Statistica Sinica **12**(2002), 273-296

RATIONAL STATISTICAL DESIGN OF ANTISENSE OLIGONUCLEOTIDES FOR HIGH THROUGHPUT FUNCTIONAL GENOMICS AND DRUG TARGET VALIDATION

Ye Ding

New York State Department of Health

Abstract: RNAs are versatile molecules that are involved in many important cellular activities including protein synthesis, antisense hybridization, RNA-RNA interactions, and RNA-protein interactions. Computational prediction of RNA higher-order structure is important because crystal structures have been determined only for a few RNA molecules. Statistical algorithms have been recently developed and shown to have advantages for RNA folding prediction. A statistical algorithm is presented for the energy model of base pair stacking, and a very important algorithm for more realistic energy rules is described. These algorithms demonstrate how statistical thinking can be successfully adopted for RNA secondary structure prediction to overcome inherent limitations in mathematical algorithms. For the determination of gene function, antisense techniques promise to offer a high-throughput platform. We illustrate how an approach based on statistical algorithms can be used for the rational design of antisense oligonuleotides (oligos). In the post-genomic era, DNA expression arrays and single-nucleotide polymorphisms (SNPs) promise to enable the prediction of gene functions and the identification of candidate genes for disease phenotypes. Functional predictions will eventually require experimental validation. An antisense approach is well suited for these high throughput applications to keep pace with rapid accumulation of genomic information, DNA expression array data, and SNP databases. The full realization of the promise of antisense technology can be greatly aided by an adequate integration of computational approaches and experimental techniques.

Key words and phrases: Antisense design, drug target validation, functional genomics, RNA folding.

1. Introduction

RNA plays a variety of important functional roles that include catalysis, RNA splicing, regulation of transcription, and translation. These roles are carried out at specific RNA structural sites, often through molecular interactions or conformational change. Hence, the function of an RNA molecule is determined by its secondary and tertiary structures. To date, crystal structure has been determined only for four RNA molecules: yeast phenylalanine transfer RNA, hammerhead ribozyme, the P4-P6 domain of the *Tetrahymena* group I intron,

and the hepatitis delta virus ribozyme (for a review, see Ferré-D'Amaré and Doudna (1999)). RNA tertiary interactions involve secondary structure elements and are substantially weaker than secondary interactions. Thus, to a large extent, the free energies in secondary structure represent the thermodynamics of RNA folding. The tendency for RNA folding to be primarily driven by secondary structure features is a tremendous advantage for structural and functional studies on RNAs. Furthermore, computational RNA tertiary structure prediction without experimental information is an intractable problem, and the thermodynamics of tertiary interactions have not been well characterized. For these reasons, computational algorithms have focused on RNA secondary structure prediction in the last several decades.

In this article, we first give an overview of the major algorithms for RNA secondary structure prediction with one RNA sequence. For the energy model of base pair stacking, we present a statistical sampling algorithm for RNA folding prediction. We also describe an algorithm for more realistic energy rules. These algorithms overcome inherent limitations in mathematical algorithms. We illustrate the application of a statistical approach to the rational design of antisense oligonucleotides. We discuss the feasibility and basis for high-throughput antisense applications to functional genomics and drug target validation.

2. Existing Algorithms for RNA Secondary Structure Prediction

A recent authoritative review on calculating nucleic acid secondary structure using one or multiple homologous sequences is given by Zuker (2000). In this section, we focus on major methods for RNA secondary structure prediction using one RNA sequence. The secondary structure of an RNA is defined by a list of base pairs. Favorable free energies (negative-valued) are assigned to base pair stacks, and destabilizing free energies (positive-valued) are given to loops of various types and size. The individual free energies are assumed to be additive in the calculation of total free energy of a specified secondary structure. The complete set of the latest free energies by the well-accepted "Turner rules" is given in Xia et al. (1998) and Mathews, Sabina, Zuker and Turner (1999). Base pair stacking represents an important step toward realistic characterization of RNA folding thermodynamics. For example, a stack between two G•C pairs with the G base of the exterior base pair on the 5' end has a free energy of -3.3 kcal/mole. The free energy of a helix with m base pairs is the sum of $(m-1)$ stacking energies.

Earlier work on RNA secondary structure prediction focused on mathematical algorithms. The goal of these algorithms for RNA folding prediction is free energy minimization, i.e., finding the secondary structure (or structures) with

the lowest total free energy. A more in-depth description of the problem can be found in Zuker (1989a, b). The most important are the algorithms for the popular RNA folding software *mfold* developed by Zuker (Zuker and Stiegler (1981), Zuker (1989a, b)). These algorithms predict optimal structure through free energy minimization based on thermodynamic parameters developed by Turner and coworkers (Mathews et al. (1999), Xia et al. (1998)). For Xlo 5S rRNA, the optimal structure and all types of secondary structural elements are illustrated in Figure 1. There are several sources for uncertainty in the predictions. There is uncertainty in free energy parameters for destabilizing loops because experimental data are difficult to obtain, and there is also uncertainty in stacking energy parameters (Freier et al. (1986)). The discrete free energy models and the assumed free energy additivity are simplifications. Also, the folding problem is ill-conditioned, i.e., very slight deviations in the energy parameters or the sequence can lead to substantial differences in the optimal folding. The suboptimal foldings from mfold are intended to mitigate the ill-conditioned nature of the folding problem. However, because of the inherent mathematical nature of the algorithm design, these suboptimal foldings do not yield a statistically valid sample of the probable structures. Other limitations of this treatment have also been documented (Wuchty, Fontana, Hofacker and Schuster (1999)).

Probabilistic approaches provide a means to address uncertainty. McCaskill (1990) pioneered this approach for RNA secondary structure prediction. He presented a partition function method to compute the exact base pair probabilities. The probabilities are displayed in a box plot qualitatively similar to the energy dot plot for suboptimal foldings (Jacobson, Zuker and Hirashima (1987), Jacobson and Zuker (1993)). Despite its elegance, this algorithm does not generate any secondary structure.

Based on a starting energy model for base pair stacking, a Bayesian treatment of the problem showed its promise to address the limitations of these algorithms (Ding and Lawrence (1999)). More specifically, the Bayesian algorithm was able to address the need for a representation of the full ensemble of probable structures, and it enabled statistical inferences on all variables in the problem, including free energy parameters, and number of destablizing loops. One also saw that the specification of the free energy parameters could be relaxed through prior assignment as a way to mitigate the ill-conditioned nature of the RNA folding problem. This approach has generated great interest for its potential to solve problems that had been considered intractable (Zuker (2000, p.310)).

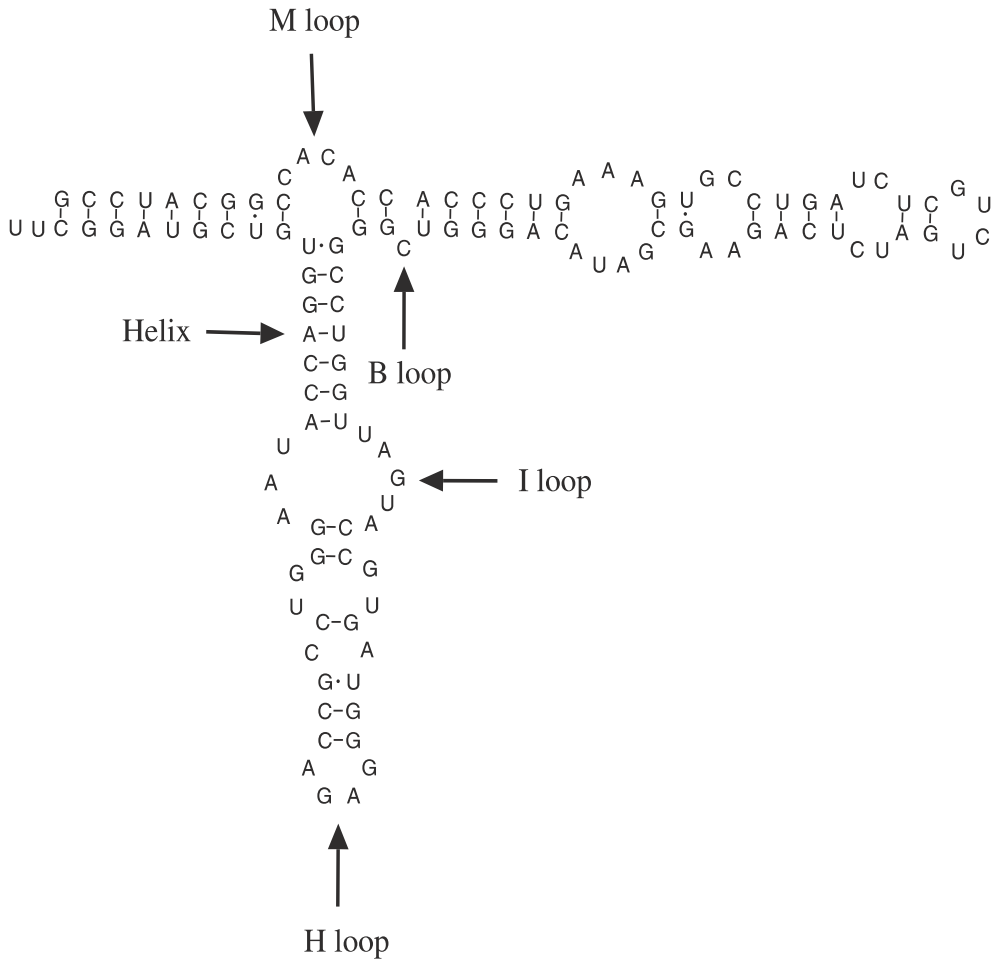

Figure 1. The minimum free energy structure for Xlo 5S rRNA and all types of secondary structural elements: helix (formed by stacked base pairs), bulge loop (B loop), interior loop (I loop), hairpin loop (H loop), and multi-branched loop (M loop).

3. New Statistical Algorithms for Sampling RNA Secondary Structures

To demonstrate how statistical approaches can be useful for RNA folding prediction, we present in detail an algorithm for the stacking energy model. This differs from our earlier algorithm (Ding and Lawrence (1999)) in that structure sampling is the focus here and Bayesian modeling is not involved. We also describe a very important algorithm for more sophisticated and realistic energy rules and discuss its significant features.

3.1. Computing partition functions

For an RNA molecule of n ribonucleotides, denote the sequence from the ith ribonucleotide from the 5' end to the jth ribonucleotide by $R_{ij} = r_i r_{i+1} \cdots r_j$, $1 \leq i,\ j \leq n$, where r_i = A, C, G, or U. Let I_{ij} be a secondary structure on R_{ij} that meets the usual constraints of unknotted structure and at least three intervening bases between any base pair. For structures under the constraints, let IP_{ij} be a structure on R_{ij} with the ends constrained to form a base pair. The partition functions restricted to R_{ij} are defined as:

$$u(i,j) = \sum_{I_{ij}} \exp[-E(R_{ij}, I_{ij})/RT], \tag{1}$$

$$up(i,j) = \sum_{IP_{ij}} \exp[-E(R_{ij}, IP_{ij})/RT], \tag{2}$$

where $E(R_{ij}, I_{ij})$ is the free energy for structure I_{ij}, R is the gas constant, T is the absolute temperature, and kcal/mol/RT=1.6625. In deriving recursions with stacking energies, we consider the following mutually exclusive and exhaustive cases for any fragment $R_{ij} = r_i \cdots r_j$: (a) r_j is single stranded; (b) r_j and r_i form a base pair $r_i - r_j$; (c) r_j and r_k form a base pair $r_k - r_j$, $i < k < j$. Then the recursions for the partition functions are as follows:

$$u(i,j) = u(i,j-1) + up(i,j) + \sum_{i<k<j} u(i,k-1)up(k,j), \tag{3}$$

$$up(i,j) = \exp(-E([i \cdot j/(i+1)] \cdot [(j-1)/RT])up(i+1,j-1)) \tag{4}$$

where $E([i \cdot j/(i+1)] \cdot [(j-1)/RT])$ is the stacking energy between the adjacent base pairs $r_i - r_j$ and $r_{i+1} - r_{j-1}$. We start the computation with boundary values for short fragments and proceed to longer ones using the recursion. For $1 \leq i \leq j \leq i+3 \leq n, u(i,j) = 1, up(i,j) = 0$; for $j = i+4 \leq n, u(i,i+4) = 1, up(i,i+4) = 1$; and $u(i+1,i) = 0, 1 \leq i \leq n$.

3.2. Sampling RNA secondary structures

The recursions for partition functions correspond to probabilities for sampling. For a fragment R_{ij}, the last nucleotide r_j can be single stranded, or can base pair with r_i, or can base pair with r_k $(i < k < j)$. The corresponding probabilities for these cases are $P_{jj} = u(i,j-1)/u(i,j)$, $P_{ij} = up(i,j)/u(i,j)$, $P_{kj} = u(i,k-1)up(k,j)/u(i,j)$, $i < k < j$. Each numerator corresponds to a term in the recursion (3) for $u(i,j)$, the denominator of all the probabilities that sum up to 1. If the ends of the fragment are known to form a base pair then the probabilities for stacking or non-stacking corresponding to the recursion (4) for

$up(i,j)$ are

$$P_{Sij} = \frac{\exp(-E([i \cdot j/(i+1)] \cdot [(j-1)/RT])up(i+1,j-1)}{up(i,j)}, \tag{5}$$

$$P_{NSij} = \frac{u(i+1,j-1) - up(i+1,j-1)}{up(i,j)}. \tag{6}$$

When non-stacking is given, i.e., r_{i+1} and r_{j-1} do not form a base pair, the probability of a single stranded r_{j-1} and the probability of a base pair $r_k - r_{j-1}$ $(i+1 < k < j-1)$ are

$$Q_{NS(j-1)(j-1)} = \frac{u(i+1,j-2)}{u(i+1,j-1) - up(i+1,j-1)}, \tag{7}$$

$$Q_{NSk(j-1)} = \frac{u(i+1,k-1)up(k,j-1)}{u(i+1,j-1) - up(i+1,j-1)}, \qquad i+1 < k < j-1. \tag{8}$$

A secondary structure is drawn recursively as follows: starting with R_{1n}, draw single-stranded r_n or a base pair according to probabilities P_{nn}, P_{1n}, and $P_{kn}, 1 < k < n$; for a new fragment R_{ij}, if base pair $r_i - r_j$ was not sampled previously, then sample with $P_{jj}, P_{ij}, P_{kj}, i < k < j$; if $r_i - r_j$ was sampled, then we sample by P_{Sij}, P_{NSij} for stacking or non-stacking; stacking implies a sampled base pair $r_{i+1} - r_{j-1}$, the interior base pair of the stack; when non-stacking is sampled, we then sample a single-stranded r_{j-1} or a base pair $r_k - r_{j-1}$ with probabilities $Q_{NS(j-1)(j-1)}, Q_{NSk(j-1)}$. During this process, single-stranded nucleotides and exterior stacking base pairs are removed from further involvement in sampling, and the sampling terminates when all remaining fragments are shorter than five bases, the minimum length needed for forming a base pair.

For more complicated free energy rules, we have developed an extended algorithm. The forward step of this algorithm is a recursive algorithm for partition functions. This recursive algorithm extends the work of McCaskill (1990) by including single base stacking energies and other up-to-date free energy parameters. The backward step takes the form of a sampling algorithm: the sampling probabilities are computed using the partition functions from the forward step. The extended algorithm accommodates the up-to-date free energy rules and parameters developed by Turner's group (Mathews et al. (1999), Xia et al. (1998)) with the exception of coaxial stacking.

The Boltzmann distribution in statistical mechanics gives the probability of a secondary structure I at equilibrium as $(1/U)\exp[-E(I)/RT]$, where $E(I)$ is the free energy of the structure, R is the gas constant, T is the absolute temperature, and U is the partition function for all admissible secondary structures of the RNA sequence. Direct discrete sampling is not feasible because the number of all possible structures grows exponentially with the length of the sequence.

The extended algorithm samples secondary structures *exactly* and *rigorously* according to the Boltzmann distribution, i.e., it can generate a statistical sample of any desired size from the Boltzmann ensemble of secondary structures. Exact sampling with Boltzmann probabilities is a significant feature rarely attainable in computational biology. While the Boltzmann distribution statistically characterizes high-dimensional folding states of a molecule, the ability to sample exactly and rigorously from this distribution depends on the tractability of the problem. All the major algorithms reviewed here require at most cubic execution time. Furthermore, the significance of this capability depends on the credibility of thermodynamic parameters. Most of the Turner parameters are estimated from chemical melting experiments and are well accepted by the RNA community. For protein folding prediction, Monte Carlo approximations are necessary and have been limited to relatively short polypeptides.

The sampling process is similar to the traceback algorithm employed in dynamic programming algorithms (Zuker and Stiegler (1981), Zuker (1989a, b)), but it differs in that the base pairing is randomly sampled with conditional probabilities computed with partition functions rather than chosen to yield a minimum free energy structure. In other words, the traditional mathematical algorithms pick a folding path according to the minimum energy principle, and the suboptimal folding scheme by Zuker (1989a) selects multiple paths without regard to a probabilistic framework, while the sampling process presents a random folding path according to a probabilistic scheme based on statistical mechanics principles. Because the probability of a structure decreases exponentially with increasing free energy, the structure with highest frequency in the sample is most likely the minimum free energy structure. When long interior loops (e.g., size $>$ 30 nt) are disallowed, the forward step of the algorithm is cubic. The sampling step of the algorithm is stochastically quadratic in the worst case; thus it can quickly generate a large number of secondary structures.

4. Statistical Prediction of Potent Antisense Targets and Rational Statistical Design of Antisense Oligos

4.1. Antisense technology

More than two decades ago, it was recognized that an oligodeoxynucleotide can bind to a messenger RNA (sense strand) through complementary base pairing to block its translation (Zamecnik and Stephenson (1978)). Antisense oligomers are short synthetic oligonucleotides of (usually) 10-25 bases in length. Oligodeoxynucleotides (DNA oligos) are usually used because they are more stable than RNAs in a cellular environment. In cells, RNAs are degraded by cellular enzymes (RNases) after completing their functions. Chemical

modifications are used to further improve the stability of DNA oligos. DNA oligos are very inexpensive to make. Furthermore, the intramolecular interaction for a DNA oligo is usually weaker than that for the RNA oligo of same base composition, a characteristic that is favorable for intermolecular interaction for DNA-RNA hybridization. Over the years since this pioneering finding, it has been proved that antisense oligonucleotides are able to modulate gene expression in both prokaryotes and eukaryotes (Vanhée-Brossollet and Vaquero (1998)). In 1998, Vitravene (Isis Pharmaceuticals, Carlsbad, CA, USA) became the first antisense drug approved by the Food and Drug Administration (FDA). Vitravene is used to treat cytomegalovirus (CMV) retinitis in AIDS patients.

The high specificity of antisense oligos is a result of complementary base pairing. A single base mismatch can result in a change in binding affinity by as much as 500-fold (Crooke (1999)). For this reason, the specificity can be demonstrated by making a few mismatches in the oligomer (mismatched control) or randomly shuffling the bases (scrambled control) to repeat the experiment and observe the change in expression level. This level of specificity is difficult, if not impossible in some cases, to achieve by compounds for inhibition of protein function. Furthermore, by taking advantage of subtle differences in the genetic sequences of closely related genes, antisense oligos could discriminate and inhibit the function of an individual gene in the family. From cellular RNA population to genomic level, the reported estimates for the minimal length of an oligo to ensure its statistical uniqueness range from 11 to 17 bases (Crooke (1999)). Oligos of 18 to 20 bases are often used in applications. However, the near-completion of human genome sequencing and subsequent annotation provide the opportunity for more accurate estimates. It is important to note that non-antisense effects can sometimes occur in a cellular environment. For example, oligomers with four contiguous guanosine residues can have non-specific effects, because the G-quartet can form tetrads, which can stack to form tetraplexes with seemingly very high affinity for heparin-binding protein (Stein (1999)). Thus, it is advisable not to use oligos with G-quartets or any other motif known to cause non-antisense effects.

Figure 2 demonstrates the mechanism of antisense inhibition of gene translation. After hybridization between the antisense oligomer and the targeted mRNA, the mRNA is degraded by RNase H, an enzyme that recognizes DNA-RNA duplex and cleaves the RNA strand. When the target is the 5' untranslated region (5' UTR) or the AUG initiation codon, the ribosome is blocked by the DNA-RNA hybrid. In either case, there is no protein product. Despite recent progress reported by Mir and Southern (1999), many subtle aspects of the molecular mechanisms of antisnese hybridization have not been well elucidated. Nevertheless, antisense hybridization can be simply viewed as a two-step process.

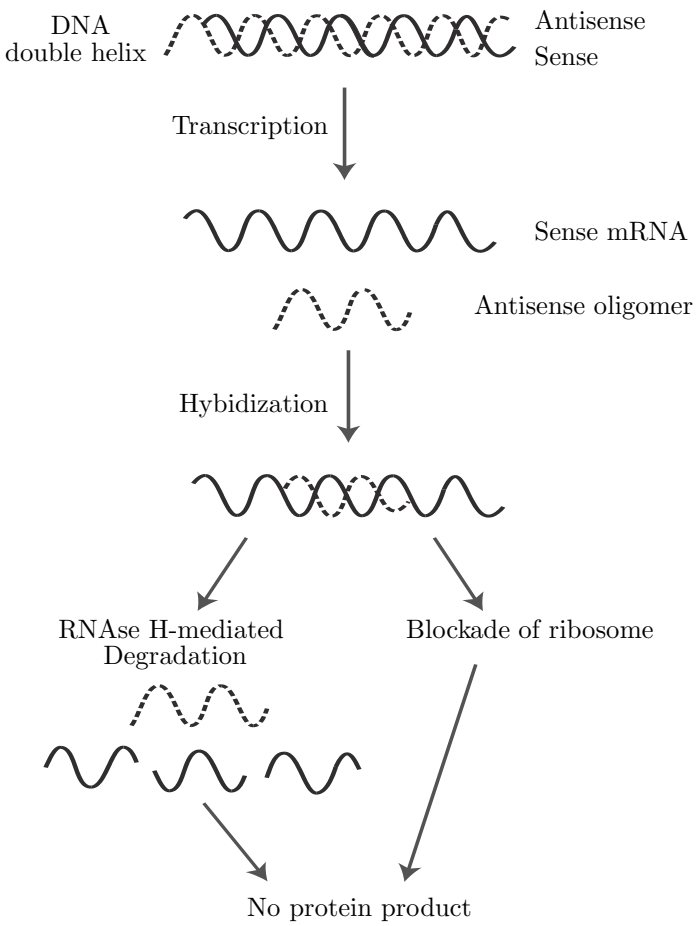

Figure 2. Inhibition of gene expression by an antisense oligomer.

It starts at an unstructured nucleation site and then unwinds the adjacent helix by a process called "zippering" (Figures 3.1, 3.2). The process stops when it meets an energy barrier such as the end of the helix or a sharp turn in the folded RNA (Milner, Mir and Southern (1997)). For rabbit β-globin mRNA (589 nt), two oligomers BG1 and BG2 reported in Milner, Mir and Southern (1997) were found to be effective by *in vitro* translation study. These two oligomers are complementary to bases C46-U62 and bases A51-C67 of the mRNA. Figure 3.2 shows that a hairpin loop in the predicted optimal structure is the common nucleation site. In general, an accessible single-stranded region is necessary for antisense activity (Lima, Monia, Ecker and Freier (1992)). A single-stranded region can be any of the types of loops illustrated in Figure 1, or it can be a free-dangling end or an unpaired segment connecting two folding domains.

Antisense inhibition can be measured either by an RNA assay or a protein assay. For the RNA assay, the Northern blot is the main experimental technique for detecting the change in mRNA level after antisense treatment. The degradation of the mRNA by RNase H cleavage will cause a decrease in the

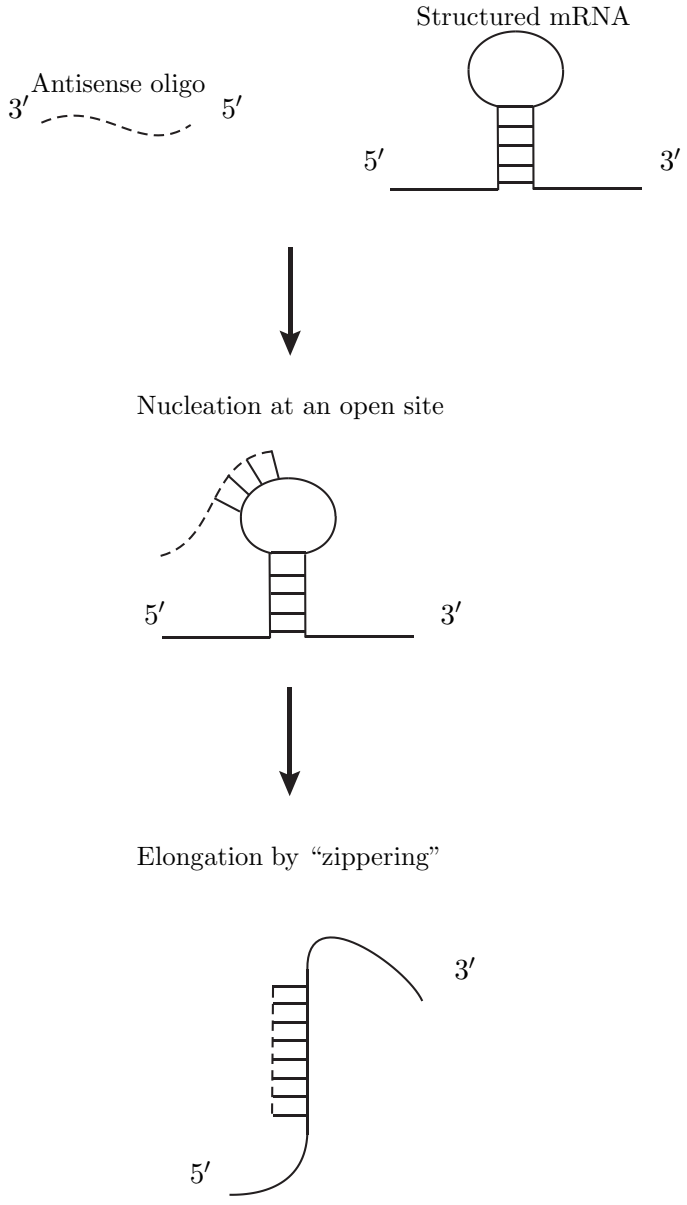

Figure 3.1. A simple two-step process for hybridization formation: nucleation at an unstructured site and elongation by "zippering" for unwinding the adjacent helix.

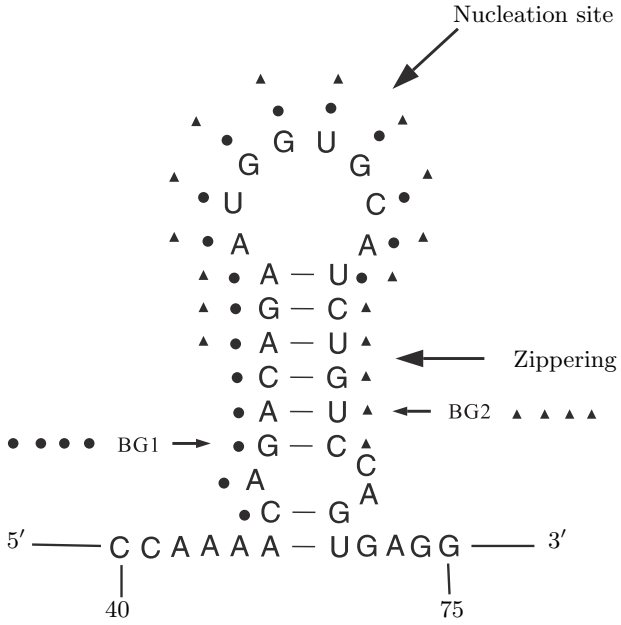

Figure 3.2. Two effective antisense oligomers BG1 and BG2 for rabbit β-globin. The common nucleation site is a hairpin loop with eight unpaired bases.

mRNA level (Figure 2). The amount of protein synthesized by the mNRA is often determined by Western blot. The major obstacle for successful antisense inhibition is the identification of effective target sites on the target mRNA. Experimental approaches for target site selection are generally laborious, time consuming, and expensive. The conventional "gene walk" or "shot gun" approach empirically tests a large number of oligomers targeted at various regions of the mRNA. This method requires a great deal of luck to work. Because mRNA structure is not considered, the success rate is only 2-5% (Sohail and Southern (2000)). Milner, Mir and Southern (1997) proposed a combinatorial oligonucleotide array approach. We discuss this technique in Subsection 5.3. Ho et al. (1996, 1998) proposed using semi-random antisense libraries for RNase H mapping accessible sites. This method requires sequencing of new RNA ends resulting from RNase H cleavage to identify the cleavage sites. Furthermore, this approach does not take advantage of the primary sequence or the secondary structure of the mRNA. In view of the limitations and difficulties with experimental approaches, computational approaches can be valuable for efficient experimental design and for expediting antisense screening.

It has been reported that a small hairpin loop of four unpaired bases can initiate antisense activity (Asano, Niimi, Yokoyama and Mizobuchi (1998)). Loops of smaller size are not considered adequate, because stable base pair stacking is

difficult to establish in the nucleation step when only ≤ 3 bases are available for base pairing. For BG1 and BG2 in Figure 3, there are eight unpaired bases in the hairpin loop targeted by both oligomers. This explains why the two oligomers are equally effective. Thus, it is important to assess the chance that a segment of four consecutive bases is entirely single-stranded. Neither *mfold* nor the McCaskill algorithm can address this need. Not surprisingly, there has been limited success in antisense design using the optimal structure from *mfold* (Sohail and Southern (2000)). However, this problem is easily addressed by our sampling approach, as detailed in the next subsection.

4.2. Probability profiles for predicting single-stranded bases and segments and effective antisense targets

Ding and Lawrence (2001) proposed the construction of a probability profile for the prediction of single-stranded regions in RNA secondary structure and for the identification of effective antisense sites. From recursively derived partition functions for an RNA sequence of n bases, McCaskill (1990) presented recursions for marginal base pairing probability: P_{ij} =Prob(base i and base j form a pair), and then the probability that base i is unbound (i.e., single-stranded) is $q_i = 1 - \Sigma_{(i+1)\leq j\leq n} P_{ij} - \Sigma_{1\leq j\leq i} P_{ji}$. As emphasized by McCaskill, the base pair binding probabilities are not locally determined by the RNA sequence; rather, they reflect a sum over all equilibrium-weighted structures in which the chosen base pair occurs. Therefore, probabilities $\{q_i\}$ statistically describe the antisense hybridization potential for every nucleotide in the sequence. Alternatively, the sampling method presents a means to estimate qi with the sampling frequency for the unbound base i. This avoids the cubic algorithm required to compute the probabilities analytically. A probability profile is then displayed by plotting $\{q_i\}$ against the nucleotide position.

However, probabilities $\{q_i\}$do not provide a suitable means to assess the potential of a sequence to be single-stranded and available for hybridization. More specifically, for a fragment from base i to base j, Q_{ij}, the probability of the fragment being single-stranded, is not simply the product of individual probabilities $\{q_m\}, i \leq m \leq j$, because independence is invalidated by the nearest-neighbor interactions. However, a probabilistic measure of the hybridization potential of a sequence can be obtained from a sample of secondary structures. Because the sample is representative of the Boltzmann ensemble of secondary structures, the fraction of the sample in which all the nucleotides in the sequence are single-stranded provides an unbiased estimate of the probability of the sequence being single-stranded. For all successive overlapping sequences of width W, the sampling estimate for the probability that a sequence is single-stranded can be plotted against the first nucleotide of the sequence for a probability profile of

single-stranded sequences with width W. Based on the rule of thumb of at least four unpaired bases (Asano, Niimi, Yokoyama and Mizobuchi (1998), Zhao and Lemke (1998)) for the nucleation step of antisense hybridization, we set $W = 4$ for antisense application (Figures 4.1, 4.2).

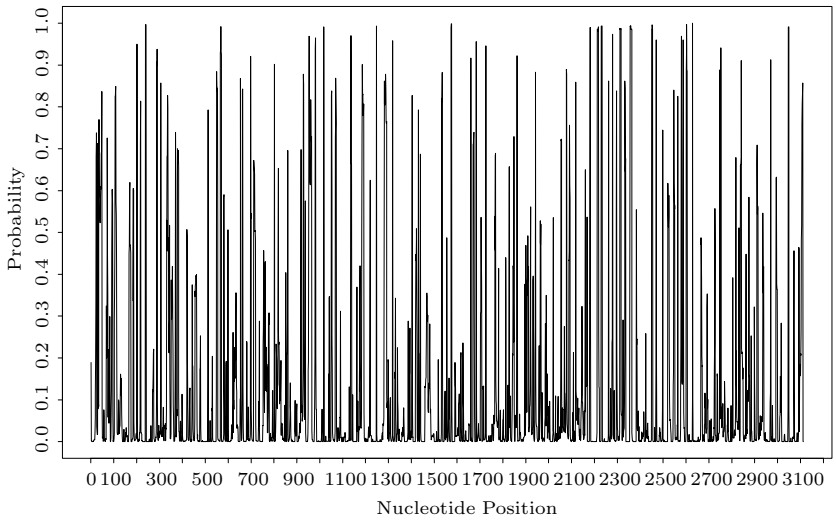

Figure 4.1. The complete probability profile for single-stranded segments of four consecutive nucleotides (segment width=4) estimated by 1,000 sampled secondary structures for *E. coli lac*Z mRNA.

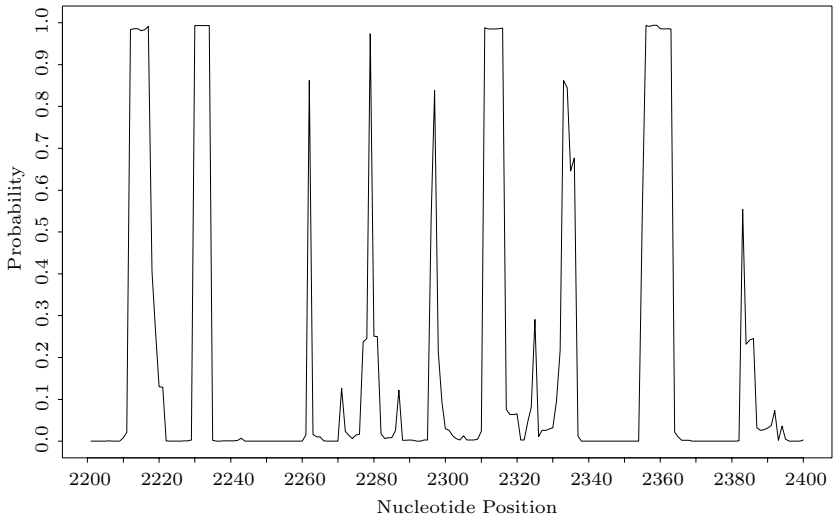

Figure 4.2. The portion of the profile from nt 2200 to nt 2400. The first high peak from the left is targeted by oligomer 4 in Table 2. Two relatively wide and high peaks on the right are targeted by oligomer 5 and oligomer 6 in Table 2.

For the predictions of single-stranded regions in phylogenetic structures of representative RNA sequences, we found that the probability profile offers substantial improvement over the minimum free energy structure from *mfold* (Ding and Lawrence (2001)). In an application to rabbit β-globin mRNA, there is a statistically significant correlation (correlation coefficient = 0.597, and P value = 0.0147) between hybridization potential predicted by the probability profile and the degree of translation inhibition reported for antisense oligonucleotides from *in vitro* experiments. There is a lack of such correlation (correlation coefficient=0.155, and P value=0.567) for the minimum free energy structure computed by mfold. These findings exemplify the advantages of the statistical sampling approach to RNA folding prediction. They also suggest that the probability profile approach is valuable for the identification of effective antisense target sites and the rational design of antisense oligos. We illustrate these utilities in the following subsection.

4.3. Antisense design for *lac*Z mRNA

Mycobacterium tuberculosis (Mtb), one of the world's most important pathogens, is responsible for millions of new cases of tuberculosis every year. Genetic manipulation is difficult, because mycobacteria grow slowly and have undeveloped methods of genetic exchange. For these reasons, molecular characterization for mycobacteria has lagged behind that for other bacteria species. Antisense technique, as an alternative to genetic manipulation, has showed promise for functional studies of Mtb genes in recent years (Rapaport, Levina, Metelev, Zamecnik (1996), Harth et al. (1999)). A fluorescence-based detection of *lac*Z reporter gene expression was recently developed by Wadsworth Center researchers (Rowland et al. (1999)). *E. coli lac*Z, which codes for the enzyme β-galactosidase, is one of the most commonly used reporter genes in mycobacteria. When *lac*Z is fused to a gene control region, the expression of the gene can be monitored through the detection of β-galactosidase.

Before pursuing studies on Mtb genes of interest, we first perform experimental testing of antisense oligomers for the inhibition of *lac*Z expression. For the 3113-nt mRNA of *E. coli lac*Z, the first 38 nt form the untranslated region (5'-UTR), and the rest is the coding region. The entire mRNA is folded by our algorithm, and Figure 4.1 presents the probability profile for the complete mRNA. The profile reveals 20 or so "well-determined" high antisense potential sites per kilobase. For a focused examination, Figure 4.2 shows an example, the portion of the profile for nt 2200 through nt 2400. For this example, the selection of antisense oligomers consists of two steps. First, the entire profile is scanned for the complete list of primary antisense sites listed in Table 1. Although many of these

sites could be good antisense targets, the number of the sites is larger than desired for this study. These sites are further screened in the next step. To maximize the chance for antisense hybridization, we select sites with relatively wide and high probabilities, or sites with narrower, high peaks in close proximity (Ding and Lawrence (2001)). For antisense oligomers, those with three or more Gs in a row are not used, to reduce non-specific affect (Burgess et al. (1995)). Furthermore, oligomers with substantial self-complementary regions are avoided, because stable intra-molecular structure within the antisense oligomer may hinder hybridization with the target mRNA. Based on the computer scanning in the first step and on these empirical rules, 10 antisense oligomers targeting various region of the mRNA are designed (Table 2). There are no more than three base pairs in the optimal structure by DNA *mfold* server (http://bioinfo.math.rpi.edu/∼mfold/dna) for any of the recommended oligomers. These short helices are relatively weak for the length of the oligomers. Alternatively, we could also compute the mean number of base pairs using our sampling algorithm and set a threshold for oligomer selection.

Table 1. Primary antisense sites and complementary sequences (CS) on *E. coli lac*Z mRNA.

Site no.	Starting base	Position	Ending base	position	CS (length)	Peak probability
1	A	48	A	51	UAAU (4)	0.837
2	C	106	U	112	GGGUUGA (7)	0.849
3	C	201	A	205	GACUU (5)	0.950
4	C	218	U	221	GAAA (4)	0.814
5	A	239	C	244	UCUUCG (6)	0.997
6	A	288	C	293	UGACAG (6)	0.938
7	A	306	U	309	UUGA (4)	0.857
8	C	335	C	338	GUAG (4)	0.827
9	G	549	A	556	CUUAAACU (8)	0.884
10	A	566	C	573	UAAAAAUG (8)	0.992
11	A	654	A	657	UACU (4)	0.868
12	C	662	U	667	GUAAAA (6)	0.843
13	U	698	A	701	AUGU (4)	0.921
14	U	802	G	805	AUGC (4)	0.902
15	A	928	C	933	UUUUGG (6)	0.878
16	A	953	C	958	UUAGGG (6)	0.969
17	A	961	U	964	UAGA (4)	0.817
18	U	980	U	985	ACUUGA (6)	0.965
19	A	1018	C	1021	UUCG (4)	0.991
20	U	1051	A	1055	AACUU (5)	0.838

Table 1 (Con't) Primary antisense sites and complementary sequences (CS) on *E. coli lac*Z mRNA.

Site no.	Starting base	Position	Ending base	position	CS (length)	Peak probability
41	A	2262	A	2265	UGGU (4)	0.862
42	U	2279	U	2282	AAAA (4)	0.974
43	U	2297	U	2300	AUUA (4)	0.838
44	A	2311	C	2319	UUAAAUUGG (9)	0.988
45	U	2333	U	2337	AGAAA (5)	0.862
46	A	2356	A	2366	UAUUUUUUGUU (11)	0.994
47	U	2450	U	2457	AUUGCGGA (8)	0.996
48	G	2470	A	2473	CCUU (4)	0.960
49	A	2547	A	2550	UAAU (4)	0.840
50	G	2564	G	2567	CACC (4)	0.825
51	A	2580	A	2583	UUUU (4)	0.969
52	U	2587	U	2591	AUAAA (5)	0.960
53	A	2601	A	2608	UUUUGGAU (8)	0.997
54	A	2629	U	2632	UUUA (4)	1.000
55	G	2747	A	2750	CGUU (4)	0.888
56	A	2752	A	2755	UUUU (4)	0.941
57	A	2842	A	2845	UUUU (4)	0.911
58	C	2970	G	2973	GUGC (4)	0.913
59	A	3047	C	3052	UCAUAG (6)	0.991
60	A	3108	A	3113	UUUAUU (6)	0.857

In a primary site of k nucleotides (k=CS length), every one of the $(k-3)$ segments of four nucleotides has a predicted probability of 0.8 or higher. The peak probability is the maximum of the probabilities for all the segments of four nucleotides in the site. In the probability profile, these sites correspond to peaks with a probability of 0.8 or higher.

Four of the designed oligomers have been tested by in *vitro* experiments for *lac*Z translation. The readout from a fluorescence reader gives the measurement of translation (Rowland et al. (1999)) which is used to compute the percentage of inhibition. The inhibition rate is defined as (expression level without oligos expression level with oligos)/(expression level without oligos)x100%. Oligomers 1 and 2 resulted in moderate inhibitions of 37.5% and 29.3%, respectively. Oligomers 3 and 4 produced high inhibitions of 75.5% and 74.7%, respectively. In comparison with the 2-5% success rate for finding effective oligomers by the gene-walk approach, the preliminary experimental results are quite encouraging.

Table 2. Antisense oligomers designed with the probability profile for *E.coli lac*Z mRNA.

Oligomer ID (primary site no.)	Start base and end base on mRNA	Antisense oligomer (length) 5' ⟶ 3'
1	A35-A51	TAATCATGGTCATAGCT (17-mer)
2 (9 &10)	G549-C573	GTAAAAATGCGCTCAGGTCAAATTC (25-mer)
3 (23)	A1185-C1198	GCGTTAAAGTTGTT (14-mer)
4 (39)	A2212-A2226	TCACACTGAGGTTTT (15-mer)
5 (44)	A2311-A2323	TGGCGGTTAAATT (13-mer)
6 (46)	A2356-G2372	CAGCAGTTGTTTTTTAT (17-mer)
7 (47)	U2450-G2463	CGACCCAGGCGTTA (14-mer)
8 (53)	A2601-G2612	CCGGTAGGTTTT (12-mer)
9 (59)	A3047-G3057	CCGCCGATACT (11-mer)
10 (60)	U3099-A3113	TTATTTTTGACACCA (15-mer)

For a primary antisense site (Table 1), the underlined segment is the complement of the CS in the site. The regions targeted by the oligomers all have relatively wide probability peaks. Oligomer 2 targets two disjoint peaks corresponding two underlined segments. Oligomer 1 targets start codon ^{39}AUG^{41}, oligomer 10 targets the 3' end of mRNA coding region.

5. High Throughput Antisense Applications to Functional Genomics and Drug Target Validation

5.1. Antisense technique for functional genomics

Functional genomics, the determination of the function of DNA sequences on a genomic scale, is a fast-growing field in biotechnology. For the current estimate of 30,000 − 40,000 genes in the human genome (International Human Genome Sequencing Consortium (2001), Venter et al. (2001)), definitive functions have been assigned to less than a thousand of these genes (Thompson (1999)). The number of transcripts and distinct proteins is estimated to be 90,000 or more (Galas (2001)), because a single gene can give rise to multiple transcripts, and thus multiple proteins, by means of alternative splicing or alternative translation initiation and termination sites. While each technique for the determination of gene function has its own strengths and weaknesses, antisense oligonucleotides receive the most favorable score on all attributes, in a comprehensive comparison of all techniques (Bennett and Cowsert (1999)). These attributes include broad applicability, usage of primary sequence, time, cost and resource requirement, chance of success, relevance to human disease, and the possibility the technique will result in a drug product. Inactivation of the gene is the classical approach

to gene-function assignment in higher organisms. At the DNA level, a gene can be inactivated by mutagenesis, gene knockout, or dominant negative; at the mRNA level, a gene can be down regulated by antisense oligos, synthetic ribozymes, or the newer RNA interference (RNAi) technique (Bass (2000), Bosher and Labouesse (2000)); antibodies can interfere at the protein level. Although for simpler animals such as *C. elegans*, mutagenesis and RNA interference have been successful (Fire et al. (1998)), mutagenesis has not been shown to be feasible in mammals. Although the field of RNAi is still in its infancy, there is optimism about its potential for application to mammals based on the finding that RNAi may operate in vertebrates at the earliest stages of development (Wiany and Zenicka-Goetz (2000)). Gene knockout is difficult for many organisms, e.g., *M. tuberculosis*. For *Drosophila*, after more than two decades of frustration, the first successful gene knockout was reported recently (Rong and Golic (2000)). Routine gene knockout in mammals has been performed only in mice. In addition to the lengthy duration (usually a year or longer), mammalian gene knockout often leads to embryonic lethal phenotype, providing very little information about the gene function. For mice, antisense strategy has been demonstrated to inhibit gene expression *in utero*, permitting the stage-specific analysis of gene function and identification of secondary phenotypes (Driver et al. (1999)). This technique is expected to be applicable to other mammalian species (Thompson (1999)). At relatively low cost, antisense not only offers high specificity of gene-expression

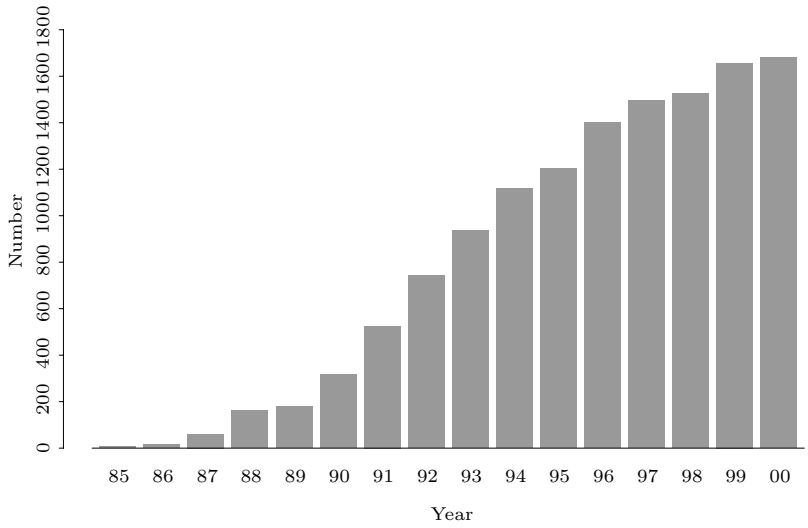

Figure 5. Annual number of antisense papers in PubMed since 1985, obtained with key words "antisense" and the year of publication in PubMed search.

inhibition and rapid detection of antisense effects, but it also enables determination of gene function in adult animals by bypassing potentially lethal embryonic stage. In recent years, antisense has gained increasing attention from research labs in both academia and industry. This is evidenced by the explosion of antisense papers in PubMed (Figure 5). In the post-genomic era, antisense stands out as an important technique that has the potential to meet the need of large-scale functional genomics.

5.2. Antisense technique for drug target validation

The antisense technique is also a very important tool for drug target validation. This is well illustrated by the most favorable scores on all attributes in a comparison of drug target validation techniques (Bennett and Cowsert (1999)). Thousands of new potential therapeutic targets have emerged from human genome sequencing. The selection and validation of molecular targets are of paramount importance for drug development in the new millennium (Ohlstein, Ruffolo and Elliott (2000)). Although phenotypes of many diseases are well known, the identification of the genes responsible for these phenotypes is a major challenge in the drug development process.

An antisense oligonucleotide, by specifically blocking the synthesis of a prospective protein drug target, provides a fast, inexpensive, and often definitive assessment of the biological effect achieved by a drug targeted against that protein. Antisense technology offers a rational alternative to the typical strategy of designing small molecules for the inhibition of a particular gene, which requires substantially more information than does antisense design. Furthermore, the interactions between many small molecules and multiple members of a gene family can confound the assessment of a gene as a drug target (Taylor, Wiederholt and Sverdrup (1999)).

5.3. High-throughput antisense applications

DNA microarrays are one of the latest breakthroughs in experimental molecular biology, allowing the measurement of gene-expression patterns of tens of thousands of genes in parallel (Brown and Botstein (1999), Lockhart and Winzeler (1996)). The emergence of expression arrays signals the dawn of the era of functional genomics. DNA expression arrays can provide important clues to gene function. The gene expression matrix allows comparison of expression profiles between genes and different samples. Similar expression behavior (e.g., similar change in expression level under similar conditions) suggests that the genes are likely to be co-regulated or possibly functionally related. Clustering gene expression data can group together genes of known similar function (Eisen, Spellman, Brown and Botstein (1998)). Genes with unknown function can be assigned tentative functions or a role in a biological process, based on the known function of genes in the same cluster. Single-nucleotide polymorphisms (SNPs) promise to

propel forward pharmacogenomics, the emerging field concerned with the dissection of the genetic basis of disease and therapeutic response. SNPs enable studies of association between a SNP and risk of a disease or drug response (McCarthy and Hilfiker (2000), Brookes (1999)). The associations are valuable for the identification of candidate genes for disease phenotype. The eventual determination of the functions of genes inferred from expression arrays and SNP databases will require experimental analysis in a systematic and high-throughput fashion to keep pace with the fast-growing genome, expression array, and SNP databases. Antisense technology is well suited for this endeavor.

For these high-throughput applications, the screening process for selecting effective antisense target sites must be efficient. Experimental approaches for finding effective antisense oligonucleotides are expensive and time-consuming, and are usually limited to a region of the mRNA. The combinatorial DNA-RNA oligonucleotide array technique is a promising experimental method (Milner, Mir and Southern (1997), Southern, Mir and Shchepinov (1999)). However, it has two drawbacks. First, the number of possible oligomers up to a preset length is huge for an mRNA. Secondly, large mRNAs can be hampered by their bulky size from approaching the oligomers densely distributed on the array surface (Southern, Mir and Shchepinov (1999)). Thus, use of selective oligomers designed by comprehensive computational screening provides a means for reducing an unnecessarily large number of oligomers. We anticipate that the combination of a reliable computational approach and experimental techniques such as the DNA-RNA arrays should prove to be the best strategy for realizing the great promise of antisense techniques (Figure 6).

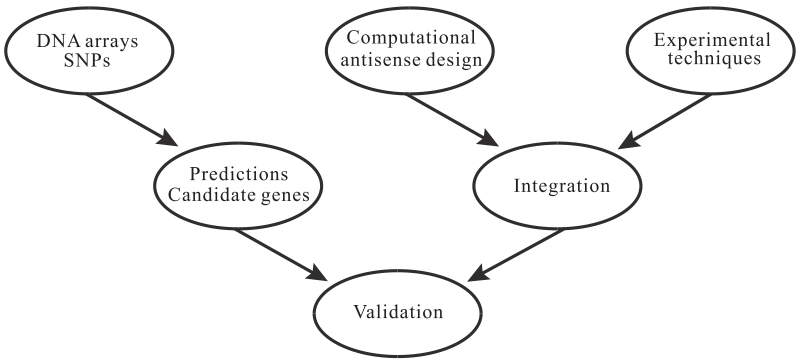

Figure 6. A potential high-throughput antisense framework for functional genomic and drug target validation. Systematic statistical analysis of DNA expression arrays and SNP databases can provide the basis for high-throughput functional analysis. Integration of computational antisense design and experimental techniques such as oligonucleotide arrays presents a rational, efficient, and high-throughput platform for antisense oligonucleotide screening.

6. Conclusions

The Bayesian inference approach has the potential to answer RNA folding questions that are unresolvable by mathematical algorithms. For example, the current free energy parameters for multibranched loops are heuristic estimates without any experimental support. By treating these parameters as random variables, and by using RNA sequences with phylogenetic structures deduced from comparative studies, Bayesian inferences and estimates for these parameters can be made under suitable modeling assumptions. Bayesian modeling may also permits a sensitivity analysis that would indicate which parameters are important for folding and which can be ignored or crudely estimated (Zuker (2000)).

The two-step process of nucleation and elongation is a simplifying representation of molecularly and energetically complex hybridization phenomena. For example, a single-base addition can drastically change the degree of inhibition. These factors partly explain the frustration over the elusiveness of a statistical model that could predict antisense efficacy with accuracy. Successful statistical modeling in genomic applications must adequately address the underlying biology. Potential areas for further statistical contributions to antisense research include the identification of factors that are correlated with antisense inhibition, and the building of statistical models for improved prediction of antisense efficacy. The former needs to focus on secondary and tertiary features of mRNA. Some correlation studies on primary structure motifs have been reported (Matveeva et al. (2000), Tu, Cao, Zhou and Israel (1998)). Potential RNA-protein interaction can also dictate the accessibility of an antisense target; however, addressing this issue is far beyond existing computational means. Improvement in statistical modeling needs to await a better understanding of the subtlety of the molecular mechanisms for hybridization.

Identification of effective antisense targets is important for the design of antisense oligonucleotides for antisense drug development, functional genomics, and drug target validation. In the post-genomic era, traditional tools for functional genomics and drug target validation can no longer keep pace with new sequence information rapidly accumulated from various genome projects. The antisense technique has emerged as an important tool for these applications both *in vitro* and *in vivo*. Fast-growing databases for DNA expression arrays and SNPs invite functional inference and predictions of candidate genes. These analyses present the basis for high-throughput antisense applications. The promise of antisense technology can only be fully realized with a proper integration of computation methods and experimental techniques.

Acknowledgements

The author is grateful to Kathleen McDonough of the Wadsworth Center for preliminary testing of antisense oligos for inhibition of *in vitro* translation of

*E. coli lac*Z. The Computational Molecular Biology and Statistics Core at the Wadsworth Center is acknowledged for providing computing resources for this work. The author is also grateful to an anonymous referee for suggestions to improve the presentation of the paper.

References

Asano, K., Niimi, T., Yokoyama, S. and Mizobuchi, K. (1998). Structural basis for binding of the plasmid ColIb-P9 antisense Inc RNA to its target RNA with the 5'-rUUGGCG-3' motif in the loop sequence. *J. Biol. Chem.* **273**, 11826-11838.

Bass, B. L. (2000). Double-stranded RNA as a template for gene silencing. *Cell* **101**, 235-238.

Bennett, C. F. and Cowsert, L. M. (1999). Antisense oligonucleotides as a tool for gene functionalization and target validation. *Biochim. Biophys. Acta.* **1489**, 19-30.

Bosher, J.M. and Labouesse, M. (2000). RNA interference: genetic wand and genetic watchdog. *Nat. Cell Biol.* **2**, 31-36.

Brookes, A. J. (1999). The essence of SNPs. *Gene* **234**, 177-186.

Brown, P.O. and Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. Nat. Genet. **21**, 33-37.

Burgess, T. L., Fisher, E. F., Ross, S. L., Bready, J. V., Qian, Y. X., Bayewitch, L. A., Cohen, A. M., Herrera, C. J., Hu, S. S. and Kramer, T. B., et al. (1995). The antiproliferative activity of c-myb and c-myc antisense oligonucleotides in smooth muscle cells is caused by a nonantisense mechanism. *Proc. Natl. Acad. Sci.* **92**, 4051-4055.

Crooke, S. T. (1999) Molecular mechanisms of action of antisense drugs. *Biochim. Biophys. Acta.* **1489**, 31-44.

Ding, Y. and Lawrence, C. E. (1999). A Bayesian statistical algorithm for RNA secondary structure prediction. *Computers and Chemistry* **23**, 387-400.

Ding, Y. and Lawrence, C. E. (2001). Statistical prediction of single-stranded regions in RNA secondary structure and application to predicting effective antisense target sites and beyond. *Nucleic Acids Res.* **29**, 1034-1046.

Driver, S. E., Robinson, G. S., Flanagan, J., Shen, W., Smith, L. E., Thomas, D. W. and Roberts, P. C. (1999). Oligonucleotide-based inhibition of embryonic gene expression. *Nat. Biotechnol.* **17**, 1184-1187.

Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.* **95**, 14863-14868.

Ferré-D'Amaré and A. R., Doudna, J. A. (1999). RNA folds: insights from recent crystal structures. *Annu. Rev. Biophys. Biomol. Struct.* **28**, 57-73.

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E. and Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegant. *Nature* **391**, 806-811.

Freier, S. M., Kierzek, R., Jaeger, J. A., Sugimoto, N., Caruthers, M. H., Neilson, T. and Turner D. H. (1986). Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl. Acad. Sci.* **83**, 9373-9377.

Galas, D. J. (2001). Making sense of the sequence. *Science* **291**, 1257-1260.

Harth, G., Zamecnik, P. C., Tang, J. Y., Tabatadze, D. and Horwitz, M. A. (2000). Treatment of Mycobacterium tuberculosis with antisense oligonucleotides to glutamine synthetase mRNA inhibits glutamine synthetase activity, formation of the poly-L-glutamate/glutamine cell wall structure, and bacterial replication. *Proc. Natl. Acad. Sci.* **97**, 418-423.

Ho, S. P., Bao, Y., Lesher, T., Malhotra, R., Ma, L. Y., Fluharty, S. J. and Sakai, R. R. (1998). Mapping of RNA accessible sites for antisense experiments with oligonucleotide libraries. *Nat. Biotechnol.* **16**, 59-63.

Ho, S. P., Britton, D. H., Stone, B. A., Behrens, D. L., Leffet, L. M., Hobbs, F. W., Miller, J. A. and Trainor, G. L. (1996). Potent antisense oligonucleotides to the human multidrug resistance-1 mRNA are rationally selected by mapping RNA-accessible sites with oligonucleotide libraries. *Nucleic Acids Res.* **24**, 1901-1907.

International Human Genome Sequencing Consortium (IHGSC) (2001). Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921.

Jacobson, A. B., Zuker, M. and Hirashima, A. (1987). Comparative studies on the secondary structure of the RNAs of related RNA Coliphages. In *Molecular Biology of RNA: New Perspectives* (Edited by M. Inouye and B. S. Dudock), 331-354. Academic Press, San Diego, CA.

Jacobson, A. B. and Zuker, M. (1993). Structural analysis by energy dot plot of a large mRNA. *J. Mol. Biol.* **233**, 261-269.

Lima, W. F., Monia, B. P., Ecker, D. J. and Freier, S. M. (1992). Implication of RNA structure on antisense oligonucleotide hybridization kinetics. *Biochemistry* **31**, 12055-12061.

Lockhart, D. J. and Winzeler, E. A. (2000). Genomics, gene expression and DNA arrays. *Nature* **405**, 827-836.

Mathews, D. H., Sabina, J., Zuker, M. and Turner, D. H. (1999). Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. *J. Mol. Biol.* **288**, 911-940.

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.

McCarthy, J. J. and Hilfiker, R. (2000). The use of single-nucleotide polymorphism maps in pharmacogenomics. *Nat. Biotechnol.* **18**, 505-508.

McCaskill, J. S. (1990). The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* **29**, 1105-1119.

Milner, N., Mir, K. U. and Southern, E. M. (1997). Selecting effective antisense reagents on combinatorial oligonucleotide arrays. *Nat. Biotechnol.* **15**, 537-541.

Mir, K. U. and Southern, E. M. (1999) Determining the influence of structure on hybridization using oligonucleotide arrays. *Nat. Biotechnol.* **17**, 788-792.

Ohlstein, E. H., Ruffolo, R. R. Jr. and Elliott, J. D. (2000). Drug discovery in the next millennium. *Annu. Rev. Pharmacol. Toxicol.* **40**, 177-191.

Rapaport, E., Levina, A., Metelev, V. and Zamecnik, P. C. (1996). Antimycobacterial activities of antisense oligodeoxynucleotide phosphorothioates in drug-resistant strains. *Proc. Natl. Acad. Sci.* **93**, 709-713.

Rong, Y. S. and Golic, K. G. (2000). Gene targeting by homologous recombination in Drosophila. *Science* **288** , 2013-2018.

Rowland, B, Purkayastha, A., Monserrat, C., Casart, Y., Takiff, H. and McDonough, K. A. (1999). Fluorescence-based detection of lacZ reporter gene expression in intact and viable bacteria including Mycobacterium species. *FEMS Microbiol. Lett.* **179**, 317-325.

Sohail, M. and Southern, E. M. (2000). Selecting optimal antisense reagents. *Adv. Drug. Deliv. Rev.* **44**, 23-34.

Southern, E., Mir, K. and Shchepinov, M. (1999). Molecular interactions on microarrays. *Nat. Genet.* **21** (1 Suppl), 5-9.

Stein, C. A. (1999). Two problems in antisense biotechnology: in vitro delivery and the design of antisense experiments. *Biochim. Biophys. Acta.* **1489**, 45-52.

Taylor, M. F., Wiederholt, K. and Sverdrup, F. (1999). Antisense oligonucleotides: a systematic high-throughput approach to target validation and gene function determination. *Drug Discov. Today* **4**, 562-567.

Thompson, J. D. (1999). Shortcuts from gene sequence to function. *Nat. Biotechnol.* **17**, 1158-1159.

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**, 25125-31.

Vanhée-Brossollet, C. and Vaquero, C. (1998). Do natural antisense transcripts make sense in eukaryote? *Gene* **211**, 1-9.

Venter, J. C. et al. (2001). The sequence of the human genome. *Science* **291**, 1304-1351.

Wianny, F. and Zernicka-Goetz, M. (2000). Specific interference with gene function by double-stranded RNA in early mouse development. *Nat. Cell. Biol.* **2**, 70-75.

Wuchty, S., Fontana, W., Hofacker, I. L. and Schuster, P. (1999). Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers* **49**, 145-165.

Xia, T., SantaLucia, J. Jr, Burkard, M. E., Kierzek, R., Schroeder, S. J., Jiao, X., Cox, C., and Turner, D. H. (1998). Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry* **37**, 14719-35.

Zamecnik, P. C. and Stephenson, M. L. (1978). Inhibition of Rous sarcoma virus replication and cell transformation by a specific oligodeoxynucleotide. *Proc. Natl. Acad. Sci.* **75**, 289-294.

Zuker, M. (2000). Calculating nucleic acid secondary structure. *Curr. Opin. Struct. Biol.* **10**, 303-10.

Zuker, M. (1989a). On finding all suboptimal foldings of an RNA molecule. *Science* **244**, 48-52.

Zuker, M. (1989b). The use of dynamic programming algorithms in RNA secondary structure prediction. In *Mathematical Methods for DNA Sequences* (Edited by M. S. Waterman), 159-184. CRC Press, Boca Raton, FL.

Zuker, M. and Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.* **9**, 133-148.

Bioinformatics Lab, Division of Molecular Medicine, Wadsworth Center, New York State Department of Health, Albany, NY 12201-0509, U.S.A.
E-mail: yding@wadsworth.org

(Received March 2001; accepted October 2001)