BMC Bioinformatics

Methodology article

Selection of antisense oligonucleotides based on multiple predicted target mRNA structures

Xiaochen Bo†, Shaoke Lou†, Daochun Sun, Wenjie Shu, Jing Yang and Shengqi Wang*

Address: Beijing Institute of Radiation Medicine, 27 Taiping Road, Beijing 100850, P R China

Email: Xiaochen Bo - boxc@bmi.ac.cn; Shaoke Lou - lousk@163.com; Daochun Sun - sdcwin@163.com; Wenjie Shu - shuwj@bmi.ac.cn; Jing Yang - jingyang0511@sina.com; Shengqi Wang* - sqwang@bmi.ac.cn

* Corresponding author †Equal contributors

Published: 09 March 2006

*BMC Bioinformatics*2006, **7**:122 doi:10.1186/1471-2105-7-122

Received: 12 July 2005
Accepted: 09 March 2006

This article is available from: http://www.biomedcentral.com/1471-2105/7/122

Abstract

Background: Local structures of target mRNAs play a significant role in determining the efficacies of antisense oligonucleotides (ODNs), but some structure-based target site selection methods are limited by uncertainties in RNA secondary structure prediction. If all the predicted structures of a given mRNA within a certain energy limit could be used simultaneously, target site selection would obviously be improved in both reliability and efficiency. In this study, some key problems in ODN target selection on the basis of multiple predicted target mRNA structures are systematically discussed.

Results: Two methods were considered for merging topologically different RNA structures into integrated representations. Several parameters were derived to characterize local target site structures. Statistical analysis on a dataset with 448 ODNs against 28 different mRNAs revealed 9 features quantitatively associated with efficacy. Features of structural consistency seemed to be more highly correlated with efficacy than indices of the proportion of bases in single-stranded or double-stranded regions. The local structures of the target site 5' and 3' termini were also shown to be important in target selection. Neural network efficacy predictors using these features, defined on integrated structures as inputs, performed well in "minus-one-gene" cross-validation experiments.

Conclusion: Topologically different target mRNA structures can be merged into integrated representations and then used in computer-aided ODN design. The results of this paper imply that some features characterizing multiple predicted target site structures can be used to predict ODN efficacy.

Background

Antisense oligonucleotides (ODNs) have served as powerful tools during the post-genome era. They provide an important approach to sequence-specific knockdown of gene expression, offering significant advantages over gene knockout techniques in respect of cost, time and resource requirements, and have therefore been widely used for determining gene function, validating drug targets and elucidating pathways [1,2]. ODNs also have potential as novel therapeutic agents for various diseases; several anti-

sense compounds have been evaluated in clinical trials with promising results [3].

However, even with careful design, only a small proportion of ODNs against a given RNA effectively suppress the target gene in living cells [4]. It is commonly accepted that the identification of accessible sites in the target RNA is of great importance in designing ODNs. Various experimental approaches to the identification of promising local target sites have been described in recent years [5-10]. There has also been much interest in computational approaches to ODN design, which have advantages over experimental methods in terms of throughput, cost and efficiency. Several approaches to efficacy prediction have been proposed for rational selection of ODN target sites [11-14].

Among the factors that influence the activity of a given ODN, the local secondary structures of the target mRNA are very significant in determining *in vitro* efficiency [5,15-17] and are therefore particularly important in current ODN design strategies [18-20]. Local target site structures have also been used as the basis of rational design for other kinds of nucleic acids drugs such as antisense RNAs [21], catalytic RNAs [22] and ribozymes [23]. However, the term "structure" in these studies refers to "single computational predicted structure", not the real structure of the target mRNA; RNA secondary structure is difficult to determine experimentally.

Many RNA secondary structure prediction algorithms have been proposed during the past 20 years. Since the thermodynamically most stable structure of a molecule is generally the one with the minimum free energy (MFE), the initial aim of these prediction methods is to determine the MFE structure [24]. Several MFE structure searching algorithms have been described and are widely used in related research [25,26], especially in ODN target selection. However, partly because of the relatively low reliability of individual target mRNA structure predictions, researchers have often drawn inconsistent conclusions about favorable local structure motifs. The results obtained by Lima *et al*. [18] and Thierry *et al*. [19] indicated that single-stranded hairpin loops in RNA were the best target sites, whereas the studies by Laptev *et al*. [20] suggested that ODNs targeted to sequences predicted to form clustered double-stranded structures in RNA transcripts had the best potential.

It is also possible to consider conformations close to the energy minimum, and algorithms for calculating suboptimal structures within certain energy limits have been proposed [27,28]. The popular RNA secondary structure prediction program MFold now provides results over a range of free energies, mitigating the uncertainty of MFE prediction. Although multiple predicted structures are apparently more reliable, the MFE structure of the target mRNA is still used as the only structural basis in some ODN research. The main difficulty may lie in how to use these foldings simultaneously, since they can be topologically very different.

Studies on ensembles of target structures in ODNs design date back to Jaroszewski *et al*. [29], who considered the 30 lowest-energy computer-simulated structures of rabbit β-globin mRNA qualitatively. In some thermodynamic models, multiple predicted target structures have been merged into the form of free energy [30,31]. The earliest work on computational ODN design based on the original forms of multiple predicted target mRNA structures was perhaps that of Patzel *et al*. [17]. Five structures with low energy were predicted and aligned for a given sequence stretch, and ODN sequences were chosen if potentially favourable local structural elements occurred in all five. *In vitro* experiments showed that this theoretical protocol increased the statistical probability of identifying local target sites accessible to ODN sequences [17,32]. Another way to explore the original forms of optimal and suboptimal mRNA structures simultaneously, which is probably more straightforward, is to merge them into a single-stranded probability profile (SSPP), $\mathbf{P} = \{p_i\}$, $1 \le i \le n$, where p_i is the probability that base i is single-stranded. Actually, algorithms for predicting single-stranded regions in RNA secondary structures have long been of interest, since such regions play many important roles in RNA-RNA, RNA -DNA and RNA-protein interactions [33]. The SFold web server [34] can now directly output the SSPP of an RNA molecule instead of definite individual structures. Ding and Lawrence [33] presented a method for predicting accessible sites in the SSPP of rabbit β-globin mRNA, obtained by summing statistical samples of probable secondary structures. Their results showed a significant correlation between the predicted hybridization potential and the degree of inhibition of *in vitro* translation. Some researchers regard this method as the most successful [11,12].

The original RNA structural information is used in essentially different ways in the two methods described above. In the method based on structure alignment, favorable structural elements are identified by base pairing patterns, which can be illustrated as graphs. The role of secondary structures in this method is similar to its role in earlier studies of ODN design based on the target mRNA MFE structure. The success of this method relies mainly on the greatly increased reliability of structural elements. However, in the method based on SSPP, the RNA structures resemble a special time series rather than molecular "structures" in the usual sense. Base pairing patterns, or topological features, can hardly be explored in SSPP. The common ground between these two methods is the

Table 1: Summary of antisense target genes and their predicted structures used in this study

Accession	Description	No. structures	No. ODNs
X62295	Rattus mRNA for vascular type-1 angiotensin II receptor.	50	36
XM_051583	Homo sapiens v-raf-1 murine leukemia viral oncogene homolog 1 (RAF1), mRNA	50	31
M14758	Homo sapiens P-glycoprotein (PGY1) mRNA	50	22
NM_004996	Homo sapiens ATP-binding cassette, sub-family C (CFTR/MRP), member 1 (ABCC1), transcript variant 1, mRNA	50	14
M24283	Human intercellular adhesion molecule-1 (ICAM-1)	50	66
X52479	Human PKC alpha mRNA for protein kinase C alpha	37	19
NM_001078	Homo sapiens vascular cell adhesion molecule 1 (VCAM1), transcript variant 1, mRNA	50	35
XM_057446	Homo sapiens selectin E (endothelial adhesion molecule 1) (SELE), mRNA.	50	11
M30640	Human endothelial leukocyte adhesion molecule 1 (ELAM1) mRNA, complete cds	50	4
NM_000877	Homo sapiens interleukin 1 receptor, type I (IL1R1), mRNA.	50	20
M31585	Mouse (clone lambda-c5e) intercellular adhesion molecule 1 (ICAM-1) mRNA, complete cds	39	8
BC036531	Homo sapiens collagen, type I, alpha 1, mRNA (cDNA clone MGC:33668 IMAGE:5264710)	50	19
NM_010784	Mus musculus midkine (Mdk), mRNA	17	4
M15077	P.pyralis (firefly) luciferase gene, complete cds	39	8
X03484	Human mRNA for raf oncogene	50	20
X14805	Mus musculus mRNA for DNA methyltransferase 1	50	8
BC005976	Homo sapiens ras homolog gene family, member A, mRNA	23	13
M10843	Rabbit beta-globin mRNA	26	24
U45880	Human X-linked inhibitor of apotosis protein XIAP mRNA	36	6
AF015950	Human telomerase reverse transcriptase mRNA	50	5
NR_001566	Homo sapiens telomerase RNA component (TERC) on chromosome 3	23	5
M34309	Human epidermal growth factor receptor (HER3) mRNA, complete cds.	50	22
NM_004507	Homo sapiens HUS1 checkpoint homolog (S. pombe) (HUS1), mRNA.	33	11
AJ278710	Escherichia coli 23S rRNA gene, strain K12 DSM 30083T	50	7
X03363	Human c-erb-B-2 mRNA	50	3
M10988	Human tumor necrosis factor (TNF) mRNA	26	4
NM_000791	Homo sapiens dihydrofolate reductase (DHFR), mRNA.	50	7
NM_001168	Homo sapiens baculoviral IAP repeat-containing 5 (survivin) (BIRC5), mRNA	29	5
NM_013642	Mus musculus dual specificity phosphatase 1 (Dusp1), mRNA	37	8
AF025846	Co-reporter vector pRL-TK, complete sequence	50	4

emphasis on the role of single-stranded regions in determining target accessibility. In the SSPP of rabbit β-globin mRNA, Ding and Lawrence found a significant correlation between the peak value of SSPP and the degree of inhibition of translation. The "well-chacterized" single-stranded regions were revealed by high probability peaks in the profile [33], while in the systematic alignment of multiple predicted target mRNA secondary structures, large (>10 nt) consecutive sequence stretches not involved in base pairing were regarded as favorable structural motifs [17]. Since these two methods were only evaluated on a single target mRNA, further research is needed on a broad range of target genes.

The purpose of this article is to systematically explore the methods for computational selection of ODN target sites based on features defined in multiple predicted structures of the target mRNA. In our approach, the predicted mRNA structures were first merged into integrated representations. Efficacy-associated features were then screened from a set of features defined on these representations. The potential of neural networks for predicting efficacy on the basis of these features was also validated.

Results

Dataset

Three ODN databases have been reported: ODNBase [35], AOdb [12] and an unnamed database with experimental data from Isis Pharmaceuticals [36]. We have also developed a database named AOBase [37] (NAR molecular biology database collection entry number 781) for both the selection and design of ODNs. Currently, it stores 705 ODNs from the published literature tested against transcripts of 54 different target genes. Since no homogeneous database is publicly available, we perforce used a heterogeneous collection of measurements made by different researchers using different experimental techniques as our dataset. Four hundred and forty-eight ODNs against 28 different mRNAs were collected from AOBase to construct this dataset; 54.2% of them had been tested at protein level and the others at mRNA level. The data selection criteria were similar to those used in other ODN efficacy prediction studies [11-13]: (a) at least 4 ODNs were tested under the same experimental conditions; (b) ODN efficacies were presented as percentages of the control target gene expression level; (c) virus targets were excluded; (d) ODNs targeting to the translational initiation site were

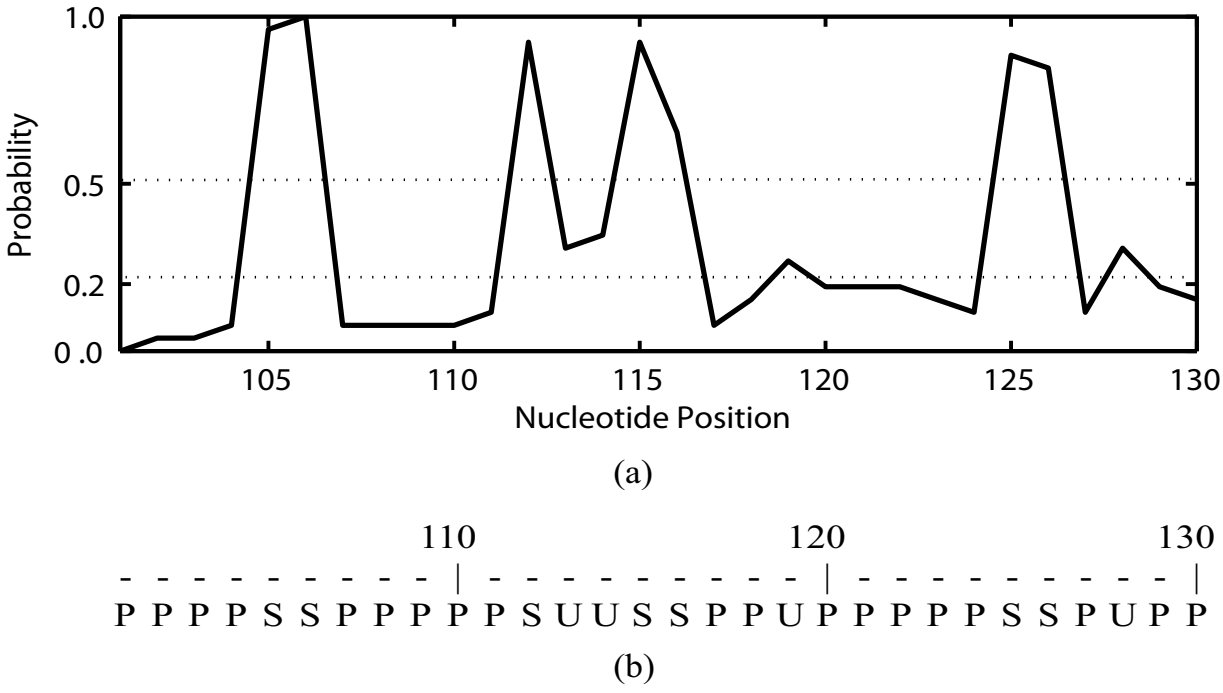

Figure 1
Two representations of multiple predicted structures of rabbit β-globin mRNA (G101-G130). (a) Single-stranded probability profile; (b) 'SUP' representation.

excluded, since regions surrounding the initiation codon are generally considered to be free of secondary structure [38]. To keep in line with most of the research on drug design, the ODN efficacies in our dataset were transformed into [100%-(% of control expression)].

RNA folding calculation times have been greatly reduced in recent years because of faster computers and improved algorithms. The MFold web server [39] can now fold 6000 bases for a batch job, which meets the need of full-length mRNA structure prediction in most cases and is therefore used in this study. Because the number of predicted suboptimal RNA secondary structures increases exponentially as the folding energy increases [40], only structures within 5 percent of the computed minimum free energy were taken into consideration. The upper bound on the number of simultaneously predicted structures was set to 50 to avoid the high computational cost of long RNA sequences. These settings were the default settings of the MFold web server. Table 1 is a brief summary of the dataset.

Integrating multiple predicted target mRNA secondary structures

In this study, two methods were used to represent the multiple predicted local structures of target sites synthetically. All the predicted local structures were first merged into an SSPP, which is easily calculated from the ss-count file in the MFold output. For a more illustrative representation of the multiple predicted structures, the SSPP was further transformed to a "single-stranded/pair/uncertain" sequence (SUP representation) $\mathbf{S} = \{s_i\}$, where s_i = '**S**' if base i is single-stranded, s_i = '**P**' if base i is paired with another base, and s_i = '**U**' if it is uncertain whether base i is single-stranded. The thresholds suggested by Ding and Lawrence [33] were used to map SSPP $\{p_i\}$ into the SUP representation $\{s_i\}$, giving

$$S_i = \begin{cases} \text{'S'}, & p_i > 0.5 \\ \text{'U'}, & 0.5 \geq p_i > 0.2 \\ \text{'P'}, & p_i \leq 0.2 \end{cases} \qquad (1)$$

Table 3: Parameters derived from the SUP sequence representation

Parameter	Definition
f_{NS}	Number of bases in single-stranded region
f_{NP}	Number of bases in double-stranded region
f_{PS}	Percentage of bases in single-stranded region to the length of ODN
f_{PP}	Percentage of bases in double-stranded region to the length of ODN
f_{CS}	Maximum length of consecutive subsequence in single-stranded region
f_{CP}	Maximum length of consecutive subsequence in base pairing
f_{5S}	Maximum length of consecutive subsequence in single-stranded region counting from 5' terminal
f_{5P}	Maximum length of consecutive subsequence in base pairing counting from 5' terminal
f_{3S}	Maximum length of consecutive subsequence in single-stranded region counting from 3' terminal
f_{3P}	Maximum length of consecutive subsequence in base pairing counting from 3' terminal
f_{SC}	Structure consistency, $f_{SC} = \frac{1}{n-1}\sum_{i=1}^{n-1} E(S_i, S_{i-1})$, where $E(x,y) = \begin{cases} 1, & x = y \neq 'U' \\ 0, & x = 'U',\ or\ y = 'U' \\ -1 & x \neq y, x \neq 'S', y \neq 'P' \end{cases}$

SUP representation loses a lot of structural information in comparison to graphical illustration or dot-parenthesis notation of RNA secondary structure and therefore cannot be used to explore the whole RNA structure. However, for RNA local structural analysis, especially of very RNA short regions, SUP gives a competent simplified representation. Figure 1 illustrates part of these two representations (101–130 nt) of rabbit β-globin mRNA structure.

Selection of efficacy-associated features

The first important step in computational design based on multiple predicted mRNA structures is to find the efficacy-associated features in the SSPP and SUP representations of the target sites. Since the data structures of these two linear representations of multiple predicted structures are very different from graphical illustrations of RNA molecules, the topological features known to be correlated with efficacy must be redefined. However, new representations also afford opportunities to discover novel efficacy-associated features.

A set of features characterizing the local multiply-predicted target mRNA secondary structures was derived. Seven of these features were defined on the SSPP representation (listed in Table 2) while the other eleven were defined on the SUP sequence representation (listed in Table 3). The size of the local target, n, in the definition of features is equal to the length of the ODN.

The mean of all single stranded probabilities within a given target site, f_{mean}, indicates the probability that the target site is single-stranded. The maximum value, f_{max}, has also been used for this purpose [33]. $f_{impulse}$, can be viewed as a relative peak value compared to the mean. The other statistics, f_{rms}, f_{peak}, f_{wave}, and $f_{difference}$, describe the structural consistency of the target site.

Numerical features defined on the SUP sequence are directly derived from research results and from empirical rules about target site selection based on local structure. Features f_{NS}, f_{NP}, f_{PS}, and f_{PP}, give an overall description of target structure, while f_{5S}, f_{5P}, f_{3S} and f_{3P} emphasize the local structure of the target site termini. Factors f_{CS} and f_{CP} are derived to confirm whether the occurrence of consecutive subsequences in single-stranded or helical regions is correlated with efficacy, as explored by Patzel *et al*. [17].

Table 2: Parameters derived from the SSPP representation

Parameter	Definition
f_{mean}	Mean, $f_{mean} = \frac{1}{n}\sum_{i=1}^{n} p_i$
f_{rms}	Root mean square, $f_{rms} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - f_{mean})^2}$
f_{max}	Maximum, $f_{\max} = \max_{i=1,\cdots,n}\{p_i\}$
$f_{impulse}$	Impulse factor, $f_{impulse} = \frac{f_{\max}}{f_{mean}}$
f_{peak}	Peak factor, $f_{peak} = \frac{f_{\max}}{f_{rms}}$
f_{wave}	Wave factor, $f_{wave} = \frac{f_{rms}}{f_{mean}}$
$f_{difference}$	Mean of difference, $f_{difference} = \frac{1}{n-1}\sum_{i=1}^{n-1}\lvert p_i - p_{i+1}\rvert$

Table 4: Correlations between features and efficacy

Parameter	Pearson Correlation	Spearman Correlation	Kendall Correlation
f_{mean}	-0.086	-0.055	-0.087
f_{rms}	-0.150**	-0.100**	-0.147**
f_{max}	-0.099*	-0.113**	-0.155**
$f_{impulse}$	0.040	0.039	0.060
f_{peak}	0.124**	0.083**	0.125**
f_{wave}	-0.030	-0.017	-0.025
$f_{difference}$	-0.094*	-0.034	-0.051
f_{NS}	-0.087	-0.057	-0.082
f_{NP}	-0.045	-0.043	-0.061
f_{PS}	-0.073	-0.050	-0.075
f_{PP}	-0.040	-0.040	-0.057
f_{CS}	-0.062	-0.037	-0.053
f_{CP}	-0.012	-0.012	-0.019
f_{5S}	0.031	0.012	0.016
f_{5P}	-0.009	-0.039	-0.055
f_{3S}	-0.050	-0.011	-0.016
f_{3P}	-0.036	-0.030	-0.039
f_{SC}	-0.064	-0.045	-0.066

**. Correlation is significant at the 0.01 level
*. Correlation is significant at the 0.05 level

Absolute numbers of bases appear in the definitions of eight features defined on the SUP representation, viz. f_{NS}, f_{NP}, f_{CS}, f_{CP}, f_{5S}, f_{5P}, f_{3S} and f_{3P}. Since the ODN lengths in the dataset are not uniform, it is necessary to determine whether these features are bound up with or limited by the size of local target. Figure 2(a) shows the distribution of ODN lengths in the dataset, which range from 10 nt to 22 nt. Most of the ODNs were 20 nt long. The dataset was divided into groups according to ODN length. The mean values of these features were calculated for each group and are shown in Figure 2(b), which indicates no obvious relationships between these features and target size.

Two types of indices, efficiency prediction potential and classification potency, were used to measure the suitability of these parameters for rational ODN design. The efficacy prediction potential was evaluated by calculating the correlation between the features and efficacy, using Pearson linear correlation, Spearman rank correlation and Kendall rank correlation. The classification potency was evaluated by exploring the performance of Fisher linear discriminators, using the feature as the single independent variable. The performance was measured as specificity $S_p = \frac{T_n}{T_n + F_p}$ and sensitivity $S_e = \frac{T_p}{T_p + F_n}$. Two different efficacy threshold values, 50% and 75%, were used to distinguish between positive and negative cases in our dataset, since these indices depend on threshold. Features matching at least one of the following two criteria were selected as efficacy-associated: (a) statistically significant correlation ($p < 0.05$) with efficacy; and (b) high specificity (≥0.7) or high sensitivity (≥0.7) in distinguishing between active and inactive ODNs.

The correlation between parameters and efficacy is presented in Table 4. Only four features defined on SSPP, i.e. f_{rms}, f_{max}, f_{peak} and $f_{difference}$, correlated strongly with efficacy. Table 5 compares the Fisher discrimination results for each parameter and different thresholds, indicating that f_{rms}, f_{max}, f_{peak}, $f_{difference}$, f_{PP}, f_{CS}, f_{CP}, f_{5S} and f_{3S} can be used to distinguish between active and inactive ODNs according to our criteria.

The most noteworthy finding is that ODN efficacy seems not to rely greatly on the degree of single-strandedness in its target site, as suggested in previous publications [18-20], since f_{mean}, f_{NS} and f_{PS} show neither sufficient correlation with efficacy nor good performance in identifying active ODNs. The lengths of consecutive single-stranded regions in the target site, which are characterized by f_{CS}, prove useful for identifying active ODNs. This result is partly consistent with the conclusion drawn by Patzel *et al* [17]. In contrast to the conclusion of Ding and Lawrence [33], although f_{max} is revealed to be efficacy-associated, the peak value of the target site SSPP correlates negatively with efficacy.

The helical region in the target site appears to be more important, as suggested by Laptev [20], because features f_{PP} and f_{CP} satisfy our selection criteria for ODN classification. From the analysis, it is obvious that the structural consistency features, f_{rms}, f_{peak}, and $f_{difference}$, are more important in target site selection. But this should not be

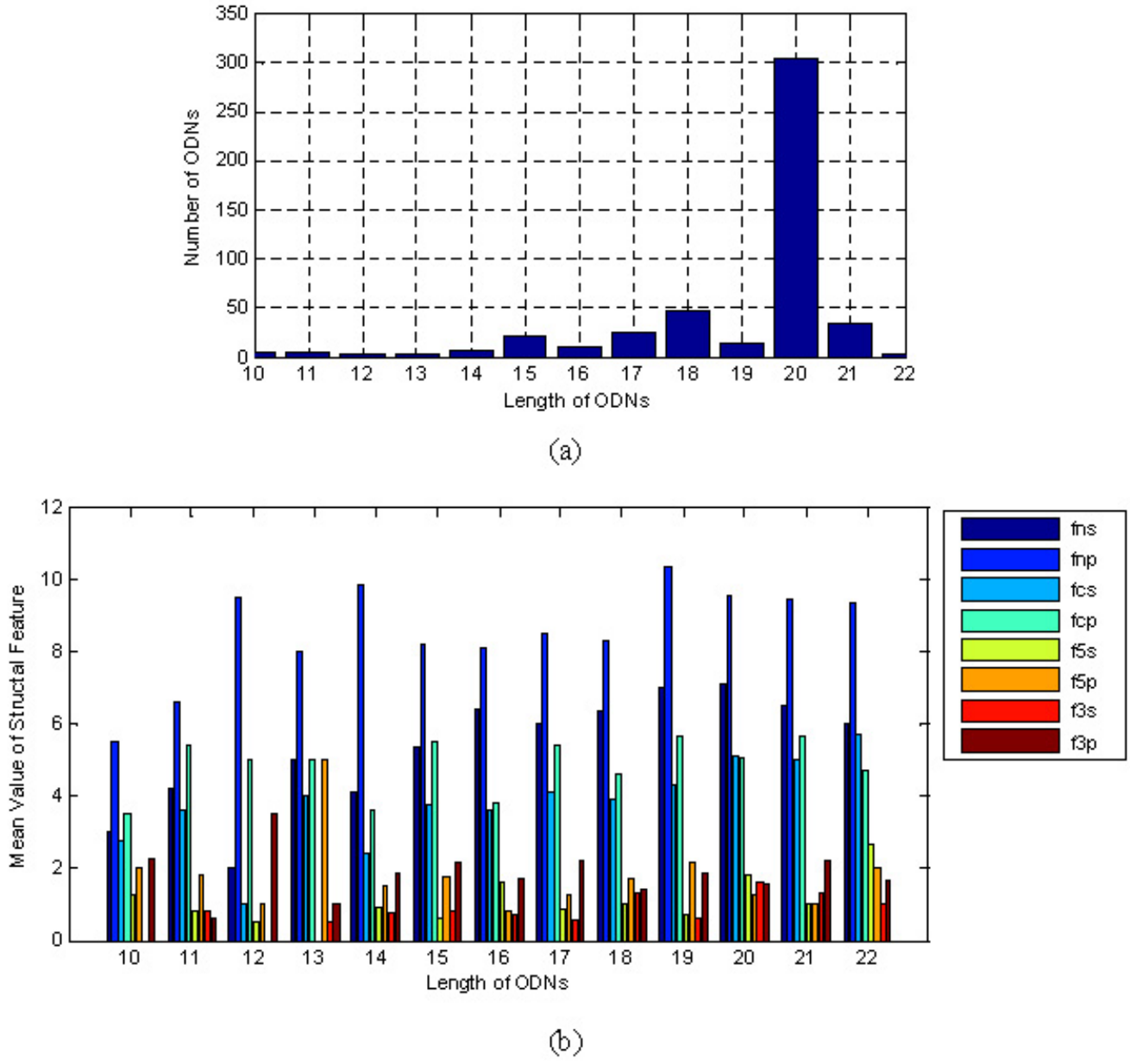

Figure 2
The distribution of ODN length and length-limited features. (a) The distribution of ODN lengths in the dataset; (b) Mean values of some features of ODNs with different lengths.

interpreted as implying simple correspondences between structural consistency and efficacy.

ODN efficacy may be closely associated with the local structures of the 5' and 3' termini of the target sites. Fisher classifiers using factors f_{5S} and f_{3S} gave high specificity or sensitivity in ODN discrimination.

Although some features are efficacy-associated, the relationship between structural factors and efficacy is highly complex. No single feature has been found to correlate highly with efficacy, and no feature is reliable on its own for distinguishing active from inactive ODNs. Two feature sets defined on the SSPP and SUP representations of the target site are selected as inputs of efficacy-predicting neu-

Table 5: Performance of Fisher linear discriminators for each parameter

Parameter	Threshold = 50%		Threshold = 75%	
	S_e	S_p	S_e	S_p
f_{mean}	0.56	0.53	0.65	0.51
f_{rms}	0.58	0.54	0.50	0.54
f_{max}	0.33	0.73	0.60	0.72
$f_{impulse}$	0.37	0.67	0.48	0.38
f_{peak}	0.50	0.61	0.56	0.59
f_{wave}	0.42	0.58	0.63	0.43
$f_{difference}$	0.52	0.50	0.52	0.50
f_{NS}	0.56	0.50	0.56	0.48
f_{NP}	0.44	0.51	0.52	0.49
f_{PS}	0.54	0.52	0.58	0.51
f_{PP}	0.49	0.45	0.70	0.54
f_{CS}	0.56	0.47	0.73	0.45
f_{CP}	0.43	0.63	0.74	0.40
f_{5S}	0.31	0.71	0.73	0.30
f_{5P}	0.33	0.63	0.58	0.37
f_{3S}	0.29	0.73	0.65	0.29
f_{3P}	0.34	0.64	0.50	0.37
f_{SC}	0.50	0.50	0.60	0.52

. high specificity (≥ 0.7) or high sensitivity (≥ 0.7)

ral networks: $F_{SSPP} = \{f_{rms}, f_{max}, f_{peak}, f_{difference}\}$ and $F_{SUP} = \{f_{PP}, f_{CS}, f_{CP}, f_{5S}, f_{3S}\}$.

Efficacy predicting using neural networks

To assess the ability of selected features to predict efficacy, two neural network models were constructed, one for features defined on the SSPP and the other for features derived from the SUP sequence representation of the target structure.

Previous studies have shown that cross-validation is important for estimating accuracy [11-14]. Since ODNs always have similar properties if they are near each other on the same gene or are measured in the same study, the network training process should be completely independent of the test data [12,13]. In this research, cross-validation was done by the "minus-one-gene" (-gene) [13] approach. ODNs targeting to 8 mRNAs (listed in Table 6) were selected alternately from the dataset for testing, while the remainder, assayed in the same studies, were used as the training set. The test mRNA selection criteria were: (a) more than 15 different target sites were tested; (b) the efficacy of at least one ODN was greater than 75%.

Sixteen neural networks for efficacy prediction were tested in our cross-validation experiments. The network group N_{SSPP} ($N_{SSPP}1 \sim N_{SSPP}8$) took F_{SSPP} as inputs, and the N_{SUP} group ($N_{SUP}1 \sim N_{SUP}8$) took F_{SUP} as the input parameter set. The outputs of all these networks met the condition of convergence within 100 training cycles.

Several methods have been used to measure the accuracy of ODN predictors [11-14]. To obtain rounded assessments for the aforementioned neural networks, two different types of indices were computed: (1) specificity S_P, sensitivity S_e and accuracy $Acc = \frac{T_n + T_p}{T_p + T_n + F_p + F_n}$ calculated using fixed threshold values, as mentioned above in the account of feature selection; (2) the receiver operating characteristics (ROC) curve [41], which is a plot of S_e versus 1 - S_P at different thresholds. The ROC area was calculated as a quantitative indicator of the ability of the network to classify. The cutoff efficacy value used to distinguish positive from negative ODNs in the cross-validation test was 75%.

The performances of the neural networks are listed in Table 7. The specificities, S_P, of all the networks in these two groups are greater than the related sensitivities, S_e. This performance is beneficial for ODN design, since users will only be interested in candidates with high predicted efficacy in practical applications [14]. The ROC curves of the 16 networks tested on ODNs targeting to 8 different mRNAs are shown in Figure 3. The best ROC curve areas were obtained in cross-validation experiment 7 (network $N_{SSPP}7$ and $N_{SUP}7$), which used the data from Matveeva *et al*. [6] as test set. The average ROC area for N_{SUP} is 0.77. The average for N_{SUP} is 0.73, which is little lower.

Table 6: Dataset for cross-validation experiments

Networks	Accession of test gene	Number in train set	Number in test set
$N_{SSPP}1$ and $N_{SUP}1$	X62295	412	36
$N_{SSPP}2$ and $N_{SUP}2$	XM_051583	417	31
$N_{SSPP}3$ and $N_{SUP}3$	M14758	426	22
$N_{SSPP}4$ and $N_{SUP}4$	M24283	356	66
$N_{SSPP}5$ and $N_{SUP}5$	NM_001078	379	35
$N_{SSPP}6$ and $N_{SUP}6$	NM_000877	428	20
$N_{SSPP}7$ and $N_{SUP}7$	X03484	428	20
$N_{SSPP}8$ and $N_{SUP}8$	M10843	424	24

Table 7: The performances of two groups of networks in cross-validation experiments

Networks	Se	Sp	Acc	ROC area	Networks	Se	Sp	Acc	ROC area
$N_{sspp}1$	0.50	0.97	0.92	0.91	$N_{sup}1$	0	0.94	0.83	0.60
$N_{sspp}2$	0.33	0.96	0.90	0.75	$N_{sup}2$	0	0.93	0.84	0.69
$N_{sspp}3$	0	0.93	0.59	0.71	$N_{sup}3$	0	1	0.64	0.65
$N_{sspp}4$	0	1	0.88	0.66	$N_{sup}4$	0.5	0.86	0.82	0.66
$N_{sspp}5$	0	1	0.71	0.69	$N_{sup}5$	0	1	0.71	0.74
$N_{sspp}6$	0	0.94	0.85	0.81	$N_{sup}6$	0	1	0.9	0.89
$N_{sspp}7$	0	1	0.70	0.98	$N_{sup}7$	0.17	0.93	0.70	0.89
$N_{sspp}8$	0	1	0.58	0.63	$N_{sup}8$	0.40	0.86	0.67	0.71

. high specificity (≥ 0.7) or high sensitivity (≥ 0.7)

Discussion

Compared with most other bioinformatics research problems, studies on computer-aided ODN design are far from "data rich". Moreover, the data collected from the published literature are variable owing to the diversity of experimental methods. To provide a more reliable basis for feature-mining and predictor development, one focus of future work will be on enlargement of the dataset. A large dataset with quality control will make the analysis and cross-validation of grouped homogeneous subsets possible, and therefore make the ODN design systems more reliable.

Another "data poor" limitation in our study and related research [6,17,29] is that not all possible target RNA structures are taken into account. As pointed out by Mathews, an ideal way to integrate the predicted RNA structures would be to compute a partition function, which sums the contributions of all structures weighted by their Boltzmann probabilities [44]. However, the determination of a partition function has $O(N^3)$ computational complexity [45], so this method is practicable only for short RNA sequences. Several studies have been done on the estimation of partition function with lower computational cost [44,46-48]. The Vienna RNA secondary structure prediction server [49] can now compute the partition function of RNA up to 5000 bases for batch jobs. One implication of this study that warrants further investigation is ODN design using the partition function of the target mRNA, which is based on more reliable structural information.

The factors influencing the potential of an ODN are complex and so far poorly understood. Although this paper focuses on the relationship between ODN efficacy and target site structure, we do not ignore other factors that have been shown to influence efficacy, such as chemical properties, DNA-RNA duplex stability, sequence motifs, metabolic properties of target mRNA, *etc.* [4]. We do believe that as more factors are considered in ODN efficacy prediction, the more reliable the target site selection becomes.

Conclusion

This paper presents a method, based on multiple predicted target mRNA structures, for reducing the uncertainty of structure prediction in ODN design. Several efficacy-associated features characterizing the integrated structure of the target site have been discovered. The structural consistency features of the target seem to be correlated with efficacy. In contrast, some features of favorable ODN targets reported in previous research, which emphasized single-stranded regions, were found to correlate weakly with efficacy. In addition, the local structures of the 5' and 3' termini were shown to be important in target site selection.

Neural network efficacy predictors using features defined on integrated structures as inputs have been shown to perform well, implying that these features can also be used for other forms of efficacy prediction such as Bayesian statistics (BS), multiple linear regression (MLR), decision tree (DT) and support vector machine (SVM).

Methods

After preliminary experiments, feed-forward network architecture with a hidden layer containing 20 nodes was applied to each network. The input neurons used a logarithmic sigmoid (tan-sigmoid) activation function; the output neurons used a hyperbolic tangent sigmoid (log-sigmoid) activation function. The weights and bias values of the networks were updated according to the Levenberg-Marquardt optimization algorithm [42], which appears to be the fastest method for training a moderate-size feed-forward neural network [43]. Matlab® Neural Network Toolbox 4.0.3 was used for all neural network implementation.

Authors' contributions

SW guided the project. XB and SW conceived of the study. XB wrote program, analyzed the results and drafted the manuscript. SL and DS helped in dataset construction. WS and JY helped in analysis and discussion, gave useful comments.

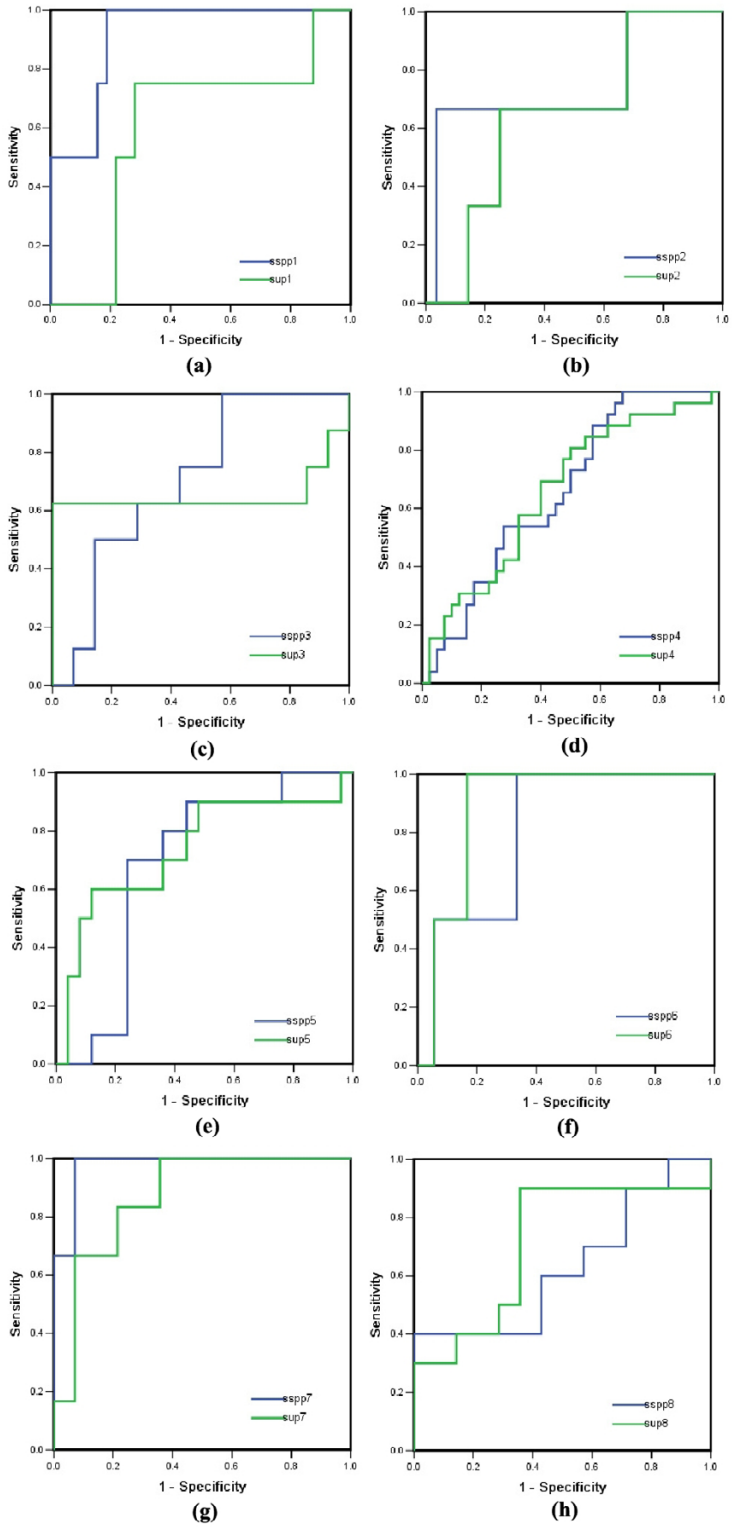

Figure 3
ROC curves for efficacy-predicting neural networks. ROC curves are shown for networks (a) $N_{SSPP}1$ and $N_{SUP}1$; (b) $N_{SSPP}2$ and $N_{SUP}2$; (c) $N_{SSPP}3$ and $N_{SUP}3$; (d) $N_{SSPP}4$ and $N_{SUP}4$; (e) $N_{SSPP}5$ and $N_{SUP}5$; (f) $N_{SSPP}6$ and $N_{SUP}6$; (g) $N_{SSPP}7$ and $N_{SUP}7$; (h) $N_{SSPP}8$ and $N_{SUP}8$.

Acknowledgements

This work was supported by grants from the National Nature Science Foundation of China (No.30171111), the National High Technology Research and Development Program of China (863 Program) (No. 2003AA234031) and the Special Funds for Major State Basic Research Program of China (973 Program) (No. 2004CB518904).

References

1. Taylor MF, Wiederholt K, Svetdrup F: **Antisense oligonucleotides: a systematic high-throughput approach to target validation and gene function determination.** *Drug Discov Today* 1999, **4:**562-567.
2. Flaherty KT, Stevenson JP, O'Dwyer PJ: **Antisense therapeutics: lessons from early clinical trials.** *Curr Opin Oncol* 2001, **13:**499-505.
3. Crooke ST: **An overview of progress in antisense therapeutics.** *Antisense Nucleic Acid Drug Dev* 1998, **8:**115-122.
4. Far RK, Nedbal W, Sczakiel G: **Concepts to automate the theoretical design of effective antisense oligonucleotides.** *Bioinformatics* 2001, **17:**1058-1061.
5. Ho SP, Bao Y, Lesher T, Malhotra R, Ma LY, Fluharty SJ, Sakai RR: **Mapping of RNA accessible sites for antisense experiments with oligonucleotide libraries.** *Nature Biotechnology* 1998, **16:**59-63.
6. Matveeva OV, Felden B, Tsodikov A, Johnston J, Monia BP, Atkins JF, Gesteland RF, Freier SM: **Prediction of antisense oligonucleotide efficacy by *in vitro* methods.** *Nature Biotechnology* 1998, **16:**1374-1375.
7. Matveeva O, Felden B, Audlin S, Gesteland RF, Atkins JF: **A rapid *in vitro* method for obtaining RNA accessibility patterns for complementary DNA probes: correlation with an intracellular pattern and known RNA structures.** *Nucleic Acids Res* 1997, **25:**5010-5016.
8. Milner N, Mir KU, Southern EM: **Selecting effective antisense reagents on combinatorial oligonucleotide arrays.** *Nature Biotechnology* 1997, **15:**537-541.
9. Allawi HT, Dong F, Ip HS, Neri BP, Lyamichev VI: **Mapping of RNA accessible sites by extension of random oligonucleotide libraries with reverse transcriptase.** *RNA* 2001, **7:**314-327.
10. Zhang HY, Modn J, Zhou D, Xu Y, Thonberg H, Liang Z, Wahlestedt C: **mRNA accessibility site tagging (MAST): a novel high throughput method for selecting effective antisense oligonucleotides.** *Nucleic Acid Res* 2003, **31:**e72.
11. Camps-Valls G, Chalk AM, Serrano-Lopez A, Martin-Guerrero JD, Sonnhammer ELL: **Profiled support vector machine for antisense oligonucleotide efficacy prediction.** *BMC Bioinformatics* 2004, **5:**135.
12. Chalk AM, Sonnhammer ELL: **Computational antisense oligo prediction with a neural network model.** *Bioinformatics* 2002, **18:**1567-1575.
13. Giddings MC, Shah AA, Freier S, Atkins JF, Gesteland RF, Matveeva OV: **Artificial neural network prediction of antisense oligodeoxynucleotide activity.** *Nucleic Acids Res* 2002, **30:**4295-4304.
14. Sætrom Pål: **Predicting the efficacy of short oligonucleotides in antisense and RNAi experiments with boosted genetic programming.** *Bioinformatics* 2004, **20:**3055-3063.
15. Scherer LJ, Rossi JJ: **Approaches for the sequence-specific knockdown of mRNA.** *Nature Biotechnology* 2003, **21:**1457-1465.
16. Vickers TA, Wyatt JR, Freier SM: **Effects of RNA secondary structure on cellular antisense activity.** *Nucleic Acids Res* 2000, **28:**1340-1347.
17. Patzel V, Steidl U, Kronenwett R, Haas R, Sczakiel G: **A theoretical approach to select effective antisense oligodeoxyribonucleotides at high statistical probability.** *Nucleic Acids Res* 1999, **27:**4328-4334.
18. Lima WF, Monia BP, Ecker DJ, Freier SM: **Implication of RNA structure on antisense oligonucleotide hybridization kinetics.** *Biochemistry* 1992, **31:**12055-12061.
19. Thierry AR, Rahman A, Dritschilo A: **Overcoming multi drug resistance in human tumor cells using free and liposomally encapsulated antisense oligodeoxynucleotides.** *Biochem Biophys Res Commun* 1993, **190:**952-960.
20. Laptev AV, Lu Z, Colige A, Prockop DJ: **Specific inhibition of expression of a human collagen gene (COL1A1) with modified antisense oligonucleotides.** *Biochemistry* 1994, **33:**11033-11039.
21. Sczakiel G, Homann M, Rittner K: **Computer-aided search for effective antisense RNA target sequences of the human immunodeficiency virus type 1.** *Antisense Res Dev* 1993, **3:**45-52.
22. Denman RB: **Using RNAFOLD to predict the activity of small catalytic RNAs.** *Biotechniques* 1993, **15:**1090-1095.
23. James W, Cowe E: **Computational approaches to the identification of ribozyme target sites.** *Methods Mol Biol* 1997, **74:**17-26.
24. Higgs PG: **RNA secondary structure: physical and computational aspect.** *Quarterly Reviews of Biophysics* 2000, **33:**199-253.
25. Zuker M, Stiegler P: **Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information.** *Nucleic Acids Res* 1981, **9:**133-148.
26. Dumas JP, Ninio J: **Efficient algorithms for folding and comparing nucleic acid sequences.** *Nucleic Acids Res* 1982, **10:**197-206.
27. Zuker M: **On finding all suboptimal foldings of an RNA molecule.** *Science* 1989, **244:**48-52.
28. Yamamoto K, Kilamura Y, Yoshikura H: **Computation of statistical secondary structure of nucleic acids.** *Nucleic Acids Res* 1984, **12:**335-346.
29. Jaroszewski JW, Syi JL, Ghosh M, Ghosh K, Cohen JS: **Targeting of antisense DNA: comparison of activity of anti-rabbit beta-globin oligodeoxyribonucleoside phosphorothioates with computer predictions of mRNA folding.** *Antisense Res Dev* 1993, **3:**339-348.
30. Walton SP, Stephanopoulos GN, Yarmush ML, Roth CM: **Thermodynamic and kinetic characterization of antisense oligodeoxynucleotide binding to a structured mRNA.** *Biophys J* 2002, **82:**366-377.
31. Mathews DH, Burkard ME, Freier SM, Wyatt JR, Turner DH: **Predicting oligonucleotide affinity to nucleic acid targets.** *RNA* 1999, **5:**1458-1469.
32. Scherr M, Rossi JJ, Sczakiel G, Patzel V: **RNA accessibility prediction: a theoretical approach is consistent with experimental studies in cell extracts.** *Nucleic Acids Res* 2000, **28:**2455-2461.
33. Ding Y, Lawrence CE: **Statistical prediction of single-stranded regions in RNA secondary structure and application to predicting effective antisense target sites and beyond.** *Nucleic Acids Res* 2001, **29:**1034-1046.
34. Ding Y, Chan CY, Lawrence CE: **Sfold web server for statistical folding and rational design of nucleic acids.** *Nucleic Acids Res* 2004, **32:**W135-W141.
35. Giddings MC, Matveeva OV, Atkins JF, Gesteland RF: **ODNBase – a web database for antisense oligonucleotide effectiveness studies.** *Bioinformatics* 2000, **16:**843-844.
36. Matveeva OV, Mathews DH, Tsodikov AD, Shabalina SA, Gesteland RF, Atkins JF, Freier SM: **Thermodynamic criteria for high hit rate antisense oligonucleotide design.** *Nucleic Acids Res* 2003, **31:**4989-4994.
37. **AOBase** [http://www.bioit.org.cn/ao/aobase]
38. Sohail M, Southern EM: **Selecting optimal antisense reagents.** *Advanced Drug Delivery Reviews* 2000, **44:**23-34.
39. Zuker M: **Mfold web server for nucleic acid folding and hybridization.** *Nucleic Acids Res* 2003, **31:**3406-3415.
40. Wuchty S, Fontana W, Hofacker IL, Schuster P: **Complete suboptimal folding of RNA and the stability of secondary structures.** *Biopolymers* 1999, **49:**145-165.
41. Hanley J, McNeil BJ: **The meaning and use of the area under a receiver operating characteristic (ROC) curve.** *Radiology* 1982, **143:**29-36.
42. Hagan MT, Menhaj M: **Training feedforward networks with the Marquardt algorithm.** *IEEE Transactions on Neural Networks* 1994, **5:**989-993.
43. Demuth H, Beale M: *Neural Network Toolbox MathWorks Inc* 2004.
44. Mathews DH: **Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization.** *RNA* 2004, **10(8):**1178-1190.
45. McCaskill JS: **The equilibrium partition function and base pair probabilities for RNA secondary structure.** *Biopolymers* 1990, **29:**1105-1119.
46. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P: **Fast folding and comparison of RNA secondary structures.** *Monatsh Chem* 1994, **125:**167-168.

47. Fekete M, Hofacker IL, Stadler PF: **Prediction of RNA base pairing probabilities on massively parallel computers.** *J Comput Biol* 2000, **7:**171-182.
48. Ding Y, Lawrence CE: **A Bayesian statistical algorithm for RNA secondary structure prediction.** *Comput Chem* 1999, **23:**387-400.
49. Hofacker IL: **Vienna RNA secondary structure server.** *Nucleic Acids Res* 2003, **31:**3429-3431.