INNOVATION & PRECISION
IN NUCLEIC ACID SYNTHESIS
IDT®
INTEGRATED DNA TECHNOLOGIES

Identification of motifs that significantly associate with antisense sequence activity

Andrew S. Peek, Meredith L. Patterson, Justin Garretson, Yuan Lin,
Jeffrey A. Manthey, and Yihe Wu
Bioinformatics
Integrated DNA Technologies, Inc.

Introduction

Antisense oligonucleotides are able to modulate gene expression levels by a post-transcriptional mechanism involving an RNA:DNA heteroduplex and the cleavage of the RNA by RNase H (1,2). This sequence-specific hybridization between an antisense molecule and its target can, in principle, allow the inhibition of any target mRNA without affecting closely related genes. One approach for identifying antisense sequences that knockdown a target gene to desired levels is to screen many candidate antisense molecules to find the most active antisense sequences. In addition to antisense screening by trial and error, several computational approaches have been developed to improve the hit rate of selecting active antisense sequences (3-14). These models all associate some factors of an antisense sequence with the antisense sequence's activity (sequence ~ activity). Methods that identify factors to incorporate into a sequence ~ activity model provide the opportunity not only provide the opportunity to improve the predictive models and avoid factors that unnecessarily introduce complexity, but also are important in contributing to a more complete understanding of the antisense knockdown mechanisms of action.

The factors responsible for antisense activity have been shown to involve oligonucleotide stability, oligonucleotide secondary structure, target sequence structure and accessibility, bioavailability, cell type, and nucleotide sequence motifs (9,14-16). When designing antisense sequences to effectively knockdown gene expression or attempting to predict the activity of an antisense oligonucleotide, it is crucial to know the factors and their association with activity. Previous studies of sequence motifs that associate with antisense activity in a sequence ~ activity model have identified a small number of motifs that significantly associate with activity (8,13), but a systematic survey of a broad range of motif lengths associating with sequence activity in a publicly available dataset has not been performed. The objective of this study is to identify nucleotide sequence motifs that are non-randomly associated with antisense activity in order to more accurately build a sequence ~ activity model for antisense molecules.

Materials and Methods

Database

We chose two criteria for selecting sequences for inclusion into the antisense sequence activity database. First, the antisense sequence contained a complete phosphorothioate backbone. No mixed or chimeric sequences were included. Second, the cellular levels of antisense activity needed to be measured by either a direct mRNA or protein product method. Antisense sequences and their associated activities were obtained from a previously described antisense database(17), or the USPTO database (http://www.uspto.gov/). The antisense sequence activity dataset can be obtained from the authors.

Motif

A word is a series of symbols from an alphabet juxtaposed, and motifs are subwords (or substrings) that comprise a word. For example, using the alphabet (18T) the word of length 4, *TTGC*, contains 2 motifs of length 3, *TTG* and *TGC*, and 3 motifs of length 1, *C*, *G* and *T*. The total number of possible motifs is then N^L, where N is the number of characters in the alphabet and L is the motif length.

Calculations and statistics

The *t*-test and chi-square test probability calculations were performed and graphs were plotted with the R statistical package (http://www.R-project.org). Directional graphs were generated with the GraphViz package (http://www.research.att.com/sw/tools/graphviz/). The Weka machine-learning package (http://www.cs.waikato.ac.nz/ml/weka) was used to build decision trees to predict sequence effectiveness.
Change in equilibrium Gibbs free energies (δG) were calculated by the nearest neighbor method (19) for change in enthalpy (δH) and entropy (δS)

(1) $\delta G = \delta H - (\delta S \times 298.15°K)$

and the bit-wise information content was calculated by the method described by Shannon for negative entropy.

(2) $-\sum_i p_i \log_2 p_i$

Programs written by the authors implemented the algorithms and other calculations described.

Results

We assembled a database of 3913 antisense sequences with associated cellular activities for knocking down their target gene. The activities ranged from 0.0 to 1.0, where an activity of 0.0 indicates complete inhibition of the target gene and 1.0 is no difference in target activity when compared to the appropriate control. Not surprisingly, the distribution of activities is asymmetrically skewed ($g_1 = -0.520$, H_0: $g_1 = 0$, $t_s = 13.3$, $P < 1 \times 10^{-10}$), with fewer sequences having activities near 0.0, and many sequences with activities at or near 1.0. Also, the activity distribution of the 3913 activities is highly platykurtic ($g_2 = -0.936$), with fewer sequence activities near the mean than when compared to a normal distribution. Of the 3913 sequences, 1130 had an activity of 1.0, and even the remaining 2783 sequence activities were skewed (g_1 = -0.233, H_0: g_1 = 0, t_s = 5.02, $P = 5 \times 10^{-7}$), with fewer sequences having activities near 0.0 when compared to a normal distribution. Furthermore, the remaining 2783 activities were also platykurtic (g_2 = -0.873).

To help identify sequence motifs that are associated with effective (activities nearer 0.0) versus ineffective (activities near 1.0) antisense sequences, we divided sequences into 2 groups: the 944 high effectiveness sequences with activities ranging from [0.0, 0.5) and the 2924 low effectiveness sequences with activities from [0.5, 1.0]. We then tried to identify unique motifs that discretely discriminate between more effective and less effective sequences. Looking for sequence motifs, unique for one group or another, can result in one of 4 possible outcomes for a single motif. A motif can occur only within high effectiveness sequence population (group A), only within low effectiveness sequence population (group C), within both high and low effectiveness populations (group B), or finally, within neither population (group D). The total number of possible motifs (group E) is E = A + B + C + D.

Motifs of lengths 1 through 3 were not discretely distributed between effective and ineffective sequences (Table 1). One sequence motif of length 4 was discretely distributed only in the ineffective sequences, specifically the tetra-nucleotide sequence motif “GGGG”. Motifs of length 5 were further able to discriminate ineffective sequences, and 21 motifs were found uniquely within ineffective antisense sequences. Furthermore, increasing motif lengths allowed for increasing discrimination between effective and ineffective sequences, but as motif length increases, it is difficult to know whether these motifs are being associated with sequence activity due to some actual relationship to activity, or simply due to the increasing number of single occurrences in either the high-or-low-effectiveness categories and the relatively small (3913) number of initial sequence activity observations.

Table 1. Length distribution of motifs in effective and ineffective antisense sequence populations A = effective sequences, B = effective and ineffective sequences, C = ineffective sequences, D neither effective nor ineffective sequences, E = all possible motifs

A	B	C	D	E	length
0	4	0	0	4	1
0	16	0	0	16	2
0	64	0	0	64	3
0	255	1	0	256	4
0	1003	21	0	1024	5
37	3219	768	72	4096	6
912	5763	6209	3500	16384	7
4039	4548	16595	40354	65536	8
6554	2229	22654	230707	262144	9
7136	1140	23347	1016953	1048576	10

In order to examine whether there are any physical properties that can be used to identify sequence motifs that occur discretely in high-effectiveness or low-effectiveness sequences, we calculated the equilibrium change in Gibbs free energy of hybridization (δG) of the motif categories in Table 1. These are shown in Table 2. Overall, the δGs of each motif length category associated only with high-effectiveness sequences (group A) are uniformly (sign test, $P = 0.03125$) and significantly more negative when compared with motifs found only in low-effectiveness sequences (group C) (5 paired *t*-tests, maximum $P = 2 \times 10^{-6}$), for the motifs lengths 6 through 10. Additionally, comparing the δGs between motifs found only in effective sequences (group A) against all motifs (group E) showed a significantly more negative value in effective sequences (5 paired *t*-tests, maximum $P = 0.006$). Finally the δGs in motifs found only in low-effectiveness sequences (group C) are overall more positive than all motifs (group E), but these differences are not significant for motif lengths 9 and 10. These patterns are consistent with previous observations demonstrating an overall δG difference between high-and-low-effectiveness antisense sequences, but additionally suggests that this observation is not only true for the entire antisense sequence, but also for the motifs that comprise an antisense sequence.

Table 2. Length distribution of motifs in effective and ineffective antisense sequence populations δG in units cal mol^{-1}, Standard Deviation given in parentheses.

A	B	C	D	E	length
-	-	-	-	-	1
-	335.8	-	-	335.8	2
	(18107)			(18107)	
-	-1304	-	-	-1304	3
	(13623)			(13623)	
-	-3012	-4330	-	-3017	4
	(6875)			(6902)	
-	-4701	-4352	-	-4693	5
	(1291)	(91128)		(2905)	
-6731	-6466	-5938	-6493	-6370	6
(14894)	(545)	(5853)	(69883)	(1097)	
-8300	-8277	-7811	-8018	-8047	7
(2227)	(276)	(1184)	(1767)	(386)	
-10049	-10074	-9621	-9693	-9723	8
(408)	(385)	(698)	(205)	(129)	
-11781	-11886	-11360	-11388	-11400	9
(142)	(404)	(516)	(47)	(41.7)	
-13522	-13681	-13064	-13073	-13076	10
(225)	(2641)	(594)	(13)	(13)	

To further examine the physical properties of motifs that occur only in high-effectiveness or low-effectiveness sequences, we calculated the Shannon entropies of the categories in Table 1. These calculations are shown in Table 3. Overall, the high-effectiveness motifs (group A) had greater bit-wise information content when compared with motifs found only in low effectiveness sequences (group C). Statistical significance is weak for lengths 6, 7 and 8 (*t*-tests, $P = 0.06, 0.04, 0.01$), but more apparent with lengths 9 and 10 (*t*-tests, $P = 1.0 \times 10^{-6}, 2.9 \times 10^{-4}$). Patterns of bit-wise information content are less clear when comparing motifs found only in high-or-low-effectiveness sequences to all motifs (group E) and no trend is apparent.

Table 3. Length distribution of motifs in effective and ineffective antisense sequence populations Entropy, Standard Deviation given in parentheses

A	B	C	D	E	length
-	-	-	-	-	1
-	1.5 (0.1406)	-	-	1.5 (0.1406)	2
-	3.254 (0.1655)	-	-	3.254 (0.1655)	3
-	5.245 (0.1079)	0	-	5.225 (0.1066)	4
-	7.375 (0.0191)	5.550 (1.4668)	-	7.337 (0.0525)	5
10.264 (0.0457)	9.501 (0.0129)	9.654 (0.0567)	9.172 (1.168)	9.531 (0.0221)	6
12.156 (0.0481)	11.512 (0.0062)	11.667 (0.0060)	12.235 (0.0427)	11.761 (0.0084)	7
13.957 (0.0092)	13.438 (0.0005)	13.664 (0.0037)	14.209 (0.0050)	14.003 (0.0029)	8
15.947 (0.0008)	15.398 (0.0088)	15.710 (0.0009)	16.320 (0.0011)	16.255 (0.0010)	9
17.997 (0.0008)	17.477 (0.0553)	17.799 (0.0017)	18.535 (0.0003)	18.512 (0.0003)	10

Some motifs are able to discriminate high-effectiveness and low-effectiveness sequences, and these motifs have distinct physical properties in their δGs and bit-wise information content, but even with a dataset of 3913 sequences, motifs of length 6 become rare in the dataset (group D). Increasing motif lengths became increasingly rare in the dataset and by motif length 8 the majority of motif observations are not seen in the dataset. In order to more closely examine motif frequency distributions in high effectiveness and low effectiveness sequences, we asked whether any motifs were overly distributed in high-effectiveness versus low-effectiveness sequences. To address this question, we developed a Monte Carlo procedure to compare the frequency distribution of a motif in the original dataset partitioned into high-activity and low-activity groups to a randomized sequence activity dataset. The randomized dataset contained the same number of sequence and activity pairs as the original dataset, and the randomized sequences were drawn from the original dataset's sequence length and base content distributions. The randomized activities were drawn from the original dataset's activity distribution. By repeating this method 10^5 times, we built a null distribution of a motif's abundance in high-effectiveness and low-effectiveness sequences, and asked whether the observations made in the original dataset were typical for the motif or if the distribution was biased towards high- or low-effectiveness sequences.

We used this Monte Carlo method to examine the 1364 motifs of lengths 1 through 5. The probability of their overabundance in high-effectiveness sequences by chance is displayed in figure 1. 156 motifs occur in high-effectiveness and 213 motifs occur in low-effectiveness sequence more than expected when compared to datasets drawn from the same underlying distributions ($\alpha = 0.05$, two-tailed). While 156 and 213 are both large proportions of the total number of motifs, 11% and 15% respectively, the occurrence of motifs across antisense effectiveness are indistinguishable from normally distributed (figure 1). The overall base distributions in motifs associated with high-effectiveness sequences deviates from random, with cytosines higher than expected by chance (C=151, T=95, G=93, A=78). The motifs are shown in Table 4. In contrast, the bases in motifs associated with low-effectiveness sequences contain a higher proportion of adenines (A= 185, T=157, G=107, C=84). These motifs are shown in Table 5. These observations and motifs are consistent with previous observations concerning antisense sequences, suggesting this method may be useful in identifying motifs significantly associated with sequence activity, either positively or negatively (8,13).

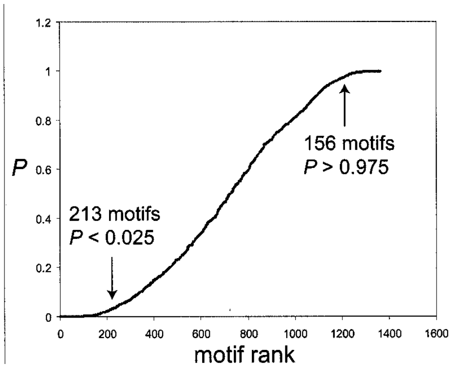

Fig. 1. Antisense activity association of 1364 sequence motifs.

Table 4. 156 Motifs Significantly Associated with high effectiveness antisense sequences, grouped by length

1	2	3	4				5				
C	CA	ACC	AACC	CACC	GCCA	TACC	AAACG	CAACC	CCGTG	GAACG	TACCA
	CC	CAC	AACG	CACT	GCCC	TCCC	AACCA	CAAGC	CCTAC	GACCG	TACCC
	CG	CCA	ACCA	CATC	GCGT	TCCG	AACCC	CACCA	CCTCC	GATAG	TCCAC
	CT	CCC	ACCC	CCAC	GTCC	TCGT	AACGA	CACTC	CCTGT	GCAAG	TCCCC
	GC	CCG	ACCT	CCAT		TCTC	ACCAT	CATCC	CCTTG	GCCAC	TCCCG
	TC	CCT	ACTC	CCCA		TGCT	ACCCT	CATCG	CGACC	GCCTC	TCCCT
		CGC	AGCC	CCCC		TGTC	ACCGG	CATGT	CGACG	GCGAC	TCCGC
		CGT	AGCT	CCCG		TGTG	ACCTA	CCAAC	CGATG	GCGCT	TCCGT
		CTC	ATCG	CCCT		TTGC	ACGAA	CCACC	CGCAA	GCGTC	TCGTC
		GCC		CCGC			ACGCA	CCACG	CGCCC	GCGTT	TCTCG
		GCT		CCTC			ACGCG	CCACT	CGCCT	GGCCA	TGATA
		TCC		CGCC			ATACC	CCATC	CGCGA	GGCGT	TTGCG
		TGC		CGCT			ATCCC	CCCAA	CGGTA	GGCTA	TTGGC
				CGTC			ATCGC	CCCAC	CGTAC	GGGCC	
				CGTG			ATCTC	CCCAT	CGTCC	GGTCC	
				CTCC				CCCCA	CGTGA	GTCAC	
								CCCCC	CGTGC	GTCCC	
								CCCGC	CTACC	GTGAG	
								CCCTC	CTCAG	GTGTG	
								CCCTG	CTCCC		
								CCCTT	CTCCG		
								CCGAC	CTCGT		
								CCGCC	CTCTC		
								CCGCG	CTGTG		
								CCGCT	CTTGC		
								CCGGG			

To further determine whether this method identifies motifs that have properties identified previously in high- and low-effectiveness sequences, we again examined the physical properties of the identified motifs. The 156 motifs abundant in high-effectiveness sequences had a lower δG when compared to the 213 motifs in low-effectiveness sequences (ave δG = -4503.26, sd = 76.07 versus ave δG = -3062.22, sd = 40.69 in the two groups respectively; $t = 16.4$, $P < 1 \times 10^{-37}$). The distribution of motif

Table 5. 213 Motifs Significantly Associated with low activity antisense sequences, grouped by length and sequence

1	2	3	4		5				
A	AA	AAA	AAAA	GAAA	AAAAA	ATAAA	CGTGT	GGGGT	TCGAC
	AT	AAG	AAAG	***GACG***	AAAAG	***ATAAC***	CTATA	GGTAG	TCTAG
	GA	AAT	AAAT	GATT	***AAACA***	ATAAT	***CTATC***	GGTTA	***TCTTA***
	GG	AGG	***AACA***	***GCGG***	AAAGA	***ATACG***	CTATT	GTAAG	TGCGA
	TA	ATA	AAGA	GGAA	AAAGT	ATAGA	CTCAC	GTAGG	TGGAT
	TT	ATT	AATA	***GGAC***	AAATA	ATAGG	CTGGG	GTGGG	TGGGG
		GAA	***AATC***	GGGA	AAATG	ATATA	CTTAC	GTGTA	TTAAA
		GGA	AATG	GGGG	AAATT	ATATT	CTTCG	GTTAT	TTAAC
		GGG	AATT	GTTA	***AACAA***	***ATCCG***	GAAAA	TAAAA	TTAAT
		TAA	***ACAA***	TAAA	AACTA	ATCGA	GAAAG	***TAAAC***	TTACA
		TAG	ACTA	TAAT	AAGAA	ATTAA	GAATA	TAAAG	TTACG
		TAT	AGAA	***TACG***	AAGAT	ATTAG	***GACGA***	TAAAT	TTACT
		TTA	AGAT	TACT	***AAGCG***	ATTAT	GAGAA	TAATA	TTAGT
		TTT	AGGA	TAGA	AATAA	ATTGA	GATAA	***TAATC***	TTATA
			AGGG	TAGG	AATAG	ATTTA	***GATCA***	TAATG	TTATT
			ATAA	TATA	AATAT	ATTTG	GATTA	***TAATT***	***TTCGA***
			ATAT	TATT	***AATCA***	ATTTT	***GCGCG***	TACGT	TTCTT
			ATTA	***TCAA***	AATCT	***CAAAT***	GCGGC	TACTA	TTTAA
			ATTT	TCTA	AATGG	CAATA	GCGGG	TACTT	***TTTAC***
			CAAT	TCTT	AATTA	CAATT	GGAAA	TAGAA	TTTAT
			CATA	TGGG	AATTG	CATTA	GGAAT	TAGAT	TTTTA
			CGGC	TTAA	AATTT	CATTT	***GGACA***	TAGTT	TTTTT
			CTAT	***TTAC***	***ACAAA***	CCGGC	GGACG	***TATAA***	
				TTAT	ACAAT	CGAGC	GGATC	TATAT	
				TTTA	ACATA	CGAGT	GGCGG	TATGA	
				TTTT	ACTAC	CGATA	GGGAC	TATTA	
					ACTAT	CGCTA	GGGCG	TATTG	
					AGAAA	CGGGG	GGGGA	TATTT	
					AGGGA	CGGTT	***GGGGC***	TCAAA	
					AGGGG	CGTAA	GGGGG	TCAAT	

δGs from all motifs, as well as motifs associated with high and low effectiveness, is consistent with previous conclusions of longer discretely distributed motifs, with more negative δGs associating with high-effectiveness antisense sequences.

Shorter length motifs can be subwords of longer motifs, and we examined whether any of the 213 motifs identified in low-effectiveness sequences were subwords of the 156 motifs in high-effectiveness sequences, and vice versa. Of the 156 motifs in high-effectiveness sequences, 67 did not contain subwords from the motifs found in low-effectiveness sequences. Conversely, of the 213 motifs in low-effectiveness sequences, 129 did not contain subwords from the motifs found in high-effectiveness sequences. The majority of cross-population subwords removed resulted from the single base motifs, “C” from the high-effectiveness motifs and “A” from the low-effectiveness motifs. Again, these subword-unique motifs are consistent with previous observations for physical properties. The 67 motifs (ave δG = -4249.69, sd = 17301) unique in high-effectiveness sequences had a lower δG when compared to the 129 motifs (ave δG = -2392.67, sd = 2575) unique in low-effectiveness sequences ($t = 13.1$, $P < 1 \times 10^{-22}$). However, when comparing the physical properties of these unique motifs with all motifs of length 1 through 5 (ave δG = -4143.2, sd = 1010), there is no significant difference in δG between the 67 motifs unique in high-effectiveness sequences ($P = 0.43$). By contrast, the 129 unique motifs have a significantly higher δG when compared to all motifs of lengths 1 through 5 ($t = 29.2$, $P < 1 \times 10^{-50}$).

The motifs identified by this Monte Carlo method are not a random sample from all possible motifs. Relationships between motifs of distinct lengths can be represented as a directional graph, where a connection between motif nodes is made when a motif is a subword of another motif. An example of such a graph is presented in figure 2. Graph connectivity in the 156 and 213 motifs found in effective sequences and in the 67 and 129 motifs found in ineffective sequences is significantly higher than connectivities for any motif dataset randomly chosen from all possible motifs of lengths 1 through 5.

The distribution of motifs across an antisense sequence may play a role in the sequence’s effectiveness. To examine whether the identified motifs deviated from a null expectation of uniform distribution across a sequence, we divided the original sequence strings into 2 regions: an inner region and an outer region. The region within the sequence was then used as the first axis of a 2 x 2 contingency analysis, and the second axis was sequence activity. Analyses were performed using several criteria to divide region and activity, but the results presented, with region divided at 7 nucleotides and activity divided at 0.5 were typical.

First, the distribution of motifs associated with effective antisense sequences was examined. 30 of the 156 motifs found in high-effectiveness sequences had numerical counts of 5 or greater in each cell of the 2 x 2 contingency table. Of these 30 motifs, 14 (8.9% of 156) had significant deviations from the expected uniform distribution by the chi-square test, incorporating Bonferonni corrections on the chi-square for multiple comparisons. All 14 of these significant deviations were consistent where the proportion of motifs in effective versus ineffective sequences was greater in the outer region of the

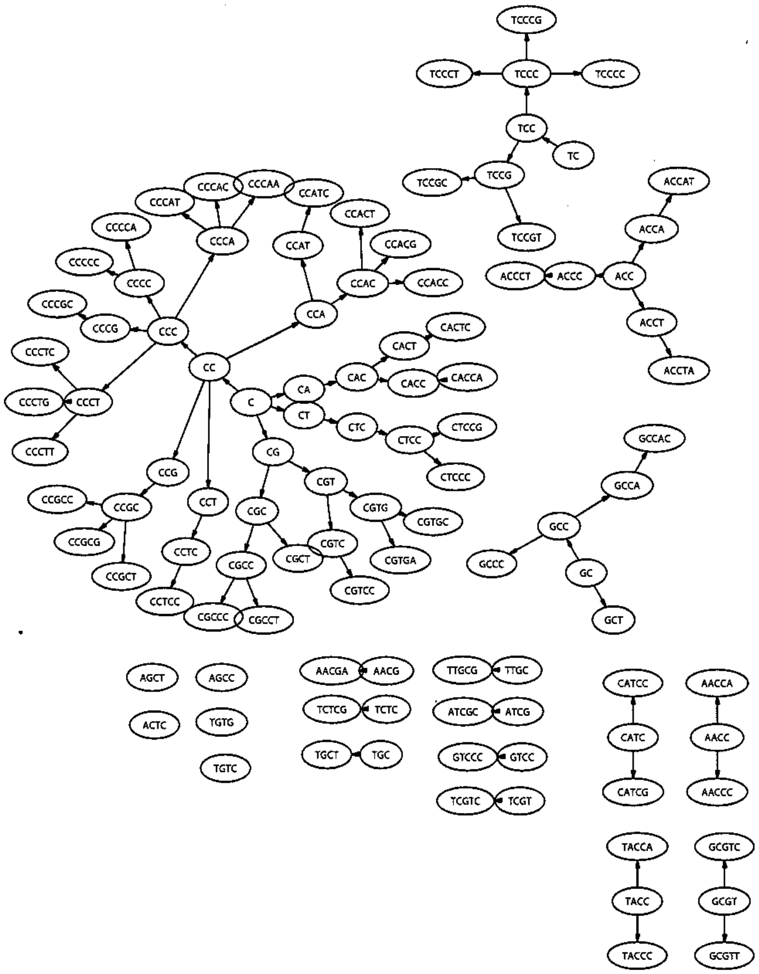

Fig. 2. Connectivity among the 156 nucleotide motifs significantly associated with high activity antisense sequences grouped by 5' sequence.

antisense when compared to the inner region. By contrast, 18 of the 213 motifs found in low-effectiveness sequences had numerical counts of 5 or greater, and of these, 11 (5.1% of 213) had significant deviation from the expected uniform distribution. Again, all 11 of these deviations were together consistent and opposite in direction from the pattern found with effective motifs. The proportion of motifs in effective versus ineffective sequences was greater in the inner region when compared to the outer region. In summary, more of the effective antisense sequences tend to have "good" motifs towards the outer regions of an effective antisense, while "bad" motifs are more tolerated towards the inner region of an effective antisense. Additional examinations for nucleotide-specific positional effects and strand asymmetry effects did not show any obvious general trends (results not shown).

Finally, we built a C4.5 decision tree (20) to predict the effectiveness of a novel sequence based on the number of "good" and "bad" motifs it contained. We determined that the motif variables alone are a poor way of discriminating sequence effectiveness; after training a model over all 3913 instances, a 10-fold cross-validation produced 2953 correctly classified instances (75.46%) and 960 incorrectly classified instances (24.53%).

The vast majority of correctly classified instances were ineffective sequences, and as a result, the model overclassifies toward "ineffective"; of the 967 effective sequences, only 39 were classified correctly. By contrast, only 32 ineffective sequences were misclassified as effective. However, this tree has only two branches (at "good-motifs > / <= 19" and "good-motifs > / <= 29") and three leaves -- it only classified sequences with more than 29 "good" motifs as effective. Introducing more data about each sequence -- for example, δG, the number of "good" motifs appearing toward the outside of the sequence, the number of "bad" motifs appearing toward the interior of the sequence, and the bit-wise information content -- will provide more decision-making criteria and a more accurate model of antisense effectiveness.

Discussion

Our intentions were to examine sequence motifs and determine if they were non-randomly associated with antisense sequences and their activities with a dataset of phosphorothioate oligonucleotides. In the course of this goal, we developed a Monte Carlo method to systematically examine the distribution of non-discretely associated motifs, and discovered novel patterns of motif association and distribution within antisense sequences. This method was able to uncover sequence motifs that have previously been identified with antisense sequence effectiveness, such as TCCC (8,13), CCAC, ACTC, and GCCA (13). However, the motif CTCT has been previously identified (13) as associating with effective antisense sequences, but we did not identify this motif with this method and dataset. Furthermore, this method uncovers motifs that have previously been identified with antisense sequence ineffectiveness, such as GGGG, AAA and TAA (13). Another motif, ACTG, which has been previously identified (13) with ineffective antisense, was not identified with this method and dataset. In addition to these motifs that have been previously identified, these methods and dataset were able to find several hundred more motifs that appear to be non-randomly associated with antisense sequence activity.

Sequence motifs that associate with antisense activities appear not to be drawn at random from all possible motifs by their relationships as subwords, thermodynamic properties or information content. This observation should not be surprising, since it is apparent that biological systems are certainly not random, but it does suggest that these methods are able to identify some properties of antisense molecules that may influence the biological systems where these sequences act. While we use statistical significance to help identify patterns, we would like to mention that statistical and biological significance do not necessarily associate. We have found a significant association of motif thermodynamics with activity, and this observation is congruent with previous observations that have shown an overall association between antisense activity and thermodynamic properties. However, from our observations we may provide some further insight into this thermodynamic association. Based on the overall thermodynamic differences between “good” and “bad” motifs, and more clearly the thermodynamic differences between unique motifs in “good”, “bad” and all motifs, the association of sequence effectiveness with thermodynamic stability may be less related to choosing motifs with high stability than to avoiding motifs with low stability. Furthermore, we have identified an overall pattern for motif distributions across an antisense molecule. Specifically, we have found that “good” motifs appear to be more effective when on the ends and that “bad” motifs appear to be less deleterious when centrally located in an antisense molecule.

The rational design of nucleic acid molecules to perform target gene knockdown, such as RNaseH mediated antisense or RISC medicated RNA interference, requires some knowledge of the how these nucleic acid molecules function within their biological context. Data mining to explore biological datasets is one method to discover previously unidentified patterns that may suggest likely biological mechanisms of action. For example, the motifs that we identify as “good” may provide some site of interaction for protein binding or may provide some preferential feature for higher RNaseH activity. Also, the identification of an inner and outer symmetry in antisense sequences may suggest some pattern of molecular recognition involving a difference between the core and the ends of an antisense molecule. Furthermore, the identification of patterns that associate with sequence activity can be used in the selection of candidate molecules for further biological screening. Thereby, information from many experiments can be combined and used to determine criteria for discriminating between effective and ineffective antisense, and these criteria can be used as rules to potentially increase candidate antisense molecule hit rates.

References

1. Crooke, S.T. (1999) Molecular mechanisms of action of antisense drugs. *Biochim Biophys Acta*, **1489,** 31-44.
2. Crooke, S.T. (1998) Molecular mechanisms of antisense drugs: RNase H. *Antisense Nucleic Acid Drug Dev*, **8,** 133-134.
3. Patzel, V., Steidl, U., Kronenwett, R., Haas, R. and Sczakiel, G. (1999) A theoretical approach to select effective antisense oligodeoxyribonucleotides at high statistical probability. *Nucleic Acids Res*, **27,** 4328-4334.
4. Giddings, M.C., Shah, A.A., Freier, S., Atkins, J.F., Gesteland, R.F. and Matveeva, O.V. (2002) Artificial neural network prediction of antisense oligodeoxynucleotide activity. *Nucleic Acids Res*, **30,** 4295-4304.
5. Chalk, A.M. and Sonnhammer, E.L. (2002) Computational antisense oligo prediction with a neural network model. *Bioinformatics*, **18,** 1567-1575.
6. Jayaraman, A., Walton, S.P., Yarmush, M.L. and Roth, C.M. (2001) Rational selection and quantitative evaluation of antisense oligonucleotides. *Biochim Biophys Acta*, **1520,** 105-114.
7. Walton, S.P., Stephanopoulos, G.N., Yarmush, M.L. and Roth, C.M. (1999) Prediction of antisense oligonucleotide binding affinity to a structured RNA target. *Biotechnol Bioeng*, **65,** 1-9.
8. Tu, G.C., Cao, Q.N., Zhou, F. and Israel, Y. (1998) Tetranucleotide GGGA motif in primary RNA transcripts. Novel target site for antisense design. *J Biol Chem*, **273,** 25125-25131.

9. Song, H.F., Tang, Z.M., Yuan, S.J. and Zhu, B.Z. (2000) Application of secondary structure prediction in antisense drug design targeting protein kinase C-alpha mRNA and QSAR analysis. *Acta Pharmacol Sin*, **21,** 80-86.
10. Smith, L., Andersen, K.B., Hovgaard, L. and Jaroszewski, J.W. (2000) Rational selection of antisense oligonucleotide sequences. *Eur J Pharm Sci*, **11,** 191-198.
11. Sczakiel, G., Homann, M. and Rittner, K. (1993) Computer-aided search for effective antisense RNA target sequences of the human immunodeficiency virus type 1. *Antisense Res Dev*, **3,** 45-52.
12. Scherr, M., Rossi, J.J., Sczakiel, G. and Patzel, V. (2000) RNA accessibility prediction: a theoretical approach is consistent with experimental studies in cell extracts. *Nucleic Acids Res*, **28,** 2455-2461.
13. Matveeva, O.V., Tsodikov, A.D., Giddings, M., Freier, S.M., Wyatt, J.R., Spiridonov, A.N., Shabalina, S.A., Gesteland, R.F. and Atkins, J.F. (2000) Identification of sequence motifs in oligonucleotides whose presence is correlated with antisense activity. *Nucleic Acids Res*, **28,** 2862-2865.
14. Matveeva, O.V., Mathews, D.H., Tsodikov, A.D., Shabalina, S.A., Gesteland, R.F., Atkins, J.F. and Freier, S.M. (2003) Thermodynamic criteria for high hit rate antisense oligonucleotide design. *Nucleic Acids Res*, **31,** 4989-4994.
15. Vickers, T.A., Koo, S., Bennett, C.F., Crooke, S.T., Dean, N.M. and Baker, B.F. (2003) Efficient reduction of target RNAs by small interfering RNA and RNase H-dependent antisense agents. A comparative analysis. *J Biol Chem*, **278,** 7108-7118.
16. Kandimalla, E.R., Manning, A., Lathan, C., Byrn, R.A. and Agrawal, S. (1995) Design, biochemical, biophysical and biological properties of cooperative antisense oligonucleotides. *Nucleic Acids Res*, **23,** 3578-3584.
17. Giddings, M.C., Matveeva, O.V., Atkins, J.F. and Gesteland, R.F. (2000) ODNBase--a web database for antisense oligonucleotide effectiveness studies. Oligodeoxynucleotides. *Bioinformatics*, **16,** 843-844.
18. Alt, M., Renz, R., Hofschneider, P.H. and Caselmann, W.H. (1997) Core specific antisense phosphorothioate oligodeoxynucleotides as potent and specific inhibitors of hepatitis C viral translation. *Arch Virol*, **142,** 589-599.
19. SantaLucia Jr., J. (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci.*, **95,** 1460-1465.
20. Quinlan, J.R. (1993) *C4.5: Programs for machine learning*. Morgan Kaufmann, San Francisco.