BIOINFORMATICS ORIGINAL PAPER
Vol. 27 no. 13 2011, pages 1788–1797
doi:10.1093/bioinformatics/btr276

Structural bioinformatics
Advance Access publication April 29, 2011

A detailed investigation of accessibilities around target sites of siRNAs and miRNAs

Hisanori Kiryu[1,*], Goro Terai[2], Osamu Imamura[3], Hiroyuki Yoneyama[3], Kenji Suzuki[4] and Kiyoshi Asai[1,5]

[1]Department of Computational Biology, Faculty of Frontier Science, The University of Tokyo, Chiba 277-8561, [2]INTEC Systems Institute, Inc., Biobusiness Division, Tokyo 136-0075, [3]Stelic Institute & Co., Tokyo 106-0044, [4]Division of Gastroenterology and Hepatology, Graduate School of Medical and Dental Sciences, Niigata University, Niigata 951-8510 and [5]Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Tokyo 135-0064, Japan

Associate Editor: Ivo Hofacker

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 show 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 show 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 show 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 ATP $\longrightarrow$ ADP+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

*To whom correspondence should be addressed.

$P_{\mathrm{acc}}(s_a)$ of segment s_a in an RNA sequence x of length N by

$$P_{\mathrm{acc}}(s_a)=\sum_{\sigma\in\mathcal{S}(s_a)}\exp\left(-E(\sigma,x)/\mathrm{RT}\right)/Z(x) \qquad (1)$$

$$Z(x)=\sum_{\zeta\in\mathcal{S}_0}\exp\left(-E(\zeta,x)/\mathrm{RT}\right)$$

where $\mathcal{S}_0$ is the set of all possible secondary structures of x, $\mathcal{S}(s_a)$ is the set of all secondary structures that have no paired bases in segment s_a of sequence x, $E(\sigma,x)$(kcal/mol) is the energy of secondary structure σ on x, which is calculated by the Turner energy model (Mathews *et al.*, 1999), $R=1.9872\times10^{-3}$(kcal/(mol·K)) is the gas constant and $T=310.15$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, $\mathcal{S}(s_a)$ 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, $P_{\mathrm{acc}}(s_a)$ would represent the probability that a single RNA molecule has accessible segment s_a.

Recent studies have shown that the techniques that combine all the possible secondary structures are highly effective in predicting secondary structures (Ding and Lawrence, 1999, 2001, 2003; Ding *et al.*, 2005; Do *et al.*, 2006; Hamada *et al.*, 2009b; Kiryu *et al.*, 2007; Mathews, 2006). These studies have shown that the behavior of RNA molecules is more accurately described by including the probabilistic fluctuation of secondary structures (Carvalho and Lawrence, 2008; Ding and Lawrence, 2003; Hamada *et al.*, 2009a). Hence, we expect that Equation (1) more accurately defines the accessible regions of real RNA molecules compared to defining them from a single predicted secondary structure.

To date, only a few programs can compute the accessibility based on the energy model. Sfold (Ding and Lawrence, 2003; Ding *et al.*, 2004) computes the accessibilities using posterior sampling techniques. The time complexity of the algorithm is $\mathcal{O}(NW^2+MNW)$ 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 $(1/M)$. 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 $\mathcal{O}(N^3)$. Since $\mathcal{O}(N^3)$ time is prohibitive to apply to long mRNAs, it cannot compute the accessibilities 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 S ($W\le S\le N$) with complexities of $\mathcal{O}(NS^2)$ in time and $\mathcal{O}(N+S^2)$ 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 $S=N$, which is again prohibitive in terms of complexity for long RNA sequences. Furthermore, the current RNAplfold implementation results in overflow errors when 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 'Raccess', which calculates exactly the accessibility of all the segments of a fixed length for a given RNA sequence. Raccess 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 W. The computational complexity is $\mathcal{O}(NW^2)$ in time and $\mathcal{O}(N+W^2)$ in space. Raccess computes the inside and outside variables by using the logarithms of Boltzmann factors, so it is robust to overflow errors and returns correctly for sequences longer than $N=100$ kb.

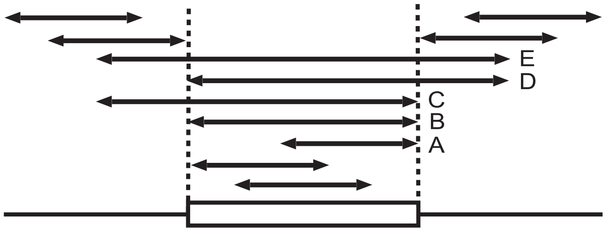

Fig. 1. Comparison of accessible segments s_a considered in this article and those in the previous studies. The white box represents the (putative) binding region of siRNAs or miRNAs. The patterns of accessible segments are represented by the arrows. The patterns investigated in the previous studies are as follows. Patterns A, B: Tafer *et al.* (2008). Pattern B: Lu and Mathews (2008a); Shao *et al.* (2007). Patterns B, C, D, E: Kertesz *et al.* (2010). All the patterns in the figure are investigated in this article.

Using Raccess, we first investigate the basic properties of the accessibilities. We show how the accessibility depends on the GC composition, the length of s_a and the maximal span. We also show that the computed accessibilities are relatively insensitive to the choice of the maximal span. We then apply Raccess 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 3′-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 x_a is the middle position of access segment s_a, and the access length l_a is the length of that segment. The accessibility $P_{\mathrm{acc}}(s_a)$ 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 $\Delta E_{\mathrm{acc}}(s_a) = -RT\log(P_{\mathrm{acc}}(s_a))$ is the thermodynamic work needed to dissociate all the paired bases within the access segment s_a. The partition function $Z(x)$ 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.

$$S \longrightarrow L$$

$$P \longrightarrow a_{<} L a_{>}$$

$$L \longrightarrow aL|M$$

$$R \longrightarrow Ra|P$$

$$M \longrightarrow MR|\epsilon$$

where ϵ is the null terminal symbol, a is a terminal symbol corresponding to an unpaired nucleotide and $a_{<}$ and $a_{>}$ 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 multifurcations 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 k-loops enclosed by the backbone and hydrogen bonds of base pairs. A k-loop with $k>0$ 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 $P \to L$ transitions. The leftmost segment of unpaired bases is emitted by $L \to L$ transitions. The other unpaired bases are emitted by $R \to R$ transitions. Each $M \to MR$ 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.

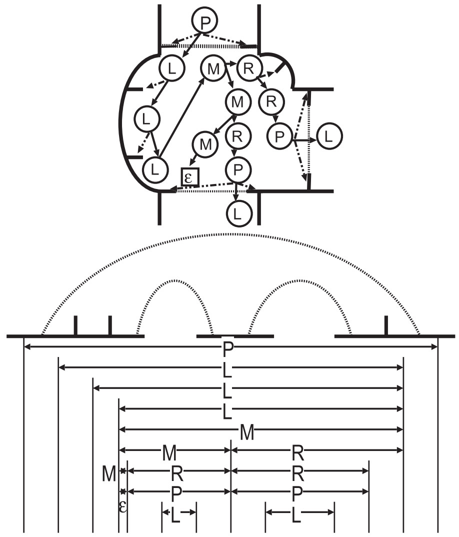

Fig. 2. The parse tree of a three-loop in our SCFG model. The solid arrows represent state transitions. The broken arrows represent base emissions (top). The solid lines represent an RNA structure. The dotted lines represent base pairs. Derivation of the structure (bottom). The base pairs are represented by arcs. For each state, the sequence range generated by the state is shown by a horizontal arrow. State transitions occur from top to bottom.

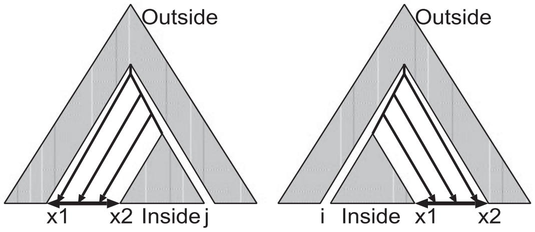

Fig. 3. The numerator of Equation (1) in our SCFG model. The inside and outside variables represent the sum of the contributions from all the parse trees in the shaded regions.

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

$$(L \longrightarrow L)^{l_a} \text{ and } (R \longrightarrow R)^{l_a}$$

where $(A \longrightarrow A)^n$ indicate n successive $A \longrightarrow A$ transitions. Hence, the sum of all the parse trees that includes these transitions can be represented using the inside $\alpha_s(i,j)$ and outside $\beta_s(i,j)$ variables:

$$p_L(s_a) = \sum_{x_2 \le j \le N} \beta_L(x_1-1,j) \cdot \prod_{k=x_1}^{x_2} t(L \longrightarrow L) e_L(k) \cdot \alpha_L(x_2,j)$$

$$p_R(s_a) = \sum_{0 \le i \le x_1-1} \beta_L(i,x_2) \cdot \prod_{k=x_1}^{x_2} t(R \longrightarrow R) e_R(k) \cdot \alpha_L(i,x_1-1)$$

where x_1 and x_2 represent the first and the last positions of s_a, respectively, $e_s(i)$ represents the emission probability of state s at position i, and $t(s \longrightarrow s')$ represents the transition probability for transition $s \longrightarrow s'$. 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 $Z(x)$:

$$P_{\mathrm{acc}}(s_a) = (p_L(s_a) + p_R(s_a))/Z(x).$$

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 $R \longrightarrow L$, then the transitions, $L \longrightarrow L \longrightarrow L$, $L \longrightarrow R \longrightarrow L \longrightarrow L$, and $L \longrightarrow R \longrightarrow R \longrightarrow R \longrightarrow L \longrightarrow R \longrightarrow L$ all emit three contiguous unpaired bases as left emissions. In this case, the computation of the accessibilities includes the summation of all such emission patterns and is practically impossible. As another example, if we add the transition $M \longrightarrow L$ 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.

2.2 Accessibility formula for the energy model

To derive the accessibility formula for the energy model, we use the grammar named Rfold that was developed in a previous paper to calculate the base pairing probabilities under the maximal span constraint (Kiryu *et al.*, 2008). Rfold fully uses the advantage of the maximal span constraint and computes base pairing probabilities and secondary structures of a given RNA sequence with $\mathcal{O}(NW^2)$ time and $\mathcal{O}(N+W^2)$ space, which reduce to familiar $\mathcal{O}(N^3)$ time and $\mathcal{O}(N^2)$ space when there is no constraint on the maximal span (i.e. $W=N$). We have reproduced the Rfold grammar in the Supplementary Material. The unambiguity of the Rfold grammar is described in detail in the Supplementary Material of the original article and we do not repeat it here.

The Rfold model has five types of unpaired base emission rules that correspond to external, multi (left emission), multi (right emission), hairpin and internal/bulge loops. The energy model assigns them distinct partial free energies. In the Rfold model, external loops are emitted by the Outer state as right emissions: $\text{Outer} \longrightarrow \text{Outer} \cdot a$. Multi loops are emitted by two states Multi and Multi2, which emit unpaired bases on the left and right, respectively: $\text{Multi} \longrightarrow a \cdot \text{Multi}$, $\text{Multi2} \longrightarrow \text{Multi2} \cdot a$. Hairpin and internal/bulge loops do not have specific emission states, and they are emitted by the following transition rules of the StemEnd state, which represent the closing base pair of the loop model of secondary structures: $\text{StemEnd} \longrightarrow c_{k1}\text{Stem}c_{k2}$ $(k1+k2 \geq 1)$ and $\text{StemEnd} \longrightarrow c_k$ $(k \geq 3)$, where c_k represents a string of length k. Owing to the grammatical structure, the complete list of all the patterns that emit contiguous sequences of unpaired bases of length l_a is (see Fig. 4):

$$(\text{Outer} \longrightarrow \text{Outer})^{l_a}$$

$$(\text{Multi} \longrightarrow \text{Multi})^{l_a}$$

$$(\text{Multi2} \longrightarrow \text{Multi2})^{l_a}$$

$$\text{StemEnd} \longrightarrow c_{k1} \cdot \text{Stem} \cdot c_{k2} \ (k1 \geq l_a \text{ or } k2 \geq l_a)$$

$$\text{StemEnd} \longrightarrow c_k \ (k \geq l_a)$$

Therefore, the numerator of Equation (1) is obtained by summing the parse trees that include one of these patterns, which can also be computed using inside–outside variables as shown graphically in Figure 4. The explicit formulas are described in the Supplementary Material. Note that because of the maximal span constraint, inside–outside variables are computed only for the regions $\{(i,j)|1 \leq i < j \leq L, |j-i| \leq W\}$ in the figure. We find the accessibilities at each computation of the outside variables and immediately output them to file. We need only $\mathcal{O}(N)$ memory for the Outer state, and $\mathcal{O}(W^2)$ memory for the other states.

We implemented the above algorithm as the software package Raccess. For a given set of access lengths l_{ak} $(k=1,\cdots K)$ and a maximal span W, Raccess computes the accessibilities for all the possible access segments (x_a, l_{ak}) $(x_a=1,\ldots,N, k=1,\ldots,K)$ in a single run of inside–outside algorithms. For a modest size of $K=10 \sim 100$, the computation is dominated by the inside–outside computations, so the time complexity is given by $\mathcal{O}(NW^2)$. The space complexity is $O(N+W^2)$ for fixed K as in the case of base pairing probabilities (Kiryu *et al.*, 2008).

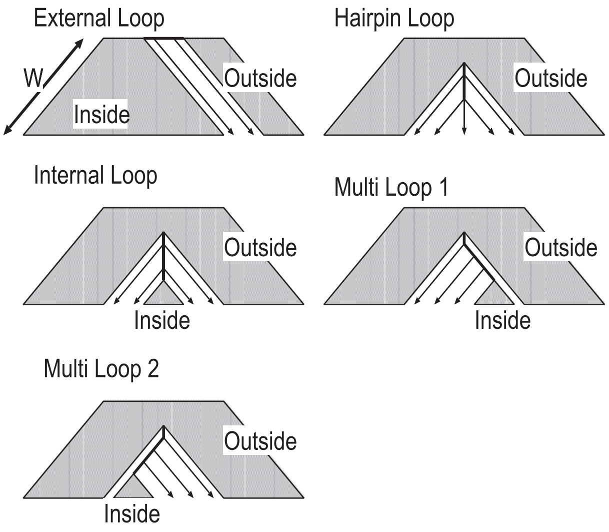

Fig. 4. The numerator of Equation (1) for the Rfold model. The explicit formulas for these figures are presented in the Supplementary Material.

As described in the Section 1, Raccess has the advantage over RNAplfold that it can calculate the exact accessibilities for long sequences, which is useful for avoiding the artificial effects caused by the ends of the sliding windows of RNAplfold. On the other hand, a major drawback to the current implementation of Raccess compared with RNAplfold is speed: currently, Raccess is 2–10 times slower than RNAplfold for similar parameter settings. This is mainly due to the substantial number of log-sum function calls needed to calculate the inside–outside variables accurately. We leave the reimplementation of Raccess as a faster program to the future.

3 DATASET AND DATA PROCESSING

3.1 Comparison of run time and memory usage

The experiments were performed on an AMD Opteron 854 Quad core 2.8 GHz with 64 GB RAM and CentOS release 5.3. We used RNAplfold of the Vienna RNA package version 1.8.4, a binary program of Sfold version 2.2 and OligoWalk version 1.0. The test sequences were generated randomly with equal A, C, G, U probabilities. As for Sfold, the number of sampled structures were set to its default value (1000). If a program had no options to constrain the maximal span or the sequence window, it was run with its default setting with respect to that option. The command line options are shown in the Supplementary Material.

3.2 Dependence on GC composition

We generated several random RNAs of length 10 000 with fixed GC biases and computed accessibilities for them. We used only the accessibility values for subsequences lying at least 1000 bases from the sequence ends in order to eliminate boundary effects.

3.3 siRNA dataset

For siRNA efficacy, we used the dataset of Huesken *et al.* (2005) produced using a dual reporter system on HeLa cells to measure the repression levels of target gene expressions. The dataset contains the efficacies of 2431 distinct siRNAs targeting 34 mRNAs derived from human and rat. The efficacies z_{eff} are represented by the amount of mRNA expression reduction and take values between 0 (no effect) and 100 (completely repressed). This dataset is the largest single dataset available so far. It has been a gold standard 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; Although this experiment essentially measures the efficacies for free-living mRNAs, real mRNAs might form complexes with other RNAs and proteins. Therefore, it may be possible that the best siRNA that completely degrades the target mRNAs in this experiment fails for the nascent mRNAs in a certain condition. Still, the knowledge obtained from this idealized experiment should be useful to understand the basic mechanisms of siRNA functions.

3.4 miRNA datasets

For miRNA efficacy, we used the results of a microarray experiment by Rodriguez *et al.* (2007), which measured the changes of gene expressions, before and after the knock down of miR-155 of mouse by RNAi. The efficacy z_{eff} in this case is given by $z_{\text{eff}}(i)=\log\left(g_A(i)/g_B(i)\right)$, where $g_B(i)$ and $g_A(i)$ are the mRNA concentrations of the i-th gene before and after the miR-155 knock down, respectively. We also used the microarray data of Lim *et al.* (2005), where the authors measured gene expression changes after the transfection of miRNAs (miR-1, miR-124) in HeLa cells. The efficacy z_{eff} in this case is given by $z_{\text{eff}}(i)=\log\left(g_B(i)/g_A(i)\right)$, where $g_B(i)$ and $g_A(i)$ are the mRNA concentrations of the i-th gene before and after the miRNA transfection, respectively. In both cases, the putative target genes are defined by all the genes that have 7mer seed sequences in their $3'$-UTR sequences. There are numerous papers that the existence of 6,7-mer seed sequence is the most important feature of the mRNAs targeted by miRNAs (Lim *et al.*, 2005). Usually, we observe clear statistical evidence that the mRNAs with seed sequences are down(up) regulated after over(under) expression of the miRNA. Since there are no comprehensive dataset that identifies the binding sites at single nucleotide resolution, we used the putative binding sites. We expect that if the number of false positives are too large, then no statistical significance will be observed. On the other hand, if there is any statistical significance, then we consider that it suggests the nature of true binding sites.

3.5 siRNA off-target dataset

The dataset of Huesken *et al.* (2005) only measures the repression of target mRNAs for a large number of siRNA species, and it does not measure off-target effects of those siRNAs. To investigate off-target effects of siRNA transfection, we conducted a microarray experiment where human fibroblast cells (CCD-18co) were transfected by a 27mer siRNA (positive-siRNA) targeting the 'carbohydrate (N-acetylgalactosamine 4-sulfate 6-O) sulfotransferase 15' (CHST15) gene. This siRNA is currently in a preclinical study and is being developed for the treatment of Crohn's disease stricture, which occurs in 30% of Crohn's disease patients, and is the leading cause of repeat surgery. CHST15 has been shown to be overexpressed at sites of fibrosis in Crohn's disease patients, and upregulation of CHST15 correlates with the severity of fibrosis in animal models of colitis with intestinal fibrosis (Imamura, 2008; Kai *et al.*, 2007; Suzuki *et al.*, 2010).

The gene expression is compared with those of a negative control [the same fibroblast cells transfected by a different 27mer siRNA (negative-siRNA)]. The siRNA in the negative control was derived from a green fluorescent gene that did not have a significant match (i.e. no contiguous 19mer match within hamming distance 1) with any of the human RefSeq genes. We calculated the efficacy by the formula $z_{\text{eff}}(i)=\log\left(g_P(i)/g_N(i)\right)$, where $g_P(i)$ and $g_N(i)$ represent the mRNA concentrations of the i-th gene after the transfection of positive-siRNA (25 nM) and negative-siRNA (5 nM), respectively. The details of the microarray experiment are described in 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 s_a which was specified by an access length l_a and an access position x_a relative to the (putative) binding site, we ranked the dataset by the accessibility of s_a. 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:

$$t=\frac{\mu_1-\mu_2}{\sqrt{(s_1^2/n_1)+(s_2^2/n_2)}} \tag{2}$$

where μ_i and s_i are the mean and the SD of the efficacy in group i, respectively, and n_i is the size of the group i. 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 those 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, Raccess is the only program that returns correctly within 10 h. RNAplfold is a few times faster than Raccess 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 10 000 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

	(N,W,S)	(100,100,100)	(1k,100,100)	(1k,100,1k)	(1k,1k,1k)	(10k,100,100)	(10k,100,1k)	(10k,100,10k)
Run Time	Raccess	0.19 s	3.67 s	3.67 s	48.47 s	38.80 s	38.80 s	38.80 s
	RNAplfold	0.10 s	1.25 s	10.71 s	19.64 s	12.96 s	x	x
	Sfold	2.34 s	47.28 s	47.28 s	3 m28 s	–	–	–
	OligoWalk	0.10 s	1 m20 s	285 m23 s	285 m23 s	15 m16 s	–	–
Memory (MB)	Raccess	16.2 MB	16.2 MB	16.2 MB	99.9 MB	16.6 MB	16.6 MB	16.6 MB
	RNAplfold	8.5 MB	11.3 MB	71.9 MB	71.9 MB	25.9 MB	x	x
	Sfold	3.3 MB	26.9 MB	26.9 MB	26.9 MB	–	–	–
	OligoWalk	39.5 MB	47.2 MB	147 MB	147 MB	84.7 MB	–	–

The numbers in the brackets on the title row represent (N: sequence length, W: maximal span, S: sequence window size), respectively. Symbol '–' indicates that the program does not return results within 10 h. Symbol 'x' indicates that the program terminates within 10 h but does not return any result due to overflow errors.

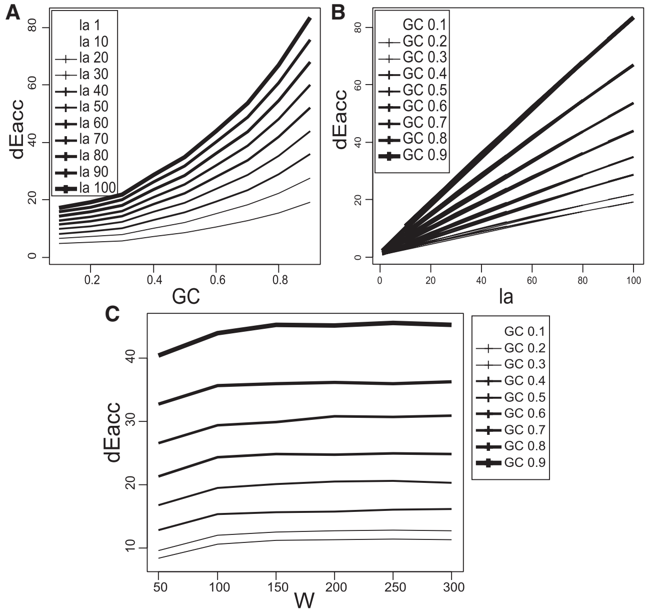

Fig. 5. (**A**) GC composition dependence of the accessibility for various accessible lengths. (**B**) Accessible length l_a dependence of the accessibility at various GC compositions. The maximal span W is fixed at 100 in both figures. (**C**) Dependence of the accessibility on the maximal span W for various GC compositions. Accessible length l_a is fixed at 50 in this figure.

4.2 Dependence on GC composition

Figure 5 shows how the access energy ΔE_{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 ΔE_{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.

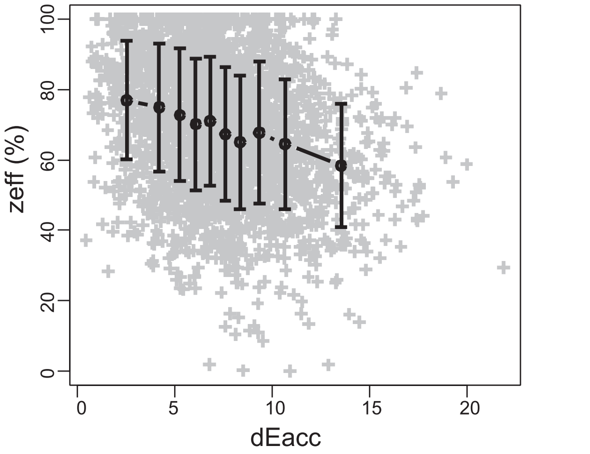

Fig. 6. siRNA efficacy–accessibility correlation. The x-axis shows the access energy. The y-axis shows the reduced fraction of target expressions. The gray points are the scatter plot for siRNA–target pairs. The circles are the mean values of z_{eff} over the subdataset separated into 10 equally sized bins based on their ΔE_{acc} values. The length of the vertical lines represents 1 SD. The maximal span W, access length l_a and access position x_a are set at 400, 16 and 0, respectively. The Spearman's rank correlation coefficient is −0.26, the corresponding P-value is 1.5×10^{-38} (two sided t-test).

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 ΔE_{acc}, which suggests that the accessibility is an important determinant of efficacy.

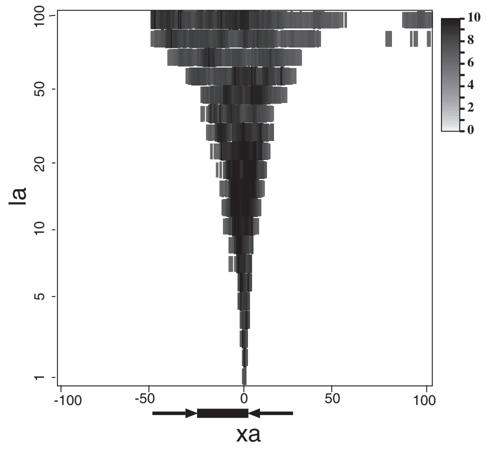

Fig. 7. Density plot of siRNA efficacy–accessibility correlations. The x-axis is the access position x_a relative to the 3′-end of each binding site. The 19mer binding region is shown as the thick bar between the arrows. The y-axis is access length l_a in log scale. The value of t given by Equation (2) is indicated by shading according to the density scale shown. Only t-values exceeding 6 which corresponds to $P \approx 10^{-9}$ in the normal approximation are shown for clarity. The maximal span W is set at 400.

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 $t_{\max}$ than the other W values. Hence, we set $W = 400$ in the following sections.

4.4 miRNA datasets

Figure 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, bearing 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 $x_a \sim 0$, but there are broad regions ($l_a = 70 \sim 100$) downstream of the putative binding site, where accessibility is positively correlated with efficacy. Interestingly, all the figures have a weakly blotted region around $x_a = 40 \ldots 60$, $l_a = 10 \ldots 20$. They might be reflecting the mechanisms of the miRNA functions such as protein binding and specific conformations of the mRNAs downstream of the binding site.

Table 2. Dependence of siRNA efficacy–accessibility correlations on the maximal span

W	r_s	e_p	$t_{\max}$	t_{mean}
5	−0.18	19.7	8.3	2.9
10	−0.22	25.8	9.8	3.3
50	−0.19	19.5	9.7	3.0
100	−0.22	28.0	10.0	3.1
200	−0.25	34.2	11.3	3.5
400	−0.26	37.8	11.5	3.3
600	−0.22	27.4	10.1	3.3
800	−0.25	33.5	10.8	3.2
1000	−0.23	28.8	10.5	2.8
2000	−0.23	30.7	10.8	2.6
4000	−0.23	30.7	10.2	2.7

W: maximal span, r_s: Spearman's rank correlation for the scatter plot of Figure 6 with different W. e_p: exponent of P-value such that $P\text{-value} = 10^{-e_p}$. P-values are calculated for r_s using two-sided t-test. $t_{\max}$ and t_{mean}: maximal and mean t-value of the density plot of Figure 7 with different W.

4.5 siRNA off-target dataset

As described in the Methods section, an siRNA targeting a glycogene (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 GAUGAAU 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 ΔE_{acc} is averaged over access positions x_a in the range $[-5, 5]$). The genes whose expressions 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

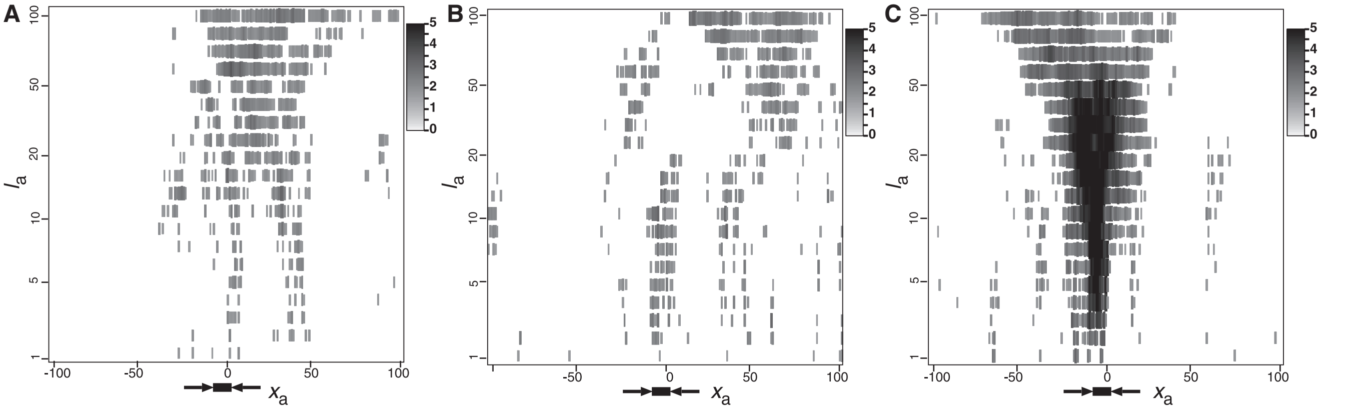

Fig. 8. Density plot of miRNA efficacy–accessibility correlations. The x-axis is the access position x_a relative to the 3′-end of each putative binding site. The 7 mer putative binding region is shown as the thick bar between the arrows. The y-axis is the access length l_a in log scale. The value of 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 (GUGCCUU) 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.

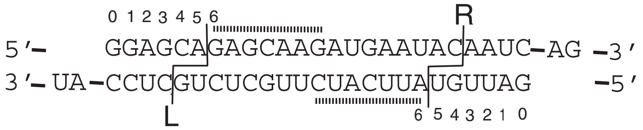
0 1 2 3 4 5 6
R
5′– GGAGCAGAGCAAGAUGAAUACAAUC–AG –3′
3′– UA–CCUCGUCUCGUUCUACUUAUGUUAG –5′
L
6 5 4 3 2 1 0

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

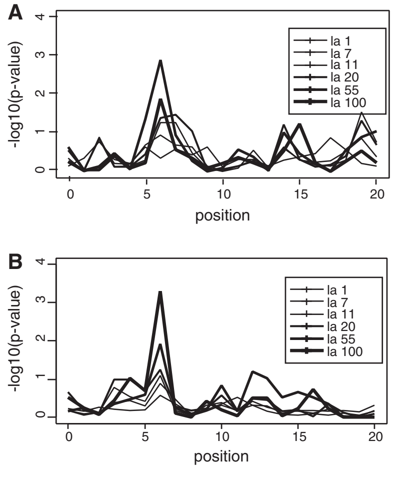

Fig. 10. Plot of P-value for each 7 mer subsequence. The x-axis is the 3′-end position of the 7 mers in the siRNA sequences. The y-axis is the value of $-\log_{10}$(P-value), which are calculated from the mean t values averaged over access positions x_a in the range [−5, 5]. Access lengths $l_a = 1,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

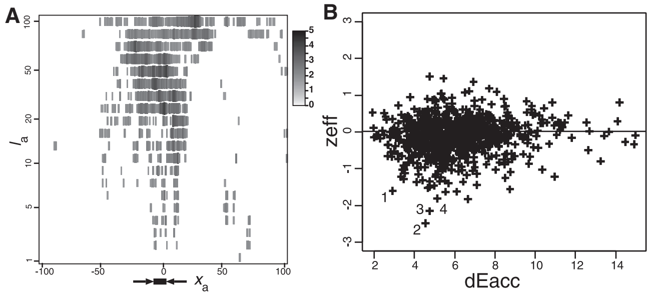

Fig. 11. Efficacy–accessibility correlations of the siRNA off-target dataset. (**A**) Density plot of t-values for the seed GAUGAAU corresponding to the 'R' type Dicer cleavage. The value of t given by Equation (2) is indicated by shading according to the density scale shown. Note that scale bars of 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.

growth factor. The repression of these genes might cause side effects in the therapeutic application of the siRNA.

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 $\mathcal{O}(NW^2 + MNW)$ (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

Table 3. The potential off-targets of the siRNA

ID	z_{eff}	ΔE_{acc}	Gene symbol	Gene name
1	-1.61	2.92	TWISTNB	TWIST neighbor
2	-2.49	4.55	IGF2BP3	Insulin-like growth factor 2 mRNA binding protein 3
3	-2.15	4.75	OXTR	Oxytocin receptor
4	-1.81	5.14	PECI	Peroxisomal D3,D2-enoyl-CoA isomerase

The ID corresponds to the numbers in Figure 11.

designing efficient siRNAs, trans-cleaving 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 structure space. It means the computed probabilities are inaccurate for such sequences.

OligoWalk uses a version of McCaskills DP algorithm adapted to the cases where the RNA sequences are subject to accessibility constraints. It has time complexity $\mathcal{O}(N^3)$. 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 $\mathcal{O}(NS^2)$ (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.

Raccess has time complexity $\mathcal{O}(NW^2)$. It can compute the accessibility values for long sequences without using sequence windows nor causing overflow errors. On the other hand, it is a few times slower than RNAplfold in many cases. It does not predict efficient siRNAs.

5.2 Limitation of our study

There are a few limitations in the experimental data as well as the analyses used in this study. As for the siRNA efficacy data (Huesken *et al.*, 2005), the number of target mRNAs is only 34, which is a small sample of the human genes. Hence, it is possible the results obtained from them might not generalize to other mRNAs. Regarding our analytical method, it should be noted that the accessibility values between neighboring or overlapping accessible segments are similar to each other. It indicates that the t-values among close points in the density plots are not independent and correlated. It is also noted that even when the computed P-values are quite significant, the underlying efficacy data are subject to high variability as shown in Figure 6. Further, since these analyses is based on observational studies, the correlation of Figure 6 could be subject to confounding and unreported correlations [see Efron (2004)]. However, none of these limitations easily explains the pattern of the t-value densities around the 5′-end of binding sites observed in Figure 7. Since the high t-value regions are roughly symmetric around position zero, it is unlikely that they are caused by any sequence bias of the siRNA sequences. The fact that the high t-values occur for accessible segment with length up to 100 indicates that they are not simply caused by the local sequence features around position zero. These are the reasons that we believe that Figure 6 reflects the roles of the accessibility in the siRNA mechanisms. On the other hand, the results are somewhat weaker for the miRNA dataset and siRNA off-target dataset (Lim *et al.*, 2005; Rodriguez *et al.*, 2007; Suzuki *et al.*, 2010). Since only a few number of miRNAs and siRNA are examined due to unavailability of data, it is difficult to extract general rules for the accessibility–efficacy correlations with confidence. Furthermore, since the binding sites used for the density plots are only putative, the results are subject to statistical noise caused by the false binding sites.

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 siRNAs of efficacy > 70% with high performance (sensitivity 22.7% and specificity 96.5%). It will be interesting to investigate whether our exhaustive calculation further improves the performance of siRNA design tools.

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 efficacy and accessibility around 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. Computation of 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 *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 mRNA 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 Contract Development Program of Japan Science and Technology Agency.

Conflict of Interest: none declared.

REFERENCES

Bernhart,S.H. *et al.* (2006) Local RNA base pairing probabilities in large sequences. *Bioinformatics*, **22**, 614–615.

Busch,A. *et al.* (2008) IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*, **24**, 2849–2856.

Carvalho,L.E. and Lawrence,C.E. (2008) Centroid estimation in discrete high-dimensional spaces with applications in biology. *Proc. Natl Acad. Sci. USA*, **105**, 3209–3214.

Chen,K. *et al.* (2009) Reexamining microRNA site accessibility in Drosophila: a population genomics study. *PLoS One*, **4**, e5681.

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

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

Ding,Y. and Lawrence,C.E. (2003) A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res.*, **31**, 7280–7301.

Ding,Y. *et al.* (2004) Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic Acids Res.*, **32**, W135–W141.

Ding,Y. *et al.* (2005) RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA*, **11**, 1157–1166.

Do,C. *et al.* (2006) CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, **22**, e90–e98.

Donald,W.Z. (2004) A note on preliminary tests of equality of variances. *Br. J. Math. Stat. Psychol.*, **57**, 173–181.

Efron,B. (2004) Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Am. Stat. Assoc.*, **99**, 96–104.

Gredell,J.A. *et al.* (2008) Impact of target mRNA structure on siRNA silencing efficiency: a large-scale study. *Biotechnol. Bioeng.*, **100**, 744–755.

Hamada,M. *et al.* (2009a) Prediction of RNA secondary structure using generalized centroid estimators. *Bioinformatics*, **25**, 465–473.

Hamada,M. *et al.* (2009b) Predictions of RNA secondary structure by combining homologous sequence information. *Bioinformatics*, **25**, i330–i338.

Hofacker,I.L. and Tafer,H. (2010) Designing optimal siRNA based on target site accessibility. *Methods Mol. Biol.*, **623**, 137–154.

Huesken,D. *et al.* (2005) Design of a genome-wide siRNA library using an artificial neural network. *Nat. Biotechnol.*, **23**, 995–1001.

Imamura,O. (2008) siRNA-mediated Erc gene silencing suppresses tumor growth in Tsc2 mutant renal carcinoma model. *Cancer Lett.*, **268**, 278–285.

Kai,Y. *et al.* (2007) Treatment with chondroitinase ABC alleviates bleomycin-induced pulmonary fibrosis. *Med. Mol. Morphol.*, **40**, 128–140.

Kertesz,M. *et al.* (2007) The role of site accessibility in microRNA target recognition. *Nat. Genet.*, **39**, 1278–1284.

Kertesz,M. *et al.* (2010) Genome-wide measurement of RNA secondary structure in yeast. *Nature*, **467**, 103–107.

Kim,D.H. *et al.* (2005) Synthetic dsRNA Dicer substrates enhance RNAi potency and efficacy. *Nat. Biotechnol.*, **23**, 222–226.

Kiryu,H. *et al.* (2007) Robust prediction of consensus secondary structures using averaged base pairing probability matrices. *Bioinformatics*, **23**, 434–441.

Kiryu,H. *et al.* (2008) Rfold: an exact algorithm for computing local base pairing probabilities. *Bioinformatics*, **24**, 367–373.

Lim,L.P. *et al.* (2005) Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, **433**, 769–773.

Lu,Z.J. and Mathews,D.H. (2008a) Efficient siRNA selection using hybridization thermodynamics. *Nucleic Acids Res.*, **36**, 640–647.

Lu,Z.J. and Mathews,D.H. (2008b) Fundamental differences in the equilibrium considerations for siRNA and antisense oligodeoxynucleotide design. *Nucleic Acids Res.*, **36**, 3738–3745.

Lu,Z.J. and Mathews,D.H. (2008c) OligoWalk: an online siRNA design tool utilizing hybridization thermodynamics. *Nucleic Acids Res.*, **36**, W104–W108.

Mathews,D. *et al.* (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.

Mathews,D.H. (2006) Revolutions in RNA secondary structure prediction. *J. Mol. Biol.*, **359**, 526–532.

Matveeva,O. *et al.* (2007) Comparison of approaches for rational siRNA design leading to a new efficient and transparent method. *Nucleic Acids Res.*, **35**, e63.

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

Morten,W.F. and Leiv,S. (2009) The WilcoxonâŁ"MannâŁ"Whitney test under scrutiny. *Stat. Med.*, **28**, 1487–1497.

Rodriguez,A. *et al.* (2007) Requirement of bic/microRNA-155 for normal immune function. *Science*, **316**, 608–611.

Shao,Y. *et al.* (2007) Effect of target secondary structure on RNAi efficiency. *RNA*, **13**, 1631–1640.

Suzuki,K. *et al.* (2010) Trial of new anti-Crohn's disease stricture therapy by using siRNA. *J. Japan. Soc. Gastroenterol.*, **107**, A292.

Tafer,H. *et al.* (2008) The impact of target site accessibility on the design of effective siRNAs. *Nat. Biotechnol.*, **26**, 578–583.

Zuker,A.M. *et al.* (1999) Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide. In Barciszewski,J. and Clark,B.F.C., eds, *RNA Biochemistry and Biotechnology.* NATO Science Series, Kluwer Academic Publishers, The Netherlands, pp. 11–43.