RNA 二级结构预测的支持向量机模型研究

何静媛　何中市　邹东升
（重庆大学计算机学院　重庆 400044）

摘　要　RNA 二级结构预测问题是生物信息学的一个研究重点，本文主要利用支持向量机(SVM)模型来研究 RNA 二级结构预测问题。通过改进 NSSEL 标签[4]，形成了能表示平面伪结结构的 E-NSSEL 标签，该标签作为 SVM 模型输出端的类别标识，因此，测试序列经过 SVM 模型预测后得到相应的 E-NSSEL 序列，该序列可以恢复为二级结构。此算法能有效地解决传统算法中存在的时间复杂性的问题和长链分子的预测问题。

关键词　SVM，E-NSSEL 标签，RNA 二级结构，平面伪结

The Model Research of Support Vector Machines in the RNA Secondary Structure Prediction

HE Jing-Yuan　HE Zhong-Shi　ZOU Dong-Sheng
(College of Computer Science, Chongqing University, Chongqing 400044)

Abstract　One of the most important research areas in bioinformatics is RNA secondary structure prediction. This paper introduces a new representation of the RNA structure information by extending NSSEL labels[4], the new labels (E-NSSEL) can express plane pseudoknots. A SVM model is presented to predict RNA secondary structures based on E-NSSEL labels, using this algorithm, test sequences can be converted to E-NSSEL sequences which can be revert to RNA secondary structures. Experiment shows that this model can solve too long computational time problem of traditional algorithms, at the same time, it can predict long RNA sequences which are difficult with traditional folding algorithms.

Keywords　SVM, E-NSSEL labels, RNA secondary structure, Plane pseudoknots

1　引言

RNA 二级结构的预测是分子生物学领域中的重要研究课题。虽然利用 x 射线可以直接对 RNA 分子的结构进行测定，但这种方法昂贵、低效，无法对所有的 RNA 分子进行测定，因此利用计算机进行结构预测成为该领域发展的必然趋势。事实上，使用计算学的方法来预测 RNA 的二级结构已经有 30 多年的研究历史了，迄今为止，比较典型的预测算法有最小自由能算法[1]、随机上下文无关(SCFG)算法等[2]。这些算法普遍具有的缺点就是计算时间复杂性高，预测长链的 RNA 分子较困难。

随着计算机技术的发展，机器学习的方法在很多领域得到了应用，目前，已经有一些算法如 SCFG、遗传算法[3]等来预测 RNA 的二级结构，取得了不错的成效，但是，使用人工神经网络来解决此问题的文献相当有限。目前还未见到将支持向量机算法(SVM)应用到该领域的文献报告。文[4]中的方法较好地解决了 BP 神经网络中输出端的结构描述问题，但是该方法不能对包含伪结结构的 RNA 分子进行预测，同时由于 BP 神经网络自身的局限，算法较难达到理想精度。本文针对 RNA 平面伪结的基本特征，结合 SVM 算法的优势，提出了一个高效的、基于 SVM 的 RNA 二级结构预测算法模型，可以较好地解决平面伪结和长链分子的预测问题。

2　RNA 二级结构简介

RNA 为单链分子，它通过自身回折使得可以彼此配对的碱基相遇，形成氢键，同时形成双螺旋结构，这些双螺旋结构称为“茎”，不能配对的碱基区形成无规环被排斥在双螺旋结构之外。

RNA 的二级结构可以用图 1 描述。

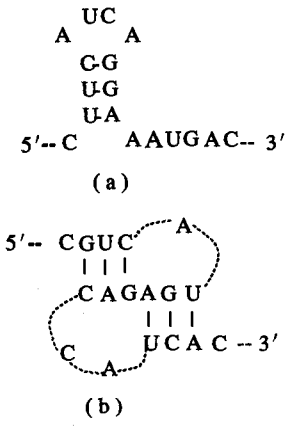


图 1　RNA 二级结构图

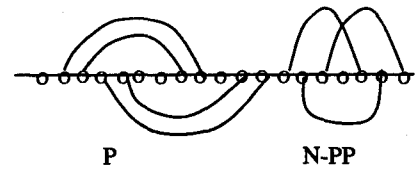


图 2　伪结点示意图

图 1(a)是典型的 stem-loop 结构，这种结构包括发夹环、内环、突起等类型；图(b)是一个平面伪结的示意图，我们也可

何静媛　博士研究生，主要从事模式识别、生物信息学方向的研究。

以这样来描述伪结：若在序列中存在基对〈i,j〉与〈ic,jc〉，若满足 $i<ic<j<jc$(这里的 i、j、ic、jc 表示在碱基在分子链中的位置)，则该序列存在伪结结构。该结构可以用 Feyman 图式表示，图 2 是伪结点的 Feyman 示意图，其中 P 结构是平面伪结结构，而 N-PP 为非平面伪结结构，因为该结构中存在相交弧线。分析可知，平面伪结与 stem-loop 都属于嵌套结构，而非平面伪结则为非嵌套结构。

3 扩展后的 NSSEL 二级结构标签

文[4]中提出了 NSSEL(New Secondary Structure Element Label)标签，这些标签可以表达简单的二级结构特征，但是不能标识伪结结构，本文在旧标签定义的基础上，去掉部分冗余标记，增加了能表达平面伪结的标记，为了便于理解新 NSSEL 中的结构单元，首先介绍正茎和负茎的概念：几个连续的碱基配对构成一段茎(stem)。直观地看，这一段茎是由两段子序列构成的，其中一段子序列在整个序列中位置比较靠近 5′端，称之为正茎，另一段子序列比较靠近 3′端，称为负茎，扩展后的 NSSEL 标签(下文称之为 E-NSSEL 标签)定义说明如下：

1)＋Stem 表示正茎中的碱基子序列；

2)＋pseudoknots 表示靠近 5′端的伪结碱基子序列(或单个结点)；

3)－Stem 表示负茎中的碱基子序列；

4)－pseudoknots 表示靠近 3′端的伪结碱基子序列(或单个结点)；

5)Loop 表示所有未配对的碱基段。

其中每类标记代表的子序列可以使用对应的数字序列表达，如果对应上述二级结构单元，用从 1 到 5 的数字给定一个 E-NSSEL 标签，则对于一个已知二级结构的 RNA 序列，就可以用一个 E-NSSEL 标签序列来表示相应的二级结构，例如图 1 中(a)可以表示为：51115555333555555；(b)则可表示为：51115222333554445。

显然，一个 RNA 序列的二级结构可以唯一地表示为一个 E-NSSEL 序列，而由 RNA 的线性序列及其 E-NSSEL 序列，也可以唯一地恢复出该序列的二级结构。对于一个已知的 E-NSSEL 序列，只要在对其顺序遍历的时候，将＋Stem 和＋pseudoknots 分别压栈，后面遇到的－Stem 和－pseudoknots 必然和栈顶的＋Stem 或＋pseudoknots 匹配。这样就很容易把二级结构勾画出来。

4 支持向量机模型

4.1 SVM 简介

支持向量机是建立在统计学习理论和结构风险最小原理基础上的，根据有限样本信息在模型的复杂性（对特定训练样本的学习精度）和学习能力（无错误地识别任意样本的能力）之间寻求最佳折衷，以期获得最好的推广能力。统计学习理论可绕过对 RNA 分子序列中的基本结构之间的物理、化学过程的分析，直接从结果对 RNA 结构进行统计分析，同时又克服了传统统计学理论对于先验信息要求过多的限制。在利用观测数据对依赖关系进行估计时，只需知道未知依赖关系所属函数集的某些一般性质，由于支持向量机专门针对有限样本，并将问题转为一个二次型寻优问题，这使它成为本次实验的一个很好的选择。

支持向量机对不可分问题，是通过核函数将输入空间映射到高维特征空间，使得在高维空间中更好地反映 RNA 碱基序列和结构构象间的关系，示意图如图 3。

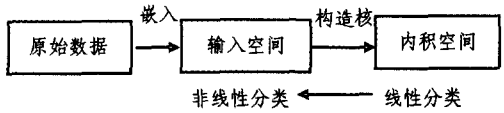


图 3 支持向量机的分类过程

由于序列中的相关性和特征在很大的程度上是未知的，因此在有限的样本中，利用适当的核函数进行特征转换，是提高 RNA 结构预测的一个有效途径。

与 BP 神经网络模型相比较，SVM 的二元分类器构建要简单得多。因为要获得一个合适的神经网络模型，需要设计者有丰富的经验，依据以往相关经验调试参数，从而使得函数不至于发散。而 SVM 模型对于以前的数据依赖并不强，这也是本文选择 SVM 模型的一个重要原因。

4.2 用 SVM 预测 RNA 的二级结构

SVM 算法是典型的二分类算法，本文涉及到的多分类算法采用的是一对多的策略。本文首先选取径向基核函数，在惩罚系数 C=100 时，系统有较好的性能。

在进行二级结构预测时，可将 RNA 序列看作与二级结构相关的信号，二级结构与碱基的长程和短程信息相关。多次实验表明窗口长度为 15 时有较好的预测效果，预测时选择一个碱基前后相邻的 7 个，共 15 个碱基作为一个框架编辑输入，通过移动中心位置得到多个样本输入。每一个窗对应一个样本，用一个 15×5b=75b 的 0—1 编码来表示这个窗的输入，编码的含义是：窗口中的每一位代表一个碱基，用一个 5b 正交的 0—1 编码来表示该碱基类型(ACGU)中的一种，除此之外，还用一位标识一条序列的结束。各碱基字符和二进制编码之间的对应关系为：

A—10000　C—01000　G—00100

U—00010　END—00001

窗口输出的则是中间位置的碱基在其分子二级结构中所属的类别(NSSEL 标签的 5 个标记中的一个)。

与神经网络类似，SVM 模型将训练样本 RNA 进行编码作为输入，对应的二级结构类别作为样本输出。学习完毕后，需要预测的序列采用同样的编码方式输入，就可以预测该序列对应的二级结构。

经过预测，当测试样本序列得到其对应的 ENSSEL 标记序列时，我们必须还要做调整和修正，主要方案如下：

(1) 序列第一个碱基预测得到的类型不可能是 3 或者 4。最后的碱基则不可能是 1 或 2。若出现错误预测，则根据堆栈操作，修改成正确标记。

(2) 当标记 2 出现时，则必须出现在标记 1 或者 5 的后面，因为伪结是由环上的碱基与环外的碱基配对形成。否则根据上下文来修正。

(3) 当 ENSSEL 标记序列进行堆栈操作，进行二级结构的恢复时，关注配对的 1 标记和 3 标记的数目是否相等(对于标记 2,4 同理)，否则尝试在相邻位置恢复。

4.3 实验数据

为了保证训练数据集中有充分的平面伪结的信息，故本文实验数据是从 EMBL 数据库(http://www.ebi.ac.uk/embl)中抽取出来的含有平面伪结结构的 RNA 序列，这些序列集可以下载(http://biology.leidenuniv.nl/～batenburg/PKBGet.html)。我们使用 35 条序列，共 1322 个碱基作为训练样本，再取出 22 条序列(684 个碱基)进行测试，应用灵敏度为 79.67、特异度为 85.43、预测精度为 83.66。同时我们使用相同的数据对文[4]中的 BP 网络进行训练和测试，结果比较

如表1所示。

表1 测试结果比较

方法	窗口长度	Stem-loop预测准确率	平面伪结预测准确率
SVM	15	80.13%	71.80%
BP网络[4]	13	70.93%	无法预测

表1中的stem-loop预测准确率定义为：预测正确的stem-loop标记数目占基本二级结构标记(不包含伪结结构标记)总数的百分比。而平面伪结预测准确率则为正确预测的平面伪结标记数目占实际平面伪结标记数目的百分比。

与文[4]所使用的BP神经网络相比，支持向量机的方法对stem-loop结构的预测准确率有了较大的提高，同时还可以较理想地预测平面伪结结构。与传统的基于确定性的动态规划算法(SCFG、最小自由能等模型)比较起来，SVM成功地解决了时间复杂度的问题，并且利用滑动窗的输入方式将不受序列长度的影响。

总结 使用SVM模型来预测RNA二级结构是一种全新的尝试。本文利用SVM算法的优势，结合RNA二级结构的特征，扩展了文[4]中的NSSEL标记的内涵，成功地实现了包含平面伪结结构的RNA二级结构预测。实验表明，该算法能够得到理想预测精度，并且能有效解决传统算法中存在的计算复杂性和长链分子的预测问题。由于未能发现非平面伪结结构的有效标记，因此本算法暂时不能实现非平面伪结结构的预测，作者将会继续深入研究，期待有进一步的发现。

参考文献

1 Zuker M. Optimal computer folding of large RNASequence using thermodynamics and auxiliary information, Nucl. Acids Res., 1981, 9:133～148
2 SakakibaraY, Brown M, Hughey R, et al. Stochastic context-free grammars for tRNA modeling[J]. Nucleic Acids Research, 1994, 22(23):5112～5120
3 Shapiro B A, Navetta J. A massively parallel genetic algorithm for RNA structure prediction. J. SuperComput 1994, 8:195～207
4 张秀苇，邓志东，宋丹丹. RNA二级结构预测的神经网络方法. 清华大学学报(自然科学版)，2006(10)
5 Gorodkin J, Stricklin S L, Stormo G D. Discovering common stem-loop motifs in unaligned RNA sequences. Nucleic Acids Res., 2001, 29:2135～2144
6 HolleyL H, KarplusM. 基于神经网络的蛋白质二级结构预测[J]. 生物物理学，1989, 86:152～156
7 Eddy S R, Durbin R. RNA sequence analysis using covariance models. Nucleic Acids Research, 1994, 22(1):2079～2088
8 VAPNIKVN著. 统计学习理论的本质[M]. 张学工，译. 北京：清华大学出版社，2000

(上接第169页)

简算法，我们确定了16条规则数和主条件属性。主条件属性有日期，威廉指数，DSY，开盘价，最高价，最低价，收盘价以及成交量。

根据上述的数据处理方法得到相应的神经网络模型，即：第1层的节点数为8个，第2层的节点数为16个，第3层的节点数为16个，第4层的节点数为1个。

4.3 网络的训练及测试的效果评价

4.3.1 网络的训练

网络的训练样本为上证A指数从2004.07.08到2005.06.30的相关数据，共计240个交易日。各个技术指标从指数的原始数据中用程序处理得到。训练结果的统计分析如表2所示。

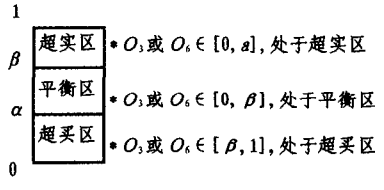
表2 网络训练结果

	最大误差	平均误差	训练周期
NO_3的网络训练结果	0.10125	0.07232	420233
NO_6的网络训练结果	0.10124	0.06412	502500

4.3.2 网络测试的效果评价

测试使用2005.07.01～2006.05.28之间的数据，共221个交易日。将样本值输入，经训练后NO_3的最大误差为0.1024，平均误差为0.0642，NO_6的最大误差为0.1140，平均误差为0.0624。

现根据若O_3和O_6两指标编制的含义，将区间[0,1]按一定的界限划分成如下图所示的三个区域



为了进一步评价所建立的网络模型对指数交易指导意义，假定在测试样本的221个交易日实行如下的交易方式：

1)投资者持有指数股票

1. O_3或$O_6\in[\alpha,1]$，处于平衡区或超卖区，说明股市行情处于盘整或见涨趋势，投资者可以继续持有股票；

2. O_3或$O_6\in[0,\alpha]$，处于超买区，说明行情即将回落，应将指标看成是卖出的信号，投资者出售股票。

2)投资者尚未持有指数股票

1. O_3或$O_6\in[0,\beta]$，行情位于平衡或买区，股市调整或回落，投资者应继续观望，不应马上介入；

2. O_3或$O_6\in[\beta,1]$，行情处于超卖区，股市即将反弹，会有一段上涨趋势，是买入的信号，投资者应买入股票。

α与β的具体数值可根据投资的不同倾向来确定。由于股市存在着交易费用问题，如经纪人佣金，印花税等，因此在一定的时间内，交易的次数很多并不意味着能给投资者带来更大的收益，具体操作要视实际情况来抉择。

结论 从大量的观察和实验数据获取知识，表达知识，推理决策规则是智能信息处理的重要任务。本文基于模糊粗糙集理论构建了新型的模糊-粗神经网络(FRNN)，该模型综合了粗糙集理论在知识获取方面的能力和模糊神经网络在数值逼近上的优势。通过粗糙集智能数据分析，可消除初始决策表中的冗余信息和噪声数据的干扰，减少了模糊神经网络中输入层和代表规则层的神经元个数，简化了神经网络的拓扑结构，减少了训练所需的计算量和时间，提高了模型的正确率。

参考文献

1 Pawalk Z. Rough sets[J]. International Journal of Computer and Information Science, 1982, 11(5):341～356
2 Hu X H, Cercone N. Learning in Relational Database: a Rough Set Approach [J]. Computational Intelligence, 1995, 11(2):323～338
3 Lingras P J. Rough Neural Networks[A]. In: Proceedings of Sixth International Conference on Information. Processing and Management of Uncertainty in Knowledge_Based Systems[C]. 1996. 1445～1450
4 Fayyad U, Shapiro G, Smyth P. The KDD Process for Extracting Usefl Knowledge From Volumes of Data. Communications of the ACM, 1996, 39(11):27～34
5 苗夺谦. 粗糙集理论中连续属性的离散化方法[J]. 自动化学报，2001, 27(3):296～302
6 刘清. Rough集及Rough推理. 科学出版社，2001
7 张文修，吴伟志. 粗糙集理论与方法. 科学出版社，2001