相似性最高, 可以首先剔除, 即删除相似性矩阵 S 对应的第一行和第 j 列, 更新相似性矩阵 S, 重复上述过程, 直到只剩 40 条序列。

通过对已知结构数据库(Rfam-seed 中的 164 组“published”结构)中配对位置空位所占比例的分析(如图 2), 89.9%的配对位置空位比例<70%。因此, 对于比对中的空位采用了下面的过滤规则: 若空位在

任一比对列中所占比例>70%，且这一列在参考结构中不构成碱基对，则将该列滤除。这样不仅降低了空位对训练模型的影响，同时有效的减少了计算量。训练样本的统计特性(包括空位过滤前后的序列相似性、比对长度和比对数量)如图 3 所示。

这里要特别注意的是训练集中正负数据的不平衡：正数据集为 1702 个，而负数据集为 745,103。为平衡正负训练集，考虑到分类器能力的大小主要是对集中在分界面附近点集的划分，在每个正数据集“附近”选出 5 个最临近负数据点，组合为完整的负训练集。临近程度通过标准化后数据点间的欧氏距离度量。此外，在模型训练中，可以通过调整正负样本的权重参数来进一步降低不平衡样本的影响。为增强模型的鲁棒性，使用了 4 轮交叉验证测试来进行模型训练。

2.2 测试数据的构造

测试集的选取标准与训练集类似，采用 49 组 Rfam 数据库的 seed 比对，比对数量从 4 到 1403，长

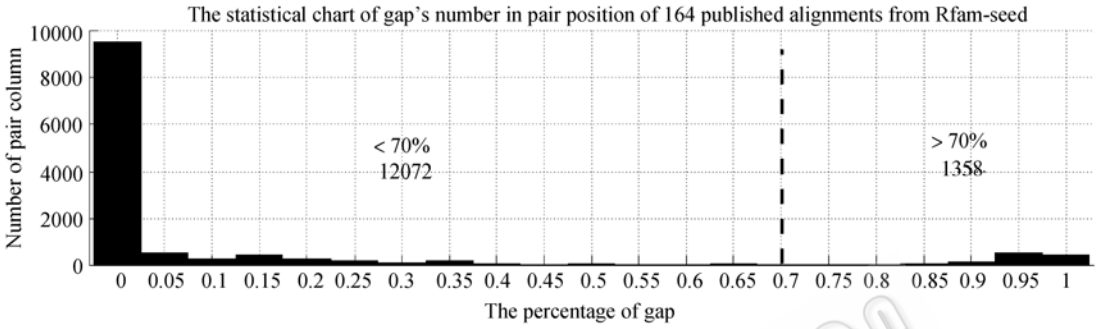

图 2 164 组 published 结构[50,51]中，配对位置空位比例统计图

Fig. 2 Statistical chart of gap's percentage in pair position of 164 published alignment from Rfam-seed[50,51]

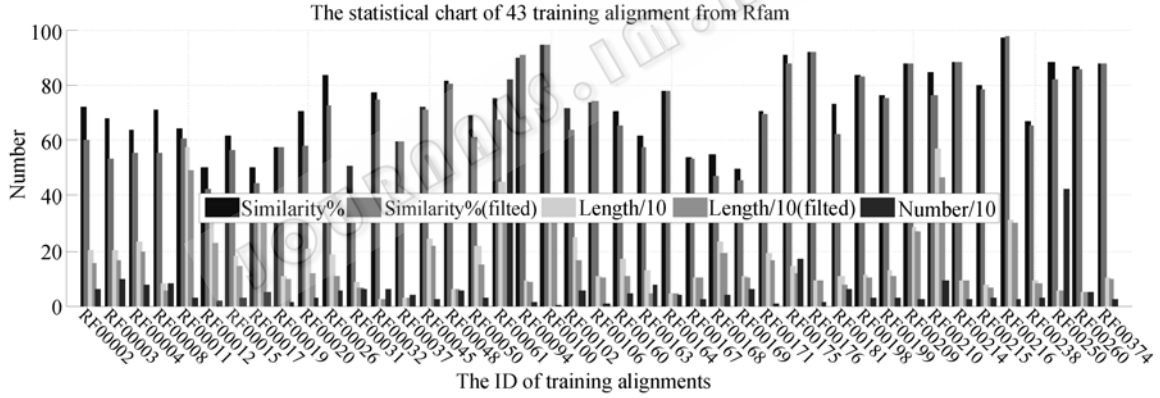

图 3 训练集[50,51]统计特性

Fig. 3 Statistical chart of 43 training alignments from Rfam[50,51]

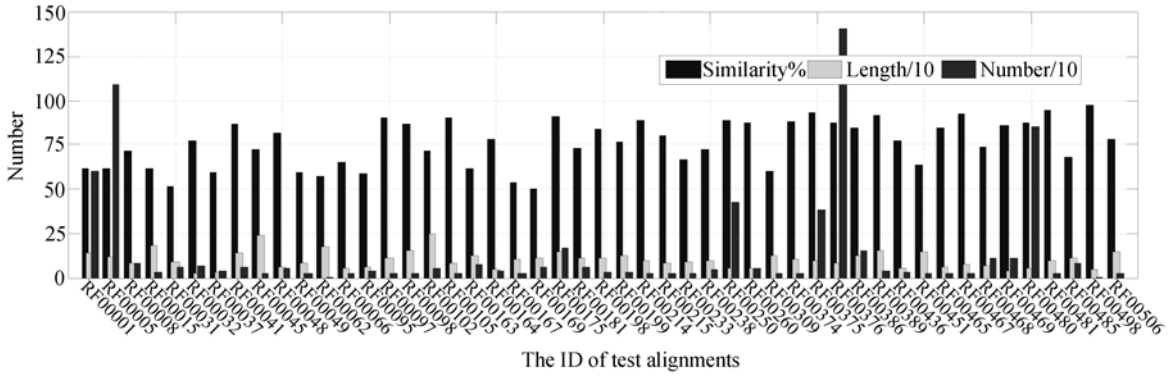

图 4 测试集[50, 51]统计特性

Fig. 4 The statistical chart of 49 test alignments from Rfam[50,51]

度从 26 到 249, 相似度从 49.7%到 97.8%。对于序列数>40 的, 按照构造测试集时采用的方法只选取 40 条代表序列, 比对中的空位不进行过滤。图 4 给出了测试集的统计特性(包括比对中序列相似性、比对长度和数量)。

3 结果及讨论

为评价预测方法的准确性, 这里采用文献[18]中的 Matthews 相关系数(Matthews correlation coefficient (*MCC*))、敏感度(Sensitivity)和特异性(Selectivity)三个指标:

$$MCC=\frac{tp\cdot tn-(fp-\xi)\cdot fn}{\sqrt{(tp+fp-\xi)(tp+fn)(tn+fp-\xi)(tn+fn)}} \quad (11)$$

$$Sensitivity=\frac{tp}{tp+fn} \quad (12)$$

$$Selectivity=\frac{tp}{tp+(fp-\xi)} \quad (13)$$

式中, *tp* 是正确预测的碱基对数, *fn* 是没有预测出的碱基对数, *fp* 是预测错误的碱基对数, ξ项表示由算法预测出但在参考结构中没有, 且不与参考结构矛盾的那些碱基对数量。

为了便于比较, 在 49 组 Rfam-seed 测试集上还分别运行了 KnetFold[21,22]、Pfold[5,19]和 RNAalifold[20]三个同类常用预测程序, 表 1 统计出各种方法的平均 MCC、特异性和敏感度。数据分为高相似性(>80%)和低相似性(<80%)两部分, *SVM-sp*/*SVM* 分别表示考虑/不考虑相似性影响因子的预测方法。

从表中数据可以看出, 相似性影响因子能有效的提高预测的准确性, 特别是对高相似性序列比对: MCC 分别提高 12.5%(高相似性, 从 0.727 到 0.818)、3.8%(低相似性, 从 0.839 到 0.871)。比对序列间的保守程度同样也影响着其他预测算法的准确性, 因为这些算法也是以共变信息为主要预测依据, 对相似度高的序列, 由于共变信息少, 预测精度都有所下降。但 RNAalifold 的结果却正好相反: 高相似性的比对预测出的结果(*MCC*=0.673)反而比低相似性的比对结果(*MCC*=0.595)要高, 可能的原因是此算法对热力学信息的依赖要大于对共变信息的依赖。图 5 分别给出了 SVM 预测结果的各项指标和比对序列相似性、长度和数量的关系。可以看出大多数预测准确性较高(>80%)的比对, 序列相似性<85%。比对长度和预测准确性间的联系不是很明显, 主要是和测试序列的选择有关。序列数量对预测精度的影响较大, 一般情况下序列越多, 预测精度越高, 这是因为序列越多, 比对提供的共变信息和热力学信息就越多, 预测就越准确(保证一定序列相似性前提下)。

假节结构, 作为 RNA 功能的一个重要部分, 由于计算复杂度较高, 通常都被大多数算法所忽略。本方法通过使用类似迭代匹配算法的茎组合规则, 能很好的探测到假节结构。为了测试方法对假节预测的有效性, 我们对 archaeal RNase P (Rfam 中的编号为 RF00373)的结构进行了预测, 结果如图 6 所示, 同时还列出了参考结构和 KnetFold 的预测结果。从图中可以看出, 算法不仅预测出了多数茎结构, 而且也预测出了两个假节结构, 图中大写字母部分。

4 结论

运用机器学习方法, 将 RNA 二级结构预测考虑为分类问题, 通过序列比对提供的互补突变信息、热力学信息和碱基组成信息, 在考虑了序列相似性

表 1 不同方法在 49 组测试比对[50,51]上的结果

Table 1 Result of various methods which performed on 49 test alignments[50,51]

Method	High similarity (>80%, 18 alignments)			Low similarity (<80%, 31 alignments)		
	MCC	Selectivity	Sensitivity	*MCC*	Selectivity	Sensitivity
SVM-sp	0.818	0.779	0.871	0.871	0.832	0.926
KnetFold[21, 22]	0.806	0.783	0.836	0.845	0.800	0.905
SVM	0.727	0.673	0.798	0.839	0.772	0.929
Pfold[5, 19]	0.719	0.649	0.816	0.752	0.671	0.869
RNAalifold[20]	0.637	0.585	0.705	0.595	0.524	0.685

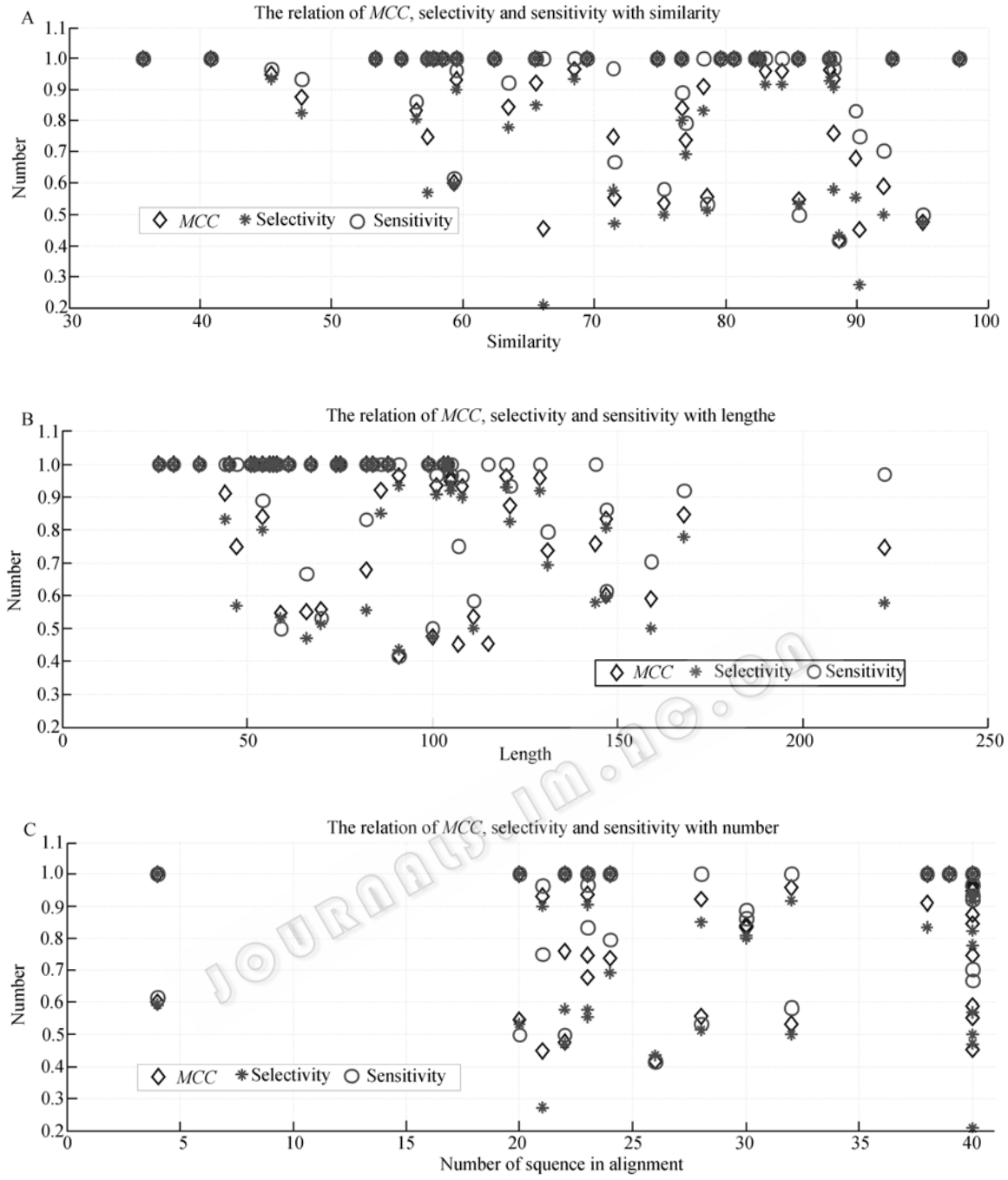

图 5 *MCC*、selectivity、sensitivity 和序列相似性(A)、比对长度(B)及序列数量(C)间的关系

Fig. 5 The relation of *MCC*, selectivity and sensitivity with the sequence similarity (A), the length (B) and the sequence Number of alignment

影响后，经过合适的训练学习，SVM 分类器很好的完成了 RNA 二级结构预测问题。和现有同类预测算法(KnetFold[21,22]、Pfold[5,19]和 RNAalifold[20])相比，预测的准确性得到了提高。同时，通过采用类似迭代匹配算法的茎组合规则实现了对简单假节的预测。

为了进一步提高预测的准确性，应该构造更完整全面的训练集，运用特征选择技术，筛选出相关性高、互补性强的特征子集，而且应该考虑到序列间的进化关系。为更全面的考虑假节的构成，可以采用多类模式分类法，将假节和普通茎结构区分开来，以实现对复杂假节的预测。

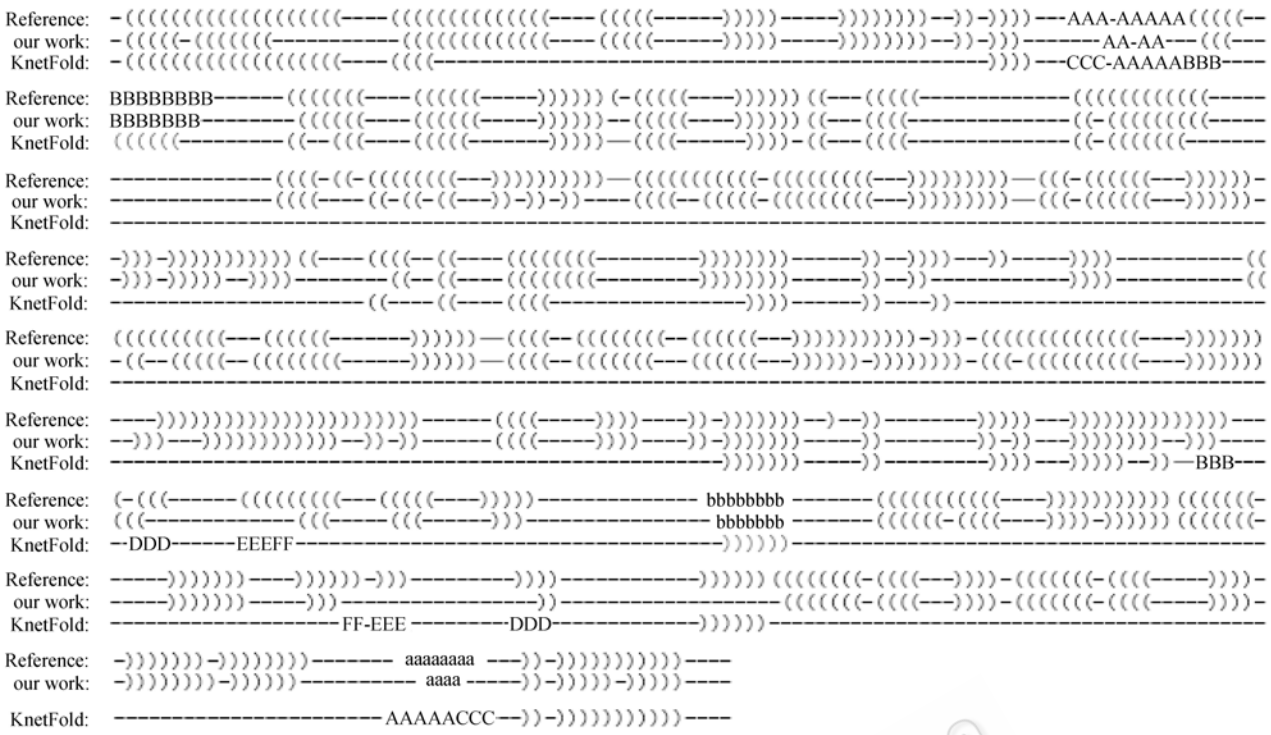

图 6 RF00373 的预测结果比较(参考结构来自 Rfam8.0[50,51], KnetFold[21,22]结果来自 KnetFold 单机版程序)

Fig. 6 Predicted result of RF00373 from various methods(our method, reference[50, 51] and KnetFold[21, 22])

REFERENCES

[1] Rivas E, Eddy SR. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *Journal of Molecular Biology*, 1999, **285**(5): 2053–2068.

[2] Zuker M. Calculating nucleic acid secondary structure. *Current Opinion in Structural Biology*, 2000, **10**(3): 303–310.

[3] Horesh Y, Doniger T, Michaeli S, *et al*. RNAspa: a shortest path approach for comparative prediction of the secondary structure of ncRNA molecules. *BMC Bioinformatics*, 2007, **8**: 366.

[4] Sakakibara Y, Brown M, Hughey R, *et al*. Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Research*, 1994, **22**(23): 5112–5120.

[5] Knudsen B, Hein J. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, 1999, **15**(6): 446–454.

[6] Searls DB. Linguistic approaches to biological sequences. *Computer Applications in the Biosciences*, 1997, **13**(4): 333–344.

[7] Ding Y, Lawrence CE. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Research*, 2003, **31**(24): 7280–7301.

[8] James BD, Olsen GJ, Pace NR. Phylogenetic comparative analysis of RNA secondary structure. *Methods Enzymol*, 1989, **180**: 227–239.

[9] Winker S, Overbeek R, Woese CR, *et al*. Structure detection through automated covariance search. *Computer Applications in the Biosciences*, 1990, **6**(4): 365–371.

[10] Eddy SR, Durbin R. RNA sequence analysis using covariance models. *Nucleic Acids Research*, 1994, **22**(11): 2079–2088.

[11] Le S-Y, Zhang K, Jacob V Maizel J. A method for predicting common structures of homologous RNAs. *Computers and Biomedical Research*, 1995, **28**(1): 53–66.

[12] Witwer C, Hofacker IL, Stadler PF. Prediction of consensus RNA secondary structures including pseudoknots. *IEEE/ACM Transactions Computational Biology and Bioinformatics*, 2004, **1**(2): 66–77.

[13] Lück R, Riesner D, Steger G. Thermodynamic prediction of conserved secondary structure: application to the RRE element of HIV, the tRNA-like element of CMV and the mRNA of prion protein. *Journal of Molecular Biology*, 1996, **258**(5): 813–826.

[14] Lück R, Gräf S, Steger G. ConStruct: a tool for thermodynamic controlled prediction of conserved secondary structure. *Nucleic Acids Research*, 1999, **27**(21): 4208–4217.

[15] Do CB, Woods DA, Batzoglou S. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, 2006, **22**(14): e90–98.

[16] Tabaska J, Cary RB, Gabow H, *et al*. An RNA folding method capable of identifying pseudoknots and base triples. *Bioinformatics*, 1998, **14**: 691–699.

[17] Ruan J, Stormo GD, Zhang W. An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics*, 2004, **20**: 58–66.

[18] Gardner PP, Giegerich R. A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics*, 2004, **5**: 140.

[19] Knudsen B, Hein J. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Research*, 2003, **31**(13): 3423–3428.

[20] Hofacker IL, Fekete M, Stadler PF. Secondary structure prediction for aligned RNA sequences. *Journal of Molecular Biology*, 2002, **319**(5): 1059–1066.

[21] Bindewald E, Shapiro BA. RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers. *RNA*, 2006, **12**(3): 342–352.

[22] Bindewald E, Schneider TD, Shapiro BA. CorreLogo: an online server for 3D sequence logos of RNA and DNA alignments. *Nucleic Acids Research*, 2006, **34**: W405–W411.

[23] Sankoff D. Simultaneous solution of the RNA folding, alignment, and proto-sequence problems. *SIAM Journal on Applied Mathematics*, 1985, **45**: 810–825.

[24] Gorodkin J, Heyer LJ, Stormo GD. Finding the most significant common sequence and structure motifs in a set of RNA sequences. *Nucleic Acids Research*, 1997, **25**(18): 3724–3732.

[25] Gorodkin J, Stricklin SL, Stormo GD. Discovering common stem-loop motifs in unaligned RNA sequences. *Nucleic Acids Research*, 2001, **29**(10): 2135–2144.

[26] Mathews DH, Turner DH. Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *Journal of Molecular Biology*, 2002, **317**(2): 191–203.

[27] Hofacker IL, Bernhart SH, Stadler PF. Alignment of RNA base pairing probability matrices. *Bioinformatics*, 2004, **20**(14): 2222–2227.

[28] Höchsmann M, Töller T, Giegerich R, *et al*. Local similarity in RNA secondary structures. *Proc IEEE Computer Society Bioinformatics Conference*, 2003, **2**: 159–168.

[29] Siebert S, Backofen R. MARMA: a server for multiple alignment of RNAs. Proceedings of the German Conference on Bioinformatics. German, 2003.

[30] Siebert S, Backofen R. MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. *Bioinformatics*, 2005, **21**(16): 3352–3359.

[31] Chiu DK, Kolodziejczak T. Inferring consensus structure from nucleic acid sequences. *Computer Applications in the Biosciences*, 1991, **7**(3): 347–352.

[32] Gutell RR, Power A, Hertz GZ, *et al*. Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucleic Acids Research*, 1992, **20**(21): 5785–5795.

[33] McCaskill JS. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 1990, **29**(6-7): 1105–1119.

[34] Layton DM, Bundschuh R. A statistical analysis of RNA folding algorithms through thermodynamic parameter perturbation. *Nucleic Acids Research*, 2005, **33**(2): 519–524.

[35] Hofacker IL. Vienna RNA secondary structure server. *Nucleic Acids Research*, 2003, **31**(13): 3429–3431.

[36] Hofacker IL, Fontana W, Stadler PF, *et al*. Fast folding and comparison of RNA secondary structures. *Monatsh Chemie*, 1994, **125**: 167–188.

[37] Vapnik VN. Statistical Learning Theory, New York: John Wiley and Sons, Inc, 1998.

[38] Karchin R, Karplus K, Haussler D. Classifying G-protein coupled receptors with support vector machines. *Bioinformatics*, 2002, **18**: 147–159.

[39] Pavlidis P, Furey TS, Liberto M, *et al*. Promoter region-based classification of genes. *Pacific Symposium Biocomputing*, 2001: 151–163.

[40] Jaakkola T, Diekhans M, Haussler D. A discriminative framework for detecting remote protein homologies. *Journal of Computational Biology*, 2000, **7(1-2)**: 95–114.

[41] Zien A, Rätsch G, Mika S, *et al*. Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics*, 2000, **16**: 799–807.

[42] Sonnenburg S, Rätsch G, Jagota A, *et al*. New methods for splice site recognition. Proceedings of the International Conference on Artificial Neural Networks, 2002.

[43] Yan K, Richard FM, Stephen RH. Classification of non-coding RNA using graph representations of secondary structure. *Pacific Symposium on Biocomputing*, 2005: 4–15.

[44] Wang JTL, Wu X. Kernel design for RNA classification using Support Vector Machines. *International Journal of Data Mining and Bioinformatics*, 2006, **1**: 57–76.

[45] Hua S, Sun Z. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. *Journal of Molecular Biology*, 2001, **308**(2): 397–407.

[46] Peek AS. Improving model predictions for RNA interference activities that use support vector machine regression by combining and filtering features. *BMC Bioinformatics*, 2007, **8**: 182.

[47] Chang CC, Lin CJ. LIBSVM: a library for support vector machines. 2001.

[48] Wuyts J, Perriere G, Van De Peer Y. The European ribosomal RNA database. *Nucleic Acids Research*, 2004, **32**(Database issue): D101–103.

[49] Cannone JJ, Subramanian S, Schnare MN, *et al*. The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics*, 2002, **3**: 2.

[50] Griffiths-Jones S. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Research*, 2005, **33**: 121–124.

[51] Griffiths-Jones S, Baterman A, Marshall M, *et al*. Rfam: an RNA family database. *Nucleic Acids Research*, 2003, **31**: 439–441.