ABSTRACT
Motivation: The importance of RNA sequence analysis has been increasing since the discovery of various types of non-coding RNAs transcribed in animal cells. Conventional RNA sequence analyses have mainly focused on structured regions, which are stabilized by the stacking energies acting on adjacent base pairs. On the other hand, recent findings regarding the mechanisms of small interfering RNAs (siRNAs) and transcription regulation by microRNAs (miRNAs) indicate the importance of analyzing accessible regions where no base pairs exist. So far, relatively few studies have investigated the nature of such regions.

Results: We have conducted a detailed investigation of accessibilities around the target sites of siRNAs and miRNAs. We have exhaustively calculated the correlations between the accessibilities around the target sites and the repression levels of the corresponding mRNAs. We have computed the accessibilities with an originally developed software package, called ‘Raccess’, which computes the accessibility of all the segments of a fixed length for a given RNA sequence when the maximal distance between base pairs is limited to a fixed size $W$. We showed that the computed accessibilities are relatively insensitive to the choice of the maximal span $W$. We have found that the efficacy of siRNAs depends strongly on the accessibility of the very 3′-end of their binding sites, which might reflect a target site recognition mechanism in the RNA-induced silencing complex. We also showed that the efficacy of miRNAs has a similar dependence on the accessibilities, but some miRNAs also show positive correlations between the efficacy and the accessibilities in broad regions downstream of their putative binding sites, which might imply that the downstream regions of the target sites are bound by other proteins that allow the miRNAs to implement their functions. We have also investigated the off-target effects of an siRNA as a potential RNAi therapeutic. We showed that the off-target effects of the siRNA have similar correlations to the miRNA repression, indicating that they are caused by the same mechanism.

Availability: The C++ source code of the Raccess software is available at http://www.ncrna.org/software/Raccess/ The microarray data on the measurements of the siRNA off-target effects are also available at the same site.

Contact: kiryu-h@k.u-tokyo.ac.jp

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on November 2, 2010; revised on April 20, 2011; accepted on April 27, 2011

1 INTRODUCTION

The energy scales of the secondary structure formations of RNA molecules are rather high compared with those of many biological interactions and processes. For example, the free energy of a single GC/CG stacking is 3.4 kcal/mol. This value is about half of the 7 kcal/mol of ATP hydrolysis $\text{ATP} \rightarrow \text{ADP} + \text{PPi}$, which is the basic energy unit of most biological processes, such as transcription and translation. Therefore, the secondary structure can strongly influence biological processes involving RNA molecules.

Recently, a few studies that measure the regulatory activities of miRNAs and siRNAs have revealed the importance of accessibility around the target regions to these functional RNAs (Kertesz et al., 2007; Shao et al., 2007; Tafer et al., 2008). An miRNA represses the expression of its target mRNAs by recognizing the ‘seed’ sequences in the target genes, which are about 7 bp in length and are complementary to the 5′-end of the miRNA sequence. The repression will be inefficient if the seed sequences are buried within a strong secondary structure of the target mRNA so that the corresponding RNA-induced silencing complex (RISC) cannot access them because of a high energy barrier (Kertesz et al., 2007). An siRNA represses the expression of the target mRNA that has the sequence complementary to the antisense strand of the siRNA. The accessibility around the target sites has also been shown to influence the efficacy of siRNAs (Gredell et al., 2008; Hofacker and Tafer, 2010; Tafer et al., 2008). Despite the importance of the accessibility of interacting RNA molecules, there have been few studies investigating the regions of RNA sequences that do not form any base pairs (Bernhart et al., 2006; Busch et al., 2008; Chen et al., 2009; Kertesz et al., 2007; Tafer et al., 2008).

In this article, we investigate the accessibilities around the (off-) target sites of siRNAs and miRNAs. We define the accessibility...
is known that the secondary structure model based on the Turner Boltzmann distribution. In this case, the gas constant and is the set of all possible secondary structures of x, S(x) is the set of all secondary structures that have no paired bases in segment x, is the energy of secondary structure \sigma on x, which is calculated by the Turner energy model (Mathews et al., 1999), }^\text{v} = 1.9872 \times 10^{-3} (\text{kcal/mol K}) is the gas constant and }^\text{v} = \text{310.15} K is the temperature. Since it is known that the secondary structure model based on the Turner energy parameters is less accurate for structures involving the distant base pairs, it is natural to restrict the maximal span of the base pairs to a fixed value W. In this case, S(x) and \mathcal{S}_0 include only the structures containing base pairs of span less than or equal to W. If the dynamics of RNA molecules were completely governed by the Turner energy model, then the conformation distribution of non-interacting RNA molecules in solution would be represented by the Boltzmann distribution. In this case, }^\text{v} p_{\text{acc}}(x \sigma) would represent the probability that a single RNA molecule has accessible segment x. Recent studies have shown that the techniques that combine all types of expectation values, their results are subject to random techniques. The time complexity of the algorithm is }^\text{v} O(NW^2+MW) where M represents the number of sampled structures. Although sampling methods are very flexible to compute various types of expectation values, their results are subject to random fluctuations, and the error of the calculated probability values is at least }^\text{v} 1/M. In particular, it is difficult to compute small probability values unless huge number of structures are sampled. As compared to the sampling algorithms, dynamic programming (DP) algorithms can take all the possible structures into account, and they can accurately compute small probability values without increasing computational costs. oligoWalk (Lu and Mathews, 2008a, b, c) uses a version of McCaskill’s DP algorithm (McCaskill, 1990) adapted to the cases where the RNA sequence are subject to accessibility constraints. The time complexity is given by }^\text{v} O(N^3) . Since }^\text{v} O(N^3) time is prohibitive to apply to long mRNAs, it cannot compute the accessibility based on the global structures as defined in Equation (1) for them. Although it can use a sequence window approach, the artificial window boundaries strongly affect the accessibility values, as described in the ‘Dependence on GC Composition’ subsection. RNAplfold (Bernhart et al., 2006) computes the mean accessibilities averaged over sequence windows of length }^\text{v} S (W \leq S \leq N) with complexities of }^\text{v} O(NS^2) in time and }^\text{v} O(N^2+S^3) in space. Such window averaging is a reasonable approach for a genomic scan where the boundaries of potential transcripts are unknown. However, to compute the accessibilities [Equation (1)] for transcripts with validated boundaries, we need to set }^\text{v} S = N , which is again prohibitive in terms of complexity for long RNA sequences. Furthermore, the current RNAplfold implementation results in overflow errors when }^\text{v} S is set to a few thousand bases, because it uses a simple scaling method for the multiplications of partial Boltzmann factors. We have, therefore, developed a software package called ‘Racccess’, which calculates exactly the accessibility of all the segments of a fixed length for a given RNA sequence. Racccess sums up the Boltzmann weights for all the possible global structures with the constraint that the maximal span of the base pairs is limited to a given size }^\text{v} W . The computational complexity is }^\text{v} O(N^2W^2) in time and }^\text{v} O(N^2+W^3) in space. Racccess computes the inside and outside variables by using the logarithm of Boltzmann factors, so it is robust to overflow errors and returns correctly for sequences longer than }^\text{v} N = 100 \text{ kb. Using Racccess, we first investigate the basic properties of the accessibility. We show how the accessibility depends on the GC composition, the length of x, and the maximal span. We also show that the computed accessibilities are relatively insensitive to the choice of the maximal span. We then apply Racccess to the mRNA sequences targeted by siRNAs and miRNAs, and compare the efficacy of their regulatory activities with the accessibilities around the target sites. As compared to the previous studies that investigated the efficacy-accessibility correlations with only limited patterns of accessible segments (Kertesz et al., 2007; Lu and Mathews, 2008a; Shao et al., 2007; Tafer et al., 2008), we computed the accessibilities for the exhaustive patterns of the accessible segments. Figure 1 shows the comparison of the patterns of the accessible segments considered in this paper and those in the previous studies. From these exhaustive calculations, we have obtained several interesting observations. In particular, we show that the efficacy of siRNAs depends strongly on the accessibility of the very 5’-end of their binding sites, which might reflect the underlying mechanisms of siRNA function. We also investigate the correlation between the accessibility and the off-target effects of an siRNA for the first time. We show that the miRNA repression and the off-target effects of the siRNA have a similar dependence on the accessibility.
2 ALGORITHMS AND IMPLEMENTATION

Both RNAplfold and Raccess compute the accessibilities based on the Turner energy model (Mathews et al., 1999). However, the algorithms they are based on are rather different: RNAplfold uses a modification of McCaskill’s algorithm (McCaskill, 1990), whereas Raccess uses a type of inside–outside algorithm associated with a defined context-free grammar (Kiryu et al., 2008). Although McCaskill’s algorithm is essentially an inside–outside algorithm, there is no symmetry between the inside and outside variables and it is difficult to apply the RNAplfold algorithm to other stochastic context-free grammars (SCFGs). Therefore, we begin by giving an illustrative example to show how to derive the formula for computing the accessibilities in the case of SCFG models. We describe how the algorithm can be applied to the energy model in the subsequent section.

We first introduce some notation to simplify the following discussion. The maximal span \( W \) is defined as the maximal distance between two sequence positions for which we consider the possibility of base pair formation. The access segment \( s_a \) is the sequence segment for which we compute whether the region is accessible or not in terms of secondary structure. The access position \( s_a \) is the middle position of access segment \( s_a \), and the access length \( l_a \) is the length of that segment. The accessibility \( P_{\text{acc}}(s_a) \) in Equation (1) is the probability that there are no paired bases in the access segment \( s_a \) in terms of the Boltzmann distribution. In other words, it is the probability that the access segment \( s_a \) is thermodynamically accessible. The access energy \( \Delta E_{\text{acc}}(s_a) = -RT\log P_{\text{acc}}(s_a) \) is the thermodynamic work needed to dissociate all the paired bases within the access segment \( s_a \). The partition function \( Z(s) \) is obtained from the usual inside algorithm. Thus, we only derive the numerator of Equation (1) in terms of the inside–outside variables below.

2.1 Accessibility formula for a SCFG

The SCFG model that we consider has five non-terminal states (S, P, L, R and M). The transition rules of the SCFG are defined as follows.

\[
\begin{align*}
S & \rightarrow L \\
P & \rightarrow a_a La_a \\
L & \rightarrow aLa_M \\
R & \rightarrow Ra_P \\
M & \rightarrow M_Ri_a
\end{align*}
\]

where \( i \) is the null terminal symbol, \( a \) is a terminal symbol corresponding to an unpaired nucleotide and \( a_a \) and \( a_s \) are the respective terminal symbols that correspond to the left and right base of a base pair. Furthermore, \( S \) represents the start state, \( P \) represents the pair emitting state, \( L \) represents the left emitting state, \( R \) the right emitting state and \( M \) represents an auxiliary state to handle multistems of stems. To explain how this SCFG parses secondary structures, we prepare some basic terms on the energy model. In this model, any secondary structure decomposes an RNA sequence into one or more \( L \)-loops enclosed by the backbone and hydrogen bonds of base pairs. A \( k \)-loop with \( k \geq 0 \) denotes a loop which composed of a closing base pair, \( k - 1 \) opening base pairs and zero or more unpaired bases. For example, a one-loop is a hairpin loop, a two-loop is either stacked base pairs or a bulge loop or an interior loop and a \( k \)-loop with \( k > 2 \) is called a multiloop. For more detail about the loop model of secondary structures, please refer to the Mfold manual (Zuker et al., 1999). For each \( k \)-loop, all the \( k \) base pairs are emitted by \( P \rightarrow k \) transitions. The leftmost segment of unpaired bases is emitted by \( L \rightarrow L \) transitions. The other unpaired bases are emitted by \( R \rightarrow R \) transitions. Each \( M \rightarrow MRi_a \) transition represents the production of an opening base pair and, possibly, flanking unpaired bases on its right. Figure 2 shows the parsing tree of a three-loop structure.

In the present case, the numerator of Equation (1) corresponds to the sum of all the probabilities of the parse trees that emit the segment \( s_a \) as unpaired bases. To compute this, we need to enumerate all the state transition patterns that emit a contiguous sequence of unpaired bases of length \( l_a \). There are only two such patterns which are given by

\[
(L \rightarrow L)^w \text{ and } (R \rightarrow R)^w
\]

where \( (A \rightarrow A)^n \) indicate \( n \) successive \( A \rightarrow A \) transitions. Hence, the sum of all the parse trees that includes these transitions can be represented using the inside \( p_l(s_a) \) and outside \( p_R(s_a) \) variables:

\[
p_l(s_a) = \sum_{s_1 \leq j < s_2} \prod_{k=1}^{s_2-s_1} \left( \sum_{i} \beta_l(i,s_{i-1}) \right) \text{ and } p_R(s_a) = \sum_{s_1 \leq j < s_2} \prod_{k=1}^{s_2-s_1} \left( \sum_{i} \beta_R(i,s_{i-1}) \right)
\]

where \( s_1 \) and \( s_2 \) represent the first and the last positions of \( s_a \), respectively, and \( \beta_l(i,s_{i-1}) \) represents the emission probability of state \( s \) at position \( i \) and \( n(s \rightarrow s') \) represents the transition probability for transition \( s \rightarrow s' \). We show these formulas graphically in Figure 3. Therefore, the accessibility of \( s_a \) is obtained by dividing the sum of the above values by the partition function \( Z(s) \):

\[
P_{\text{acc}}(s_a) = \frac{p_l(s_a) + p_R(s_a)}{Z(s)}.
\]
2.2 Accessibility formula for the energy model

To derive the accessibility formula for the energy model, we use the grammar has a simple accessibility formula. unambiguous, it is easier to enumerate all the patterns emitting unpaired bases as left emissions. In this case, the computation of the accessibility includes the summation of all such emission patterns and is practically impossible. As another example, if we add the transition then both bifurcating children emit parts of a single contiguous segment. This case also does not have a simple representation. If the grammar is unambiguous, it is easier to enumerate all the patterns emitting unpaired segments, but it is not obvious whether an unambiguous grammar generally has a simple accessibility formula.

We note that whether or not the accessibility has such simple expression strongly depends on the grammar structure. For example, if we add to the model a transition then both bifurcating children emit parts of a single contiguous segment.

To derive the accessibility formula for the energy model, we use the grammar

\[
\text{has a simple accessibility formula.}
\]

unambiguous, it is easier to enumerate all the patterns emitting unpaired bases as left emissions. In this case, the computation of the accessibility includes the summation of all such emission patterns and is practically impossible. As another example, if we add the transition then both bifurcating children emit parts of a single contiguous segment.

This case also does not have a simple representation. If the grammar is unambiguous, it is easier to enumerate all the patterns emitting unpaired segments, but it is not obvious whether an unambiguous grammar generally has a simple accessibility formula.

\[\text{has a simple accessibility formula.}\]

unambiguous, it is easier to enumerate all the patterns emitting unpaired bases as left emissions. In this case, the computation of the accessibility includes the summation of all such emission patterns and is practically impossible. As another example, if we add the transition then both bifurcating children emit parts of a single contiguous segment.

This case also does not have a simple representation. If the grammar is unambiguous, it is easier to enumerate all the patterns emitting unpaired segments, but it is not obvious whether an unambiguous grammar generally has a simple accessibility formula.

\[\text{has a simple accessibility formula.}\]

unambiguous, it is easier to enumerate all the patterns emitting unpaired bases as left emissions. In this case, the computation of the accessibility includes the summation of all such emission patterns and is practically impossible. As another example, if we add the transition then both bifurcating children emit parts of a single contiguous segment.

This case also does not have a simple representation. If the grammar is unambiguous, it is easier to enumerate all the patterns emitting unpaired segments, but it is not obvious whether an unambiguous grammar generally has a simple accessibility formula.

\[\text{has a simple accessibility formula.}\]

unambiguous, it is easier to enumerate all the patterns emitting unpaired bases as left emissions. In this case, the computation of the accessibility includes the summation of all such emission patterns and is practically impossible. As another example, if we add the transition then both bifurcating children emit parts of a single contiguous segment.

This case also does not have a simple representation. If the grammar is unambiguous, it is easier to enumerate all the patterns emitting unpaired segments, but it is not obvious whether an unambiguous grammar generally has a simple accessibility formula.

\[\text{has a simple accessibility formula.}\]

unambiguous, it is easier to enumerate all the patterns emitting unpaired bases as left emissions. In this case, the computation of the accessibility includes the summation of all such emission patterns and is practically impossible. As another example, if we add the transition then both bifurcating children emit parts of a single contiguous segment.

This case also does not have a simple representation. If the grammar is unambiguous, it is easier to enumerate all the patterns emitting unpaired segments, but it is not obvious whether an unambiguous grammar generally has a simple accessibility formula.

\[\text{has a simple accessibility formula.}\]

unambiguous, it is easier to enumerate all the patterns emitting unpaired bases as left emissions. In this case, the computation of the accessibility includes the summation of all such emission patterns and is practically impossible. As another example, if we add the transition then both bifurcating children emit parts of a single contiguous segment.

This case also does not have a simple representation. If the grammar is unambiguous, it is easier to enumerate all the patterns emitting unpaired segments, but it is not obvious whether an unambiguous grammar generally has a simple accessibility formula.

\[\text{has a simple accessibility formula.}\]

unambiguous, it is easier to enumerate all the patterns emitting unpaired bases as left emissions. In this case, the computation of the accessibility includes the summation of all such emission patterns and is practically impossible. As another example, if we add the transition then both bifurcating children emit parts of a single contiguous segment.

This case also does not have a simple representation. If the grammar is unambiguous, it is easier to enumerate all the patterns emitting unpaired segments, but it is not obvious whether an unambiguous grammar generally has a simple accessibility formula.

\[\text{has a simple accessibility formula.}\]

unambiguous, it is easier to enumerate all the patterns emitting unpaired bases as left emissions. In this case, the computation of the accessibility includes the summation of all such emission patterns and is practically impossible. As another example, if we add the transition then both bifurcating children emit parts of a single contiguous segment.

This case also does not have a simple representation. If the grammar is unambiguous, it is easier to enumerate all the patterns emitting unpaired segments, but it is not obvious whether an unambiguous grammar generally has a simple accessibility formula.

\[\text{has a simple accessibility formula.}\]

unambiguous, it is easier to enumerate all the patterns emitting unpaired bases as left emissions. In this case, the computation of the accessibility includes the summation of all such emission patterns and is practically impossible. As another example, if we add the transition then both bifurcating children emit parts of a single contiguous segment.

This case also does not have a simple representation. If the grammar is unambiguous, it is easier to enumerate all the patterns emitting unpaired segments, but it is not obvious whether an unambiguous grammar generally has a simple accessibility formula.

\[\text{has a simple accessibility formula.}\]

unambiguous, it is easier to enumerate all the patterns emitting unpaired bases as left emissions. In this case, the computation of the accessibility includes the summation of all such emission patterns and is practically impossible. As another example, if we add the transition then both bifurcating children emit parts of a single contiguous segment.

This case also does not have a simple representation. If the grammar is unambiguous, it is easier to enumerate all the patterns emitting unpaired segments, but it is not obvious whether an unambiguous grammar generally has a simple accessibility formula.
dataset for testing siRNA design tools since it was published. There are several experiments for the siRNA efficacy dataset, but the number of tested siRNAs are at least a few times smaller (see Matveeva et al. (2007), for example). Although the set of 34 mRNAs is a very tiny portion of >10000 mRNAs in the animal genomes, the diversity of the efficacies among different siRNAs targeting the same mRNAs indicates that a significant portion of efficacy can be explained by local features. Therefore, to investigate the local determinants affecting the efficacy, this small number of mRNAs will not be a critical limiting factor. One possible problem of this dataset is that the experimental conditions might be too idealized to represent real mRNAs; not be a critical limiting factor. One possible problem of this dataset is that local efficacy can be explained by local features. Therefore, to investigate the mRNAs in the animal genomes, the diversity of the efficacies among different example). Although the set of 34 mRNAs is a very tiny portion of the Supplementary Material. The microarray dataset is available from our web site.

3.6 Efficacy–accessibility correlation
We categorized the dataset into two groups based on the accessibilities around the target sites. For a given access segment $u_i$ which was specified by an access length $l_a$ and an access position $n_i$ relative to the (putative) binding site, we ranked the dataset by the accessibility of $u_i$. We labeled the most accessible 30% of data points as group 1 and the others as group 2. Then we computed Welch’s $t$-statistic, which tests the difference of the efficacy distribution between two groups:

$$ t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} $$

where $\mu_1$ and $\mu_2$ are the mean and the SD of the efficacy in group 1, respectively, and $n_1$ is the size of the group 1. A value $t > 0$ indicates that more accessible segments are more effective targets. Because the size of the dataset is rather large, the null distribution for $t$ can be approximated by the standard normal distribution. This approximation is used when we mention $P$-values in Figures 7, 8, 10 and 11.

There are three major statistics for testing the difference of means (or medians) between two groups of data: Student’s $t$-statistic (ST), Wilcoxon–Mann–Whitney $U$ statistic (WMW) and Welch’s $t$-statistic (WT). Each of these methods is based on specific assumptions about the null distribution: ST assumes that two groups of data are sampled from an identical normal distribution; WMW assumes two groups of data are sampled from an identical but arbitrary-shape distribution; WT assumes two groups of data are sampled from two distinct normal distributions with equal mean. Since it is difficult to assume that our datasets exactly conform to these assumptions even in the null case, we have to use them under the existence of certain violations of basic assumptions. There are numerous statistics papers that investigate to what extent these methods are robust under such violations. Generally, the statistical power does not suddenly disappear with a tiny violation from the assumptions, but it rather degrades gradually with increase of deviation from them (Donald, 2004; Morten and Leiv, 2009). A recent article recommends WT, which we have used in the following, as the first choice method for testing differences between group-wise means (Morten and Leiv, 2009).

In the following, we do not consider the multiple corrections for $P$-values, since it is very difficult to find an appropriate method for multiple correction, due to the strong correlations of the $P$-values among neighboring accessible segments. Still, the obtained $P$-values are very useful, as they represent the relative strengths of the efficacy–accessibility correlations taking the datasizes and variances into account.

4 RESULTS AND DISCUSSION

4.1 Comparison of run time and memory usage
Table 1 shows the run time and memory usages for several patterns of sequence length, maximal span and sequence window size. The accessible length $l_a$ is fixed to 20 in these computations. Among the programs, Raccass is the only program that returns correctly within 10 h. RNApIfold is a few times faster than Raccass in many cases, but it does not work for large sequence windows due to overflow errors. Sfold does not return within 10 h for the 100000 base sequence. OligoWalk is slower than the other programs. The memory usage shows that all the programs use only moderate memory space when they return correctly within the time limit.
Table 1. Comparison of run time and memory usage

<table>
<thead>
<tr>
<th></th>
<th>(N,W,S)</th>
<th>(100,100,100)</th>
<th>(1k,100,100)</th>
<th>(1k,1k,1k)</th>
<th>(10k,100,100)</th>
<th>(10k,100,1k)</th>
<th>(10k,100,1k)</th>
</tr>
</thead>
</table>
|        | (N: sequence length, W: maximal span, S: sequence window size), respectively. Symbol ‘–’ indicates that the program does not return results within 10h. Symbol ‘x’ indicates that the program terminates within 10h but does not return any result due to overflow errors.

4.2 Dependence on GC composition

Figure 5 shows how the access energy $\Delta E_{acc}$ depends on the GC composition, the access length $l_a$ and the maximal span $W$. Figure 5A shows that the access energy increases with GC composition and their slopes are more steep with increasing $l_a$. Figure 5B shows that $\Delta E_{acc}$ is linearly dependent on the access length $l_a$. Figure 5C shows that, if $W > 100$, the accessibility is only weakly dependent on the maximal span $W$, as compared to the strong dependence on the GC composition. This insensitivity to $W$ is a favorable feature which is not satisfied by other statistics, such as the base pairing probability matrix where the number of highly probable base pairs constantly increases with $W$. The predicted secondary structure also changes considerably with $W$. In this sense, it may be said that the accessibility is a better statistic of the energy model than the base pairing probability matrix or the predicted secondary structure.

We have also investigated the dependence of the accessibility on the distance to the end points of the sequence and on the distant nucleotide contents of the sequence. The results are shown in the Supplementary Material, where we show that the accessibilities are affected by the positions of the end points and by the distant nucleotides if they are located within a distance several times as large as the maximal span $W$.

4.3 siRNA efficacy dataset

Figure 6 shows the scatter plot of siRNA–target data at particular $W$, $l_a$ and $x_a$ values. Although the variances are very large, the mean value of efficacy $z_{eff}$ steadily decreases with the access energy $\Delta E_{acc}$, which suggests that the accessibility is an important determinant of efficacy.
Figure 7 shows the main result of the article, namely the correlation between the accessibility and the siRNA efficacy for various access lengths and access positions relative to the binding sites. Here, we can clearly see a strong positive correlation between the efficacy and the access segments on the vertical line of the downstream end of the binding site ($x_a=0$). The blotted regions are roughly symmetrical around $x_a=0$, and moreover, almost all the corresponding access segments include the base at $x_a=0$. This may indicate the underlying mechanism of the siRNA function: the accessibility of this position may be essential for the RISC complex to recognize the target site and to implement its cleavage function. Previous studies have considered the efficacy–accessibility correlation in the context of static or equilibrium aspects of siRNA function, where they have compared the free energy difference between bound and unbound states of siRNA–mRNA pairs (Lu and Mathews, 2008a; Shao et al., 2007; Tafer et al., 2008). However, our results suggest that the accessibility is more related to the dynamical or kinetic aspects of the siRNA mechanism, that is, accessible target sites might help siRNA binding by lowering the activation (initiation) energy barrier of the reaction.

Table 2 shows the dependence of siRNA efficacy–accessibility correlations on the maximal span $W$. Although the accessibility values are relatively insensitive to $W$ as shown in Figure 5C, $W=400$ shows better correlation $r_s$ and maximal $t$-value $t_{\text{max}}$ than the other $W$ values. Hence, we set $W=400$ in the following sections.

4.4 miRNA datasets

Figures 8 shows the density plots of $t$-values for the miRNA datasets. Compared to Figure 7, the statistical significances appear much weaker. This may be due to the non-specificity of 7 mer sequence, being a high level of noise, and the smaller size of the dataset (a few hundred points compared to 2431 for siRNA). We can, however, still find some interesting features in these figures. The miR-124 dataset (Fig. 8C) shows strong correlations between the efficacy and the accessibility around the putative binding site, which appears similar to the siRNA case. On the other hand, the miR-155 and miR-1 datasets show weak $t$-values at $x_a=0$, but there are broad regions ($Q_s=70$–100) downstream of the putative binding site, where accessibility is positively correlated with efficacy. Interestingly, all the figures have a weakly blotted region around $x_a=40...60$, $W=10...20$. They might be reflecting the mechanisms of the miRNA functions such as protein binding and specific conformations of the miRNAs downstream of the binding site.

4.5 siRNA off-target dataset

As described in the Methods section, an siRNA targeting a glycosyorechon (CHST15) are transfected into human fibroblast cells. The length of the siRNA is 27 mer instead of the 21mer of conventional siRNAs. Such longer siRNAs are called Dicer-substrate siRNAs (DsiRNAs) (Kim et al., 2005). They are first cleaved by Dicer to form 21mer siRNAs, which then implement their usual RNAi function. Previous studies (Kim et al., 2005) have shown that transfection of DsiRNAs is far more efficient than transfecting 21mer siRNAs directly. The sequences of the siRNA are shown in Figure 9. There are two possible patterns (L, R) that produce a 21mer siRNA with a single Dicer slicing. Each cleavage pattern is expected to show off-target effects specific to the seed sequence at the cleavage site. We computed the $t$-values for all the 7 mer subsequences of siRNAs and the corresponding $P$-values are plotted in Figure 10. Position 6 of the sense and antisense strand have the largest $P$-values, which implies that both the ‘L’ and ‘R’ cleavage patterns show off-target effects. The density plot of $t$-values for this seed GAUGAAUU corresponding to ‘R’ type cleavage (Fig. 11A) is very similar to the plots for the miRNAs, especially that for the miRNA knock down dataset (Fig. 8A). This indicates that the siRNA off-target effect occurs by the same mechanisms as miRNA repression. Figure 11B shows the accessibility–efficacy plot for each gene ($l_a=16$, $W=400$ and $\Delta E_{\text{acc}}$ is averaged over access positions $x_a$ in the range $[-5,5]$). The genes whose expression might be most strongly affected by the off-target effects are shown in Table 3. These include TWIST neighbor protein, oxytocin receptor and insulin-like
Target accessibility of siRNA and miRNA

Fig. 8. Density plot of miRNA efficacy-accessibility correlations. The x-axis is the access position $x_a$ relative to the 3′-end of each putative binding site. The 7-mer putative binding region is shown as the thick bar between the arrows. The y-axis is the access length $l_a$ in log scale. The value of $t$ given by Equation (2) is indicated by shading according to the density scale shown. Only $t$-values exceeding 1.64 which corresponds to $p \approx 0.05$ in the normal approximation are shown. The maximal span $W$ is set to 400. (A) miR-155 (seed AGCAUUA) knock down dataset. (B) miR-1 (ACAUUCC) transfection dataset. (C) miR-124 (GUCCCUU) transfection dataset. Note that scale bars of these plots are [0,5], which is different from [0,10] of Figure 7. Therefore, the significances are much lower than those in Figure 7.

Fig. 9. siRNA sequence and the potential Dicer cleavage sites. The potential cleavage patterns L and R that produce 21-mer siRNA are shown as solid lines. The seed 7-mer regions created by these cleavage patterns are indicated by the dotted lines.

Fig. 10. Plot of $P$-value for each 7-mer subsequence. The x-axis is the 3′-end position of the 7-mers in the siRNA sequences. The y-axis is the value of $-\log_2(P$-value), which are calculated from the mean $t$ values averaged over access positions $x_a$ in the range $[-5, 5]$. Access lengths $l_a = 1, 7, 11, 20, 55, 100$ (from thin to thick lines) are shown. (A) the sense strand. (B) the antisense strand.

Fig. 11. Efficacy-accessibility correlations of the siRNA off-target dataset. (A) Density plot of $t$-values for the seed GAUGAAU corresponding to the ‘R’ type Dicer cleavage. The value of $t$ given by Equation (2) is indicated by shading according to the density scale shown. Note that scale bars of these plots are [0,5], which is different from [0,10] of Figure 7. Therefore, the significances are much lower than those in Figure 7. (B) accessibility-efficacy plot in which each dot corresponds to a gene having seed GAUGAAU. The x-axis is the mean access energy averaged over access positions $x_a$ in the range $[-5, 5]$. The y-axis is the value of $\log_2$ (fold change) for gene expressions after siRNA transfection. Four possible off-targets of siRNA are numbered.

5 FURTHER DISCUSSION

5.1 Comparison of different tools

In this subsection, we summarize the notable features of the used programs. Sfold is based on the posterior sampling algorithm and has time complexity $O(N^2 M + M W^2)$ ($N$: sequence length, $W$: maximal span, $M$: number of sampled structures). It is useful for computing various types of expected values other than the accessibility. Sfold is an integrated software package which has various options for growth factor. The repression of these genes might cause side effects in the therapeutic application of the siRNA.
Table 3. The potential off-targets of the siRNA

<table>
<thead>
<tr>
<th>ID</th>
<th>ΔE_{acc}</th>
<th>Gene symbol</th>
<th>Gene name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1.61</td>
<td>TWISTNB</td>
<td>TWIST neighbor</td>
</tr>
<tr>
<td>2</td>
<td>-2.49</td>
<td>IGF2BP5</td>
<td>Insulin-like growth factor 2 mRNA binding protein 3</td>
</tr>
<tr>
<td>3</td>
<td>-2.15</td>
<td>OXTR</td>
<td>Oxytocin receptor</td>
</tr>
<tr>
<td>4</td>
<td>-1.81</td>
<td>PECI</td>
<td>Peroxisomal ID3,2-ensyl-CoA isomerase</td>
</tr>
</tbody>
</table>

The ID corresponds to the numbers in Figure 11.

designing efficient siRNAs, trans-clearing ribozymes, nucleic acid probes, etc. The disadvantages of Sfold are that the results are subject to random fluctuations and the errors of the calculated probability values are at least 1/M. For a long sequence, sampled structures are not a good representation of the Boltzmann distribution in the large sequence space. It means the computed probabilities are inaccurate for such sequences.

OligoWalk uses a version of McCaskill's DP algorithm adapted to the cases where the RNA sequences are subject to accessibility constraints. It has time complexity $O(N^3)$. It does not only compute the accessibilities but also predicts efficient siRNAs by combining the accessibility with other features. The high accuracy of predictions are confirmed in Lu and Mathews (2008a). As a tool to compute the accessibilities, it is slower than other programs. Further, since the maximal span is always equal to the sequence window size, the computed accessibility values are subject to boundary effects for long RNA sequences.

RNAplfold has time complexity $O(N^2)$ ($S$: sequence window size). It is the fastest program in many cases and appropriate for a genomic scan. When sequence window size $S$ is small, the accessibility values are subject to boundary effects. When $S$ is as large as one thousand bases, RNAplfold causes overflow errors.

Raccoon has time complexity $O(N^3W^3)$. It can compute the accessibility values for long sequences without causing overflow errors. Raccoon has recently been developed to compute local accessibilities of RNA sequences using a support vector machine. Although the correlation of each of those features is only up to 0.35, it predicts miRNAs of efficacy > 70% with high performance (sensitivity 22.7% and specificity 96.5%).

5.3 Using accessibility for efficient siRNA design

Previous studies (Lu and Mathews, 2008a; Tafer et al., 2008) have used the accessibility as one of the major features to predict efficient siRNAs. In particular, OligoWalk (Lu and Mathews, 2008a) combines several sequence and thermodynamic features including the accessibility using a support vector machine. Although the correlation of each of those features is only up to 0.35, it predicts miRNAs of efficacy > 70% with high performance (sensitivity 22.7% and specificity 96.5%).

6 CONCLUSION

We have conducted a detailed investigation of accessibilities around the target sites of siRNAs and miRNAs. We first developed an algorithm for computing local accessibilities of RNA sequences, and implemented it as a software package called ‘Raccess’. Raccess exhaustively calculates the exact accessibilities based on an energy model under a maximal span constraint of base pairs. Using Raccess, we investigated how accessibility depends on the GC content, the access length and the maximal span. We showed that the computed accessibilities are relatively insensitive to the choice of the maximal span. Then, we computed for both siRNAs and miRNAs the correlations between accessibility and the off-target effects of the target sites. We found that the efficacy of siRNAs depends strongly on the accessibility of the very 3'-end of their binding sites, which might reflect the mechanism for target site recognition by the RISC complex. We also showed that the efficacies of some miRNAs have positive correlations with accessibilities in broad regions downstream of their putative binding sites, which might imply that the downstream regions of the target sites are bound by other proteins, allowing the miRNAs to implement their function. We also investigated the off-target effects of an siRNA which is being investigated as a potential RNA medicine. We showed that the off-target effects of the siRNA have similar correlations to the miRNA activities, indicating that the off-target effects of the siRNA are caused by the same mechanism as the target repression of the miRNAs.

We only investigated the relation between accessibility and siRNA/miRNA efficacy. However, accessibility is expected to influence every type of biological process involving RNA molecules. It will be interesting to investigate the accessibilities of general mRNAs and other functional RNA molecules.
accessibilities has many potential biotechnological applications other than efficient siRNA design. For example, accessibility can be an important factor in the design of efficient probes for \textit{in situ} hybridization. Some portion of the positional inhomogeneity of read tag counts observed in RNA-seq data can be also attributed to the local accessibilities of \textit{m}RNA molecules. Investigating to what extent the computed accessibilities are useful in these applications will also be interesting.

**ACKNOWLEDGEMENTS**

The authors thank their colleagues who worked on the project for discussion and comments. Computations were performed using the supercomputing facilities at the Human Genome Center, University of Tokyo (http://sc.hgc.jp/shirokane.html).

**Funding:** ‘Functional RNA Project’ funded by the New Energy and Industrial Technology Development Organization (NEDO), Japan: a Grant-in-Aid for Young Scientists (21700330); a Grant-in-Aid for Scientific Research (A) (22240031); grants from the Development Program of Japan Science and Technology Agency.

**Conflict of Interest:** none declared.

**REFERENCES**


