BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 8 2012, pages 1093–1101
doi:10.1093/bioinformatics/bts097

Structural bioinformatics

Advance Access publication February 28, 2012

Rchange: algorithms for computing energy changes of RNA secondary structures in response to base mutations

Hisanori Kiryu[1,*] and Kiyoshi Asai[1,2]

[1]Department of Computational Biology, Faculty of Frontier Science, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8561 and [2]Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology (AIST), 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: Measuring the effects of base mutations is a powerful tool for functional and evolutionary analyses of RNA structures. To date, only a few methods have been developed for systematically computing the thermodynamic changes of RNA secondary structures in response to base mutations.

Results: We have developed algorithms for computing the changes of the ensemble free energy, mean energy and the thermodynamic entropy of RNA secondary structures for exhaustive patterns of single and double mutations. The computational complexities are $\mathcal{O}(NW^2)$ (where N is sequence length and W is maximal base pair span) for single mutations and $\mathcal{O}(N^2W^2)$ for double mutations with large constant factors. We show that the changes are relatively insensitive to GC composition and the maximal span constraint. The mean free energy changes are bounded $\sim 7-9$ kcal/mol and depend only weakly on position if sequence lengths are sufficiently large. For tRNA sequences, the most stabilizing mutations come from the change of the 5′-most base of the anticodon loop. We also show that most of the base changes in the acceptor stem destabilize the structures, indicating that the nucleotide sequence in the acceptor stem is highly optimized for secondary structure stability. We investigate the 22 tRNA genes in the human mitochondrial genome and show that non-pathogenic polymorphisms tend to cause smaller changes in thermodynamic variables than generic mutations, suggesting that a mutation which largely increases thermodynamic variables has higher possibility to be a pathogenic or lethal mutation.

Availability and implementation: The C++ source code of the Rchange software is available at http://www.ncrna.org/software/rchange/

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

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 26, 2011; revised on February 19, 2012; accepted on February 21, 2012

1 INTRODUCTION

Measuring the effects of base mutations has long been a standard method for studying relationships between structural RNAs and their biological functions. Molecular biologists mutate some nucleotides of the RNAs of interest in order to alter the structures, and draw various conclusions by observing whether the mutations disrupt biological functions [see, for example, (Warf *et al.*, 2009)]. Observing mutation patterns in genome evolution has also been important for detecting evolutionarily conserved functional RNAs and evolutionarily conserved structural RNAs show special types of mutation patterns that preserve their secondary structures. Several studies have used such patterns as a major feature for finding novel structural RNAs from multiply aligned genomic sequences (Knudsen and Hein, 2003; Pedersen *et al.*, 2004; Rivas and Eddy, 2001; Washietl *et al.*, 2005). Investigating the structural changes due to base mutations has potentially important applications in genome-wide association studies, where many disease-associated single-nucleotide polymorphisms (SNPs) are located outside the protein coding regions in the human genome (Halvorsen *et al.*, 2010; International HapMap Consortium, 2005; Manolio, 2010). It is possible that part of the disease associations of these SNPs can be explained by altered structures of the non-coding RNAs that are transcribed around those regions. Despite the widespread use of mutational analyses and the potential relevance to human diseases, very few studies have developed methods for systematically computing the changes in RNA secondary structures in response to base mutations (Barash and Churkin, 2011; Churkin and Barash, 2006, 2008; Shu *et al.*, 2006; Waldispuhl *et al.*, 2008, 2009).

In our previous paper, we have developed algorithms for computing the thermodynamic variables for a given sequence (Kiryu and Asai, submitted for publication) [see also the earlier works (Clote *et al.*, 2005; Miklos *et al.*, 2005) which developed algorithms for the mean free energy computation]. In this article, we present algorithms for computing the changes in the ensemble free energy, mean free energy and the thermodynamic entropy of secondary structures for all the possible single and double mutations, assuming that the RNA secondary structures follow the Boltzmann distribution $P(\zeta|x)$ of the Turner energy model (TEM) (Mathews *et al.*, 1999):

$$P(\zeta|x) = \exp\left(-\Delta G(\zeta,x)/RT\right)/Z(x)$$
$$Z(x) = \textstyle\sum_{\zeta\in\Omega} \exp\left(-E(\zeta,x)/RT\right)$$

where $\Delta G(\zeta,x)$(kcal/mol) represents the difference of the Gibbs energies $G(\zeta,x)-G(\zeta_0,x)$ between a structure ζ and the structure ζ_0 that contains no base pairs, Ω is the set of all possible secondary structures of x, $R=1.9872\times10^{-3}$ 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 predicting secondary

[*]To whom correspondence should be addressed.

structures of long RNA sequences (Mathews *et al.*, 1999) and the base pair predictions between distant nucleotide positions are particularly inaccurate for such long RNAs (Doshi *et al.*, 2004), it is natural to restrict the maximal span of the base pairs to a fixed value W. The maximal span constraint avoids predicting large number of spurious base pairs between distant positions while still allowing global structure prediction. When the maximal span constraint is applied, Ω includes only structures containing base pairs of span $\leq W$.

Using the thermodynamic model, the mean free energy $U(x)$ and the macroscopic entropy $S(x)$ in energy units are defined by

$$S(x) = -RT\sum\nolimits_{\zeta\in\Omega} P(\zeta|x)\log(P(\zeta|x)) \qquad (1)$$

$$U(x) = \sum_{\zeta\in\Omega} P(\zeta|x)\Delta G(\zeta,x)$$

$$F(x) = -RT\log(Z(x)) = U(x)-S(x)$$

Here, $U(x)$ is the mean free energy difference of RNA secondary structure, $F(x)$ is the ensemble free energy difference $(G(x)-G(\zeta_0,x))$ between the state of the secondary structure ensemble and the unstructured state ζ_0. Our algorithms compute the differences $dF(x^*,x)=F(x^*)-F(x)$, $dU(x^*,x)=U(x^*)-U(x)$ and $dS(x^*,x)=S(x^*)-S(x)$ for all the possible single and double mutants x^* of sequence x. The computational complexities are $\mathcal{O}(NW^2)$ (where N is sequence length and W is maximal base pair span) for single mutations and $\mathcal{O}(N^2W^2)$ for double mutations with large constant factors.

RNAmute (Churkin and Barash, 2006, 2008) computes the minimum free energies and the structural changes for all the possible single mutations by running the RNAfold program in the Vienna RNA package (Hofacker, 2003) for each RNA mutant. RNAmute can also compute k-point mutations that are likely to cause large structural alterations by using the RNAsubopt and RNAfold programs of the Vienna package. The computational complexities are $\mathcal{O}(N^4)$ for single mutations and $\mathcal{O}(Ne^{CN})$ $(C>0)$, which is the complexity of RNAsubopt, for k-point mutations. RDMAS (Shu *et al.*, 2006) also uses RNAsubopt and RNAfold to predict the most deleterious single mutations that cause large structural changes. They compute suboptimal structures $\Gamma(x^*)$ within a fixed range of free energy for each of $3N$ possible single point mutants x^* and returns 'deleteriousness' $D(\zeta_{\mathrm{MFE}}(x)\Gamma(x^*))$ defined by a Boltzmann average of structural distances $d(\zeta_1,\zeta_2)$:

$$D(\zeta_{\mathrm{MFE}}(x),\Gamma(x^*)) = \sum_{\zeta^*\in\Gamma(x^*)} P(\zeta|x)d(\zeta_{\mathrm{MFE}}(x),\zeta^*)$$

$$\zeta_{\mathrm{MFE}}(x) = \mathrm{argmin}_{\zeta\in\Omega}E(\zeta,x)$$

where $\zeta_{\mathrm{MFE}}(x)$ is the minimum free energy (MFE) structure of the original sequence x. RDMAS provide a few options for the distance measure $d(\zeta,\zeta^*)$ including free energy difference $|E(\zeta,x)-E(\zeta^*,x^*)|$, edit distance between secondary structures, and difference between topological indices of secondary structures. The complexity for this procedure is $\mathcal{O}(Ne^{CN})$. Since both RNAmute and RDMAS use either MFE structures or suboptimal structures of a fixed energy range, they do not take into account all the possible secondary structures. On the other hand, our algorithm exactly sums the contributions from all the possible secondary structures in a polynomial time. RNAmutants (Waldispuhl *et al.*, 2008, 2009) calculates all the k-point mutant partition functions $Z_k(x)$ (for $k\leq K$ with a given K) which are given by the sum of all the Boltzmann factors of all the possible k-point mutants:

$$\begin{aligned} Z_k(x) &= \sum\nolimits_{x^*\in\mathcal{M}(k,x)}\sum\nolimits_{\zeta\in\Omega}\exp\big(-E(\zeta,x^*)/RT\big) \\ &= \sum\nolimits_{x^*\in\mathcal{M}(k,x)} Z(x^*) \end{aligned}$$

where $\mathcal{M}(k,x)$ is the set of all the k-point mutants of sequence x. The complexity for computing the partition functions is $\mathcal{O}(K^2N^3)$. While we are generally interested in the individual effects of each mutation x^*, k-point partition functions $Z_k(x)$ mix up all the contributions of the k-point mutants x^* and it is not possible to extract individual partition functions $Z(x^*)$ from $Z_k(x)$ with any simple modification to the RNAmutants algorithm. In their paper, the authors used a statistical sampling method that samples mutated sequences x^* with structures from the k-point Boltzmann distribution associated with $Z_k(x)$. However, the meaning of the sampled sequences is difficult to understand from the statistical physics viewpoint; it is equivalent to a thermodynamic system where time scales of base mutations are as fast as those of RNA folding. Furthermore, sampling methods are very inefficient for long RNA sequences with large structure spaces (Kiryu *et al.*, 2011). On the other hand our algorithm exhaustively computes the individual partition functions $Z(x^*)$ of all the $k=1,2$ points mutants x^*. Finally, none of those studies have computed the differences of thermodynamic quantities such as the ensemble free energy, mean free energy and entropy for each individual mutation.

In the following sections, we first derive our algorithms for a simple stochastic context-free grammar (SCFG). Next we derive similar algorithms for the more complex case in which the emission probabilities depend on the neighboring bases; this mimics the complexities encountered in the energy model. We then describe how to extend the algorithms in these simple cases to the full energy model. In Section 4, we show the run time and required memory size for several parameter settings. We then investigate basic properties of the thermodynamic changes, such as the dependencies on the sequence lengths and the GC compositions, and the choices of the maximal base pair span. We also investigate the dependencies of the thermodynamic changes on the mutated base positions. Finally, we investigate the characteristics of the thermodynamic changes of the tRNA family.

2 ALGORITHMS AND IMPLEMENTATION

2.1 Simple stochastic context-free grammar

The SCFG model that we consider has a single non-terminal state (S) and five transition rules (P, L, R, B, E).

$$P: S \longrightarrow a_< S a_>$$

$$L: S \longrightarrow aS$$

$$R: S \longrightarrow Sa$$

$$B: S \longrightarrow SS$$

$$E: S \longrightarrow \epsilon$$

where $(a_<,\ a_>)$ represents a base pair, a represents an unpaired base and ϵ represents a null terminal symbol. Although this grammar is an ambiguous grammar, we can still define and compute the partition function $Z_0(x) = \sum_\xi P(x,\xi)$ and entropy $S_0(x) = -\sum_\xi P(\xi|x)\log P(\xi|x)$, where the sum is over all the possible parse trees ξ. If a grammar is unambiguous, then the space of the parse trees has one to one correspondence to the space of the secondary

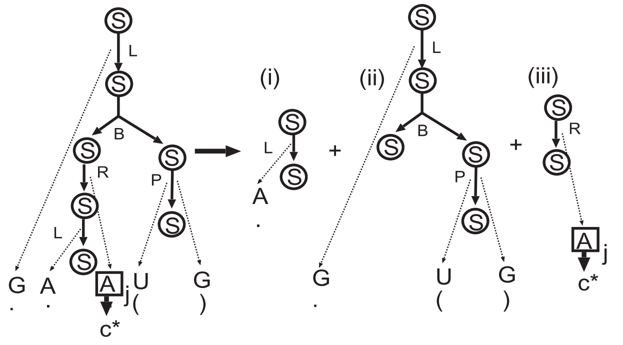

Fig. 1. An example parse tree. When the base A at position j (rectangle) is mutated, we decompose the tree into the inside part (i), the outside part (ii) and the transition that emits the mutating base (iii).

structures and the corresponding probability distribution is defined over the secondary structure space. Although our implementation 'Rchange' uses an unambiguous grammar called the Rfold grammar (see below and the Supplementary Material), the following algorithm works independently of the ambiguity of grammars. To compute the entropy of a single mutant for this model, we must first compute the partition function of the mutated sequence. If the emission probabilities $e_P(a_<,a_>)$, $e_L(a)$ and $e_L(a)$ are independent of the neighboring bases, as is usually the case, then the partition function can easily be computed. For example, we consider a mutant x^* in which the original base x_j at position j is replaced with base c^* (Fig. 1). To compute the partition function $Z(x^*)$, we sum the probabilities $P(x,\zeta)$ over all the parse trees ζ. Each parse tree contains exactly one transition that emits base c^* at position j, and we split the parse tree into two parts that are upstream and downstream of the transition. These partial parse trees are independent of the base at j, hence they also appear in the original sequence x. By collecting these partial parse trees and reorganizing the sum using the inside and outside variables, we obtain the formula

$$
\begin{aligned}
Z(x^*)= & \sum_{j+1<k\leq N} \alpha(j+1,k-1)e_P(c^*,x_k)t(P)\beta(j,k) \qquad (2)\\
& + \sum_{1\leq k<j-1} \alpha(k+1,j-1)e_P(x_k,c^*)t(P)\beta(k,j)\\
& + \sum_{j<k\leq N} \alpha(j+1,k)e_L(c^*)t(L)\beta(j,k)\\
& + \sum_{1\leq k<j} \alpha(k,j-1)e_R(c^*)t(R)\beta(k,j)
\end{aligned}
$$

where $t(r)$ represents the transition probability associated with transition rule r, c^* represents the mutated base, and $\alpha(u,v)$ and $\beta(u,v)$ $(u\leq v)$ represent the inside and outside variables of the original sequence x (Fig. 2). The computational complexity of this calculation is $\mathcal{O}(N)$. Hence, the computation of all the single mutations ($\mathcal{O}(N)$) is dominated by the inside–outside algorithms and the total complexity is $\mathcal{O}(N^3)$.

In our previous paper, we developed a method to calculate the entropy, the mean free energy, and the energy variance of secondary structures by using dynamic programming (DP) algorithms (Kiryu and Asai, submitted for publication). Briefly, we introduced copies of the inside variables, $\alpha_1(i,j)$. [These were denoted by $\alpha'(i,j)$ in the previous paper but that symbol has a different meaning in this article.] These copies represent sums of tree scores as in the ordinary inside variables except that each parse tree score contains a single log transition score insertion [such as $\log(e_P(c,c')t(P))$, $\log(e_L(c)t(L))$, etc.] at some transition. As there is no temperature for SCFG models, we use a definition of entropy that is slightly different from Equation (1): the entropy $S(x)=-\sum_\zeta P(\zeta|x)\log P(\zeta|x)$ is given by $-(Z_1(x)/Z(x))+\log Z(x)$, where $Z_1(x)=\alpha_1(1,N)$ [for a detailed explanation see (Kiryu and Asai, submitted for publication)]. In the present case, we compute $Z_1(x^*)$ for each single-point mutant x^* by using $\alpha_1(i,j)$ and $\beta_1(i,j)$ (which are log-score inserted outside variables analogous to $\alpha_1(i,j)$):

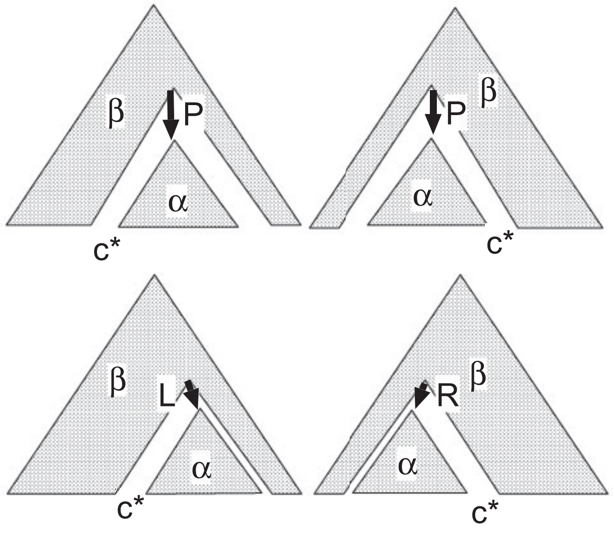

Fig. 2. Schematic diagram representing the four terms that are required for calculating the mutant partition function $Z(x^*)$ [Equation (2)].

$$
\begin{aligned}
Z_1(x^*)= & \sum_{j+1<k\leq N}\sum_{u+v+w=1} \alpha_u(j+1,k-1)\tau_P^v(c^*,x_k)\beta_w(j,k)\\
& + \sum_{1\leq k<j-1}\sum_{u+v+w=1} \alpha_u(k+1,j-1)\tau_P^v(x_k,c^*)\beta_w(k,j)\\
& + \sum_{j<k\leq N}\sum_{u+v+w=1} \alpha_u(j+1,k)\tau_L^v(c^*)\beta_w(j,k)\\
& + \sum_{1\leq k<j}\sum_{u+v+w=1} \alpha_u(k,j-1)\tau_R^v(c^*)\beta_w(k,j)
\end{aligned}
$$

where $u,v,w\in\{0,1\}$, $\alpha_0(i,j)=\alpha(i,j)$, $\beta_0(i,j)=\beta(i,j)$, $\tau_P^v(a,b)=e_P(a,b)\,t(P)[\log e_P(a,b)t(P)]^v$, $\tau_L^v(a)=e_L(a)t(L)[\log e_L(a)t(L)]^v$ and $\tau_R^v(a)=e_R(a)\,t(R)[\log e_R(a)t(R)]^v$. The entropy of the mutated sequence is given by $S(x^*)=-(Z_1(x^*)/Z(x^*))+\log Z(x^*)$. Once the partition function $Z(x^*)$ and entropy $S(x^*)$ are known the ensemble free energy $F(x^*)$ and mean free energy $U(x^*)$ can be computed using Equation (1).

We next consider the computation of the partition function $Z(x^{**})$ for a double mutant x^{**} in which bases i and j $(i<j)$ are replaced with c^{**} and c^*, respectively. To compute $Z(x^{**})$, we simply replace the original sequence x with x^* and compute the inside and outside variables, $\alpha'(i,j)$ and $\beta'(i,j)$, for sequence x^* and apply Equation (2) for the partition function of double mutant x^{**} with the replacements $x^*,j,c^*,\alpha,\beta \rightarrow x^{**},i,c^{**},\alpha',\beta'$. We can economize the calculation of α' and β' by using the fact that parts of the inside and outside variables of sequence x^* are the same as the corresponding variables of sequence x. In Figure 3, the inside and outside variables of x^* are the same as those of x in regions (b–c) and (a), respectively. Furthermore, it is unnecessary to compute the outside variables of x^* in region (b) to compute $Z(x^{**})$ for an arbitrary i $(i<j)$. The computational complexity is $\mathcal{O}(N^4)$ since we have to apply the outside algorithm [which has complexity $\mathcal{O}(N^3)$] to $3N$ single mutations, where N is the input sequence length. Note that there are three possible 1-point mutations for each base. For example, A→C, A→G and A→U are the possible mutations for base A.

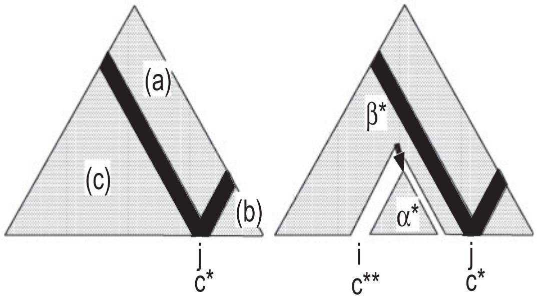

Fig. 3. Schematic illustration of the computation of the partition functions for double-point mutants. Left: Regions of the DP matrix. We compute the inside–outside variables α' and β' after mutating the j-th base. Right: Compute the double mutant partition function using α' and β'.

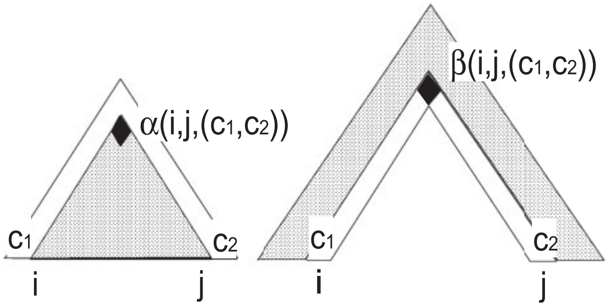

Fig. 4. The inside and outside variables $\alpha(i,j,(c_1,c_2))$ and $\beta(i,j,(c_1,c_2))$ which depend on pairs of bases (c_1,c_2) outside of the parsing subsequences.

2.2 Emission probabilities dependent on neighboring bases

We now consider the case in which the emission probabilities depend on the neighboring bases: $e_P(x_i,x_j|x_{i-1},x_{i+1},x_{j-1},x_{j+1})$, $e_L(x_i|x_{i-1},x_{i+1})$ and $e_R(x_i|x_{i-1},x_{i+1})$. Although such a model does not belong to the category of generative models, it approximates the essential complexities of the energy model. In such a case, the value of the inside variable $\alpha(i,j)$ computed for the original sequence x depends on the bases x_{i-1} and x_{j+1}, unlike in the previous subsection. Similarly, the outside variable $\beta(i,j)$ depends on x_i and x_j. We cannot apply Equation (2) since the mutation at position j alters the values of the neighboring inside and outside variables. It is in principle possible to derive a formula similar to Equation (2) that uses only the inside and outside variables which are two bases away from position j and hence are independent of the mutation of the base at j. However, the corresponding formula would be much more complicated since it would include three layers of DP recursions, corresponding to the emissions of bases x_{j-1}, x_j and x_{j+1}. Instead, we introduce seven copies of the inside variables $\alpha(i,j,(c_1,c_2))$, where (c_1,c_2) represents one of seven types of nucleotide pairs at positions $(i-1,\, j+1)$ that differ by at most one base from the original nucleotide sequence. For example, if $(x_{i-1},\, x_{j+1})$ = (A,G) as in Figure 4, then (c_1,c_2) is one of the set {(A,G),(A,U),(A,A),(A,C),(C,G),(G,G),(U,G)}. Then $\alpha(i,j,(c_1,c_2))$ is defined to be the inside variable on the condition that the bases at $i-1$ and $j+1$ are c_1 and c_2, respectively. Similarly, we introduce seven copies of the outside variables $\beta(i,j,(c_1,c_2))$, where (c_1,c_2) represents one of seven types of nucleotide pairs at positions (i,j) that differ by at most one base from the original bases. Accordingly, $\beta(i,j,(c_1,c_2))$ is defined to be the outside variable on the condition that the bases at i and j are c_1 and c_2, respectively. Once these variables are calculated, we can apply Equation 2 by choosing the appropriate $\alpha(i,j,(c_1,c_2))$ and $\beta(i,j,(c_1,c_2))$ depending on the mutated base.

$$
\begin{aligned}
Z(x^*) = & \sum_{j+1<k\leq N} \alpha(j+1,k-1,(c^*,x_k))e_P(c^*,x_k|j,k)t(P)\beta(j,k,(c^*,x_k)) \\
& + \sum_{1\leq k<j-1} \alpha(k+1,j-1,(x_k,c^*))e_P(x_k,c^*|k,j)t(P)\beta(k,j,(x_k,c^*)) \\
& + \sum_{j<k\leq N} \alpha(j+1,k,(c^*,x_{k+1}))e_L(c^*|j)t(L)\beta(j,k,(c^*,x_k)) \\
& + \sum_{1\leq k<j} \alpha(k,j-1,(x_{k-1},c^*))e_R(c^*|j)t(R)\beta(k,j,(x_k,c^*))
\end{aligned}
$$

where $e_P(a,b|x_{u-1},x_{u+1},x_{v-1},x_{v+1})$, $e_L(a|x_{u-1},x_{u+1})$ and $e_R(a|x_{u-1},x_{u+1})$ are denoted by $e_P(a,b|u,v)$, $e_L(a|u)$ and $e_R(a|u)$, respectively. The variables $\alpha(i,j,(c_1,c_2))$ and $\beta(i,j,(c_1,c_2))$ are computed by the usual inside–outside algorithms with special care taken to match the outside nucleotide pairs (c_1,c_2). For example, the inside algorithm is

$$
\alpha(i,j,(c_1,c_2)) = \sum \begin{cases}
\alpha(i+1,j-1,(x_i,x_j))e_P(x_i,x_j|c_1,x_{i+1},x_{j-1},c_2)t(P) \\
\alpha(i+1,j,(x_i,c_2))e_L(x_i|c_1,x_{i+1})t(L) \\
\alpha(i,j-1,(c_1,x_j))e_R(x_j|x_{j-1},c_2)t(R) \\
\sum_{i\leq k<j}\alpha(i,k,(c_1,x_{k+1}))\alpha(k+1,j,(x_k,c_2))t(B)
\end{cases}
$$

The outside algorithm is similar to the inside algorithm and is shown in the Supplementary Material. The computational complexity of the inside–outside algorithms is simply seven times larger than that of the usual inside–outside algorithms and is given by $\mathcal{O}(N^3)$.

2.3 Application to the energy model

We use the grammatical formulation called the Rfold model, which was introduced in (Kiryu *et al.*, 2008), to derive the mean free energy and the entropy differences for TEM. The Rfold model is based on an unambiguous grammar that can generate all the secondary structures without duplication, and its transition rules are capable of accommodating the score system of TEM. 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 (Kiryu *et al.*, 2008). The model fully uses the advantage provided by the maximal span constraint and computes the inside–outside algorithms in $\mathcal{O}(NW^2)$ time and $\mathcal{O}(N+W^2)$ space (where N is sequence length, and W is the maximal span of base pairs). Although the Rfold model is not an SCFG model, it is usually straightforward to derive DP algorithms analogous to those in SCFG models (Kiryu *et al.*, 2008, 2011; Kiryu and Asai, submitted for publication).

In the inside–outside algorithms of the Rfold model, the dangle and terminal mismatch energies of TEM (Zuker *et al.*, 1999) correspond to dependencies on nucleotides outside of the DP cells. Fortunately, these dependencies do not extend >1 base on either side of the current DP cell. Therefore, we only need six additional DP matrices, just as described in the previous section, to accommodate arbitrary single-point mutations of DP cells that are immediately outside of the current cell. To compute the entropy and mean free energy, we double the DP matrices for computing the log-score inserted tree scores. Consequently, the memory size required for the DP matrices in the case of single-point mutations is 14 times as large as that in the usual inside–outside algorithms, although the complexities are the same $\mathcal{O}(N+W^2)$ (Kiryu *et al.*, 2008). In the case of double-point mutations, we use $\mathcal{O}(N+LW^2)$ memory (where L is the maximal distance of mutated base pairs) to keep track of the mutant inside–outside variables α', β'. The computation times are dominated by the bifurcation calculations and there are $7\times 3=21$ times as many of those as there are in the ordinary inside–outside algorithm. The factor of 7 corresponds the number of nucleotide pairs (c_1,c_2) for computing the inside–outside variables $\alpha(i,j,(c_1,c_2))$ and $\beta(i,j,(c_1,c_2))$, and 3 corresponds to the patterns of bifurcations that mix log-score inserted DP matrices: $\sum_k \alpha_u(i,k)\cdot\alpha_v(k,j)$, $\sum_k \alpha_u(j,k)\cdot\beta_v(i,k)$ and $\sum_k \alpha_u(k,i)\cdot\beta_v(k,j)$ (for $u+v\leq 1$). Despite this, the

Table 1. Run time and memory usage

(*N*, *W*, *L*)	Run time (m:s)	Memory (MB)
(100, 100, 0)	0:8	31
(1k, 100, 0)	3:11	113
(1k, 1k, 0)	33:9	2659
(10k, 100, 0)	34:37	118
(100, 100, 10)	8:12	31
(100, 100, 50)	17:23	31
(100, 100,100)	20:41	31
(1k, 100, 10)	263:1	113
(1k, 100, 50)	625:40	124
(1k, 100, 100)	1061:2	137

N is sequence length, *W* is maximal span and *L* is the maximal distance between two mutated nucleotides. $L=0$ indicates that only single mutations are computed. k=kilo.

computational complexity is still $\mathcal{O}(NW^2)$ for single-point mutations, and $\mathcal{O}(NW^2L)$ for double-point mutations. We have described the full algorithm in the Supplementary Material.

We have implemented the algorithms in C++ in a software package called ‘Rchange’. The source code is freely available from our download site (http://www.ncrna.org/software/rchange/).

3 DATASET AND DATA PROCESSING

We measured the run time and memory usage for random RNA sequences. The experiments were performed on a PC with an Intel Quad Core Xeon E5450 3.0 GHz processor and 32 GB of memory. Random RNA sequences of varying lengths and GC compositions were generated using a Ruby script. The tRNA sequences are taken from the Rfam database version 10 (Griffiths-Jones *et al.*, 2003). We used only the manually curated seed alignment of 967 sequences and its accompanying structure annotation.

We have obtained non-pathogenic polymorphisms and pathogenic mutations from MITOMAP (Ruiz-Pesini *et al.*, 2007). The non-pathogenic polymorphism are considered as mutations from the reference genome, which can be regarded as a collection of dominant alleles (Ruiz-Pesini *et al.*, 2007). The background distributions of $dF/|F|$, $dU/|U|$ and $dS/|S|$ (sample size $n=4142$ for each) are defined by the distributions for all the possible single-point mutations in the 22 mitochondrial tRNA genes except the non-pathogenic polymorphisms ($n=238$) and the pathogenic mutations ($n=209$) of MITOMAP.

In our previous work (Kiryu and Asai, submitted for publication), we have shown that the scales of mean energies $U(x)$ are generally about 10 times higher than $S(x)$. It means $F(x)=U(x)-S(x)$ and $U(x)$ are well correlated. Consistently with this finding, we found $dF=dU-dS$ and dU are very well correlated. Therefore, we only show the figures for dU in the following section. We have shown the comparison between $dF=dU-dS$ and dU in the Supplementary Material.

4 RESULTS AND DISCUSSION

4.1 Comparison of run time and memory usage

Table 1 shows the run time and memory usage of the Rchange program. As described earlier, Rchange is at least 21 times slower than the usual inside–outside algorithms. The table shows that it is difficult to apply Rchange to genome scale sequences. Because of the complexities of the algorithm, the optimization level of our implementation is rather primitive at present. The run time can be improved by devising data structures and orders of computations, though it will require a considerable work. We would like to leave more elaborate optimization to make Rchange a faster program to the future.

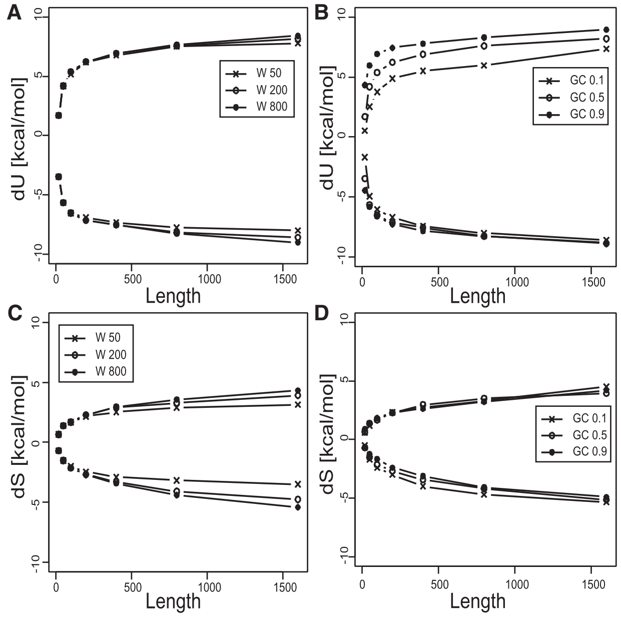

Fig. 5. The upper and lower bounds of the changes in the mean free energies (**A**, **B**) and entropies (**C**, **D**) which are caused by single-point mutations with varying maximal spans *W* (A, C) and GC contents (B, D). (A, C) are computed with $W=400$, whereas (B, D) are computed with GC = 0.5.

4.2 Scales of mean free energy and entropy changes

Figure 5 shows the effects of single-point mutations on the mean free energies *U* and the entropies *S*. For each random sequence generated with a specified length and GC composition, we ran Rchange to obtain the differences *dU* and *dS* for all the single-point mutations and we selected the largest and smallest values from these 3*N* difference values. We plotted the mean values of these upper and lower bounds for 100 random sequences with the same length and GC composition. The upper bound values indicate the largest energy or entropy increase caused by a single mutation; the lower bound indicates the largest energy or entropy decrease caused by a single mutation. These four figures show that the effect of varying GC compositions or maximal spans *W* do not have much influence on the differences. The mean free energy change *dU* (Fig. 5A and B) reaches a bound of 7–9 kcal/mol in magnitude at around $N=200$, while the entropy change *dS* (Fig. 5C and D) gradually increases in magnitude as the sequence length increases. Interestingly, the magnitudes of the entropy changes are about half of those of the mean free energies, which is a much higher ratio than the typical ratio (about one-tenth) for the total entropies and mean free energies (Kiryu and Asai, submitted for publication). This indicates that the entropy changes might affect in the *in vivo* processes of mutations (such as RNA editing) on RNA transcripts. Since the dynamic range of single point mutations are bounded independently of the sequence length, base mutations have considerable impact on the thermodynamics of less stable RNAs. As the thermodynamic variables *F*, *U* and *S* roughly proportional to the sequence length (Kiryu and Asai, submitted for publication), smaller RNAs are

Table 2. Ranked list of mutation types that are frequently observed to cause the maximal differences of the thermodynamic values U and S

Type	Mutation	Enrichment	Mean change
U max increase	G→A	6.2	+6.4
	G→C	3.5	+5.9
	G→U	3.2	+6.1
U max decrease	U→G	4.4	−7.1
	A→G	4.0	−7.3
	A→C	3.4	−7.1
S max increase	G→U	3.1	+2.1
	C→G	2.4	+2.8
	G→A	2.3	+2.1
S max decrease	C→G	2.9	−2.6
	A→G	2.8	−2.8
	U→G	2.5	−3.1

'Enrichment' indicates how often each specific mutation caused the maximal change of U and S, compared with the background frequency. 'Mean change' is the average change of the thermodynamic values in kcal/mol. All the random sequences are of length 200 and are generated with GC = 0.5. The maximal span was set to 400.

more influenced by the mutations. We have analyzed the effects of single point mutations on other Rfam families such as microRNAs (let-7) and snoRNAs (SNORA1, SNORD73) in the Supplementary Material. We have shown that any mutation to the stems of the stable RNAs (let-7 and SNORA1) generally causes a increase of energies whereas less stable RNAs have a large number of stabilizing mutations, implying selective pressures for the secondary structure stability vary considerably from family to family.

Table 2 shows a ranked list of the mutation patterns that were frequently observed to cause the maximal changes of the thermodynamic values U and S in the random sequences. We computed the mutation patterns that caused the maximal changes of U and S for each sequence, and counted the number of occurrences for 100 random sequences. We ranked the patterns by the enrichment—the ratio of the frequency of occurrence of pattern $c_1 \to c_2$ to a simple background frequency given by $\{(\text{total count of } c_1 \text{ in all the sequences}) \times 1/4\}$. The table indicates the maximal increases are achieved by mutating G to the other bases in many cases and the maximal decreases of the thermodynamic values are achieved by mutating a base into G. We have shown nucleotide dependencies and structural preferences of the thermodynamic change distributions in the Supplementary Material.

4.3 Position dependence of mean free energy changes

Figure 6 shows the position dependence of dU. For each position of random sequences of lengths 20 (A, C) and 200 (B, D), we obtain the upper and lower bound for energy changes. Figure 6A and B show the mean values of these upper and lower bounds for 10 000 random samples. Figure 6C and D show the standard deviations (SDs) of dU for each position calculated from the 10 000 random sequences. For shorter sequences (A, C), we observe two peaks around positions 5 and 15 which indicate the dominance of single hairpin structures over other structural configurations in short sequences. This position dependence rapidly disappears as sequence length increases, and we find no position dependence except the very small values of dUs at the ends of sequences (B, D). Figure 6B and D show that mean values are ∼ 1–2 kcal/mol whereas the SD is ∼ 2 kcal/mol, indicating the strong sequence dependence of the energy changes.

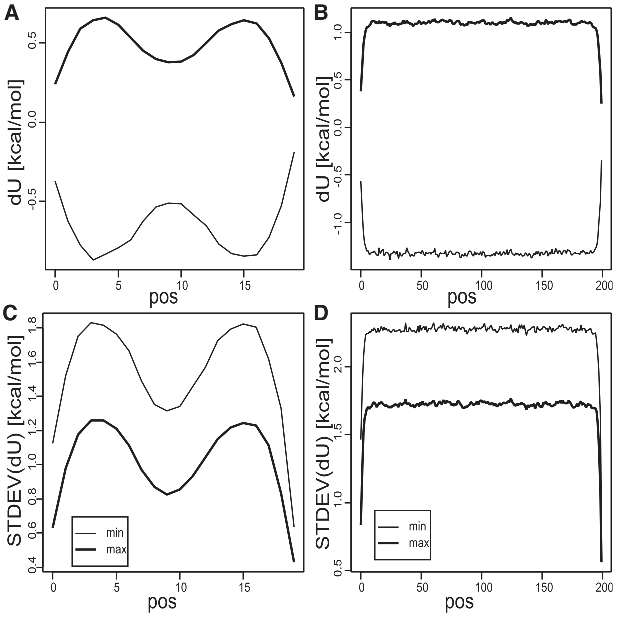

Fig. 6. The position dependence of dU for random sequences. The figures show the means (**A**, **B**) and the SDs (**C**, **D**) of the upper bound (thick lines) and lower bound (thin lines) at each position for random RNA sequences of length 20 (A, C) and 200 (B, D). The random sequences were generated with GC = 0.5. The maximal span W was set to 400.

4.4 Effects of single-point mutations on tRNA

Figure 7 shows the position dependence of the upper and lower bounds of the relative changes $dU/|U|$ and dS/S for single-point mutations of tRNA sequences. For each position of each tRNA sequence in the Rfam seed alignment, we calculated the upper bound and bound of $dU/|U|$ and dS/S. The average values of these bounds are plotted for all the seed sequences using the alignment information. We removed alignment columns with many gaps (> 30% of the number of the sequences). The figure shows that mutations in the stem regions can destabilize the structure more than those in non-stem regions. Mutations in the acceptor stem (the closing stem) influence the increase of $dU/|U|$ the most, and a change of >20% of the mean free energy can be produced by only a single mutation. On the other hand, the increase of dS/S is most affected by the T arm (the right most stem aside from the acceptor stem) on average, which might indicate that an increased probability of disruption of the T arm stem may lead to a large fluctuation of these regions, including the variable loop between the anticodon stem and the T arm stem; however, the differences from the peaks of the other stems are small. The plots of the lower bounds show sharp peaks at the unpaired bases that are next to the stem regions, which indicates that changes that extend the stem length stabilize the secondary structure the most. The upper figure shows that the secondary structures are most stabilized by changing the 5′ end of the anticodon loop (the loop in the center stem). This figure also shows that any changes in the acceptor stem increase

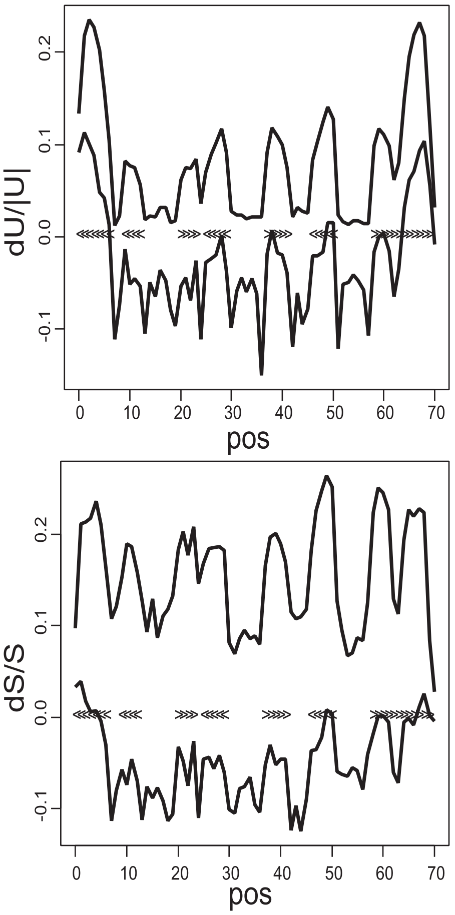

Fig. 7. The position dependence of the upper bound and lower bound of the relative changes $dU/|U|$ and dS/S for single-point mutations of tRNA sequences. We have averaged the upper bounds and the lower bounds over the 967 tRNA sequences using the seed alignment of the Rfam database. The brackets '<' and '>' shown at $y=0.0$ represent the common secondary structure of the tRNA sequences.

the mean free energy, which indicates that the sequences of the acceptor stem may be strongly optimized for secondary structure stability.

4.5 Effects of double-point mutations on tRNA

Figure 8 shows the position dependence of the upper and lower bounds of the relative changes $dU/|U|$ and dS/S for double-point mutants of tRNA sequences. The upper right triangles show that the largest increases of $dU/|U|$ and dS/S can be obtained by mutating nucleotide pairs that are both located in the stem regions. The destabilizing effects on U are larger if one mutating base is on the acceptor stem than if both nucleotides are on mean free stems, while the influence on entropies S is similar regardless of which stem regions the pairs are located in. The lower bound of $dU/|U|$ (the lower left triangle in Fig. 8A) shows that the double mutation of a base at a stem boundary and a base in a loop region can cause $>30\%$ decrease of the mean energy. On the other hand, most of the double mutations that include any base in the acceptor stem increase dU. There are no $dU>0$ regions (or green regions in the lower left triangles) inside the acceptor stem, even at the intersections of different stem regions in which any double mutations are expected to destabilize both of the stem structures. This indicates that the D arm, anticodon and T arm stems are not very stable compared with the acceptor stem and there are considerable fluctuations in their secondary structures. At present, it is not clear whether this represents biophysical properties of tRNAs or is simply a limitation of TEM.

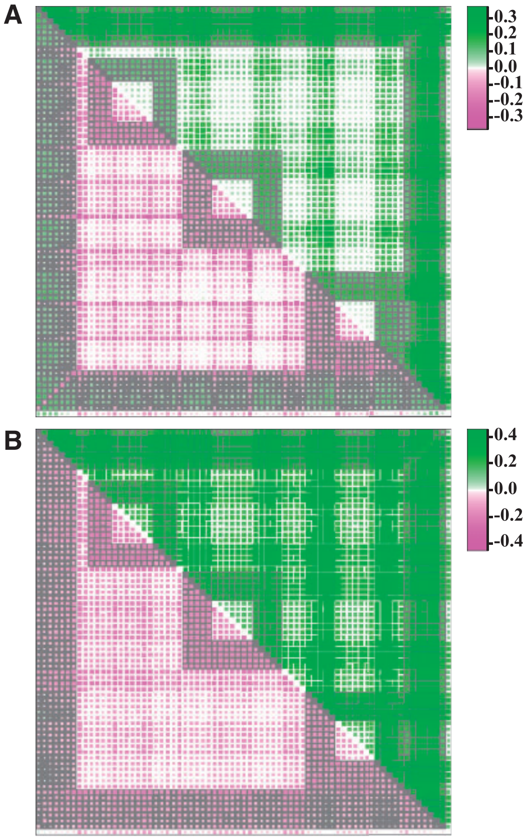

Fig. 8. These colored figures show the position dependence of the upper and lower bound of the relative changes $dU/|U|$ (**A**) and dS/S (**B**) for double-point mutations of tRNA sequences. For each density plot, the upper right (lower left) triangle shows the density plot of the mean upper (lower) bound of $dU/|U|$ or dS/S for all the non-trivial double-point mutations. We have averaged the upper bounds and the lower bounds over the 967 tRNA sequences using the seed alignment of the Rfam database. The positive and negative changes are displayed using green and magenta colors, respectively. The gray colored background indicates the annotated stem regions of tRNAs.

4.6 Mutations in human mitochondrial tRNAs

The human mitochondrial genome involves 22 tRNA genes. Mitochondrial DNA is known to have about 10 times higher mutation rate than nuclear DNA, and the mutations in the tRNA genes cause various mitochondrial diseases such as myopathy, deafness and Alzheimer's disease (Zifa *et al.*, 2007). We compared the distributions of $dF/|F|$, $dU/|U|$ and dS/S for non-pathogenic polymorphisms and pathogenic mutations with the background distribution. Figure 9 shows the dS/S distribution for three mutation types. It shows that non-pathogenic polymorphisms are depleted

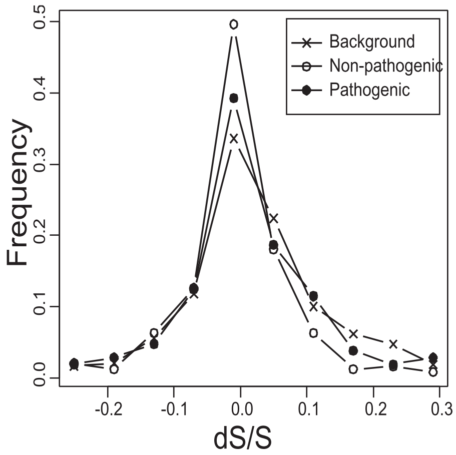

Fig. 9. dS/S distributions of mitochondrial tRNA genes. Each distribution is normalized to sum to one. Crosses, the background distribution; open circles, non-pathogenic mutations; filled circles, pathogenic mutations.

Table 3. P-values of Wilcoxon rank sum tests.

	Non-pathogenic		Pathogenic	
Variable	Lowered	Centered	Lowered	Centered
$dF/\|F\|$	0.017	5.8×10^{-5}	0.48	0.034
$dU/\|U\|$	1.6×10^{-5}	1.1×10^{-6}	0.090	0.059
dS/S	6.2×10^{-7}	2.0×10^{-8}	0.066	0.031

'Lowered' tests whether the $dX/|X|$ distribution has a smaller median and 'Centered' tests whether $|dX/X|$ distribution has a smaller median, than the background distribution ($X=F,U,S$).

from the large dS/S region and concentrated around $dS/S=0$ as compared with the background mutations. Table 3 shows the P-values that test the biases of the distributions. We do not find any significant bias for the pathogenic mutations. It is reasonable since the seriousness of symptoms vary widely from disease to disease and the carriers of pathogenic mutations are still viable with all their sickness as opposed to the cases of lethal mutations, which may obscure the relationships between the tRNA functions and the pathogenic mutations. On the other hand, the non-pathogenic polymorphisms tend to cause smaller changes, suggesting that a mutation that largely increases either dS/S, $dU/|U|$ or $dF/|F|$ tends to be a pathogenic or lethal mutation. It is interesting that entropy change dS/S has greater significance than the energy changes $dU/|U|$ and $dF/|F|$, which might indicates that the structural fluctuations are more related to tRNA dysfunction than the energetic instabilities.

5 CONCLUSION

In this article, we have developed algorithms for computing changes in the ensemble free energy, mean free energy and the thermodynamic entropy of RNA secondary structures for all possible patterns of single and double mutations. The computational complexities are $\mathcal{O}(NW^2)$ (where N is sequence length and W is maximal base pair span) for single mutations and $\mathcal{O}(N^2W^2)$ for double mutations with relatively large constant factors. We have shown that the changes in the mean free energy and entropy are relatively insensitive to GC composition and the maximal span constraint. The mean energy changes are bounded at ∼7–9 kcal/mol in magnitude and there is only a small position dependence if sequence lengths are sufficiently large. For tRNA sequences, the most stabilizing mutations come from changes of the $5'$-most base of the anticodon loop. We have also shown that most base changes in the acceptor stem destabilize the structures, indicating that the nucleotide sequence in the acceptor stem is highly optimized for secondary structure stability. We have analyzed the effects of single point mutations on other Rfam families such as microRNAs (let-7) and snoRNAs (SNORA1, SNORD73) in the Supplementary Material, where we have shown that any mutation to the stems of stable RNAs generally causes a increase of energies whereas less stable RNAs have stabilizing mutations in most regions, implying selective pressures for the secondary structure stability vary considerably from family to family. We have investigated human mitochondrial tRNAs and found that non-pathogenic polymorphisms tend to cause smaller changes in thermodynamic variables as compared with a background distribution, suggesting that a mutation which largely increases thermodynamic variables has higher possibility to be a pathogenic or lethal mutation.

There are several applications of our algorithms. As described in the Introduction, our algorithms can be directly used in a mutation analysis to design the most stabilizing and destabilizing mutations. They will also be useful for investigating the correlations between the evolutionary variations of RNA genes in genome sequences and the changes in their structures. The algorithms can also be of interest for investigating the disease associations of SNPs located in non-coding RNA genes. Although we only considered mutations within the standard four bases 'ACGU' in this article, similar techniques will be used for the structural analysis of A-to-I RNA editing if there exists a complete set of nearest-neighbor parameters for the inosine residue [see Watkins and SantaLucia (2005) for DNA case].

ACKNOWLEDGEMENTS

The authors thank their colleagues who worked on the project for discussions and comments.

Funding: This study was supported by the 'Functional RNA Project' funded by the New Energy and Industrial Technology Development Organization (NEDO) of Japan, a Grant-in-Aid for Young Scientists (21700330) and a Grant-in-Aid for Scientific Research (A) (22240031). Computations were performed using the supercomputing facilities at the Human Genome Center, University of Tokyo (http://sc.hgc.jp/shirokane.html).

Conflict of Interest: none declared.

REFERENCES

Barash,D. and Churkin,A. (2011) Mutational analysis in RNAs: comparing programs for RNA deleterious mutation prediction. *Brief. Bioinform.*, **12**, 104–114.

Churkin,A. and Barash,D. (2006) RNAmute: RNA secondary structure mutation analysis tool. *BMC Bioinformatics*, **7**, 221.

Churkin,A. and Barash,D. (2008) An efficient method for the prediction of deleterious multiple-point mutations in the secondary structure of RNAs using suboptimal folding solutions. *BMC Bioinformatics*, **9**, 222.

Clote,P. *et al*. (2005) Energy landscape of k-point mutants of an RNA molecule. *Bioinformatics*, **21**, 4140–4147.

Doshi,K.J. *et al*. (2004) Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinformatics*, **5**, 105.

Griffiths-Jones,S. *et al*. (2003) Rfam: an RNA family database. *Nucleic Acids Res.*, **31**, 439–441.

Halvorsen,M. *et al*. (2010) Disease-associated mutations that alter the RNA structural ensemble. *PLoS Genet.*, **6**, e1001074.

Hofacker,I. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.

International HapMap Consortium. (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.

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

Kiryu,H. *et al*. (2011) A detailed investigation of accessibilities around target sites of siRNAs and miRNAs. *Bioinformatics*, **27**, 1788–1797.

Knudsen,B. and Hein,J. (2003) Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res.*, **31**, 3423–3428.

Manolio,T.A. (2010) Genomewide association studies and assessment of the risk of disease. *N. Engl. J. Med.*, **363**, 166–176.

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

Miklos,I. *et al*. (2005) Moments of the Boltzmann distribution for RNA secondary structures. *Bull. Math. Biol.*, **67**, 1031–1047.

Pedersen,J. *et al*. (2004) A comparative method for finding and folding RNA secondary structures within protein-coding regions. *Nucleic Acids Res.*, **32**, 4925–4936.

Rivas,E. and Eddy,S. (2001) Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, **2**, 8.

Ruiz-Pesini,E. *et al*. (2007) An enhanced MITOMAP with a global mtDNA mutational phylogeny. *Nucleic Acids Res.*, **35**, D823–D828.

Shu,W. *et al*. (2006) RDMAS: a web server for RNA deleterious mutation analysis. *BMC Bioinformatics*, **7**, 404.

Waldispuhl,J. *et al*. (2008) Efficient algorithms for probing the RNA mutation landscape. *PLoS Comput. Biol.*, **4**, e1000124.

Waldispuhl,J. *et al*. (2009) RNAmutants: a web server to explore the mutational landscape of RNA secondary structures. *Nucleic Acids Res.*, **37**, W281–W286.

Warf,M.B. *et al*. (2009) The protein factors MBNL1 and U2AF65 bind alternative RNA structures to regulate splicing. *Proc. Natl Acad. Sci. USA*, **106**, 9203–9208.

Washietl,S. *et al*. (2005) Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat. Biotechnol.*, **23**, 1383–1390.

Watkins,N.E. and SantaLucia,J. (2005) Nearest-neighbor thermodynamics of deoxyinosine pairs in DNA duplexes. *Nucleic Acids Res.*, **33**, 6258–6267.

Zifa,E. *et al*. (2007) Mitochondrial tRNA mutations: clinical and functional perturbations. *RNA Biol.*, **4**, 38–66.

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, Dordrecht, The Netherlands, pp. 11–43.