技术方法

MiRscreen:一种基于遗传算法和支持向量机的microRNA前体识别方法

侯妍妍,李 华,应晓敏*,李伍举*
(军事医学科学院基础医学研究所计算生物学中心,北京 100850)

[摘要] 目的:构建具有高敏感性和高特异性的microRNA前体(pre-miRNA)识别模型。方法:根据300例经实验验证的人pre-miRNA和300例从3′UTR折成茎环结构的片段中随机选取的阴性样本,基于支持向量机方法构建了区分pre-miRNA和pseudo pre-miRNA的分类器MiRscreen。为提高分类器的性能,我们采用遗传算法搜索影响分类器性能的2个重要参数C和γ。结果与结论:该分类器对训练集的敏感性为99.33%,特异性为100%,对剩余的91例人pre-miRNA和91例3′UTR中的pseudo pre-miRNA敏感性和特异性分别达到91.21%(83/91)和93.41%(85/91)。在除人以外的其他20种动物和病毒的1 353例pre-miRNA中,MiRscreen正确判断出其中的1 192例,敏感性达到88.10%,其中马雷克病病毒、猕猴淋巴隐病毒、EB病毒、猿猴病毒40、非洲爪蟾、狗、绵羊和猕猴共计8个物种的敏感性达到100%;在随机抽取的100条RefSeq基因折叠形成的556例pseudo pre-miRNA和随机抽取的797例人19号染色体折叠形成的pseudo pre-miRNA(共计1 353例混合阴性样本)中,MiRscreen的特异性达到85.14%(1 152/1 353)。与其他6种同类方法相比,MiRscreen在敏感性和特异性方面均具有较好的性能,分类精度最高,达到86.62%,比其他方法高6%以上;MiRscreen的AUC值达到0.938,也明显高于其他方法。

[关键词] 微RNAs;识别;遗传算法;支持向量机
[中图分类号] Q75 [文献标识码] A
[文章编号] 1000-5501(2008)03-0287-06

MiRscreen: a prediction model for microRNA precursors using genetic algorithm and support vector machines

HOU Yan-Yan, LI Hua, YING Xiao-Min, LI Wu-Ju**
(Center of Computational Biology, Institute of Basic Medical Sciences, Academy of Military Medical Sciences, Beijing 100850, China)

[Abstract] Objective: To construct a prediction model for microRNA precursors (pre-miRNAs) with high sensitivity and high specificity. **Methods:** A prediction model, MiRscreen, for microRNA precursors using genetic algorithm and support vector machines was introduced. The training dataset contained 300 human experimentally validated pre-miRNAs as positive samples and 300 pseudo pre-miRNAs as negative samples. The negative samples were randomly selected from 3′ UTR stem-loops. To improve the performance of the classifier, genetic algorithm was employed to search for *C* and γ, which were two important parameters for SVM classifiers. **Results and conclusion:** The sensitivity and specificity for the training dataset were 99.33% and 100%, respectively. For the remaining 91 human pre-miRNAs and 91 pseudo pre-miRNAs from 3′UTR, the sensitivity and specificity were 91.21% (83/91) and 93.41% (85/91), respectively. The overall sensitivity of MiRscreen for 1 353 experimentally validated animal and virus (excluding human) pre-miRNAs was 88.10% (1 192/1 353), and the sensitivity for eight species was 100%, including Marek's disease virus, rhesus lymphocryptovirus, Epstein-Barr virus, simian virus 40, *Xenopus laevis*, *Canis familiaris*, *Ovis aries* and *Macaca mulatta*. The overall specificity for the 556 pseudo pre-miRNAs from 100 randomly selected RefSeq genes and 797 pseudo pre-miRNAs randomly selected from human chromosome 19 was 85.14% (1 152/1 353). Compared with the other six miRNA classification methods proposed previously, MiRscreen is remarkable in both sensitivity and specificity on the independent test dataset. The accuracy of MiRscreen is 86.62%, which is 6% higher than that of the other methods. The AUC of MiRscreen is 0.938, also

[收稿日期] 2008-01-23
[基金项目] 国家自然科学基金资助项目(30500105, 30470411)
[作者简介] 侯妍妍(1983-),女,湖南省常德市人,在读硕士研究生,研究方向为计算生物学。
*通讯联系人,李伍举(Tel: 010-66931324, E-mail: liwj@bmi.ac.cn);应晓敏(Tel: 010-66932301, E-mail: yingxm@bmi.ac.cn)

greater than the AUC of each of the other six methods. Therefore, the presented model MiRScreen can facilitate experimental identification of pre-miRNAs.

[Key words] microRNAs; classification; genetic algorithm; support vector machines

MicroRNAs (miRNAs)是近年来发现的一类长度为 ~22 nt的内源、单链的非编码小 RNA。目前的研究表明,miRNA基因由 RNA聚合酶Ⅱ[1,2]或Ⅲ[3]转录成初级转录物(pri-miRNA),而后经 Drosha酶剪切形成长度约为 70 nt的 miRNA前体(pre-miRNA)[4,5],在传输蛋白 exportin-5的作用下由细胞核内移到细胞质中[6,7],最后经 Dicer酶进一步切割产生成熟的 miRNA[8~10](可参照文献[11])。miRNA的显著特点是前体折叠形成茎环或类似茎环的二级结构。而且,通过对 pre-miRNA的基因组定位和注释发现,miRNA主要位于基因间区或已知转录本的内含子中[12],较大比例的 miRNA呈现成簇分布的特点,且在相近或多物种中保守[13]。miRNA参与广泛的调控通路,在生物体内发挥着重要的调控功能,如调控幼虫发育时序[14,15]、细胞增殖[16]、脂肪代谢[17]、造血系统分化[18]、生殖干细胞自我更新[19]、花的发育[20]等。自 1993年第一个 miRNA——*lin*-4发现以来,到目前为止已有 4 000多个 miRNA被陆续发现,它们广泛地存在于 55个物种中[21]。尽管有研究给出人、果蝇和线虫 miRNA的数量估计,分别不超过 255、110和 120个[22~24],然而,有证据表明 miRNA的数量远远超出这一估计[25~27],还有大量的 miRNA有待发现。

miRNA的发现主要有 cDNA克隆测序和计算预测两种方法。早期 miRNA的发现主要通过 cDNA克隆测序。这种方法直接、可靠,然而很难克隆出在不同时期表达或只在特定组织或细胞系中表达的 miRNA,而且由于克隆方法固有的局限性,也很难捕获表达丰度较低的 miRNA[28,29]。近年来,通过计算预测 miRNA的方法成为 miRNA发现的另一条重要途径,其优点是不受 miRNA表达的时间和组织特异性以及表达水平的影响,可以弥补 cDNA克隆测序方法的不足[30](有关 miRNA计算预测方法可参考文献[31])。基于机器学习方法预测 miRNA是近两年出现的一类新的预测方法,与其他预测方法最大的不同在于,基于机器学习的预测方法不仅需要已知的 miRNA,还需要已知的"非 miRNA",通过阳性和阴性数据集构建区分两者的分类器。这类方法的优点是可以找出与已知 miRNA同源和非同源、保守和非保守的 miRNA。机器学习方法的引入为大规模预测 miRNA提供了新的思路。由于基因组中存在大量可折叠形成茎环结构的序列片段[26],因此,构建同时具有高敏感性和高特异性的分类器成为基于机器学习方法预测 miRNA的关键。由于支持向量机(support vector machines, SVM)方法在逼近和泛化能力方面均具有良好的性能,因而,目前大多数 miRNA预测方法采用 SVM训练分类器[32~37],也有少数预测方法采用其他机器学习方法训练分类器,如随机森林(random forest)方法[38]、隐马尔可夫模型(hidden Markov model, HMM)[39]和 Naïve贝叶斯分类器(Naïve Bayes classifier)[40]。总的说来,目前基于机器学习识别 miRNA的方法或者敏感性较高而特异性不够好[32],或者为了提高特异性,增加阴性训练样本的数量,构建有偏的分类器,从而降低了敏感性[33,34,36,39]。

考虑到 SVM分类器的性能受核函数和相关参数的影响很大,而通常采用的参数搜索方法容易陷入局部最优,因而我们提出采用遗传算法(genetic algorithm, GA)搜索 SVM的相关参数,以构建无偏、且同时具有较高敏感性和特异性的分类器。本文基于人已知 pre-miRNA和 3′UTR中折叠形成茎环结构的片段,用 128个序列和结构特征描述样本,采用 SVM方法、以径向基函数(radial basis function, RBF)为核函数,构建分类器 MiRscreen。由于 SVM分类器的性能受惩罚参数 C和 RBF核参数 γ的影响很大,因此我们采用 GA搜索近优参数 C和 γ。结果表明,通过 GA搜索 C和 γ能够提高 SVM分类器的性能;与其他分类器相比,我们构建的分类器在具有较高敏感性的同时也具有较高的特异性。

1 材料与方法

通过机器学习方法构建的分类器性能主要取决于 3个因素:①阳性与阴性训练集;②描述样本的特征;③机器学习方法。阳性与阴性训练样本需要能够很好地代表相应的数据空间,而且两者的总和也应该能够很好地代表未知数据空间,这样构建出来的分类器对未知数据才能有效地分类;描述样本的特征应尽可能地反映阳性数据与阴性数据的差别;机器学习方法则需要既有很好的逼近能力,能对训练数据得到很好的分类性能,同时也要具有很好的泛化能力,对未知数据也能得到很好的分类效果。

1.1 训练数据和测试数据

阳性数据选取 miRBase 9.0[21]的 391条经实验验证的人类 pre-miRNA,随机抽取其中的 300条作为训练集,余下的 91条作为测试集。由于 3′UTR序列与 miRNA以及绝大部分基因间区一样,不编码蛋白,而且,已发现的 miRNA中只有少数几例位于 3′UTR区[40],因而我们选择人 3′UTR序列作为阴性数据的来源。3′UTR序列下载自 UTRdb版本 22[41],采用 RNAfold[42]折叠二级结构,满足以下 3个条件的茎环结构片段作为阴性数据集:①总长度 ≥55个核苷酸;②至少 18个配对碱基对;③环长度 ≥3个核苷酸。共计获得 83 437条阴性茎环结构片段(pseudo pre-miRNA)。我们从中随机抽取 300条和 91条序列分别作为阴性训练集和测试集。

此外,我们还采用了以下 3个数据集作为独立阳性和阴性测试集:①miRBase 9.0[21]中除人以外的 20种动物和病毒的 pre-miRNA共计 1 353条序列作为独立阳性测试集;②混合独立阴性测试集(combined independent negative test set, CN)。由于基因组中既包含编码基因的区域也包含基因间区,而且基因间区中绝大部分折成茎环结构的片段不是 pre-miRNA,因此,我们从人 RefSeq基因中随机抽取 100条序列,将其中折叠形成茎环结构并满足上述 3个条件的 556个片段作为阴性测试样本,同时从人 19号染色体正链和负链中折叠形成茎环结构并满足上述 3个条件的片段中随机抽取 797个序列作为阴性测试样本,得到共计 1 353个混合独立阴性测试样本。

1.2 特征

我们采用 85个序列特征和 43个结构特征描述整个样本,具体如下:

(1)一联、二联和三联碱基组成,共计 84个;

(2) GC含量;

(3)内部环和膨胀圈的个数,内部环 膨胀圈的个数,最大内部环 膨胀圈的大小,最小内部环 膨胀圈的大小,大小分别为 1～10 nt的内部环 膨胀圈的个数,大小 ≤5 nt的内部环 膨胀圈的个数,大小为 6～10 nt的内部环 膨胀圈的个数,大小 ≥11 nt的内部环 膨胀圈的个数,所有内部环 膨胀圈大小的总和,所有内部环和膨胀圈大小的总和,环的个数,最大环的大小,最小环的大小,配对数,最低自由能,序列长度,共计 42个特征;

(4)与 1 000条保持二联碱基成分的随机序列的最低自由能的随机检验 *P*值[43]。

其中,42个结构特征是采用 RNAfold[42]折叠序列后在最低自由能结构中提取的,*P*值采用 randfold程序[43]计算。

1.3 分类器的构建

SVM方法是在统计学理论的基础上发展起来的一种有监督功能的机器学习方法,具有很好的逼近和泛化能力。而且,SVM方法在 miRNA预测和分类上也取得了较好的效果[32～34,36,37]。因此,我们选用 SVM方法构建分类器,SVM采用 libSVM 2.83[44]实现。

GA是基于生物进化过程中优胜劣汰规则与群体内部染色体信息交换机制的一种自适应启发式全局搜索算法。由于它简单、通用、鲁棒性强、不依赖于问题模型,因而在函数优化、组合优化、机器学习等众多领域得到了广泛而且成功的应用。我们采用实值编码 GA(real-coded GA, RGA)对惩罚参数 *C*和 RBF核参数 γ进行搜索,RGA采用 A I::Genetic包实现。

分类器的构建包含以下 3个步骤:①将训练集中的样本采用 128个特征描述为特征向量,而后采用 libSVM中的 svm-scale将特征向量归一化至 [-1, +1]区间;②SVM的核函数采用 RBF函数,惩罚参数 *C*和 RBF核参数 γ采用 GA搜索近优参数,$\log_2 C$的搜索空间为 [-5, 15],$\log_2 \gamma$的搜索空间为 [3, -15],初始种群为 30,交叉概率为 0.8,变异概率为 0.01,个体的适应值为该个体对训练集的 5折交叉检验精度,连续三代种群的平均适应值的波动 <1%则终止进化;③采用 libSVM中的 svm-train,根据搜索得到的 *C*和 γ训练分类模型,使用"-b 1"参数,以计算每个样本的概率估计。

2 结果与讨论

2.1 分类器的性能

我们构建的分类器 MiRscreen对训练集的分类精度达到 99.67%,对测试集的敏感性为 91.21% (83/91),特异性为 93.41% (85/91),敏感性和特异性均超过 90%。对所有人 391个经实验验证的 pre-miRNA,MiRscreen的分类精度达到 97.44% (381/391)。

在除人以外的其他 20种动物和病毒的 1 353例 pre-miRNA中,MiRscreen正确判断出其中的 1 192例,敏感性达到 88.10%,其中马雷克病病毒 (8例)、猕猴淋巴隐病毒 (16例)、EB病毒 (23例)、猿猴病毒 40(1例)、非洲爪蟾 (7例)、狗 (6例)、绵羊 (4例)和恒河猴 (18例)共计 8个物种的敏感性达到 100%,人巨细胞病毒 (11例)、果蝇 (75例)、鸡 (90例)、大鼠 (161例)和牛 (98例)共计 4个物种的敏感性超过 90%。MiRscreen对各个物种的分类精度详见表 1。

为考察 MiRscreen的分类性能是否来自测试样本与训练样本的同源性,我们采用 BLASTCLUST[45]对除人以外的其他 20种动物和病毒中被正确预测的 1 192例 pre-miRNA与训练集中的人 300例 pre-miRNA进行同源性聚类。我们发现,即使采用很宽松的同源参数 (S = 80, L = 0.5, W = 8),也只有 403例动物 pre-miRNA与人 pre-miRNA同源,其他 789例动物和病毒 pre-miRNA均不与人 pre-miRNA同源。病毒 pre-miRNA不仅不与人 pre-miRNA同源,而且也不与任何其他动物 pre-miRNA同源。这一结果说明,MiRscreen良好的分类性能并非来自 pre-miRNA的序列同源性,同时说明 MiRscreen采用的特征能够较好地区分 pre-miRNA与非 pre-miRNA。

在对混合独立阴性测试集的测试中,MiRscreen将其中的 1 152个样本判断为阴性,分类精度为 85.14% (1 152/1 353),具有较高的特异性。

2.2 与其他分类器的比较

我们将分类器 MiRscreen与其他基于机器学习方法识别保守和非保守 miRNA的分类器进行了比较,见表 1。从表 1中可以看出,在除人以外的 20个物种的 pre-miRNA中,MiRscreen在 13个物种中的敏感性最高,MiPred居次,在 11个物种中敏感性最高,其他 5个分类器则只在 0～2个物种中敏感性最高。MiRscreen和 MiPred[38] (即 MiPred_RF)对 1 353个 pre-miRNA的总体敏感性最高,均为 90%左右,两者均远高于其他 5个分类器。然而值得注意的是,MiPred在混合独立阴性测试集中的特异性是 7个分类器中最低的,仅为 45.38%,远远低于 MiRscreen和其他分类器。

从表 1中还可以看出,在对混合独立阴性测试集的测试中,ProMiR[39]和 miRabela[34]的特异性最高,均超过 90%,MiRscreen的特异性居第三,达到 85%,远高于其他 4个分类器。然而 ProMiR和 miRabela在 1 353个其他物种的 pre-miRNA中的敏感性很低,仅为 65%和 71%,远低于 MiRscreen。

综合而言,在对其他物种的 pre-miRNA和混合独立阴性测试集的测试中,尽管 MiRscreen的敏感性和特异性均不是最高的,但 MiRscreen同时具有较高的敏感性和特异性,分类精度最高,达到 86.62%,远高于 3SVM (77.61%)、miPred (78.49%)、ProMiR (80.56%)、MiPred(即 MiPred_RF) (67.96%)、BayesMiRNAfind (74.28%)和 miRabela (80.78%)。我们还绘制了 7个分类器在独立阳性和阴性测试集中的 ROC曲线,如图 1所示。MiRscreen的 AUC值 (area under the ROC)为 0.938,高于 3SVM (0.842)、miPred (0.873)、ProMiR (0.783)、MiPred_RF (0.811)、BayesMiRNAfind (0.782)和 miRabela (0.847)。

表 1 MiRscreen与其他 6个分类器对动物和病毒 pre-miRNA以及混合独立阴性测试集的分类精度

物种 (个数)	分类精度 (%)						
	MiRscreen	3SVM	miPred	ProMiR	MiPred_RF	BayesMiRNAfind	miR-abela
疱疹病毒科 α疱疹病毒亚科							
单纯疱疹病毒 1型 (1)	0	0	0	0	**100.00**	**100.00**	0
马雷克病病毒 (8)	**100.00**	75.00	62.5	50.00	62.50	50.00	87.50
疱疹病毒科 β疱疹病毒亚科							
人巨细胞病毒 (11)	90.91	63.64	81.82	63.64	**100.00**	81.82	45.45
疱疹病毒科 γ疱疹病毒亚科							
小鼠 γ疱疹病毒 68 (9)	**88.89**	**88.89**	77.78	55.56	**88.89**	44.44	22.22
猕猴淋巴隐病毒 (16)	**100.00**	81.25	93.75	56.25	**100.00**	93.75	68.75
EB病毒 (23)	**100.00**	91.30	95.65	52.17	**100.00**	82.61	78.26
卡波济肉瘤相关疱疹病毒 (13)	69.23	69.23	84.62	30.77	**92.31**	61.54	38.46
多瘤病毒科							
猿猴病毒 40 (1)	**100.00**	**100.00**	**100.00**	0	**100.00**	**100.00**	**100.00**
节肢动物门							
黑腹果蝇 (75)	**93.33**	82.67	84.00	56.00	90.67	68.00	66.67
线虫动物门							
秀丽线虫 (112)	89.29	82.14	75.89	53.57	**90.18**	58.04	67.86
扁形动物门							
涡虫 (63)	69.84	77.78	90.48	57.14	**96.83**	65.08	42.86
脊椎动物门鱼纲							
斑马鱼 (322)	87.27	81.37	88.51	58.07	**90.37**	62.11	78.57
脊椎动物门两栖纲							
非洲爪蟾 (7)	**100.00**	85.71	85.71	71.43	85.71	71.43	57.14
脊椎动物门鸟纲							
鸡 (90)	**94.44**	83.33	90.00	77.78	91.11	78.89	80.00
脊椎动物门哺乳纲							
狗 (6)	**100.00**	83.33	83.33	50.00	**100.00**	66.67	83.33
小鼠 (315)	81.27	78.73	79.05	60.95	**88.25**	73.02	62.22
大鼠 (161)	**95.65**	85.71	84.47	68.32	91.30	79.50	80.75
牛 (98)	**93.88**	86.73	80.61	68.37	91.84	74.49	75.51
绵羊 (4)	**100.00**	50.00	75.00	25.00	50.00	50.00	75.00
恒河猴 (18)	**100.00**	77.78	94.44	33.33	88.89	77.78	83.33
总数 (1 353)	88.10	84.41	83.96	65.11	90.54	69.84	70.51
混合独立阴性测试集 (1 353)	85.14	70.81	73.02	96.01	45.38	78.71	91.06
人 (391)	**97.44**	82.10	83.89	65.73	89.77	65.73	71.61

表中所示为 MiRscreen与 3SVM[32]、miPred[33]、ProMiR[39]、MiPred[38] (与 miPred区分,本文称为 MiPred_RF)、BayesMiRNAfind[40]、miR-abela[34]共 7种分类器在所有动物和病毒 pre-miRNA以及混合独立性测试集 CN中的分类精度;物种名称后面括号中的数字表示 pre-miRNA的个数;加粗显示的值是分类器在各物种中最高的分类精度

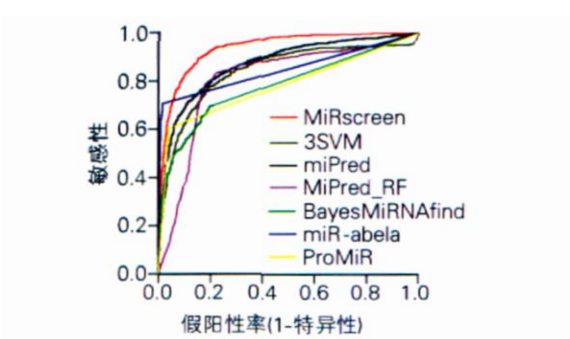

图 1 MiRscreen与其他 6个分类器在独立阳性和阴性测试集上的 ROC曲线

横坐标为假阳性率 (1特异性),纵坐标为真阳性率 (敏感性)。红色、橄榄绿、黑色、紫红、绿色、蓝色和黄色线分别为 MiRscreen、3SVM、miPred、MiPred (即 MiPred_RF)、BayesMiRNAfind、miR-abela和 ProMiR的 ROC曲线

2.3 通过 GA搜索 SVM相关参数以提高分类器的性能

对于采用 RBF核函数的 SVM而言,惩罚参数 C和 RBF核参数 γ是影响分类器性能的重要参数。不好的参数会导致 SVM无法搜索到最优分类超平面,从而大大降低分类器的性能。最优 C和 γ的确定实质是组合问题,即根据经验确定 C和 γ的搜索范围之后,在相应范围中找出一对 C和 γ使得分类器具有最优的性能。枚举所有的组合固然能够找出最优的 C和 γ,然而却要耗费大量的计算时间。目前的解决方法是先通过粗糙网格搜索确定一个更好的区域,而后在更好的区域中精细搜索,找出该区域中最好的 C和 γ组合[44]。这种搜索方法的缺点是容易在搜索的过程中陷入局部最优。为解决这一问题,我们采用 GA搜索 C和 γ。由于 GA是在多个个体中并行搜索,每个个体都通过交叉与突变生成新的个体,从而使整个进化过程能够覆盖尽可能大的组合空间,不易陷入局部最优。表 2是采用 GA和网格搜索方法 (步长分别为 1和 2)对同一组训练集搜索 C和 γ的比较,

并对同一组测试集进行测试。其中 $\log_2 C$的搜索空间为 [-5, 15], $\log_2 \gamma$的搜索空间为 [3,-15], GA的初始种群为 30,交叉概率为 0.8,变异概率为 0.01,个体的适应值为该个体对训练集的 5折交叉检验精度,连续三代种群的平均适应值的波动 <1%则终止进化。

表 2　GA与网格搜索方法搜索 C和 γ的比较

方法	C	γ	CPU时间 (t/s)	交叉检验精度 (%)	测试集 (阳性和阴性样本各 91例)		
					敏感性 (%)	特异性 (%)	精度 (%)
Grid-search 步长 =2	8	0.03125	101.50	88.6667	80.22	94.51	87.3626
Grid-search 步长 =1	8	0.00390625	367.63	88.8333	85.71	91.21	88.4615
GA	16 384	0.00048828125	322.39	88.6667	91.21	93.41	92.3077

从表 2可以看出,采用步长为 2的网格搜索方法搜索时间最短,但采用其搜索得到的 C和 γ构建的 SVM分类器对测试集的分类精度仅为 87.3626%,特异性较高但敏感性较低,也就是说找到的分类超平面倾向于将样本判断为阴性;采用步长为 1的网格搜索方法搜索时间最长,为步长为 2的网格搜索方法的 3.6倍,然而采用其搜索得到的 C和 γ构建的 SVM分类器对测试集的分类精度为 88.4615%,仅提高了 1.1%。通过 GA搜索的计算时间较步长为 1的网格搜索方法短,然而其搜索到的 C和 γ使分类器的分类精度达到 92.3077%,较步长为 1的网格搜索方法提高了近 4%,较步长为 2的网格搜索方法提高了近 5%。由此可见,通过 GA搜索 C和 γ能够提高 SVM分类器的性能,且计算时间并没有显著增加。图 2是 GA进化过程中各代个体的平均适应度 (即所有个体的平均 5折交叉检验精度)曲线,从图中可以看出, GA从第 1代到第 3代平均适应度迅速提高,在第 4和第 5代缓慢提高,经第 6代小幅波动后,从第 7代开始收敛。

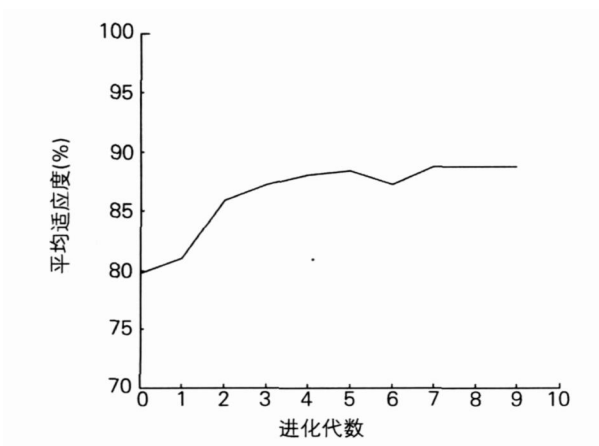

图 2　GA进化过程中各代个体的平均适应度曲线

3　结论

由于基因组中存在大量与 miRNA结构类似的片段,因而从基因组中识别出 miRNA非常具有挑战性。已有的基于机器学习的 miRNA识别方法或者敏感性高而特异性不好,或者特异性高而敏感性又不尽如人意。为构建同时具有高敏感性和高特异性的分类器,本文用 128个序列和结构特征描述序列、采用 SVM训练分类器。为提高分类器的性能,我们采用 GA搜索影响分类器性能的 2个重要参数 C和 γ。我们构建的分类器在分类精度方面超过目前已有的分类器。

当然,以目前分类器的特异性在全基因组中预测 miRNA依然会产生相当数量的假阳性。进一步的工作需要探索新的区分 miRNA与非 miRNA的特征,或者采用新的方法构建分类器,以提高特异性,同时尽可能地减少敏感性的损失。

致谢:本工作得到军事医学科学院生物医学超级计算中心的支持与帮助。

[参考文献]

[1] Cai X, Hagedorn CH, Cullen BR. Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs[J]. RNA, 2004, 10(12): 1957-1966.

[2] Lee Y, Kim M, Han J, *et al*. MicroRNA genes are transcribed by RNA polymerase II [J]. EMBO J, 2004, 23(20): 4051-4060.

[3] Borchert GM, Lanier W, Davidson BL. RNA polymerase III transcribes human microRNAs[J]. Nat Struct Mol Biol, 2006, 13(12): 1097-1101.

[4] Lee Y, Ahn C, Han J, *et al*. The nuclear RNase III Drosha initiates microRNA processing[J]. Nature, 2003, 425(6956): 415-419.

[5] Zeng Y, Yi R, Cullen BR. Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha [J]. EMBO J, 2005, 24(1): 138-148.

[6] Yi R, Qin Y, Macara IG, *et al*. Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs[J]. Genes Dev, 2003, 17(24): 3011-3016.

[7] Bohnsack MT, Czaplinski K, Gorlich D. Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs[J]. RNA, 2004, 10(2): 185-191.

[8] Ketting RF, Fischer SE, Bernstein E, *et al*. Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in *C. elegans*[J]. Genes Dev, 2001, 15(20): 2654-2659.

[9] Jiang F, Ye X, Liu X, *et al*. Dicer-1 and R3D1-L catalyze microRNA maturation in *Drosophila*[J]. Genes Dev, 2005, 19(14): 1674-1679.

[10] Lee YS, Nakahara K, Pham JW, *et al*. Distinct roles for *Dro-*

sophila Dicer-1 and Dicer-2 in the siRNA/miRNA silencing pathways[J]. Cell, 2004, 117(1): 69 - 81.

[11] Kim VN. MicroRNA biogenesis: coordinated cropping and dicing [J]. Nat Rev Mol Cell Biol, 2005, 6(5): 376 - 385.

[12] Rodriguez A, Griffiths-Jones S, Ashurst JL, *et al* Identification of mammalian microRNA host genes and transcription units[J]. Genome Res, 2004, 14(10a): 1902 - 1910.

[13] Altuvia Y, Landgraf P, Lithwick G, *et al* Clustering and conservation patterns of human microRNAs[J]. Nucleic Acids Res, 2005, 33(8): 2697 - 2706.

[14] Moss EG, Lee RC, Ambros V. The cold shock domain protein LIN-28 controls developmental timing in *C. elegans* and is regulated by the lin-4 RNA[J]. Cell, 1997, 88(5): 637 - 646.

[15] Reinhart BJ, Slack FJ, Basson M, *et al* The 21 nucleotide *let*-7 RNA regulates developmental timing in *Caenorhabditis elegans* [J]. Nature, 2000, 403(6772): 901 - 906.

[16] Brennecke J, Hipfner DR, Stark A, *et al* *Bantam* encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene *hid* in *Drosophila*[J]. Cell, 2003, 113(1): 25 - 36.

[17] Xu P, Vernooy SY, Guo M, *et al* The *Drosophila* MicroRNA Mir-14 suppresses cell death and is required for normal fat metabolism[J]. Curr Biol, 2003, 13(9): 790 - 795.

[18] Chen CZ, Li L, Lodish HF, *et al* MicroRNAs modulate hematopoietic lineage differentiation[J]. Science, 2004, 303(5654): 83 - 86.

[19] Park JK, Liu X, Strauss TJ, *et al* The miRNA pathway intrinsically controls self-renewal of *Drosophila* germline stem cells[J]. Curr Biol, 2007, 17(6): 533 - 538.

[20] Chen X. A MicroRNA as a translational repressor of APETALA2 in *Arabidopsis* flower development [J]. Science, 2004, 303(5666): 2022 - 2025.

[21] Griffiths-Jones S, Grocock RJ, van Dongen S, *et al* miRBase: microRNA sequences, targets and gene nomenclature[J]. Nucl Acids Res, 2006, 34(Suppl 1): D140 - D144.

[22] Lim LP, Glasner ME, Yekta S, *et al* Vertebrate microRNA genes[J]. Science, 2003, 299(5612): 1540 - 1540.

[23] Lai EC, Tomancak P, Williams RW, *et al* Computational identification of *Drosophila* microRNA genes[J]. Genome Biol, 2003, 4(7): R42.

[24] Lim LP, Lau NC, Weinstein EG, *et al* The microRNAs of *Caenorhabditis elegans*[J]. Genes Dev, 2003, 17(8): 991 - 1008.

[25] Berezikov E, van Tetering G, Verheul M, *et al* Many novel mammalian microRNA candidates identified by extensive cloning and RAKE analysis[J]. Genome Res, 2006, 16(10): 1289 - 1298.

[26] Bentwich I, Avniel A, Karov Y, *et al* Identification of hundreds of conserved and nonconserved human microRNAs[J]. Nat Genet, 2005, 37(7): 766 - 770.

[27] Berezikov E, Guryev V, van de BJ, *et al* Phylogenetic shadowing and computational identification of human microRNA genes[J]. Cell, 2005, 120(1): 21 - 24.

[28] Berezikov E, Cuppen E, Plasterk RHA. Approaches to microRNA discovery[J]. Nat Genet, 2006, 38(Suppl): S2 - S7.

[29] Bentwich I Prediction and validation of microRNAs and their targets[J]. FEBS Lett, 2005, 579(26): 5904 - 5910.

[30] Kim VN, Nam JW. Genomics of microRNA[J]. Trends Genet, 2006, 22(3): 165 - 173.

[31] 侯妍妍, 应晓敏, 李伍举. MicroRNA计算发现方法研究进展[J]. 遗传, 2008, 30(6): 687 - 696.

[32] Xue C, 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.

[33] Ng KLS, Mishra SK. *De novo* SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures[J]. Bioinformatics, 2007, 23(11): 1321 - 1330.

[34] Sewer A, Paul N, Landgraf P, *et al* Identification of clustered microRNAs using an *ab initio* prediction method[J]. BMC Bioinformatics, 2005, 6(1): 267.

[35] Pfeffer S, Sewer A, Lagos-Quintana M, *et al* Identification of microRNAs of the Herpesvirus family[J]. Nat Meth, 2005, 2(4): 269 - 276.

[36] Hertel J, Stadler PF. Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data[J]. Bioinformatics, 2006, 22(14): e197 - e202.

[37] Helvik SA, Snove O Jr, Saetrom P. Reliable prediction of Drosha processing sites improves microRNA gene prediction[J]. Bioinformatics, 2007, 23(2): 142 - 149.

[38] Jiang P, Wu H, Wang W, *et al* MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features[J]. Nucl Acids Res, 2007, 35(Web Server issue): W339 - W344.

[39] Nam JW, Shin KR, Han J, *et al* Human microRNA prediction through a probabilistic co-learning model of sequence and structure[J]. Nucleic Acids Res, 2005, 33(11): 3570 - 3581.

[40] Yousef M, Nebozhyn M, Shatkay H, *et al* Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier[J]. Bioinformatics, 2006, 22(11): 1325 - 1334.

[41] Mignone F, Grillo G, Licciulli F, *et al* UTRdb and UTRsite: a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs[J]. Nucleic Acids Res, 2005, 33(Database issue): D141 - D146.

[42] Hofacker IL, Fontana W, Stadler PF, *et al* Fast folding and comparison of RNA secondary structures[J]. Monatsh Chem, 1994, 125(2): 167 - 188.

[43] Bonnet E, Wuyts J, Rouze P, *et al* Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences[J]. Bioinformatics, 2004, 20(17): 2911 - 2917.

[44] Chang CC, Lin CJ. LIBSVM: a library for support vector machine [R]. 2001.

[45] Altschul SF, Gish W, Miller W, *et al* Basic local alignment search tool[J]. J Mol Biol, 1990, 215(3): 403 - 410.

(本文编辑 孙承媛)