BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 8(23)

# Prediction of antisense oligonucleotides using structural and thermodynamic motifs

**Abdul Rahiman Anusha[1]* & Vinod Chandra[1, 2]**

[1]Department of Computational Biology and Bioinformatics, University of Kerala, Thiruvananthapuram - 695581, India; [2]College of Engineering Trivandrum - 695016, Kerala, India; Abdul Rahiman Anusha - Email: anushapraveenkhan@gmail.com; Phone: 0091-471-2515556; *Corresponding author

Received September 24, 2012; Accepted October 27, 2012; Published November 23, 2012

**Abstract:**
Specific gene expression regulation strategy using antisense oligonucleotides occupy significant space in recent clinical trials. The therapeutical potential of oligos lies in the identification and prediction of accurate oligonucleotides against specific target mRNA. In this work we present a computational method that is built on Artificial Neural Network (ANN) which could recognize and predict oligonucleotides effectively. In this study first we identified 11 major parameters associated with oligo:mRNA duplex linkage. A feed forward multilayer perceptron ANN classifier is trained with a set of experimentally proven feature vectors. The classifier gives an exact prediction of the input sequences under 2 classes – oligo or non-oligo. On validation, our tool showed comparatively significant accuracy of 92.48% with 91.7% sensitivity and 92.09% specificity. This study was also able to reveal the relative impact of individual parameters we considered on antisense oligonucleotide predictions.

**Background:**
Antisense Oligonucleotides (AOs) are short sequences with 7-30 nucleotides (nt) in length designed to bind a specific region of a target messenger RNA (mRNA). Ground theory that highlights the concept of antisense strategy is the use of a complementary sequence that can inhibit the expression of a specific mRNA. The binding of Antisense nucleotides to target mRNA is accomplished by standard Watson – Crick base pair interaction. When nucleotides pair up, specific gene expression occurs via different mechanisms such as RNase-H mediated cleavage, interface with splicing, translational arrest and prevention or destabilization of the target mRNA. Once target cell receives an AO through electroporation or microinjection, this gene expression is blocked or disabled through a reversible process called gene knock-down. Clinically AOs are proved to have immense significance in therapeutical field in the treatment of virtual diseases, cancer and inflammation **[1].** In addition, this helps to exploit the study of gene function and has been proposed as a strategy for systematic use in functional genomics **[2].** Advanced studies on antisense technology as a therapeutic agent and as gene expression modulation tool was started during late 1960's and 1970's. The report on inhibition of prokaryotic gene expression and viral replication using the potential of oligodeoxynucleotides blazed the light for such a beginning. The far and wide outburst of antisense technology was witnessed in the last two decades with the discovery of RNA interference (RNAi) mechanism. It enabled deactivation or silencing of a specific gene with the advancement in automated DNA synthesis and advances in the reel of nucleic acid chemistry. Emphasizing on this AO research group, use this as a new era technology worldwide in modern drug discovery. The antisense approach pioneered the transformation of costly and time consuming traditional drug designing methods to the present day low cost pharmaceutical inventions. Antisense oligonucleotides are finally paving the way of functional genomics as the most powerful experimental tools in the design of novel pathways and new gene specific drugs **[3].**

One basic question that arises while working with antisense oligonucleotides is how to select the exact AO from a cluster against a specific target mRNA. Literatures reported various experimental and theoretical methods in oligonucleotides prediction **[4–7].** In most cases the experimental in-vivo AO screening time-line and expense stands incomparably high. At this point, necessity of computational prediction approaches got

 

intensified and later on scientists were successful in their efforts.

Computational approaches eased many tedious works like structural predictions of target mRNAs, target pairing regions and promising sites for oligo binding **[8].** Consequently, while the experimental cost and AO testing duration got subdued, the efficacy achieved seemed to be significantly fair.

***Outline of existing computational methods***

The computational methods for oligonucleotide prediction can be classified into mathematical methods, motif discovery and machine learning techniques. Among mathematical methods, two major ones are statistical studies by sampling secondary structures and computational formalization for optimal oligonucleotide microarray synthesis. Charles revealed mathematical modeling of different cellular mechanistic events while an AO associates with its target **[9].** Mathematical models were created to describe antisense activity under steady state and dynamic conditions based on AO mass action kinetics. Itsik et al. discussed a polynomial algorithm that reconstructs long DNA targets by universal oligonucleotides **[10].** A work reported that ten sequence motifs have been identified with significant correlation coefficient value for oligo activity **[11].** Discovery of motif sequences incorporated with soft computing approaches like artificial neural network enables better understanding of factors affecting oligo predictions **[12].** Infact AOs can be designed and synthesized in laboratories by motif discovery **[13].**

Soft computing approaches like ANN, SVM and HMM are popularly good classification approximation and knowledge discovery machine learning techniques for efficient AO prediction. Gustavo *et al.* proposed an SVM based AO prediction and efficacy analysis using correlation analysis, the mutual information feature selection and recursive feature elimination **[14].**

**Methodology:**

***Dataset preparation***

We collected the data from published literatures **[11, 12, 14, 15].** We used experimentally validated oligo sequences in the training set. We defined a scoring system with a range 0 – 1 to measure oligo: target mRNA binding intensity. We fixed '0.5' as cut-off score value with the consideration that oligos that fall above '0.5' tend to have stable binding and those fall below '0.5' show poor binding. We set a score value of '1' for absolute oligo: mRNA locking and '0' for no locking. Thus positive dataset survived with those successful oligos based on cut-off after filtration and classification process. We ignored those oligos that fell below cut-off score assuming poor binding.

The required negative dataset was generated after manual cross validation of sequences. Since randomly generated sample sequences could create ambiguities, they are completely avoided from the training set. Target nucleotide positions in the target sequences were changed either by insertion or deletion. Nearly 1 to 3 position changes of this kind were done per sequence in order to generate large negative samples. We then aligned all training dataset sequences. The good stand alone negative dataset was built up by removing those sequences that showed repetition and any kind of matches after sequence alignment. Thus we had a total of 423 oligo patterns out of which 180 served as positive and 243 served as negative dataset.

***Parameter calculation***

Several parameters associated with oligo:mRNA hybridization was identified and examined for its behavioral function to the duplex formation using Principle Component Analysis (PCA). We could find a set of 11 most relevant parameters as the outcome of PCA scaling results listed in **Table 1 (see supplementary material)**. We categorized these selected features under two classes – structural parameters 1 to 7 in **Table 1** and thermodynamic parameters 8 to 11 in **Table 1**. Structural Parameters show the potency of oligo binding based on local secondary structural analysis of the target and thermodynamic parameters allow a precise prediction of oligonucleotide stability. All these calculations are done using nearest neighbour (NN) method. **Table 2 (see supplementary material)** presents the thermodynamic nearest neighbor (NN) parameters for Watson- Crick base pairs in 1 M NaCl.

***Method for Generating AO***

***Sliding Window by Matrix Expansion***

Inorder to generate oligonucleotides, we used a method - sliding window by matrix expansion. The method defines a window of nucleotides with a predefined constant value 5 as window size. When this window is slided across the target sequence, oligonucleotide sequences are generated from one end to the other end. The exact complementary sequence to the mRNA subsequence is generated and is validated against its parameter cutoffs. If values are above the cutoff, subsequence is considered as an oligonucleotide and subsequently added to the oligonucleotide set. The generated oligo sequence is randomly changed to produce three oligo nucleotide matrices. Again, each of these oligo is verified against the specified criteria described in the parameter selection and resulting sets are generated accordingly.

After completing the matrix expansion, window is slides toward right to the next position of the target mRNA. Validation process is done on each oligo generated and the successful ones are added to the set. The whole process is repeated and continued till the required oligos are generated or end of the mRNA is reached.

***Training Network Architecture***

We have trained a multilayer feed forwarded Artificial Neural Network (ANN) with error back propagation algorithm for validating the generated AOs. The ANN's input layer has 11 nodes to feed 11 selected oligo features, hidden layer has 8 nodes fixed by trial and error method and output layer has 1 node to measure the desired output as a score of either 0 (low) or 1 (high). That means predictor scores 1 for an 'oligo' and 0 for 'non-oligo'. The ANN is trained with a total of 290 oligo sequences consisting of 105 positive dataset having their experimentally determined mRNA targeting activities and 185 negative dataset which are manually verified for its exclusion from the positive dataset. Thus, out of 423 patterns obtained during dataset preparation, 290 were used as training set and rest patterns were kept as testing dataset. The learning rate is set as o.2. Increasing sigmoid function is chosen as the activation function since these functions are mathematically well behaved and enables smooth transition between 0 and 1.

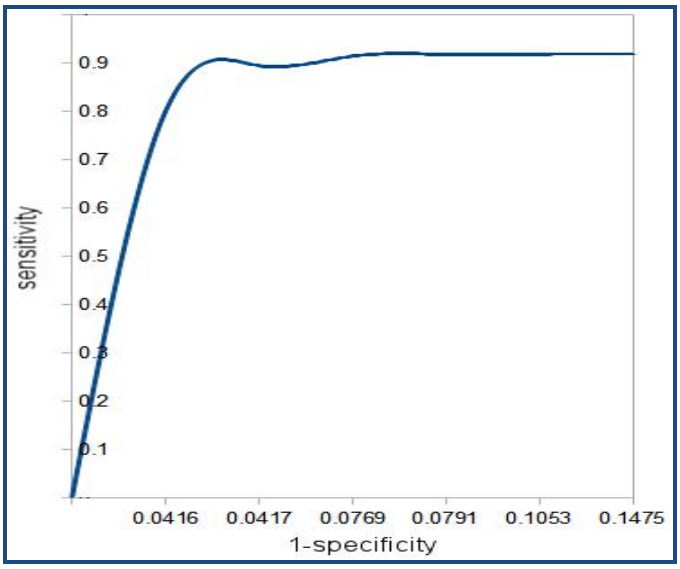


**Figure 1:** ROC Curve of learning rate validation performance.

**Results and Discussion:**

Our tool predicts antisense oligonucleotides for a given set of input sequence. In order to study the response of the trained neural network, we have conducted 2 different levels of cross validation examinations. Both validations were done using the test dataset of 133 elements composed of 78 positive and 55 negative samples that were kept aside during ANN training. Test examples were carefully verified for its exclusion from the training sequences.

The tool efficiency accounts for the measure of overall predictor performance expressed in the standard terms, sensitivity and specificity. The parameter, accuracy ranks the quality of test performance by the predictor. Sensitivity (Sn) is the ratio of true positive to sum of true positive and false negative and specificity (Sp) is the ratio of true positive to sum of true positive and false positive. Accuracy (Acc) accounts for the proportion of true data in the total testing data set. Tool performance is rated high when both sensitivity and specificity measures show high scores. Each term is expressed as shown in **supplementary material.**

***Validation with learning rates***

First level validation was carried out for different values of learning rate, $\eta$. The system performance was analyzed and compared during every training period. **Table 3 (see supplementary material)** depicts the test result of five training sessions to find ideal oligos against target mRNA. We included only those five epochs in the table 3 since they gave significant values for the prediction. The test showed enhanced performance at lower learning rates from which could conclude that our model is an efficient one. These differential effects could be best viewed with the help of ROC curves. ROC graph plots true positive rate (sensitivity) along y-axis and false positive rate (1- specificity) along x-axis. The area under the curve corresponds to the measure of model accuracy. When plotted, we could find that the maximum area coverage under the curve is 92.48%. This is considered as the best prediction result with 92.09% specificity and 91.7% sensitivity for a threshold 0.92. **Figure 1** shows ROC plot of the same. For other values of learning rates, ROC area showed decrement but promising accuracy.

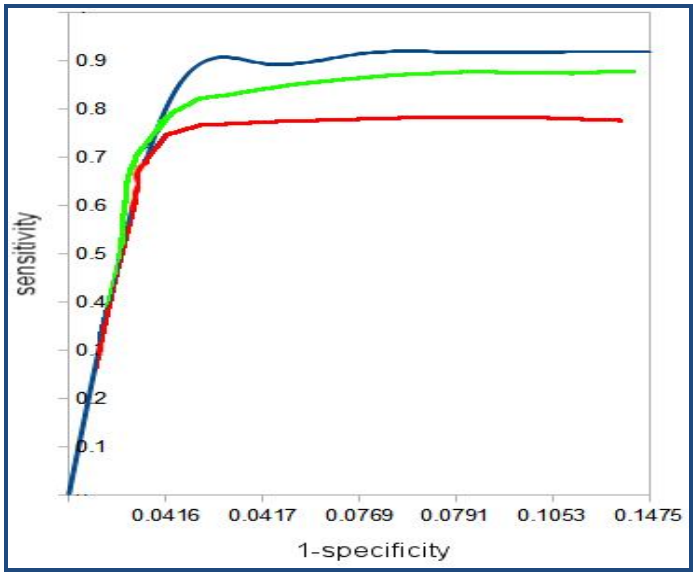


**Figure 2:** Performance comparison summary of validations (a) Blue line indicates the performance when all features are included. (b) Green line indicates the performance for one feature replacement to ANN input. (c) Red line indicates the performance for two parameter replacement to ANN input. Area under the curve

***Validation on parameter impacts***

Second level testing was performed based on the analysis of effects of selected features. The same first test examples were used for this analysis. We examined the relative impact of individual parameter on the overall system performance on a 10 x 1 basis. That is, we made changes in the ANN input vectors by replacing one known feature out of eleven in the parameter set with an unknown feature. Each time the network performance was measured with same training and testing dataset. Resultant accuracies are compared with the accuracy produced while all parameters are in the dataset (92.48%). It is seen that accuracy attained during each feature change over is notably different. Depending on this relative comparison, we ranked the influence of each parameter in the prediction process and are listed in **Table 4 (see supplementary material)**. It is obvious that thermodynamic parameters show immense impact than structural parameters in antisense oligonucleotide predictions. We carried out an evaluation to know the response of this system when combinations of only two parameters are considered. Validation yielded decreasingly varying accuracy results depending on parameter combination constrains. The summary of various performance outcomes of all validation tests are shown in **(Figure 2).**

Our study started with 2 main objectives related to antisense oligo nucleotides. One was to design a computational tool for predicting the efficiency of a given antisense oligonucleotide and second was to predict antisense oligonucleotide that binds to target mRNA with high efficiency. We availed good results for both the objectives. Meanwhile, other than these two primary objectives, we were able to produce two more results. The first is a provision to filter mRNA by specifying the value

of each parameter. Another one is to find out the value of each parameter corresponding to an oligonucleotide.

Our system performance was compared with the existing initiatives in computational AO prediction techniques based on Artificial Neural Network. A previous approach using ANN classifier generated 92% success rate. When compared with this, our work highlights enhanced performance as well as ranking of both thermodynamic and structural features associated with AO predictions.

**Conclusion:**

We have derived an antisense prediction methodology that might help to obtain highend systematic knockdown of targets. An artificial neural network classifier was trained with experimentally validated dataset. Eleven parameters were identified and fed to ANN to get optimal output. We obtained appreciable system performance in terms of sensitivity (91.7%), specificity (92.09%), and accuracy (92.48%). The role and impact of some relevant thermodynamic and structural parameters in AO predictions are estimated. Our method predicts AO's under two categories – low efficacy AO's and high efficacy AO's. Soft computing community could implement other classification techniques to maximize sensitivity. The computational technique used here within could be expanded or coupled to identify new and better oligomers to impart better translational block depending on individual properties. However, experimental validation holds the final statement in digging out the hidden potentials of antisense oligonucleotide. We believe that our work could be a reference to future researchers in this field who tries to bridge the gap between computational and experimental strategies of antisense oligonucleotide predictions.

**References:**


**[1]** Opalinska JB & Gewirtz AM, *Nat Rev Drug Discov.* 2002 **1**: 503 [PMID: 12120257]
**[2]** Dean NM, *Curr Opin Biotechnol.* 2001 **12**: 622 [PMID: 11849945]
**[3]** Huber LC *et al. Adv Drug Deliv Rev.* 2006 **58**: 285 [PMID: 16574269]
**[4]** Allawi H *et al. RNA.* 2001 **7**: 314 [PMID: 11233988]
**[5]** Ho SP *et al. Nucleic Acids Res.* 1996 **24**: 1901 [PMID: 8657572]
**[6]** Ho SP *et al. Nat Biotechnol.* 1998 **16**: 59 [PMID: 9447595]
**[7]** Milner N *et al. Nat Biotechnol.* 1997 **15**: 537 [PMID: 9181575]
**[8]** Chandra V *et al. BMC Bioinformatics.* 2010 **11 Suppl 1**: S2 [PMID: 20122191]
**[9]** Roth CM, *Biophys J.* 2005 **89**: 2286 [PMID: 16055530]
**[10]** Pe'er I *et al. Proc Natl Acad Sci U S A.* 2002 **99**: 15492 [PMID: 12429861]
**[11]** Matveeva OV *et al. Nucleic Acids Res.* 2000 **28**: 2862 [PMID: 10908347]
**[12]** Giddings MC *et al. Nucleic Acids Res.* 2002 **30**: 4295 [PMID: 12364609]
**[13]** Adams AM *et al. BMC Mol Biol.* 2007 **8**: 57 [PMID: 17601349]
**[14]** Gustavo CV *et al. BMC Bioinformatics.* 2004 **5**: 135 [PMID: 15383156]
**[15]** Xiaochen BO *et al. Nucleic Acids Res.* 2006 **34**: D664


**Edited by P Kangueane**

**Citation: Anusha & Chandra,** Bioinformation 8(23): 1162-1166 (2012)

BIOINFORMATION

open access

# Supplementary material:

### Term used in discussion:

$$\text{Sensitivity (Sn)} = \frac{TP}{(TP+FN)}$$

$$\text{Specificity (Sp)} = \frac{TP}{(TP+FP)}$$

$$\text{Accuracy (Acc)} = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

**Table 1: Selected parameters for oligo validation. Structural parameters (Numbered 1 to 7) and thermodynamic parameters (Numbered 8 to 11)**

| No | Parameter | Parameter Description |
|---|---|---|
| 1 | Length | Sequence length of oligonucleotide |
| 2 | GC Content | Percentage of guanine or cytosine residues in oligo sequence |
| 3 | AU Content | Percentage of Adenine or Uracil (Thymine) in oligo sequence |
| 4 | Molecular Weight | Sum of molecular weights of each of the nucleotides |
| 5 | Seed Score | Sum of pair scores in the seed region |
| 6 | GC Pairs | Proportion of total G:C pairs in oligo sequence |
| 7 | AU Pairs | Proportion of total A:U pairs in oligo sequence |
| 8 | Enthalpy (ΔH) | Change in enthalpy of oligo base stacking interactions adjusted for helix initiation factors in kcal / mol |
| 9 | Entropy (ΔS) | Change in entropy of oligo base stacking adjusted for helix initiation factors and for the contributions of salts to the entropy of the system in kcal $K^{-1}$ $mol^{-1}$ of interaction |
| 10 | Free Energy (ΔG) | Minimum free energy called Gibbs free energy of the sequence |
| 11 | Melting Temperatures | Temperature at which 50% of the oligonucleotide and its perfect complements are in duplex |

**Table 2: Thermodynamic parameters for nearest-neighbour melting temperature**

| Propagation Sequence | ΔH° (kcal/mol) | ΔS° (e.u.) | ΔG° (kcal/mol) |
|---|---|---|---|
| AA / TT | -7.6 | -21.3 | -1.00 |
| AT / TA | -7.2 | -20.4 | -0.88 |
| TA / AT | -7.2 | -21.3 | -0.58 |
| CA / GT | -8.5 | -22.7 | -1.45 |
| GT / CA | -8.4 | -22.4 | -1.44 |
| CT / GA | -7.8 | -21.0 | -1.28 |
| GA / CT | -8.2 | -22.2 | -1.30 |
| CG / GC | -10.6 | -27.2 | -2.17 |
| GC / CG | -9.8 | -24.4 | -2.24 |
| GG / CC | -8.0 | -19.9 | -1.84 |
| Initiation | +0.2 | -5.7 | +1.96 |
| Terminal AT penalty | +2.2 | +6.9 | +0.05 |
| Symmetry correction | 0.0 | -1.4 | +0.43 |

**Table 3: Prediction performance with respect to learning rates**

| Learning rate, η | Sensitivity (*Sn*) (%) | Specificity (*Sp*)(%) |
|---|---|---|
| 0.30 | 40.08 | 96.50 |
| 0.24 | 69.01 | 96.03 |
| 0.15 | 81.00 | 95.01 |
| 0.11 | 82.90 | 94.00 |
| 0.01 | 85.10 | 93.20 |
| 0.005 | 83.84 | 93.95 |

**Table 4: Ranking of parameters**

| Rank | Selected Parameters |
|---|---|
| 1 | Gibbs free energy |
| 2 | G:C content |
| 3 | Entropy |
| 4 | Melting Temperature |
| 5 | Seed Score |
| 6 | Enthalpy |
| 7 | G:C pair |
| 8 | A:U pair |
| 9 | A:U content |
| 10 | Molecular weight |
| 11 | Length |