Prediction of Antisense Oligonucleotide Efficacy Using Local and Global Structure Information With Support Vector Machines

Roger Craig and Li Liao*
Computer and Information Sciences, University of Delaware
Newark, Delaware 19716, USA
Email: lliao@cis.udel.edu

Absract

Designing antisense oligonucleotides with high efficacy is of great interest both for its usefulness to the study of gene regulation and for its potential therapeutic effects. The high cost associated with experimental approaches has motivated the development of computational methods to assist in their design. Essentially, these computational methods rely on various sequential and structural features to differentiate the high efficacy antisense oligonucleotides from the low efficacy. By far, however, most of the features used are either local motifs present in primary sequences or in secondary structures. We proposed a novel approach to profiling antisense oligonucleotides and the target RNA to reflect some of the global structural features such as hairpin structures. Such profiles are then utilized for classification and prediction of high efficacy oligonucleotides using support vector machines. The method was tested on a set of 348 antisense oligonucleotides of 19 RNA targets with known activity. The performance was evaluated by cross validation and ROC scores. It was shown that the prediction accuracy was significantly enhanced.

1. Introduction

Antisense oligonucleotides (AO) are typically 15-30 bases long, can bind to mRNA sequences at specific locations determined by Watson-Crick base pairing rules, and as a result, can inhibit gene expression. Designing high efficacy antisense oligonucleotides has drawn a lot of attention because of the usefulness to the study of gene regulation, and also because of potential therapeutic applications [1]. However, it is reported that only about 20% of the corresponding AOs are actually effective gene inhibitors *in vivo*, if the target site on the mRNA is randomly selected [2]. Experimental approaches to AO selection, in which a number of candidate target sites on the mRNA are chosen for testing their efficacy, are time consuming and costly.

Several computational methods have thus been developed to this end [3, 4, 5, 6]. Essentially, these methods rely on various sequential and structural features to differentiate antisense oligonucleotides with high efficacy from those with low efficacy. For example, in a paper by Matveeva *et al*, the occurrence of some motifs of 3 or 4 bases long, such as CCAC (GGGG), are reported to be correlated with high (low) activity efficacy. In the work by Giddings et al, 2002, an artificial neural network was developed to scan the AO sequences of all 256 tetranucleotide motifs, and achieved an accuracy of 53% for high efficacy AOs. In [5] by Camps-Valls et al, 2004, a support vector regressor was developed to select among a variety of features the most discriminating ones in identifying high efficacy AOs, and a success rate of 83.3% was reported for predicting AOs with activity efficacy higher than 0.75. By far, however, most of the features used in these methods are local to very short segments, such as sequence motifs or secondary structural signatures or combination of both. Features pertaining to the global properties of oligonucleotide and mRNAs are often either not included at all, or included but with significant information loss. For example, in Camps-Valls et al's work [5], hairpin quality, a kind of global feature that reflects base paring across a distance, is considered, but only encapsulated as a single value to input to the classifier. The information loss suffered from such situations is much like that the number of base 'C' in the oligonucleotide sequence does not tell how these bases are actually distributed.

In this paper, we proposed a new method that incorporates both local and global information from the sequences and secondary structure of the antisense oligonucleotides and the target mRNA. The sequences and secondary structures of AOs are profiled into features (*N*mers) that are believed to be differentiating and are input to support vector machines for classification. Similar profiling on mRNAs at the binding site is also implemented. In addition to profiling on the occurrences of *N*mers, which are

* Corresponding author.

considered local due to typically small *N*, we devised another profiling method that can capture some global features, for example, the base pairings across a long distance, and then retain adequate information to input to the classifier. By testing on a widely adopted dataset, it is shown that our method provides superior performance with a ROC score 0.973.

2. Method

In order to classify antisense oligonucleotides by their activity efficacy and further to predict high efficacy oligonucleotides, we first need to identify and extract features/attributes of the corresponding AOs and the target mRNAs that are believed to be correlated to the activity efficacy. Then we input those attributes to a classifier for classification and/or prediction. In this work, we adopted a powerful and increasing popular classifier – support vector machines [12, 13].

2.1 Data

The data used in this work is the same as in Matveeva et al (2000). The set consists of 348 oligonucleotide DNA sequences whose activities at targeting mRNA were already experimentally determined. The activity efficacy is measured as a score between 0 (low) and 1 (high). In order to evaluate the performance of the classifier, a scheme of 6 fold cross validation is adopted. That is, the set was split randomly into 6 subsets, of which 5 subsets were used for training the classifier and one subset was used for testing. In each subset, AOs with efficacy higher than 0.5 are considered as positive examples and AOs with efficacy equal to or lower than 0.5 are considered as negative examples. The learning and testing process was repeated 6 times, with the test set rotating among the 6 subsets. The average results are reported. To evaluate the method's accuracy of predicting high efficacy AOs, we follow the same convention as used in the work of Gamps-Valls et al [5] by testing on a subset of AOs with activity efficacy either higher than 0.75 or lower than 0.25. Such a dichotomy is reasonable, as in practice we are mainly interested in predicting AOs that can reduce the expression level of target mRNA by 75% or more.

2.2 Profiling

To be able to classify AOs with different activity efficacy, the first step is to identify and extract some features/attributes that are correlated with AOs' activity efficacy, and to represent these features in a format suitable to be used in support vector machines. In general, it is easy to conceive that certain such features must be embedded in the sequences and structure of oligonucleotides and target mRNAs. For example, in Matveeva et al, some motifs of size 4 such as CCAC (GGGG) are reported to be correlated with high (low) activity efficacy. One systematic way to extract such features is to profile the sequence on all *N*mers for a given *N*. Such profiling, which has been widely used, for instance to classify proteins in (Leslie et al, 2004), will generate for each AO a vector of size *D*, where *D* is the number of all possible *N*mers, and each component of the vector gives the occurrence frequency for the corresponding *N*mer.

In this work, we propose to profile both the sequences and secondary structure of AOs, as it is reported in literature that the secondary structure of oligonucleotides and mRNAs may play a very important role in the activity efficacy [4]. A similar method has been used in the work by Xue *et al* for classifying microRNA precursors [7]. Moreover, in this paper, we devised a scheme that also took into account some global features of secondary structure of AOs and target mRNAs.

RNAfold from *Vienna RNA package* (http://www.tbi.univie.ac.at/~ivo/RNA/) was used to produce the secondary structure for a given antisense oligonucleotide or mRNA. The predicted secondary structure is represented as a string of nested parentheses for paired nucleotides and dots for unpaired nucleotides.

In the Triplet profiling, the secondary structure is scanned for any of the 8 possible triplets: …, (.., .(., ..(, ((., (.(, .((, and (((. Here the orientation of parentheses is indiscriminate. To account for the correlation between the sequence and the secondary structure, the triplet is coupled with the nucleotide in the middle base of the triplet, extending the feature set to a size of 8x4 = 32. The occurrence frequencies of these 32 features are counted and used to represent the corresponding molecule, which can be either an antisense oligonucleotide or the target site of mRNA. As a straightforward generalization, we also concatenate the vectors of the AO and its target mRNA into a combined vector of dimension 64. It should be noted that we also experimented with using parenthesis-orientation sensitive triplets and quadruplets in profiling. With parenthesis orientation sensitive triplets, the profile vector dimension increases to 3x3x3x4 = 108, as at each position there are three possibilities: ., (, or). Similarly, quadruplets, either parenthesis orientation sensitive or not, will increase the profile vector dimension significantly. Our experiments indicated that such efforts did not pay off, instead, the prediction accuracy decreased slightly, see Results section for details for an explanation.

Although the Triplet profiling is useful in capturing local features, we are also interested in features that may span across the antisense oligonucleotides and/or mRNA target sites, for example, a hairpin structure (see Figure 1 panel B). However, the straightforward generalization of using *N*mers with larger value of *N* does not scale up: the profile vector dimension increases exponentially with *N*. More importantly, as mentioned above, larger *N* does not lead to the improvement, but to the decrease in performance.

In this work, we further proposed the GS profiling to incorporate some global structural information by scoring the secondary structure cumulatively. That is, we start with an initial score, say zero, and we scan the secondary structure one base at a time. At each base, we score the base according to whether it is unpaired, paired with a base downstream, or paired with a base upstream. The cumulative score, i.e., the score for the current base added to the score up to the previous base, is recorded for the current position. The cumulative scores form a score landscape, which is used to characterize the secondary structure somewhat globally. Figure 1 illustrates how this works for four typical cases using a scoring scheme as follows:

$$S(a) = \begin{cases} +1 & \text{if a is paired downstream,} \\ -1 & \text{if a is paired upstream,} \\ 0 & \text{if a is unpaired,} \end{cases}$$

where a $\in${A, C, G, U}. It can be easily seen that the cases presented in Panels B and C of Figure 1 may be indistinguishable in Triplet profiling, but have very different GS profiles. It is worth noting that while the secondary structure in Panel C is not possible for AOs, it can happen to target sites of the mRNA, which folds as a whole. A refined scoring scheme can be devised to account for the varied strength of different pairing bases, for instance, higher score is assigned to GC pair than AU pair. The score landscape thus obtained is similar in spirit to what one would get from the traceback path of Nussinov Algorithm for RNA folding [8], though they differ in an important way: unlike the score landscapes produced here, which may have show up and down trends as indicated in Figure 1 Panel B for hairpins, the traceback in Nussinov gives a score landscape monotonically increasing from left to right, since the dynamic programming algorithm maximizes the number of pairings incrementally. Because the size of a score landscape vector thus obtained is the same as the length of the antisense oligonucleotide, which varies for different AOs, we need to normalize these vectors into the same size before we can input them to the support vector machine. The size of the normalized vectors, which we refer to as GS profile, is a free parameter in our method. In this study, we have used a value 10. In addition to the length normalization (into 10 bins), the amplitudes of these vectors are also normalized, into the range [-1, 1]. The amplitude normalization becomes necessary when we later concatenate the GS profiles with the Triplet profiles to take the benefits from both – the resultant vectors will not be so skewed in terms of the amplitudes. We refer to the concatenated profiles as GS+Triplet.

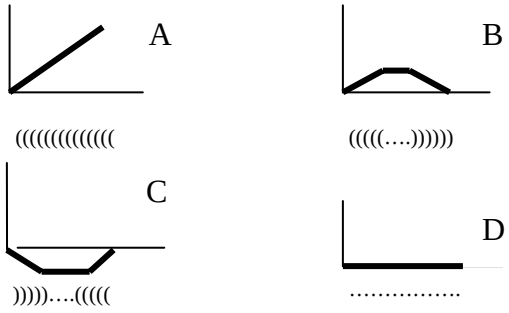

Figure 1. Schematic score landscape for four typical secondary structural features.

The profiles are then used as the input to the support vector machine for classification. The support vector machine used in this work is the implementation from the SVMLight package [9]. Three types of kernel functions – linear, polynomial and Gaussian with default parameter settings – are tested. The best performance was reported from linear kernel function.

3. Results

The performance of classification and prediction is evaluated using receiver operating characteristic (ROC) score [10]. A ROC score is the normalized area under a curve that plots the true positives as a function of false positives for varying classification thresholds. ROC scores are in the range of [0, 1], with 1 for a perfect classification.

The ROC scores for the 6-fold cross validation experiments are reported in Table 1. The row labeled with Triplet lists the ROC scores for using Triplet profiling for three cases: antisense oligonucleotide only (Oligo), target mRNA only (mRNA), and concatenation of the two (Oligo+mRNA). The row labeled with GS lists the ROC scores for the above three cases using the GS profiling. And the last row lists the ROC scores when Triplet profiling and GS profiling are used in conjunction.

It can be seen that for Triplet profiling the best performance (ROC = 0.960) was achieved when oligonucleotide and mRNA target site are combined. Although the sequences of AO and mRNA are reverse complementary, therefore do not provide new information by concatenation, the secondary structure of AO and that of its target mRNA at the binding site are almost certain different, and thus supply new information. However, such performance gain from concatenating oligonucleotide and mRNA does not hold for GS and GS+Triplet. If oligonucleotides and mRNAs are used alone, we noticed that mRNA give better performance for all three profiling methods: Triplet, GS, and GS+Triplet. Among all 9 variations considered in Table 1, the best performance (ROC = 0.973) was achieved when the GS and Triplet profiling are used in conjunction on mRNA. This performance is apparently better than that of the previous methods on the same data set [5, 6].

Table 1. ROC scores for various experiments

	Oligo	mRNA	Oligo+ mRNA
Triplet	0.935	0.955	0.960
GS	0.567	0.660	0.586
GS+Triplet	0.871	0.973	0.888

Profiling on parenthesis orientation sensitive triplets and quadruplets was also tested, but did not improve the performance, instead the performance decreased slightly (results not reported here). To a certain extent, such phenomena are not unusual. As the *N*mer's size increases, the profile vector dimension increases exponentially, and most of the *N*mers do not occur in the short sequence of AO, leaving many components of the resultant profile vector with zero value, which dilutes the information content. This is part of the consequences of so-called *dimensionality curse*, and may also be responsible for why GS+Triplet on Oligo+mRNA did not perform well as initially expected.

A closer look at the *N*mer's distribution can shed light on why the profiling on mRNAs outperforms that on oligonucleotides, and more importantly, may provide useful guidance for designing high efficacy AOs. In Figures 2 and 3, the distributions of all 32 triplets for oligonucleotides and mRNA target sites are displayed respectively. The occurrence frequency (Y-axis) of a given triplet (X-axis) is counted separately for high activity (>0.75) v.s. low activity (<0.25) oligonucleotides. On X-axis, from left to right, the first 8 triplets have nucleotide A in the middle position, the next 8 triplets have C, and the next 8 triplets have U, and the last 8 triplets have G. It can be seen that at mRNA target sites, triplets '(((' have higher occurrence frequency regardless what nucleotide is in the middle position, implying a more stable structure because more bases are paired up. When the middle nucleotide is considered, the target sites for AOs with high efficacy have about 23% chance of having a triplet '(((' with a base G in the middle, which is more than double the 11% for sites that correspond to low efficacy. On the contrary, high occurrences are observed for triplets '…' on the oligonucleotides, implying a less stable structure because fewer bases are paired up. No surprisingly, base C is the most prominent composition, due to the sequence complementarity between the AOs and the target mRNAs, because there are more Gs on mRNA as mentioned above. It is worth noting that, the difference between the frequencies of a triplet to occur in high efficacy and in low efficacy is more pronounced in mRNAs (Figure 3) than in oligonucleotides (Figure 2). This partly explains why profiling on mRNAs is more effective in discriminating high activity AOs from low activity AOs, than profiling on AOs. As the sequences are complementary between an oligonucleotide and its target mRNA, it is reasonable to think that the discriminative power of profiling on mRNAs comes from the secondary structure of mRNAs at the binding site, which, as part of the folding of the whole mRNA molecule, can be quite different from what the oligonucleotide alone can fold into. In other words, the secondary structure of mRNAs at the binding site carries some global structural information, which provides the rational behind the GS profiling.

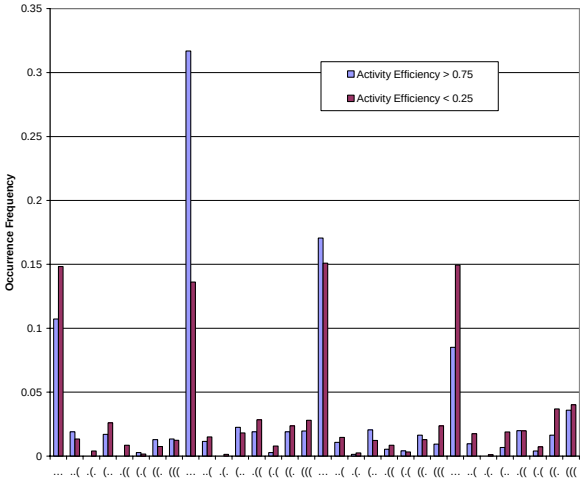

Figure 2. Distribution of occurrence frequency for triplets on oligonucleotides

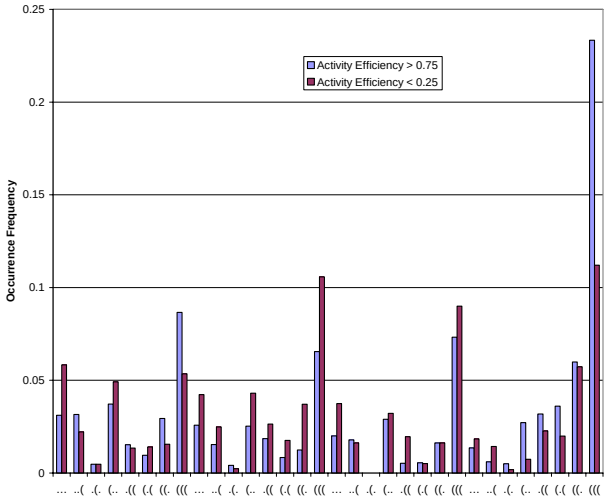

Figure 3. Distribution of occurrence frequency for triplets on mRNAs

4. Discussion

In summary, in this work, we have developed a new method to classify and predict oligonucleotides into two classes of high activity efficacy and low activity efficacy respectively, based on some local and global structural features. The results from the cross validation experiments indicated that, although these global features alone do not yield better performance, the classification accuracy was improved when they are used in conjunction with the local features. We also found that profiling on mRNAs is more discriminative than on oligonucleotides themselves in telling whether they will have high activity. What is more, the highly discriminative triplets can be useful guides in designing/selecting AOs for high activity efficacy.

Acknowledgments

This work was in part supported by a grant from the University of Delaware Research Foundation.

References

[1] Agrawal S. and Zhao Q. Antisense therapeutics in neuropharmacology. *Cur. Opi. Chem Biol* 1998; **2**:519-528.

[2] Myers K and Dean N. Sensible use of antisense: how to use oligonucleotides as research toos. *Trends Phamacol Sci* 2000; **21**:19-23.

[3] Matveeva OV, Tsodikov AD, Giddings M, Freier SM, Wyatt JR, Spiridonov AN, Shabalina SA, Gesteland RF, and Atkins JF. Identification of sequence motifs in oligonucleotides whose presence is correlated with antisense activity. *Nucleic Acids Res.* 2000; **28**:2862-2865.

[4] Ding Y and Lawrence CE. Statistical prediction of single-stranded regions in RNA secondary structure and application to prediction of effective antisense target sites and beyond. *Nucleic Acids Res*. 2001; **29**:1034-1046.

[5] Camps-Valls G, Chalk AM, Serrano-Lopez AJ, Martin-Guerrero JD, and Sonnhammer ELL. Profiled support vector machines for antisense oligonucleotide efficacy prediction. BMC Bioinformatics, 2004; 5:135.

[6] Giddings MC, Shah AA, Freier S, Atkins JF and Gesteland RF, Matveeva OV. Artificial neural network prediction of antisense oligodeoxynucleotide activity. *Nucleic Acids Res.* 2002; **30**:4295-4304.

[7] Xue C, Li F, He T, Liu G-P, Li Y, and Zhang X. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics* 2005; **6**:310.

[8] Nussinov R, Pieczenik G, Griggs JR, and Kleitman DJ. Algorithms for loop matching. *SIAM Journal of Applied Mathematics* 1978; **35**:68-82.

[9] Joachims T. Marking large-scale SVM Learning Practical. *Advances in Kernel Methods – Support Vector Learning,* B. Scholkopf and C. Burges and A. Smola (e.d), MIT Press, 1999.

[10] Gribskov M and Robinson N. Use of receiver operating characteristic analysis to evaluate sequence matching. *Computers and Chemistry*, 1996; **10**:25-33.

[11] Leslie SC, Eskin E, Cohen A, Weston J, and Noble WS. Mismatch string kernels for discriminative protein classification. *Bioinformatics* 2004; **20**:467-476.

[12] Cristianini N and Shawe-Taylor, J, *Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, UK, 2000.

[13] Vapnik V. *The Nature of Statistical Learning Theory*, Springer Verlag, 1995.