1034–1046 *Nucleic Acids Research, 2001, Vol. 29, No. 5*

Statistical prediction of single-stranded regions in RNA secondary structure and application to predicting effective antisense target sites and beyond

Ye Ding* and Charles E. Lawrence

Division of Molecular Medicine, Wadsworth Center, New York State Department of Health, Albany, NY 12201-0509, USA

Received as resubmission January 10, 2001; Revised and Accepted January 11, 2001

ABSTRACT

Single-stranded regions in RNA secondary structure are important for RNA–RNA and RNA–protein interactions. We present a probability profile approach for the prediction of these regions based on a statistical algorithm for sampling RNA secondary structures. For the prediction of phylogenetically-determined single-stranded regions in secondary structures of representative RNA sequences, the probability profile offers substantial improvement over the minimum free energy structure. In designing antisense oligonucleotides, a practical problem is how to select a secondary structure for the target mRNA from the optimal structure(s) and many suboptimal structures with similar free energies. By summarizing the information from a statistical sample of probable secondary structures in a single plot, the probability profile not only presents a solution to this dilemma, but also reveals 'well-determined' single-stranded regions through the assignment of probabilities as measures of confidence in predictions. In antisense application to the rabbit β-globin mRNA, a significant correlation between hybridization potential predicted by the probability profile and the degree of inhibition of *in vitro* translation suggests that the probability profile approach is valuable for the identification of effective antisense target sites. Coupling computational design with DNA–RNA array technique provides a rational, efficient framework for antisense oligonucleotide screening. This framework has the potential for high-throughput applications to functional genomics and drug target validation.

INTRODUCTION

In the last two decades several important approaches have been developed for the prediction of secondary structure from an RNA sequence. The popular *mfold* program, developed with dynamic programming algorithms, predicts optimal structure and suboptimal structures through free energy minimization (1–3; http://bioinfo.math.rpi.edu/~zukerm/rna/). The partition function approach by McCaskill (4) computes base pair probabilities and the binding probability for any base. A C program for this algorithm is available in a suite of RNA secondary structure software known as the Vienna RNA package. This package was developed by a theoretical chemistry group at the University of Vienna (5; http://www.tbi.univie.ac.at/~ivo/RNA/). Ding and Lawrence (6) formulated the RNA folding problem in a Bayesian statistical framework and extended the partition function method by generating a statistically representative sample of the probable structures. Heuristic Monte Carlo algorithms based on kinetics and genetic principles have also been described (7–10).

In this paper, we explore the use of a sampling approach for the prediction of single-stranded regions in an RNA molecule. While we focus on the important antisense application, single-stranded regions, particularly destabilizing loops, can play many important functional roles. These include, for example, protein binding (11,12), ribozyme binding and catalysis (13), regulation of cellular processes (14,15), pseudoknot formation and tertiary interactions for kissing hairpins, bulge–loop complexes, hairpin loop–internal loop complexes, etc. (16). For these applications, computational prediction of single-stranded regions can also be helpful for experimental design for structure probing by RNases or chemical means.

A novel regulatory mechanism was recognized about two decades ago: an oligodeoxynucleotide can bind to an mRNA through complementary base pairing to block its translation (17). The discovery that oligonucleotides can play a regulatory role in gene expression led to the development of the antisense strategy to artificially control gene expression. Although variable degrees of success have been achieved in the application of antisense methods to the research of biological phenomena and human disease treatment, it has been proven that antisense oligonucleotides (ASOs) are able to modulate gene expression in both prokaryotes and eukaryotes (18). In recent years, several antisense compounds for disease treatment have been evaluated in clinical trials with promising results, and more compounds are being evaluated in clinical trials (19,20). In 1998, Vitravene (Isis Pharmaceuticals, Carlsbad, CA) became the first antisense drug approved by the Food and Drug Administration. It is used to treat cytomegalovirus retinitis in AIDS patients.

For ASOs to be effective, the complementary target sequence on mRNA must be available for hybridization. RNA nucleotides can be inaccessible when they are sequestered in

*To whom correspondence should be addressed. Tel: +1 518 486 1719; Fax: +1 518 473 2900; Email: yding@wadsworth.org

secondary structure. The usually weaker tertiary interactions and RNA–protein interactions can also be factors that affect accessibility. The identification of regions likely to remain single-stranded in RNA secondary structure can be an important part of antisense technology.

Lima *et al.* (21) concluded that the tightest binding of ASOs occurs at target sites for which disruption of the target structure is minimal, and single-stranded regions should be selected over double-stranded regions in the consideration of target sites. Vickers *et al.* (22) directly demonstrated that mRNA structures play a significant role in determining antisense oligonucleotide efficacy *in vivo*. They concluded that discovery of active ASOs requires identification of unstructured sites in the cellular mRNA. Matveeva *et al.* (23) also found that there is a correlation between single-stranded specific probes and accessible sites for antisense targeting, but there are a few exceptions, probably due to steric hindrance that limits RNase H access. It has been speculated that duplex formation is initiated at an accessible substructure that includes a site for nucleation with unpaired bases and then propagates from the nucleation site through a 'zippering' process (24,25). A hairpin of four unpaired bases can be involved in hybrid formation (26).

A few secondary-structure-prediction based computational approaches to the evaluation of potential antisense targets have been reported. Stull *et al.* (27) presented thermodynamic indices by averaging relevant free energies of secondary structures generated from a Monte Carlo RNA folding algorithm based on an evolutionary heuristic (7,28). Because this algorithm does not guarantee the generation of a valid statistical sample of low energy structures, the most likely structure is not necessarily the lowest free energy structure.

For the genomic RNA (~9700 nt) and the complementary RNA strand of the human immunodeficiency virus type 1 (HIV-1), Sczakiel *et al.* (29) found that local folding potential can shed light on effective antisense targets. The local folding potential was computed for each of successive overlapping segments of a chosen window width (ranging from 50 to 400 nt) along the RNA chain, by folding each segment with *mfold* and computing its minimum free energy. This method was proposed by Le *et al.* (30) for assessing stable structures in HIV-1. Because long distance interactions and short term interactions between the nucleotides near the ends of the segment and the neighboring nucleotides outside the segment are ignored, this method appears to be reasonable only for relatively long window width, as it cannot address the hybridization potential of individual nucleotides or short segments.

Zhao and Lemke (31) performed a comparative analysis on 22 RNAs using *mfold*. The RNAs were previously studied for selective gene inactivation by ASOs and ribozymes, small catalytical RNA molecules that specifically bind to and cleave target RNAs. Despite limited representation of alternative structures by four or five suboptimal foldings, they found a correlation between the predicted base-pairing accessibility of the targets and the experimental efficacy of the antisense reagents. They recommend that the cleavage site for ribozymes should fall within a loop of at least four nucleotides, and one, preferably both, of the 5′- and 3′-ends of the antisense segment should fall within a single-stranded rather than a stem region. Despite the inherent difficulty in selecting a representative sample of the suboptimal foldings, James and Cowe (32) proposed addressing the hybridization potential using suboptimal foldings from *mfold* and showed that their procedure works well for the rat OX40 mRNA.

These findings lend additional support to the importance of exploring secondary structure in the selection of antisense targets. Our strategy is to focus on single-stranded regions in RNA secondary structure, in particular those of at least four consecutive unpaired bases. The Vienna package can calculate the probability of a single base being unpaired; however, it cannot address the hybridization potential of a region. This is not a problem for the sampling-based probability profile approach we propose in this work, which can overcome limitations of existing computational approaches. We illustrate our approach with applications to representative RNA sequences and an antisense application to rabbit β-globin mRNA.

MATERIALS AND METHODS

Statistical sampling of RNA secondary structures

The structure sampling algorithm of Ding and Lawrence (6) yields a representative statistical sample of secondary structures. This algorithm was based on free energies for stacking in helices. The sampling probabilities are computed using partition functions calculated in the forward step of the algorithm. For more sophisticated and realistic 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 (4) 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 computed in the forward step.

The extended algorithm accommodates the up-to-date free energy rules and parameters developed by Turner's group (33,34) with the exception of coaxial stacking. These include free energies for stacking in a helix, stacking for a terminal mismatch in a hairpin loop (size ≥4 nt) or an interior loop, and penalties for hairpin, bulge, interior and multi-branched loops. Free energies for single base stacking (dangling ends) are used for exterior and multi-branched loops. For hairpins, a bonus for UU and GA first mismatches (included in the terminal stacking data) and a bonus for G·U closure preceded by two G nucleotides in base pairs are applied, and a penalty for oligo-C loops (all unpaired nucleotides are C) are used. A table is consulted for tetraloops (hairpin loops with four unpaired nucleotides). For a bulge of 1 nt, the stacking energy of the adjacent pairs is added. For interior loops, tables for 1×1, 1×2 and 2×2 loops are consulted and a penalty for asymmetry is applied. A terminal A-U, G·U penalty is explicitly applied to exterior loop, multi-branched loops, bulges longer than 1 nt and triloops (hairpin loops with three unpaired nucleotides), while this penalty is included in the terminal stacking data for hairpin loops (size ≥4 nt) and interior loops. These free energy parameters are for 37°C and 1 M NaCl; however, this algorithm can be used with any set of nearest neighbor parameters derived for other conditions.

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. The extended algorithm samples exactly according to the Boltzmann distribution, i.e. it can generate a statistical sample of any desired size from the Boltzmann ensemble of secondary structures. The sampling process is similar to the traceback algorithm employed in the dynamic programming algorithms (1–3) but differs in that the base pairing is randomly sampled from Boltzmann probabilities rather than chosen to yield a minimum free energy structure. 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.

Probability profiles of single-stranded bases and sequences

From recursively derived partition functions for an RNA sequence of n bases, McCaskill (4) also presented recursions for marginal base pairing probability,

P_{ij} = Prob(base i and base j form a pair);

then the probability that base i is unbound, i.e. single-stranded, is

$$q_i = 1 - \Sigma_{(i+1) \le j \le n} P_{ij} - \Sigma_{1 \le j \le i} P_{ji}.$$

As emphasized by McCaskill (4), 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, $\{q_i\}$ statistically describe the antisense hybridization potential for every nucleotide in the sequence. Alternatively, the sampling method presents a means to estimate q_i 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 $\{q_m\}$, $i \le m \le j$, 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 (26,31), we set $W = 4$ for antisense application.

RESULTS

Examples of probability profiles

For single-stranded bases in *Escherichia coli* $tRNA^{Ala}$, Figure 1A demonstrates the probability profile estimated from 1000 sampled secondary structures, the probability profile computed by the Vienna RNA package, the profile indicated by the minimum free energy (MFE) structure computed with version 3.1 of *mfold* (33,34; http://bioinfo.math.rpi.edu/~mfold/rna/form1.cgi) and that indicated by the phylogenetically-determined structure (35). A sample size of 1000 was found to be adequate because the profile estimates from this sample and a larger sample of 10 000 structures were not readily distinguishable. For the unpaired individual bases, the probability profile and the profile by the MFE structure are comparable. This is generally expected because the MFE structure is the most probable structure in the sample. However, the MFE structure substantially underpredicts the width of the region around nucleotide G^{35} of the anticodon loop, while a significant portion of the sample adequately reveals the width. For the region between nucleotide G^{30} and A^{76}, the sampling approach and version 1.3.1 of the Vienna RNA package gave comparable results; however, for the region between nucleotides C^{5} and C^{25}, the sampling profile predicted the phylogenetic structure substantially better than the Vienna profile. This version of Vienna package is based on an earlier compilation of Turner's free energy parameters described by Walter *et al.* (36). It has been shown that the latest update improves the prediction of secondary structure (33). This explains the better performance by our sampling algorithm.

Figure 1B shows the probability profile of the sample for single-stranded sequences with a sequence width of 4 nt. For comparison with the phylogenetic structure, a dot with coordinates (i, 1) is shown in Figure 1B if the 4 nt sequence starting at nucleotide i is single-stranded, and a dot with coordinates (i, 0) is plotted if any of the four nucleotides is base paired. Similarly, the MFE structure is plotted. The unstructured region of the anticodon loop is missed by the MFE structure, but is revealed by the sampling profile through a peak of substantial probability. For the two sampling profiles in Figure 1A and B, not only do the single-stranded regions in the phylogenetic structure correspond well to the local peaks of the probability profiles, but also the width of the regions matches the width of the peaks with only one exception, region ^{32}AUGGCAU^{38} of the anticodon loop. The peak for this region in the phylogenetic structure is slightly narrower because two Watson–Crick pairs A^{32}–U^{38} and U^{33}–A^{37} are likely to be predicted by any free-energy based algorithm, while these two base pairs are absent in the phylogenetic structure. In Figure 1B, the peak of the sampling profile between A^{32} and U^{38} is much lower than the corresponding peak in Figure 1A because, while the single-stranded probability for each of G^{34}, G^{35} and C^{36} is >0.96, the probabilities for U^{33} and A^{37} are <0.28. Thus, for identifying a single-stranded region of at least 4 nt, a high peak in the profile of single-stranded bases can be visually misleading when the width of the peak is <4 nt. The probability profile of single-stranded sequences presents a clearer picture of potential antisense sites, because it has fewer and narrower peaks than the profile of single-stranded bases. This probability profile cannot be obtained by the Vienna RNA package or any other existing computational methods.

To further illustrate our approach, we present the probability profiles in Figure 2A–D for the following representative RNA sequences with phylogenetically-determined secondary structures: *Xenopus laevis* oocyte (Xlo) 5S rRNA (37), domain II of *E.coli* 16S rRNA (38), *E.coli* RNase P (39) and group I intron

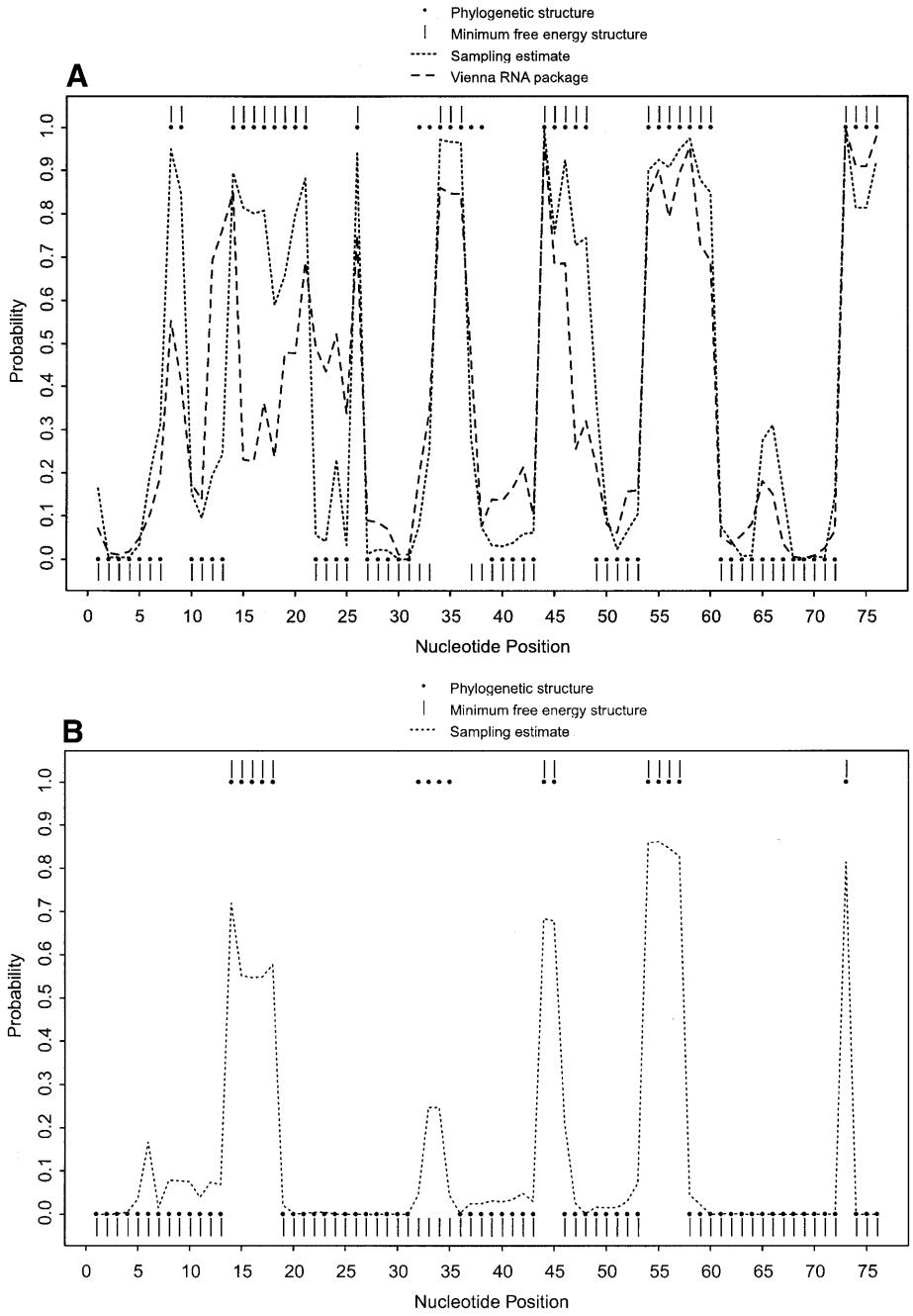
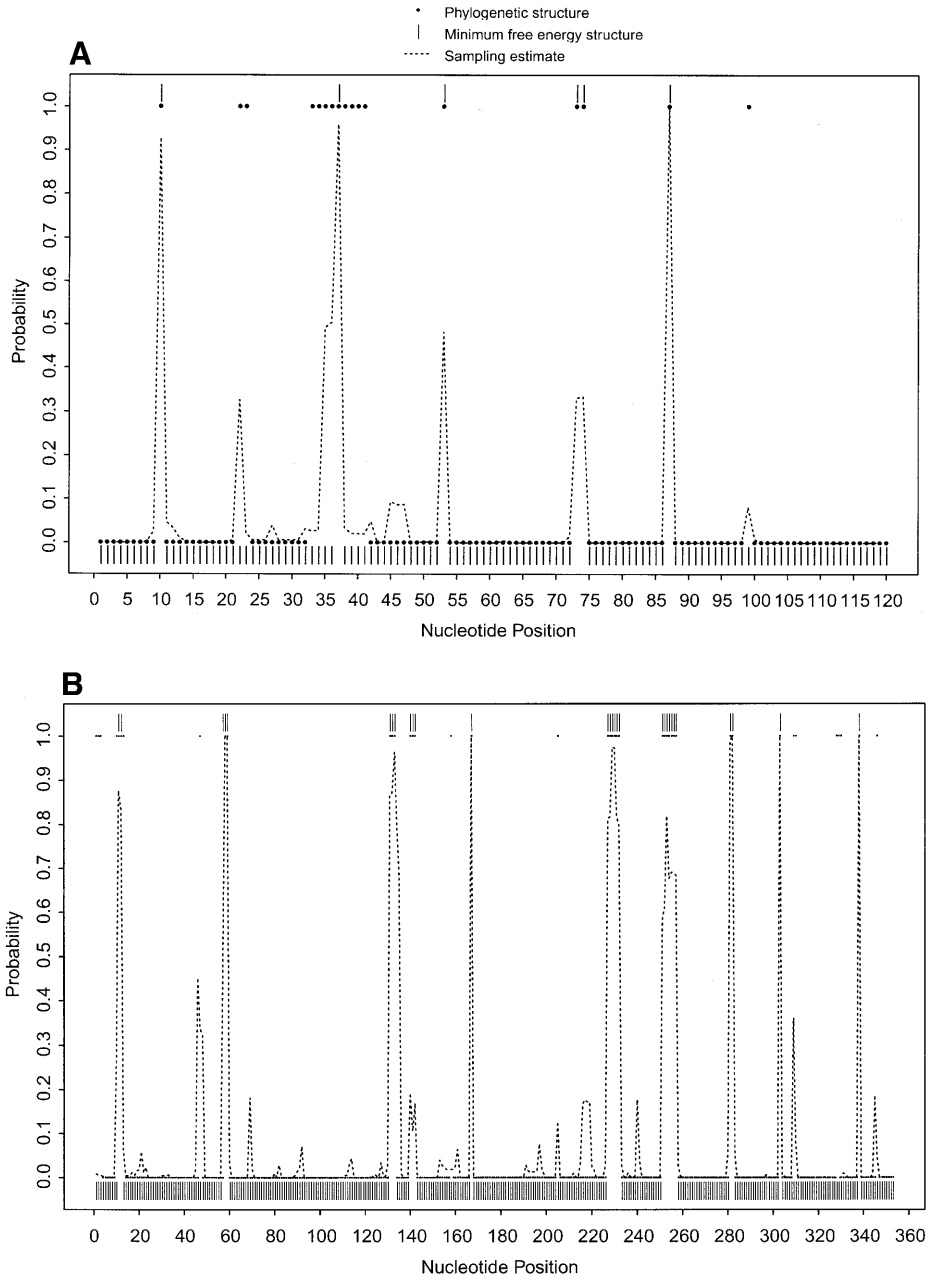


Figure 1. Probability profiles for *E.coli* $tRNA^{Ala}$, with sampling estimates computed from 1000 sampled secondary structures. (**A**) The probability profiles for single-stranded nucleotides (sequence width $W = 1$) indicated by the phylogenetic structure (large dots) and by the minimum free energy structure (vertical bars), estimated by the sampling algorithm (short dashed line) and computed by the Vienna RNA package (long dashed line). For the region between C^5 and C^{25}, the sampling estimate predicts the phylogenetic structure substantially better than the Vienna RNA package. (**B**) The probability profiles for single-stranded sequences of four consecutive nucleotides (sequence width $W = 4$) in *E.coli* $tRNA^{Ala}$ indicated by the phylogenetic structure (large dots) and by the minimum free energy structure (vertical bars) and estimated by the sampling algorithm (dashed line). The probability profile cannot be computed by the Vienna RNA package or other existing algorithms.

from 26S rRNA of *Tetrahymena thermophila* (40,41). For these sequences, phylogenetically-determined single-stranded regions correspond to peaks in the probability profile with near certainty (Fig. 2; P_C in Table 1). On the other hand, peaks with at least a maximum probability of 0.5 almost certainly point to single-stranded regions (P_{C1} in Table 1); peaks with a maximum probability between 0.2 and 0.5 have at least a 50% chance of correctly indicating single-stranded regions (P_{C2} in Table 1), whereas there is a far smaller but appreciable chance for peaks with a maximum probability <0.2 (P_{C3} in Table 1) to correctly indicate single-stranded regions. As in the case of *E.coli* $tRNA^{Ala}$, for all these RNA sequences, the probability



profile reveals more single-stranded regions in the phylogenetic structure than the MFE structure (P_I in Table 1). The substantial improvement is because the alternative structures in the sample are able to reveal structural motifs not predicted by the MFE structure. On the other hand, the motifs in the MFE structure are well reported by the sample because it is the most probable structure in the sample. The improvement is noticeably greater for *E.coli* RNase P, which has the highest percentage of nucleotides in pseudoknots, a motif not allowed by either *mfold* or our current algorithm.

The results reveal variation in the reliability of prediction among different RNAs. For free energy minimization for the prediction of RNA secondary structure, variability in the reliability of prediction for different RNAs has been well documented (33,42,43). Because our sampling algorithm is also based on free energies, it is not surprising to observe a similar phenomenon in our context. There is also substantial variability in the maximum probabilities for the peaks that correspond to single-stranded regions. Similarly, for minimum free energy prediction of secondary structure, there is variability in the reliability of predictions for different regions of a sequence (43). The summary in Table 1 indicates that single-stranded regions predicted by high probability peaks are 'well-determined' by the probability profile. In other words, these regions are highly stable and, thus, are present with high probability in a sample of probable secondary structures. For regions of lower stability, their probabilities are either moderate or low, because alternative structural motifs will be

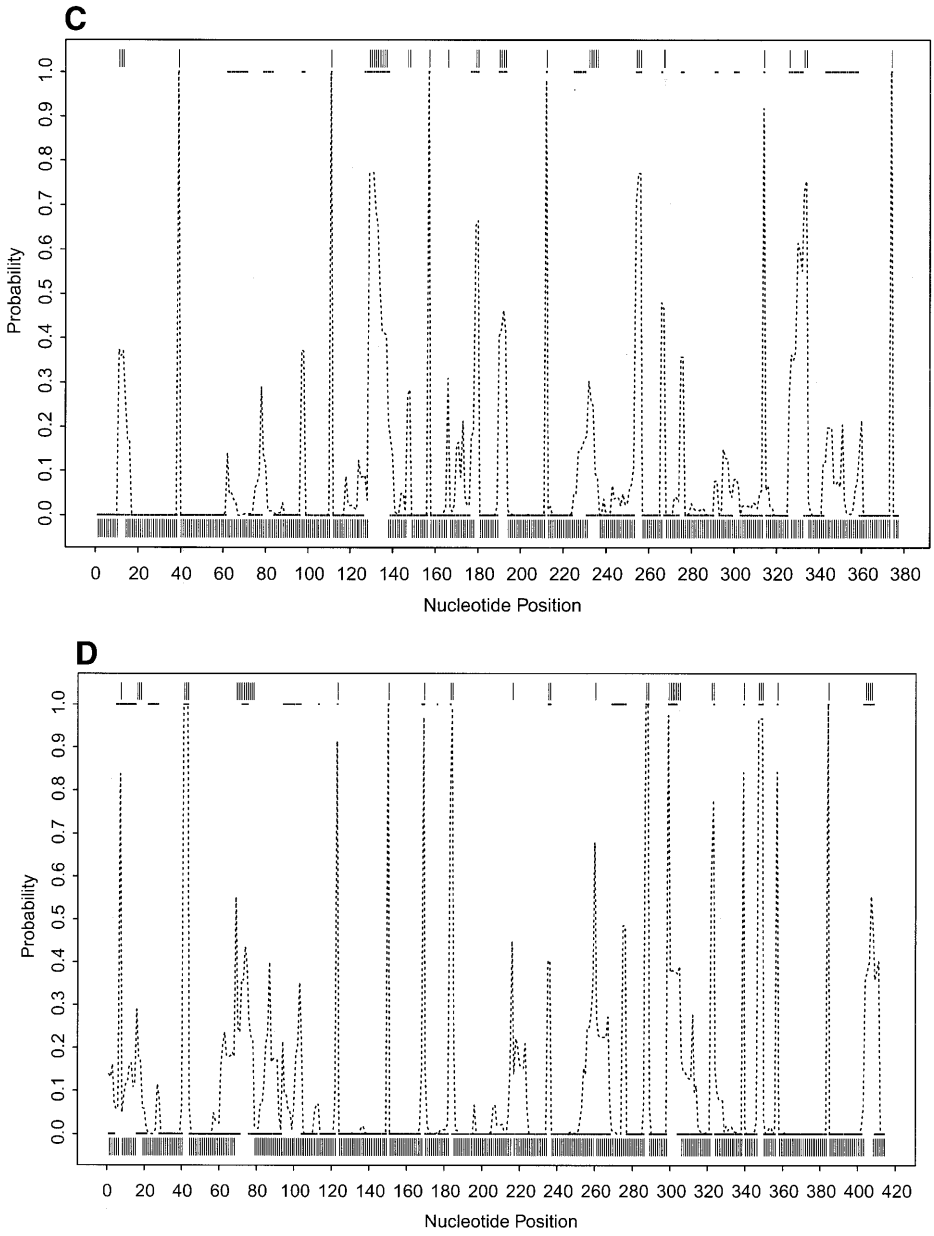


Figure 2. (Opposite and above) Probability profiles (sequence width $W = 4$) for other representative RNA sequences, with sampling estimates computed from 1000 sampled secondary structures. For *X.laevis* oocyte 5S rRNA (**A**, opposite), the large dots present the profile indicated by the phylogenetic structure, the dashed line is the sampling estimate and the vertical bars represent the minimum free energy structure. For *E.coli* 16S rRNA domain II (**B**, opposite), *E. coli* RNase P (**C**, above) and Group I intron from 26S rRNA of *T.thermophila* (**D**, above), the small solid squares (adjacent squares appear to form line segments) present the profile indicated by phylogenetic structure, the dashed line is the sampling estimate and the vertical bars represent the minimum free energy structure. For the *Tetrahymena* Group I intron, a 6 bp double-stranded region called P3 (38) in the phylogenetic structure is not considered here because of the creation of a pseudoknot. The current sampling algorithm needs to be extended to predict certain types of pseudoknots.

more likely to be present in the sample. Our sampling algorithm gives a complete statistical presentation of probable competing alternative structures. Thus, the probability profile provides a statistical delineation of single-stranded regions with varying stabilities.

Antisense application

The rabbit β-globin mRNA (589 nt, GenBank accession V00879, coding region 54–497) has been well studied for antisense inhibition of protein synthesis (24,44,45). Cazenave *et al.* (44) used an 11mer and three 17mers targeted to rabbit β-globin mRNA in a wheatgerm extract as well as in microinjected *Xenopus* oocytes. Goodchild *et al.* (45) examined the inhibition of cell-free translation by eight phosphodiester ASOs targeted to this mRNA. Milner *et al.* (24) described a combinatorial oligonucleotide array technique for hybridization assessment of oligonucleotides within a given region. For the rabbit β-globin mRNA, an array of 1938 oligonucleotides

Table 1. Correspondence between phylogenetically-determined single-stranded regions and peaks on the probability profile and improvement in predictions over minimum free energy structure

RNA sequence	Accession no.	Length (nt)	P_C (%)[a]	P_{C1} (%)[b]	P_{C2} (%)[c]	P_{C3} (%)[d]	P_I (%)[e]
Escherichia coli $tRNA^{Ala}$	X66515	76	100	100	100	0	20
Xenopus laevis oocyte 5S rRNA	K02695	120	100	100	100	25	28
Escherichia coli 16S rRNA domain II	J01695	353	82	100	50	33	29
Escherichia coli RNase P	V00338	377	100	100	58	50	40
Tetrahymena thermophila LSU group I intron	V01416	413	95	88	67	29	19

[a]P_C is the percentage of phylogenetically-determined single-stranded regions (region here is either a sequence of four consecutive nucleotides or several such sequences in a row) that correspond to peaks (regardless of the magnitude of the maximum probability) in the probability profile in Figure 2.
[b]For peaks with a maximum probability ≥0.5, P_{C1} is the percentage of these peaks that correspond to single-stranded regions.
[c]P_{C2} is the percentage of the correspondence for peaks with a maximum probability between 0.2 and 0.5.
[d]P_{C3} is the percentage of the correspondence for peaks with maximum probability <0.2.
[e]A probability profile predicts more single-stranded regions in the phylogenetic structure than the minimum free energy structure (Figs 1B and 2). P_I is the percentage of improvement in the prediction by the probability profile over the MFE structure. This is computed by the number of regions missed by the MFE structure but predicted by the probability profile divided by the total number of single-stranded regions in the phylogenetic structure (e.g., seven for Xlo 5S RNA) and multiplied by 100%.

Table 2. Comparison of inhibition of rabbit β-globin synthesis in cell-free translation systems and hybridization potential predicted by probability profile for rabbit β-globin mRNA

ASO name (length in nt)	Target sequence/site on mRNA	% Inhibition[a]	Hybridization potential	Reference
β1 (20)	A^{14}–C^{33}/5′-UTR	23 (5.2)	high	45
β2 (20)	C^{46}–G^{65}/start	61 (5.2)	high	45
β3 (20)	A^{144}–C^{163}/coding	18 (5.2)	moderate	45
β4 (20)	G^{207}–A^{226}/coding	43 (5.2)	high	45
β5 (22)	A^{1}–G^{22}/cap	67 (5.2)	high	45
β6 (23)	U^{23}–A^{45}/5′-UTR	47 (5.2)	high	45
β7 (CCC+β5, 25)	A^{1}–G^{22}/cap	75 (5.2)	high	45
β8 (β7β6, 48)	A^{1}–A^{45}/cap	89 (2.6)	high	45
β6+β7 (mixture)	A^{1}–A^{45}/cap	89 (2.6)	high	45
BG1 (17)	C^{46}–U^{62}/start	50 (0.1)	high	24
BG2 (17)	A^{51}–C^{67}/start	50 (0.5)	high	24
BG3 (15)	C^{85}–U^{95}/coding	0 (1)	low	24
17 Glo[3-19] (17)	A^{3}–A^{19}/cap	72 (0.5)	high	44
17 Glo[51–67] (17)	U^{51}–C^{67}/start	95 (0.5)	high	44
11 Glo (11)	A^{44}–A^{54}/start	65 (0.5)	high	44
17 Glo[113–129] (17)	U^{113}–G^{129} / coding	95 (0.5)	low	44

[a]Numbers in parentheses represent ASO concentration in μM.

up to a length of 17 bases was used to measure the ASO:mRNA hybridization potential. These oligonucleotides were complementary to the first 122 bases of the mRNA. Three oligomers, BG1, BG2 and BG3, were chosen for study by *in vitro* translation in wheatgerm extract and the RNase H assay.

In our analysis, the results for BG1, BG2 and BG3 are directly compared to the data from the other two groups, because all these ASOs were studied in cell-free translation systems and the percentages of translation inhibition were reported (Table 2). The inhibition percentages facilitate a quantitative comparison and assessment of the correlation between inhibition of cell-free translation and computational predictions. The qualitative array hybridization data of Milner *et al.* (24) and the computational predictions were summarized and compared separately (Table 3).

The probability profile with a sequence width of 4 nt was computed with a sample of 1000 secondary structures for the rabbit β-globin mRNA. The probability profile and the profile by the MFE structure for the region A^{1}–U^{230} are shown in Figure 3, as the ASOs in these studies were targeted to this part of the mRNA. The target sites on the mRNA, the inhibition effect in cell-free translation systems in the three studies, and the hybridization potential predicted by the probability profile are summarized in Table 2. For this analysis, the hybridization potential was assessed as high if, for the target site, there was

Table 3. Comparison of the intensity of ASO:mRNA hybridization on the oligodeoxynucleotide array and the probability profile for the first 122 bases of rabbit β-globin mRNA

Region	Hybridization intensity	Probability profile (peak feature)
A^{1}–C^{37}	not detectable	high peaks (narrow)
C^{46}–C^{60} [a]	high	high peak (wide)
A^{61}–C^{91}	weak but detectable	low
A^{76}–A^{90}	not detectable	low
C^{94}–G^{110}	moderate	moderate

[a]C^{46}–C^{60} is contained in two 16mers C^{46}–A^{61} and A^{45}–C^{60}, and three 17mers C^{46}–U^{62} (BG1), A^{44}–C^{60} and A^{45}–A^{61}. The hybridization yields for ASOs complementary to these six sequences are at least three times that of any other oligonucleotides in the array (24).

at least one peak with probability ≥0.6; the potential was considered moderate for a peak with probability between 0.3 and 0.6; and the potential was low for a site with a probability <0.3 of being partly single-stranded. For ASOs described by Cazenave *et al*. (44), the inhibition figures for wheatgerm extract were estimated from figures 3 and 7 in that report. The region A^{1}–A^{45} was targeted by five of eight ASOs described by Goodchild *et al*. (45). There are three high probability sequences in this region: A^{1}–C^{4}, A^{18}–U^{21} and U^{36}–A^{45}. They explain the predicted high hybridization potential for β5, β6, β7, β8 and β6 + β7. The moderate inhibition by β1 indicates that A^{18}–U^{21} alone is not as effective as the other two. One explanation is that the two adjacent nucleotides, C^{17} and G^{22}, are predicted to almost certainly engage in G·C pairing and, thus, they might present a substantial energy barrier for hybridization elongation by 'zippering'. The high inhibition by β8 and β6 + β7 suggests that an antisense effect can be enhanced by simultaneously targeting several high potential sites. We also found consistent results for BG1, BG2 and BG3 reported by Milner *et al*. (24). Clear inconsistency between our predictions and the observed inhibition was found for 17 Glo[113–129] of Cazenave *et al*. (44), which appears to be an exception to the rule-of-thumb of at least four unpaired bases (26,31). In the case of an effective antisense site with less than four unpaired bases, the site would not be predicted by the probability profile with a sequence width of 4 nt. On the target site of 17 Glo[113–129], the probabilities of being unpaired for U^{125} and G^{126} are 0.61 and 0.56, respectively, but the probabilities are <0.1 for adjacent bases U^{124} and G^{127}. Among many other potential reasons for poor prediction in this case could be tertiary interactions and RNA–protein interactions, and self-folding of the oligomer that are unaccounted for by the current algorithm.

If we associate low, moderate and high hybridization potential with inhibition of 0–19, 20–39 and 40–100%, respectively, then for 13 of the 16 ASOs (81%) examined here, the hybridization potential revealed by the probability profile is indicative of the antisense inhibitory effect. For all the ASOs, there is a significant correlation (P = 0.0147, correlation coefficient = 0.597) between the hybridization potential predicted by the probability profile and the degree of translation inhibition. For β1–β8, there is a substantially higher correlation (P = 0.0037, correlation coefficient = 0.882). In contrast, Stull *et al.* (27) found no significant correlations between observed inhibition and any predictive indices for β1–β8. For ASOs described by Cazenave *et al*. (44), Stull *et al.* (27) found a correlation between Dscore, one of their indices, and inhibition for oligomer concentration at 6 μM, but no significant correlation for oligomer concentrations <6 μM. The probability profile and the MFE structure give

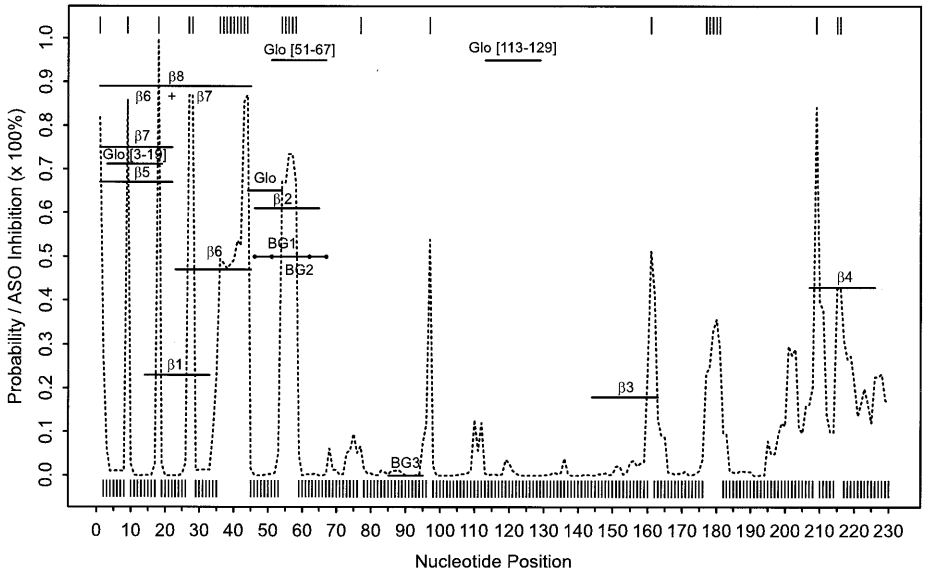


Figure 3. The probability profile for single-stranded sequences of four consecutive nucleotides (sequence width W = 4) estimated by 1000 sampled secondary structures (dashed line) and the profile indicated by the minimum free energy structure (vertical bars) for rabbit β-globin mRNA and the experimentally measured inhibition of ASOs in cell-free translation systems. The profile is shown for the region of the first 230 nt that is targeted by the ASOs. The length and binding sites of the ASOs are indicated by horizontal lines with the names of the ASOs centered above or below the lines. These lines also indicate the inhibition of translation through their position on the vertical axis. The vertical axis also shows the probability for the profile with inhibition corresponding to probability multiplied by 100%.

comparable predictions of single-stranded regions. However, without an associated measure of confidence, there is a lack of correlation between the binary prediction by the MFE structure and the degree of translation inhibition ($P = 0.567$, correlation coefficient = 0.155). This exemplifies the observation that there is limited success in using MFE structure for antisense design (46). Because the sampling profile provides a statistical measure of confidence in the predictions, it is not surprising that the profile is found to be generally indicative of the degree of translation inhibition.

For the hybridization intensity data of Milner *et al.* (24), there is very good agreement between the hybridization intensity and the probability profile for regions C^{46}–C^{60}, A^{76}–C^{90} and G^{94}–G^{110} (Table 3). The hybridization intensity for region A^{61}–C^{91} is in reasonable agreement with the probability profile. In this region, the maximum probability of a peak is ~0.1. For a peak with a maximum probability <0.2, there is an appreciable chance for the peak to correctly predict single-stranded regions (Table 1). Thus, weak hybridization is possible for low but appreciable probabilities. For region A^{1}–C^{37}, it is an intriguing contrast that the hybridization data are in disagreement both with the data described by Goodchild *et al.* (45) and with the probability profile, but the probability profile is in good agreement with the data of Goodchild *et al*. (45). The length of the oligomers, 20–45 nt discussed by Goodchild *et al.* (45), and at most 17 nt in the report of Milner *et al.* (24), offers an explanation for the conflicting results. Goodchild *et al.* (45) indicated that a greater inhibition could be obtained by covering a longer portion of the mRNA. This is evidenced by the greater inhibition of β8 or a mixture of β6 and β7 than either β6 or β7 alone (Table 2). There are several sharp peaks in the probability profile for this region. Thus, a plausible explanation from the profile is that substantially longer ASOs cover more peaks in this region and hence enhance the chance of both nucleation and propagation of duplex formation. Although oligomer length has a positive effect on translation inhibition in this case, this may not be generally true (24). We also note that the conclusion of insignificant hybridization by Milner *et al.* (24) for region A^{1}–C^{37} appears to be based on the lack of a continuous subregion with detectable hybridization. In this region, there are two isolated intensity bands in figure 1 of Milner *et al.* (24), indicating substantial hybridization at sequence positions that were also targeted by Glo[3–19] of Cazenave *et al.* (44).

The six oligomers containing bases C^{46}–C^{60} (Table 3), and BG2, β2, Glo[51–67] in Table 2 share one common feature on the profile: a relatively wide, high probability peak between A^{54} and U^{58}, with ^{54}AUG^{56} being the initiation codon. This suggests that a smooth and relatively wide peak on the probability profile can be a high potency antisense site because the chance of hybridization is improved for a wider single-stranded region.

DISCUSSION

Prediction of single-stranded regions

In this work, we have demonstrated the probability profile approach for the prediction of single-stranded regions in RNA secondary structure. The results on the test sequences show good correspondence between single-stranded regions by phylogenetic structure and predictions by the probability profile. In particular, high probability peaks correspond to single-stranded regions with high reliability. The variability in the reliability of predictions suggests that the results from the test sequences may not indicate the reliability of prediction for other RNA sequences. Poor predictions for a particular RNA or regions of the sequence can be partly attributed to uncertainties in thermodynamic parameters, pseudoknots and other tertiary interactions and RNA–protein interactions. These factors are not taken into account in the current algorithm. These caveats are associated with other efficient and free-energy based algorithms for RNA secondary structure prediction. In future work, we plan to extend the sampling algorithm to address certain types of pseudoknots.

The sampling stage of the algorithm is rather fast, although the implementation of the algorithm in Fortran 77 is not yet highly optimized. For the rabbit β-globin mRNA, it took 527 s to complete the partition function calculation, and 75 s to generate 1000 structures on a 300 MHz CPU of an Ultra 2 SPARC workstation. The computation time for the partition function calculation can be substantially improved if simpler multibranched loop evaluation is made possible, as was done for version 3.1 of *mfold* (33). This will be investigated. The sampling algorithm can be used for estimating the probability of any structural motif, such as helices, hairpin loops, bulges or interior loops, in addition to single-stranded regions.

Computational screening of effective antisense sites

For long mRNA sequences, there are many suboptimal foldings with free energies close to the minimum free energy. It has been a practical problem for antisense experimentalists to select one of the low free energy structures as the basis for antisense design. Furthermore, the suboptimal foldings from *mfold* do not guarantee a statistically unbiased sample of probable secondary structures. This makes it difficult to assign a statistical measure of confidence for predictions based on these suboptimal foldings. By summarizing a statistical sample of probable structures in a single plot, the probability profile approach overcomes these difficulties. The 'well-determined' single-stranded regions are revealed by peaks with high probabilities on the profile. While the test RNA sequences in Figures 1 and 2 probably have a single well-defined structure, it is possible that mRNAs do not. Statistical sampling of probable structures provides a suitable means to address this uncertainty. This is demonstrated by the substantial improvement in predictions over the minimum free energy structure. The sampling method also has the advantage that it does not require the generation of a huge number of all possible structures, as suggested previously (47).

The probability profile approach offers a comprehensive computational screening of the entire mRNA. For several other mRNA sequences with length ranging from 1 to 3 kb, we observed 15–20 high hybridization sites per kilobase (data not shown). These sites provide ample opportunities for rational design of antisense oligomers. An antisense oligomer is the reversed complement of a target sequence. The identification of optimal oligomers could be particularly important for antisense drug development. In applications, one can focus on sites within a particular mRNA region (e.g., coding region) of interest. In designing antisense oligomers, some basic rules are applicable for avoiding non-antisense effects and for

enhancing antisense potency. Three or more Gs in a row should be avoided (48). To minimize the possibility of binding to a non-targeted mRNA with strong sequence homology at the binding site, a BLAST search for a prospect oligomer can be performed to ensure no appreciable overlapping with other mRNAs in the experimental system (49). In particular, investigators need to be aware that translation initiation sites can have good homology in both related and non-related genes (46). To avoid stable intra-molecular structure within oligomers, oligomers that contain self-complementary regions (i.e. palindromic sequences) should not be used. Other suggestions for experimental design can be found in the literature (49–51).

The results with rabbit β-globin mRNA suggest that relatively wide, high probability peaks on the probability profile are very likely to be effective antisense sites. In some cases, relatively long oligonucleotides covering narrow high probability peaks in a region can be effective, because the chance of nucleation and propagation of duplex formation is enhanced with several single-stranded sites, despite their short length. In other cases, increasing oligomer length may not improve hybridization, because intramolecular base pairing within the oligomer may hinder hybridization (24). The finding of a significant correlation between the degree of translation inhibition in cell-free systems and predicted hybridization potential is consistent with results from *in vitro* and *in vivo* studies (21–23). This further supports the belief that the accessibility of the target site through complementary base pairing is a very important factor in determining the efficacy of an antisense oligonucleotide.

The antisense analysis is based on data from cell-free translation systems and an oligonucleotide array. A correlation between *in vitro* accessibility data and oligonucleotide intracellular activity has been documented for several RNAs (23,52–54) and a significant correlation by statistical analysis on a large number of oligonucleotides has been reported (55). Also, for oligonucleotide array, good correspondence between hybridization yield on the array and *in vivo* and *in vitro* antisense activities has been reported (23,54).

Similar to several existing procedures, our approach to antisense application is also based on mRNA secondary structure, one of several factors that can affect the accessibility and efficacy of an antisense molecule. Tertiary structure and RNA–protein interaction are not addressed. Cellular concentrations of the ASO and the mRNA are also important variables that can dictate how likely it is that the target mRNA and the antisense oligonucleotide will encounter and hybridize. Our finding reiterates the belief that secondary structure is perhaps the most important variable.

The approach presented in this work does not address the potential impact of secondary structure around single-stranded regions. In view of the multi-step process of nucleation and propagation for duplex formation (24,25), the probability profile presents a statistical delineation of the potential for nucleation. The initial pairing from the nucleation step needs to be stable enough to allow propagation (25). While the free energy gained can be computed, some empirical rule for minimum free energy gain required for propagation will be helpful for selecting sites with propagation potential from the profile. Propagation can more readily proceed by strand invasion of stems, while propagation through a single-stranded region may involve considerable reorientation of the strand (25). This striking observation suggests that it does not require too great a free energy price to invade a helically-ordered strand for propagation. It also suggests that for a sizable loop, sites on the ends of the loop may be more efficient for propagation. Furthermore, duplex propagation is affected by features in the target structure that are not well understood (25). The 'zippering' process for propagation stops when it meets an energy barrier such as the ends of stems or sharp turns in the folded RNA (24,25). These observations offer some clues as to how a quantitative score based on the profile might be developed to assess the potential of propagation. The algorithm can be expanded to include the free energy change attributed to oligomer–target duplex formation and to assess its impact on the predictions. A lack of apparent correlation between the measured free energy change and the measured hybridization has been demonstrated (24). However, correlation between the free energy change and antisense efficacy has been indicated by computational studies (27,56). With a better understanding of the contributions of structures around single-stranded regions, an extension to the present work may improve the reliability of the predictions.

Functional genomics and drug target validation

Functional genomics, the determination of the function of DNA sequence on a genomic scale, is a fast-growing field in biotechnology. Recently, the first assembly of 3.12 billion base pairs for the human genome was completed by Celera Genomics, and a working draft of the human genome by the publicly funded Human Genome Project has been finished. Definitive functions have been assigned to less than 1000 of the estimated 100 000 genes in the genome. While every technique for the determination of gene function has its own strengths and weaknesses, ASOs are given the most favorable assessment on all attributes in a comprehensive comparison of techniques (57). 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. For example, gene knockout is not broadly applicable. Routine gene knockout in mammals has been performed only in mice. In addition to the lengthy duration of the procedure, mammalian gene knockout often leads to an embryonic lethal phenotype, providing very little information about the gene function. Mutagenesis has not been shown to be feasible in mammals. 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 (58). This technique is expected to be applicable to other mammalian species (59). At relatively low cost, antisense not only offers high specificity of gene expression inhibition and rapid detection of antisense effects, but it also enables determination of gene function in adult animals by bypassing the potentially lethal embryonic stage. For functional genomics in the post-genomic era, traditional tools can no longer keep pace with new sequence information rapidly accumulated from various genome projects. The antisense technique has emerged as an important tool both *in vitro* and *in vivo*. When properly used, it has the potential to meet the need of large-scale functional genomics.

Antisense technique is also a very important tool for drug target validation. This is also well illustrated by the most

favorable assessment on all attributes in a comparison of drug target validation techniques (57). 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 (60). 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 ASO, 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. The 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 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 (20).

Complicated multi-component biological systems can be studied by using an appropriate set of ASOs to independently block the synthesis of each individual protein in the system. Antisense also promises to reveal genetic pathways through expression arrays. By antisense inhibition of protein expression and target mRNA, and through the evaluation of inhibitory effects on expression of genes on DNA arrays, insight will be gained on gene interaction and regulatory pathways (20).

High-throughput antisense applications

DNA expression arrays have emerged as major high-throughput experimental tools in the post-genomic era which allow the measurement of gene expression patterns of tens of thousands of genes in parallel (61,62). DNA expression arrays can provide important clues to gene function. Genes of similar expression behavior suggest that they are likely to be co-regulated or possibly functionally related. Indeed, statistical clustering analysis has revealed that gene expression data tend to organize genes into functional categories (63). 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) show promise in the progression of pharmacogenomics, the emerging field concerned with the dissection of the genetic basis of disease and therapeutic response. Assembly of large SNP databases has been undertaken by the next phase of the Human Genome Project (64), a non-profit SNP Consortium of 10 pharmaceutical companies and the Wellcome trust (65; http://snp.cshl.org/data/), and by several industrial efforts. SNPs enable studies of association between a SNP and risk of a disease or drug response (66,67). The associations are valuable for the identification of candidate genes for disease phenotypes.

The eventual determination of the functions of the candidate genes and confirmation of gene functional predictions based on analysis of DNA expression arrays 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. Expression array and SNP databases can provide the basis for high-throughput antisense applications to functional genomics and drug target validation.

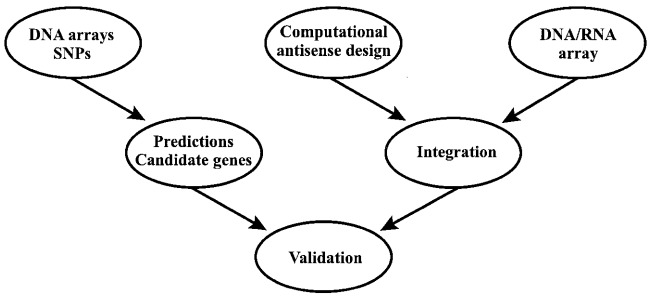


Figure 4. A potential high-throughput antisense framework for functional genomics, drug target validation and elucidation of genetic pathways. Systematic statistical analysis of DNA expression arrays and SNP databases can provide the basis for high-throughput functional analysis. Integration of computational antisense design and oligonucleotide array presents a rational, efficient, high-throughput platform for antisense oligonucleotide screening.

Experimental approaches for finding effective ASOs are expensive, time consuming and laborious, and are usually limited to a region of the mRNA. Published work suggests that, at the very best, only one in eight ASOs is effective (68). To realize the promise of antisense technique for high-throughput functional genomics and drug target validation, efficient screening for identifying active antisense target sites on the mRNA is necessary. This must be based on the combination of a high-throughput experimental platform and rational antisense design by a computational method. The combinatorial DNA–RNA oligonucleotide array technique appears to be an adequate experimental approach (25,69). With labeled transcripts, hybridization intensity can be measured and visualized (25). However, there are seemingly two practical limitations. First, the number of all 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 (69). Use of selective oligomers designed by comprehensive computational screening provides a solution. Hence, we advocate the strategy of integrating computational predictions and the array technique for a rational, efficient and comprehensive platform for ASO screening (Fig. 4). This platform needs to be extensively tested by experimentalists in both academia and industry to assess its potential for high-throughput applications.

ACKNOWLEDGEMENTS

The authors are grateful to two referees for insightful comments and critiques that led to substantial improvement of the paper. The Computational Molecular Biology and Statistics Core at the Wadsworth Center is acknowledged for providing computing resources for this work.

REFERENCES

1. Zuker,M. (1989) On finding all suboptimal foldings of an RNA molecule. *Science*, **244**, 48–52.
2. Zuker,M. (1989) The use of dynamic programming algorithms in RNA secondary structure prediction. In Waterman,M.S. (ed.), *Mathematical Methods for DNA Sequences*. CRC Press, Boca Raton, FL, pp. 159–184.

3. Zuker,M. and Stiegler,P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133–148.
4. McCaskill,J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.
5. Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhöffer,S., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatshefte f. Chemie*, **125**, 167–188.
6. Ding,Y. and Lawrence,C.E. (1999) A Bayesian statistical algorithm for RNA secondary structure prediction. *Comput. Chem.*, **23**, 387–400.
7. Martinez,H.M. (1984) An RNA folding rule. *Nucleic Acids Res.*, **12**, 323–334.
8. Mironov,A.A., Dyakonova,L.P. and Kister,A.E. (1985) A kinetic approach to the prediction of RNA secondary structures. *J. Biomol. Struct. Dyn.*, **2**, 953–962.
9. Mironov,A.A. and Lebedev,V.F. (1993) A kinetic model of RNA folding. *Biosystems*, **30**, 49–56.
10. Gultyaev,A.P., van Batenburg,F.H.D. and Pleij,C.W.A. (1995) The computer simulation of RNA folding pathways using a genetic algorithm. *J. Mol. Biol.*, **250**, 37–51.
11. Dallas,A. and Moore,P.B. (1997) The loop E–loop D region of *Escherichia coli* 5S rRNA: the solution structure reveals an unusual loop that may be important for binding ribosomal proteins. *Structure*, **5**, 1639–1653.
12. Easterwood,T.R. and Harvey,S.C. (1997) Ribonuclease P RNA: models of the 15/16 bulge from *Escherichia coli* and the P15 stem loop of *Bacillus subtilis*. *RNA*, **3**, 577–585.
13. Szewczak,A.A. and Cech,T.R. (1997) An RNA internal loop acts as a hinge to facilitate ribozyme folding and catalysis. *RNA*, **3**, 838–849.
14. Comolli,L.R., Pelton,J.G. and Tinoco,I.,Jr (1998) Mapping of a protein–RNA kissing hairpin interface: Rom and Tar-Tar*. *Nucleic Acids Res.*, **26**, 4688–4695.
15. Mirmira,S.R. and Tinoco,I.,Jr (1996) NMR structure of a bacteriophage T4 RNA hairpin involved in translational repression. *Biochemistry*, **35**, 7664–7674.
16. Nowakowski,J. and Tinoco,I.,Jr (1999) RNA structure in solution. In Stephen,N. (ed.), *Oxford Handbook of Nucleic Acid Structures*. Oxford University Press, New York, NY, pp. 567–602.
17. 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. USA*, **75**, 289–294.
18. Vanhée-Brossollet,C. and Vaquero,C. (1998) Do natural antisense transcripts make sense in eukaryote? *Gene*, **211**, 1–9.
19. Crooke,S.T. (1998) An overview of progress in antisense therapeutics. *Antisense Nucleic Acid Drug Dev.*, **8**, 115–122.
20. 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.
21. 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.
22. Vickers,T.A., Wyatt,J.R. and Freier,S.M. (2000) Effects of RNA secondary structure on cellular antisense activity. *Nucleic Acids Res.*, **28**, 1340–1347.
23. Matveeva,O., Felden,B., Audlin,S., Gesteland,R.F. and Atkins,J.F. (1997) A rapid *in vitro* method for obtaining RNA accessibility patterns for complementary DNA probes: correlation with an intracellular pattern and known RNA structures. *Nucleic Acids Res.*, **25**, 5010–5016.
24. Milner,N., Mir,K.U. and Southern,E.M. (1997) Selecting effective antisense reagents on combinatorial oligonucleotide arrays. *Nat. Biotechnol.*, **15**, 537–541.
25. Mir,K.U. and Southern,E.M. (1999) Determining the influence of structure on hybridization using oligonucleotide arrays. *Nat. Biotechnol.*, **17**, 788–792.
26. 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.
27. Stull,R.A., Taylor,L.A. and Szoka,F.C.,Jr (1992) Predicting antisense oligonucleotide inhibitory efficacy: a computational approach using histograms and thermodynamic indices. *Nucleic Acids Res.*, **20**, 3501–3508
28. Martinez,H.M. (1988) An RNA secondary structure workbench. *Nucleic Acids Res.*, **16**, 1789–1798.
29. 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.
30. Le,S.Y., Chen,J.H., Braun,M.J., Gonda,M.A. and Maizel,J.V. (1988) Stability of RNA stem–loop structure and distribution of non-random structure in the human immunodeficiency virus (HIV-I). *Nucleic Acids Res.*, **16**, 5153–5168.
31. Zhao,J.J. and Lemke,G. (1998) Rules for ribozymes. *Mol. Cell. Neurosci.*, **11**, 92–97.
32. James,W. and Cowe,E. (1997) Computational approaches to the identification of ribozyme target sites. *Methods Mol. Biol.*, **74**, 17–26
33. 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.
34. 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**, 14719–14735.
35. Sprinzl,M., Horn,C., Brown,M., Ioudovitch,A. and Steinberg,S. (1998) Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.*, **26**, 148–153.
36. Walter,A.E., Turner,D.H., Kim,J., Lyttle,M.H., Muller,P., Mathews,D.H. and Zuker,M. (1994) Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. *Proc. Natl Acad. Sci. USA*, **91**, 9218–9222.
37. Szymanski,M., Specht,T., Barciszewska,M.Z., Barciszewski,J. and Erdmann,V.A. (1998) 5S rRNA data bank. *Nucleic Acids Res.*, **26**, 156–159.
38. Gutell,R.R. (1994) Collection of small subunit (16S- and 16S-like) ribosomal RNA structures. *Nucleic Acids Res.*, **22**, 3502–3507.
39. Brown,J.W. (1999) The ribonuclease P database. *Nucleic Acids Res.*, **27**, 314.
40. Cech,T.R., Damberger,S.H. and Gutell,R.R. (1994) Representation of the secondary and tertiary structure of group I introns. *Nature Struct. Biol.*, **1**, 273–280.
41. Damberger,S.H. and Gutell,R.R. (1994) A comparative database of group I intron structures. *Nucleic Acids Res.*, **22**, 3508–3510.
42. Konings,D.A. and Gutell,R.R. (1995) A comparison of thermodynamic foldings with comparatively derived structures of 16S and 16S-like rRNAs. *RNA*, **1**, 559–574.
43. Zuker,M. and Jacobson,A.B. (1995) 'Well-determined' regions in RNA secondary structure prediction: analysis of small subunit ribosomal RNA. *Nucleic Acids Res.*, **23**, 2791–2798.
44. Cazenave,C., Loreau,N., Thuong,N.T., Toulme,J.J. and Helene,C. (1987) Enzymatic amplification of translation inhibition of rabbit beta-globin mRNA mediated by anti-messenger oligodeoxynucleotides covalently linked to intercalating agents. *Nucleic Acids Res.*, **15**, 4717–4736.
45. Goodchild,J., Carrol,E.,III and Greenberg,J.R. (1988) Inhibition of human immunodeficiency virus replication by antisense oligodeoxynucleotides. *Arch. Biochem. Biophys.*, **263**, 401–409.
46. Sohail,M. and Southern,E.M. (2000) Selecting optimal antisense reagents. *Adv. Drug Deliv. Rev.*, **44**, 23–34.
47. 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.
48. 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., 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. USA*, **92**, 4051–4055.
49. Phillips,M.I. and Zhang,Y.C. (2000) Basic principles of using antisense oligonucleotides *in vivo*. *Methods Enzymol.*, **313**, 46–56.
50. Crooke,S.T. (2000) Progress in antisense technology: the end of the beginning. *Methods Enzymol.*, **313**, 3–45.
51. Stein,C.A. (1999) Two problems in antisense biotechnology: *in vitro* delivery and the design of antisense experiments. *Biochim. Biophys. Acta*, **1489**, 45–52.
52. 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.
53. 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.
54. Southern,E.M., Milner,N. and Mir,K.U. (1997) Discovering antisense reagents by hybridization of RNA to oligonucleotide arrays. *Ciba Found. Symp.*, **209**, 38–44.
55. Matveeva,O., Felden,B., Tsodikov,A., Johnston,J., Monia,B.P., Atkins,J.F., Gesteland,R.F. and Freier,S.M. (1998) Prediction of antisense oligonucleotide efficacy by *in vitro* methods. *Nat. Biotechnol.*, **16**, 1374–1375.
56. Mathews,D.H., Burkard,M.E., Freier,S.M., Wyatt,J.R. and Turner,D.H. (1999) Predicting oligonucleotide affinity to nucleic acid targets. *RNA*, **5**, 1458–1469.
57. 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.
58. 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.
59. Thompson,J.D. (1999) Shortcuts from gene sequence to function. *Nat. Biotechnol.*, **17**, 1158–1159.
60. 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.
61. Brown,P.O. and Botstein,D. (1999) Exploring the new world of the genome with DNA microarrays. *Nature Genet.*, **21**, 33–37.
62. Lockhart,D.J., Dong,H., Byrne,M.C., Follettie,M.T., Gallo,M.V., Chee,M.S., Mittmann,M., Wang,C., Kobayashi,M., Horton,H. and Brown,E.L. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.*, **14**, 1675–1680.
63. 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. USA*, **95**, 14863–14868.
64. Brower,V. (1998) Genome II: the next frontier. *Nat. Biotechnol.*, **16**, 1004.
65. Marshall,E. (1999) Drug firms to create public database of genetic mutations. *Science*, **284**, 406–407.
66. McCarthy,J.J. and Hilfiker,R. (2000) The use of single-nucleotide polymorphism maps in pharmacogenomics. *Nat. Biotechnol.*, **18**, 505–508.
67. Brookes,A.J. (1999) The essence of SNPs. *Gene*, **234**, 177–186.
68. Stein,C.A. (1999) Keeping the biotechnology of antisense in context. *Nat. Biotechnol.*, **17**, 209–212.
69. Southern,E., Mir,K. and Shchepinov,M. (1999) Molecular interactions on microarrays. *Nature Genet.*, **21**, 5–9.