

# 基于前馈人工神经网络的 miRNA 预测

林云光, 陈月辉, 邵光亭

(济南大学 信息科学与工程学院, 山东 济南 250022)

**摘要:** microRNA (miRNA) 是一类长度约为 20 ~ 24 个核苷酸保守的非编码小分子 RNA, 如何能准确预测 miRNA 一直是生物信息学的难点之一。文中提出一种新的预测方法—粒子群优化的前馈人工神经网络预测 miRNA, 从 331 (阴性数据 168, 阳性数据 163) 个样本组成的数据集中提取每个样本的 36 维特征向量训练人工神经网络模型, 并用训练好的模型对不同的测试集进行测试, 结果表明这种方法平均预测精度达到 91.0%, 高于传统的 SVM 预测方法, 从而为 miRNA 预测提供了一个新的研究方向。

**关键词:** microRNA; 前馈人工神经网络; 粒子群优化算法

**中图分类号:** TP183

**文献标识码:** A

**文章编号:** 1673-629X(2012)05-0019-04

## Prediction of miRNA Based on Feedforward Artificial Neural Network

LIN Yun-guang, CHEN Yue-hui, SHAO Guang-ting

(School of Information Science and Engineering, University of Jinan, Jinan 250022, China)

**Abstract:** microRNA (miRNA) is a class of 20 ~ 24 long nucleotides conserved non-coding small RNA. How to predict miRNA accurately is one of the difficulties in bioinformatics. A new predicting method has been proposed in this paper, that is, particle swarm optimized feedforward artificial neural network. Use 36 feature extracted from the data set comprised of 331 samples to train the neural network model, which used to test new data-sets get a prediction accuracy up to 91.0%. This indicates that the model can be used as a new direction to predict miRNA.

**Key words:** microRNA; feedforward artificial neural network; particle swarm optimization

### 0 引言

miRNA 是一类长度约为 20 ~ 24 个核苷酸的内源性的保守的非编码蛋白质的小分子 RNA, 它通过与 miRNA 靶基因的 3' 端非编码区域互补配对来实现对靶基因的抑制或裂解, 从而达到调控基因的目的<sup>[1]</sup>。miRNA 是由具有茎环二级结构的 miRNA 前体 (pre-miRNA) 加工而成。对 miRNA 的研究有助于了解基因间的网络调控关系, 并且对基因功能的深入研究和生物进化探索等有着重要的意义, 所以 miRNA 自发现以来便成为生物信息学领域研究的热点问题<sup>[2]</sup>。

据推测实际存在的 miRNA 数量比人类已知的数量要多的多, 由于组成 miRNA 的核苷酸数量非常小, 因此如何高效地从基因序列中检测出真正的 miRNA 成为 miRNA 研究中的一个首要问题。目前预测 miRNA

的方法主要包括生物实验方法和计算识别的方法。

生物实验的方法准确率虽然高, 但费时、费力、费成本限制了该方法的应用。计算识别的方法因其低成本、高效、信息处理量大从而适合大规模的测试。多种计算方法被应用到动植物 miRNA 的预测中, 总结这些方法可以发现基本上都采取如下策略:

(1) 利用同源性搜索和在已知 miRNA 附近搜索基因簇的方法<sup>[3]</sup>。此种方法主要是利用 miRNA 及 pre-miRNA 在进化过程中的保守性, 代表性的算法如 MIRScan<sup>[4]</sup>。

(2) 其他不依赖于同源性和 miRNA 基因簇的预测方法。这些方法综合利用了 pre-miRNA 基因序列可形成潜在茎环结构、pre-miRNA 序列最小自由能、茎环结构的碱基配对数等特征进行 miRNA 预测<sup>[5]</sup>。目前此类方法中比较成功的是 XUE<sup>[6]</sup> 等人开发的 triplet-SVM 识别算法, 平均预测精度接近 89%。

文中提出了一种全新的预测方法——用粒子群优化的前馈人工神经网络来预测 miRNA, 采用类似 XUE

收稿日期: 2011-10-15; 修回日期: 2012-01-20

基金项目: 国家自然科学基金 (61070130)

作者简介: 林云光 (1982-), 男, 广西平乐人, 硕士研究生, 研究方向为计算智能、生物信息学; 陈月辉, 教授, 博 (硕) 士生导师, 研究方向为计算智能、生物信息学、金融序列预测。



## 2 预测工具与方法

### 2.1 前馈人工神经网络

人工神经网络是指模拟人脑神经系统的结构和功能,运用大量的处理单元来构造模型,解决诸如知识表达、推理学习、联想记忆,乃至复杂的社会现象等问题,它以非线性、分布式存储和并行协同处理等为特色<sup>[11]</sup>。虽然单个神经元的结构和功能简单有限,但是由大量的神经元所组成的网络系统却能模拟出极其复杂的非线性问题。已经证明只要隐层神经元个数足够多,使用单调递增可微函数的单隐层前馈神经网络能够逼近任意的连续函数<sup>[14]</sup>。

文中所使用的是前馈型人工神经网络<sup>[15]</sup>,属于有导师训练的神经网络,在有导师训练中输入向量和与其对应的目标向量构成一个训练样本。文中输入向量由 36 维组成,目标向量为 +1、-1,分别代表是或不是 pre-miRNA 前体。前馈神经网络通过目标向量与实际输出值的差值的均方差来调整网络权值从而降低整体误差,最终模拟出一个最佳预测模型用来检测未知序列是否为 miRNA 前体。

图 2 是一个标准的前馈神经网络,包括三层:一个输入层、一个隐层和一个输出层。其中隐层也可由多层组成,并且输出层也可有多个输出单元。文中输入层  $I = 36$ ,取隐层神经元个数  $J = 10$ ,输出层  $k = 1$ 。

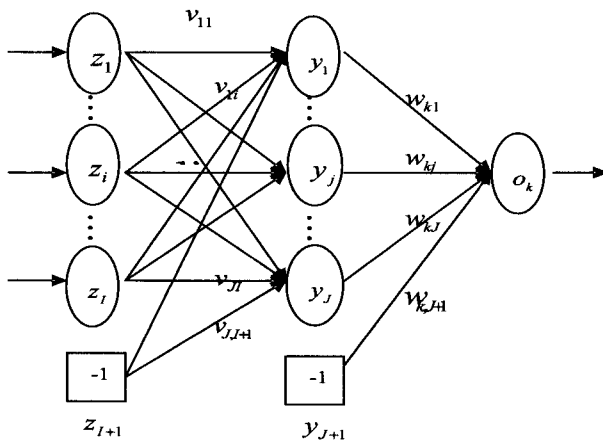


图2 前馈人工神经网络模型

对于任意给定的输入模式  $z_p$ , 输出结果如下:

$$\begin{aligned} O_{k,p} &= f_{o_k}(\text{net}_{o_k}) \\ &= f_{o_k}\left(\sum_{j=1}^{J+1} W_{kj} f_{y_j}(\text{net}_{y_j})\right) \\ &= f_{o_k}\left(\sum_{j=1}^{J+1} W_{kj} f_{y_j}\left(\sum_{i=1}^{I+1} V_{ji} Z_{i,p}\right)\right) \end{aligned} \quad (5)$$

其中  $f_{o_k}$  和  $f_{y_j}$  分别是输出单元  $o_k$  和隐层单元  $y_j$  的激活函数。 $W_{kj}$  是输出单元  $o_k$  和隐层单元  $y_j$  之间的权值。 $z_{i,p}$  是输入模式  $z_p$  的输出单元  $z_i$  的值。第  $(I+1)$  个输入单元和第  $(J+1)$  个隐层单元分别代表下一层中神经元阈值的偏置单元,一般取为 -1。

这里使用的激活函数是 Sigmoid 函数,它具有连续、光滑、严格单调等特性,是一个良好的阈值函数,因而被广泛应用于神经元的输出特性中。形式为:

$$f(x) = \frac{1}{(1 + e^{-x})} \quad 0 \leq f(x) \leq 1 \quad (6)$$

### 2.2 粒子群优化算法

人工神经网络的推理知识表示主要体现在网络连接权值的优化上,文中用粒子群优化算法 (PSO) 来优化神经网络权值。PSO 是一种以群体为基础的最佳化搜寻方法,由 James Kennedy 和 Russell Eberhart 两位学者于 1995 年时所提出<sup>[16]</sup>。群中的每个粒子在解空间中都有一个对应的位置向量  $x^i$  和速度向量  $v^i$ ,每个粒子经由适应函数的衡量而具有一个适应值,迭代过程就是朝着最优化适应值的方向发展。

$$v_{k+1}^i = \omega \times v_k^i + c_1 \times \text{rand}() \times (p_k^i - x_k^i) + c_2 \times \text{rand}() \times (p_k^g - x_k^i) \quad (7)$$

$$x_{k+1}^i = x_k^i + v_{k+1}^i \quad (8)$$

其中,  $c_1$ 、 $c_2$  为学习因子,  $\omega$  为惯性因子作用是调节解空间的搜索范围。 $p^i$  代表每个粒子的历史最优解,而  $p^g$  代表整个种群的最优解。算法循环执行直到得到一个令人满意的结果或满足终止条件为止。

## 3 对比分析

以 TR-ANN 为训练样本来优化前馈人工神经网络,并利用优化好的模型来对 TE-30-set、TE-updated-set、TE-1000-set、TE-2444-set 四个测试集进行预测,测试结果如表 1 所示:

表1 不同测试集的预测精度比较

测试集	类型	大小	预测精度	
			triplet-SVM	文中
TE-30-set	阳性	30	93.3	96.7
TE-updated-set	阳性	39	92.3	95.1
TE-1000-set	阴性	1000	88.1	94.6
TE-2444-set	阴性	2444	89.0	89.4

从表中可以看出,文中所使用的预测方法结果均优于 triplet-SVM 方法,triplet-SVM 方法的平均预测精度为 88.8%,而文中的预测平均精度达到了 91.0%。说明以 PSO 优化的前馈神经网络方法在预测 pre-miRNA 方面是有效的,在一定程度上提高了 miRNA 预测的准确度。并开拓了一条新的 miRNA 预测途径。

## 4 结束语

我们知道 SVM 的预测方法在小样本的集合上有很大的优势,具有较高的准确度。而文中通过增加了 4 维特征后用前馈神经网络提高了预测的准确度,表

明利用神经网络的方法在预测 miRNA 方面也是行之有效的。并且神经网络更擅长处理高维、大样本的数据集,随着已发现的 miRNA 数量越来越多,在大样本预测的情况下人工神经网络将体现出它独特的优势,准确度也会进一步提高。后续工作可以在两方面进行继续跟进,一是提取更多的 miRNA 相关数据集,更多的训练样本可以拟合出更精确的模型;二是改进人工神经网络,通过增加反馈、集成的方法进一步提高预测的精度,相信人工神经网络的方法在预测 miRNA 的方面会有更大的发挥。

#### 参考文献:

- [1] Bartel D P. MicroRNAs: genomics, biogenesis, mechanism and function[J]. Cell, 2004, 116(2): 281-297.
- [2] Robert J, Johnston S. A microRNA controlling left/right neuronal asymmetry in *Caenorhabditis elegans* [J]. Nature, 2003, 426(6968): 845-849.
- [3] Altuv Y, Landgraf P, Lithwick G, et al. Clustering and conservation patterns of human microRNAs[J]. Nucleic Acids Res, 2005, 33(8): 2697-2706.
- [4] Lim L P, Lau N C, Weinstein E G, et al. The microRNAs of *Caenorhabditis elegans* [J]. Genes Dev, 2003, 17(8): 991-1008.
- [5] Carter R J, Dubchak I, Holbrook S R. A computational approach to identify genes for functional RNAs in genomic sequences[J]. Nucleic Acids Res, 2001, 29(19): 3928-3938.
- [6] Xue C H, Li F, He T, et al. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine[J]. BMC Bioinformatics, 2005, 6(1): 310-310.
- [7] Griffiths-Jones S. The microRNA Registry[J]. Nucleic Acids Research, 2004, 32(1): 109-111.
- [8] Karolchik D, Baertsch R, Diekhans M, et al. The UCSC genome browser database[J]. Nucleic Acids Res, 2003, 31(1): 51-54.
- [9] Hofacker I L. Vienna RNA secondary structure server[J]. Nucleic Acids Res, 2003, 31(13): 3429-3431.
- [10] Pruitt K D, Maglott T D R. RefSeq and LocusLink: NCBI gene-centered resources [J]. Nucleic Acids Res, 2001, 29(1): 137-140.
- [11] 阎平凡, 张长水. 人工神经网络与模拟进化计算[M]. 北京: 清华大学出版社, 2005.
- [12] Bonnet E, Wuyts J, Rouze P, et al. Evidence that micro-RNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences [J]. Bioinformatics, 2004, 20(17): 2911-2917.
- [13] Zhao Dongyu, Wang Yan, Luo Di, et al. PMirP: A pre-micro RNA prediction method based on structure-sequence hybrid features [J]. Artificial Intelligence in Medicine, 2010, 49(2): 127-132.
- [14] Hornik K. Multilayer Feedforward Networks are Universal Approximators[J]. Neural Networks, 1989(2): 359-366.
- [15] 谭 营. 计算智能导论[M]. 第 2 版. 北京: 清华大学出版社, 2010.
- [16] Kennedy J, Eberhart R. Particle swarm optimization[C]//Proceedings of the 1995 IEEE International Conference on Neural Networks. [s. l.]: [s. n.], 1995: 1942-1948.

(上接第 9 页)

矩作为特征量识别运动物体,然后提出了一种从图像坐标到空间坐标的映射方法,并结合文中选用的机械臂的结构特点,提出一种运动学逆解方法。最后,在室内简单背景的环境中,实现了机械臂对运动物体的跟踪,验证了本方法。

今后的工作主要在图像处理算法、编程架构上进行改进和提高,使系统实时性更高,提高跟踪速度和精度。

#### 参考文献:

- [1] Hutchinson S, Hager G D, Corke P I. A Tutorial on Visual Servo Control[J]. IEEE Transaction on Robotics and Automation (S1042-296X), 1996, 12(5): 641-670.
- [2] 廖万辉, 李 琳, 陈 祯. 基于嵌入式的智能视觉伺服系统的研发[J]. 机电产品开发和设计, 2009, 22(3): 3-5.
- [3] 刘 伟, 王建平, 张崇巍. 一种移动机器人对运动目标的检测跟踪方法[J]. 计算机技术与发展, 2009, 19(4): 105-108.
- [4] 陈 虹, 梁文彬. 基于机器人的神经网络预测控制算法[J]. 计算机技术与发展, 2008, 18(8): 65-68.
- [5] 魏长水, 贺巧龙, 李东亮. Mirobot 足球机器人决策系统的研究[J]. 计算机技术与发展, 2008, 18(5): 31-33.
- [6] Wilson W, Hulls C, Bell G. Relative end-effector control using Cartesian position-based visual servoing[J]. IEEE Transaction on Robotics and Automation, 1996, 12(10): 684-696.
- [7] Feddema F, Mitchell O. Vision-guided servoing with feature-based trajectory generation[J]. IEEE Transaction on Robotics and Automation, 1989, 5(10): 691-700.
- [8] 赵清杰, 连广宇, 李增圻. 机器人视觉伺服综述[J]. 控制与决策, 2008, 16(6): 849-852.
- [9] 陈友东, 王田苗, 魏洪兴. 工业机器人嵌入式控制系统的研究[J]. 机器人技术与应用, 2010(5): 10-13.
- [10] 刘 欢, 魏立峰, 王 健. 机器人视觉伺服系统的标定[J]. 机器人技术, 2007, 23(4): 278-279.
- [11] 王洪斌, 吕 玲, 李 萍. 在线估计雅可比矩阵的视觉伺服控制算法[J]. 系统仿真学报, 2010, 22(12): 2934-2937.
- [12] 曾福振, 闵联营. 基于 ARM 和 Linux 的嵌入式平台的搭建[J]. 微型机与应用, 2010, 30(12): 51-53.