

Support vector machine for prediction of siRNA silencing efficacy

Wu Jiansheng Hu Minjing Zhou Tong Weng Jianhong Jiang Peng Sun Xiao

(State Key Laboratory of Bioelectronics, Southeast University, Nanjing 210096, China)

Abstract: In order to assist the design of short interfering ribonucleic acids (siRNA), 573 non-redundant siRNAs were collected from published literatures and the relationship between siRNAs sequences and RNA interference (RNAi) effect is analyzed by a support vector machine (SVM) based algorithm relied on a base-base correlation (BBC) feature. The results show that the proposed algorithm has the highest area under curve (AUC) value (0.73) of the receive operating characteristic (ROC) curve and the greatest r value (0.43) of the Pearson's correlation coefficient. This indicates that the proposed algorithm is better than the published algorithms on the collected datasets and that more attention should be paid to the base-base correlation information in future siRNA design.

Key words: short interfering ribonucleic acid (siRNA); support vector machine; base-base correlation; receive operating characteristic (ROC) curve

Short interfering RNAs (siRNAs) reduce gene expression through a process called RNA interference (RNAi), which results in sequence-specific degradation of mRNA by introducing double-stranded RNA (dsRNA) into cells^[1]. siRNAs suppress the expression of their target genes by four major steps: incorporation of an siRNA strand with an RNA induced silencing complex (RISC), activation of RICS, target mRNA recognition and target mRNA cleavage^[2–3].

Published papers have indicated that differing efficacy for functional siRNA suppresses the expression of target genes and their functionally-correlated features including mainly: ① duplex stability, ② sequence characteristics, ③ mRNA secondary structure, and ④ target site uniqueness, etc^[4]. Many algorithms that predict siRNA silencing efficacy have been published recently^[5–8]. These algorithms scored siRNAs based on the above characteristics, and all the siRNAs were classified as either effective or ineffective, typically by using a cutoff on the measured siRNA efficacy score. But small scale samples of the algorithms indicate that over-fitting is unavoidable and these algorithms may be not so robust in a larger dataset, and, moreover, the algorithms primarily focus on position conversation without considering position correlation. The support vector machine (SVM) is one of the best machine learning methods, especially in the classification of small scale

data sets^[9]. In this paper, in order to increase the probability for obtaining a siRNA which induces effective silencing of the desired gene, we perform a statistical analysis of a non-redundant dataset of functionally validated siRNAs collected from publications by SVM combined with a base-base correlation (BBC)^[10] feature.

1 Materials and Methods

1.1 Data set

In this paper, we mainly collected experimental data on the efficacy of siRNAs to silence their corresponding target gene expression from Refs. [5 – 11] including 573 non-redundant antisense-stranded siRNAs with the length of 19 bp (lack of 2 nt overhangs at the 5' and 3' terminal). The knockdown efficiencies and the corresponding target genes of 46 of them were collected from Ref. [5], 44 from Ref. [6], 108 from Ref. [7], 14 from Ref. [8], 240 from Ref. [9], 76 from Ref. [10], and 53 from Ref. [11] (8 of them are repeated, so the factual total is 573). To compare with the results of published literature^[5–8], class labels are also assigned based on the siRNAs knockdown efficiency: siRNAs that knockdown their target genes expression by 80% or more are classified as an effective group while the rest are assigned to an ineffective group. This assignment results in 194 effective and 379 ineffective siRNAs.

1.2 Feature of BBC

We have proposed a feature called BBC^[10] that reflects the base-base correlation.

Received 2006-04-28.

Foundation item: The National Natural Science Foundation of China (No. 60671018, 60121101).

Biographies: Wu Jiansheng (1979—), male, graduate; Sun Xiao (corresponding author), male, doctor, professor, xsun@seu.edu.cn.

$$T_{ij}(k) = \sum_{l=1}^k p_{ij}(l) \cdot \log_2 \left(\frac{p_{ij}(l)}{p_i p_j} \right) \quad i, j \in \{1, 2, 3, 4\} \quad (1)$$

where p_i and p_j mean the frequency of appearance of single bases i and j ; $p_{ij}(l)$ means the joint probability of the bases i and j at a distance of l ; $T_{ij}(k)$ represents the average relevance of the two-base combination with different gaps from 1 to k , it reflects a local feature of two bases with an interval of k . The BBC feature forms a 16-dimensional vector.

1.3 Support vector machine

The SVM, introduced by Vapnik^[9], is a learning algorithm for two- or multi-class classification problems and is known for its good performance.

The basic principle of SVM is: For a given data set $x_i \in \mathbf{R}^n$ ($i = 1, 2, \dots, N$) with corresponding labels y_i ($y_i = +1$ or -1 , representing the two classes to be classified), the SVM gives a decision function (classifier),

$$f(x) = \text{sgn} \left(\sum_{i=1}^N y_i \alpha_i K(x, x_i) + b \right) \quad (2)$$

where α_i are the coefficients to be learned, and K is a kernel function. Parameters α_i are trained by maximizing function,

$$\sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i,j=1}^N \alpha_i \alpha_j y_i y_j K(x_i, x_j) \quad (3)$$

where subject to $0 \leq \alpha_i \leq C$ ($i = 1, 2, \dots, N$) and $\sum_{i=1}^N \alpha_i y_i = 0$.

For this study, our method for automatical prediction of siRNA silencing efficacy is a two-stage process. The first stage is that the 16-dimensional BBC feature vectors of 573 siRNA sequences are calculated as the input of the SVM classifier. The second stage is that SVMlight version 6.01 (<http://svmlight.joachims.org/>) is used for data training and classifying. To ensure that the parameter estimation and model generation of the SVM are completely independent of the test data, the original data set was divided into two parts: set A and set B . Set A was used as a separate validation set to optimize the parameters of SVM, and the remainder sequences were put into set B for performance evaluation. The SVM models were trained by all the effective siRNAs with positive labels and all the ineffective siRNAs with negative labels. Ten-fold cross-validation was used for both parameter and accuracy estimation that is to divide the dataset of effective and ineffective siRNAs randomly

into ten subsets and then alternately use one subset for testing and the other nine sets for training. We used different kernel functions in our experiments, including the linear function, the polynomial function and the radial basis function. The best result was obtained by using the polynomial function kernel with $d = 7$.

1.4 Measurement of algorithms performance

As scoring was performed in published references^[5-8], the algorithms performance in distinguishing effective siRNAs from ineffective siRNAs is evaluated by the receive operating characteristic curve (ROC curve). The ROC curve is a useful technique for organizing classifiers and visualizing their performance^[11]. ROC graphs are two-dimensional graphs in which true positive rate (T) is plotted on the y axis and false positive rate (F) is plotted on the x axis. The true positive rate and the false positive rate of a siRNA sequence are defined by

$$T = \frac{\alpha}{\alpha + \delta} \quad (4)$$

$$F = \frac{\gamma}{\beta + \gamma} \quad (5)$$

where α , β , γ and δ are the number of true positives, true negatives, false positives and false negatives, respectively. Random guessing would generate identical false positive and true positive rates on average. Therefore, the diagonal ($y = x$) in the ROC plot is the performance of random guessing. The ROC curves move towards the upper left corner, indicating rising accuracy of performance. The area under the ROC curve can be used to characterize the performance of a classifier for siRNA sequences. The AUC (area under curve) value ranges from 0.5 to 1, and a rising AUC value indicates higher accuracy of performance.

The Pearson's correlation coefficient (r value) based on the classifiers' output values and the siRNAs knockdown efficiencies are also applied to evaluate algorithms performance. The higher r value indicates the better algorithms performance^[12].

2 Results and Discussion

We performed the classification of effective and ineffective siRNAs, using SVM combined with the BBC feature. Ten-fold cross-validation was used for obtaining the corresponding output value of each siRNA sequence and ROC graph was plotted according to these output values and siRNA knockdown efficacy (see Fig. 1).

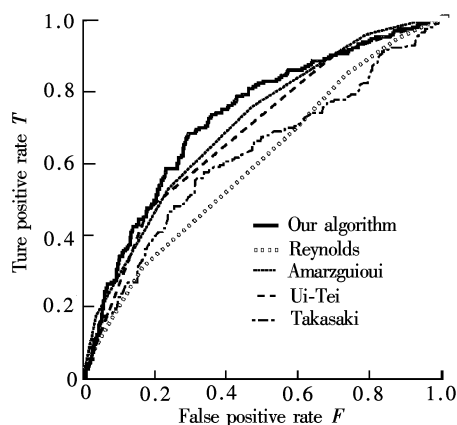


Fig. 1 ROC graphs for the Takasaki, Ui-Tei, Amarzguioui and Reynolds classifiers and our SVM based classifier on our data sets

In this paper, to compare with the results of published references, we also took the four scoring algorithms performed on our data set. Tab. 1 shows how

Tab. 1 Sequence characteristics used by the four published scoring algorithms

Position	Base	Algorithms			
		Reynolds	Ui-Tei	Amarzguioui	Takasaki
1	A		-1	-1	-3.97
	C		1	1	
	G		1	1	7.40
	U		-1	-2	-3.75
2	A			-1	
	U			-1	
3	A	1		-1	
	U			-1	
6	A			1	2.33
	C				
7	G				2.40
	U				-2.59
8	A				3.02
	G				-2.35
9	G				-2.35
	U				2.3
10	U	1			
11	C				
	G				
13	A				
	G	-1			
15	A	1			
	U	1			2.7
16	A	1			
	U	1			
17	A	1		1	
	U	1		1	
18	A	1		1	
	U	1		1	
19	A	2	1	2	
	C	-1	-1		
	G	-1	-1	-1	-2
	U	1	1	2	

various algorithms scored a siRNA based on individual nucleotides and indicates the position conversation of siRNA sequences. 1 to 19 in the first column corresponds to the 19 positions of siRNA sense strands (kicking off two nucleotides of the 5' and 3' ends caps). Scores in Tab. 1 describe base preferences at certain positions of siRNA sequences by the Amarzguioui algorithm^[5], the Reynolds algorithm^[6], the Ui-Tei algorithm^[7] and the Takasaki algorithm^[8]. More details to be illustrated are to add one score by the Reynolds algorithm and the Amarzguioui algorithm if the GC abundance of siRNA sequence lies between 30% and 52% and to reduce one score performed in the Ui-Tei classifier if any long GC stretch of more than 9 bp in length occurred in targeted siRNA sequence. The four upper scoring systems refer to sense strand, and to anti-sense strand where complementary sequence should be taken. We plot the ROC graph according to the output values by the four scoring algorithms and siRNA knockdown efficacy (see Fig. 1). Fig. 1 shows that our classifier has a higher true positive rate than the other classifiers for all false positive rate levels; that is to say, we can predict more effective siRNAs in the same false positive rate levels than that of the other algorithms.

We also calculated the classifiers' AUC value and Pearson's correlation coefficient between algorithm output and siRNA efficacy. This resulted in *r* value of 0.43, 0.26, 0.41, 0.42 and 0.30, and AUC value of 0.73, 0.65, 0.69, 0.71, 0.60 for our SVM based algorithm, the Reynolds algorithm, the Ui-Tei algorithm, the Amarzguioui algorithm and the Takasaki algorithm. The algorithm output has significant correlation with siRNA efficacy for all the five classifiers (*p* < 0.01, see Tab. 2). It is obvious that our classifier has the greatest AUC value and *r* value, and this indicates that our classifier is better than the other classifiers based on scoring.

Tab. 2 Comparison of algorithms performed on the collected datasets

Algorithms	AUC value	<i>r</i> value
Our algorithm	0.73	0.43 *
Takasaki	0.62	0.26 *
Ui-Tei	0.69	0.41 *
Amarzguioui	0.71	0.42 *
Reynolds	0.60	0.30 *

Note: * means *p* < 0.01, significant.

We believe that our algorithm has a higher performance because ① The SVM classifier performs better in the classification of small scale data sets only with several hundreds of samples than the scoring sys-

tems; ② Takasaki, Ui-Tei, Amarzguioui and Reynolds algorithms actually have high and stable performance on their own datasets^[5-8], whereas their small scale samples indicate that over-fitting is a problem with these algorithms that maybe are not very robust in a large dataset; and ③ the published methods have paid more attention to the information of position conversation but have not allowed for the correlation between different positions. The results suggest that the BBC feature we have used possibly captures more complex characteristics of effective siRNAs than those of the other algorithms, and position correlation maybe has a greater influence on siRNA efficacy, especially on target mRNA recognition, but this fact has not been proven and should still be further investigated. The RNAi field is developing rapidly, and we believe that new siRNA efficacy prediction algorithms will rise partly due to larger and better datasets.

3 Conclusion

In summary, our SVM based algorithm combined with a BBC feature that pays more attention to the correlation between different positions performs better than the other scoring algorithms in the classification of siRNA sequences. It can be explained that the BBC feature used in our algorithm captures more complex characteristics of effective siRNAs and that the SVM model has better performance in the classification of small scale samples.

References

[1] Fire A, Xu S, Montgomery M K, et al. Potent and specific

genetic interference by double-stranded RNA in *Caenorhabditis elegans* [J]. *Nature*, 1998, **391** (6669): 806 – 811.

- [2] Bernstein E, Caudy A A, Hammond S M, et al. Role for a bidentate ribonuclease in the initiation step of RNA interference [J]. *Nature*, 2001, **409** (6818): 363 – 366.
- [3] Nykanen A, Haley B, Zamore P. ATP requirements and small interfering RNA structure in the RNA interference pathway [J]. *Cell*, 2001, **107** (3): 309 – 321.
- [4] Saetrom P, Snove O Jr. A comparison of siRNA efficacy predictors [J]. *Biochem Biophys Res Commun*, 2004, **321** (1): 247 – 253.
- [5] Amarzguioui M, Prydz H. An algorithm for selection of functional siRNA sequences [J]. *Biochem Biophys Res Commun*, 2004, **316** (4): 1050 – 1058.
- [6] Reynolds A, Leake D, Boese Q, et al. Rational siRNA design for RNA interference [J]. *Nature Biotech*, 2004, **22** (3): 326 – 330.
- [7] Ui-Tei K, Naito Y, Takahashi F, et al. Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference [J]. *Nucleic Acids Res*, 2004, **32** (3): 936 – 948.
- [8] Takasaki S, Kotani S, Konagaya A. An effective method for selecting siRNA target sequences in mammalian cells [J]. *Cell Cycle*, 2004, **3** (6): 790 – 795.
- [9] Vapnik V N. *The nature of statistical learning theory* [M]. Springer, 1995.
- [10] Liu Z H, Jiao D, Sun X. Classifying genomics sequences by sequence feature analysis [J]. *Genomics Proteomics Bioinformatics*, 2005, **3** (4): 201 – 205.
- [11] Egan J P. *Signal detection theory and ROC analysis* [M]. New York: Academic Press, 1975.
- [12] Sætrom P, Snöve O J. A comparison of siRNA efficacy predictors [J]. *Biochem Biophys Res Commun*, 2004, **321** (1): 247 – 253.

基于支持向量机的 siRNA 降解效率预测

吴建盛 胡敏菁 周 童 翁建洪 江 澎 孙 啸

(东南大学生物电子学国家重点实验室, 南京 210096)

摘要: 为了辅助 siRNA 的设计, 从已发表文献中共收集到 573 个 siRNA 的实验数据, 使用基于统计学习理论的支持向量机(SVM)方法, 提取了 siRNA 序列的碱基对关联性(BBC)特征, 然后使用十倍交叉验证方法, 对 siRNA 沉默目标基因的效率进行了预测。结果表明, 基于支持向量机, 选用多项式核作为核函数的算法具有最高的 AUC 值(0.73, ROC 曲线图)和最高的 r 值(0.43, Pearson 相关系数分析), 优于以前基于打分的算法。结果说明, 在以后的 siRNA 的设计中应该更多关注碱基之间的关联信息。

关键词: siRNA; 支持向量机; 碱基对关联性; ROC 曲线
中图分类号: Q52