Hagio *et al. BMC Bioinformatics* 2018, **19**(Suppl 1):38
https://doi.org/10.1186/s12859-018-2018-4

BMC Bioinformatics

RESEARCH Open Access

Capturing alternative secondary structures of RNA by decomposition of base-pairing probabilities

Taichi Hagio[1], Shun Sakuraba[1], Junichi Iwakiri[1], Ryota Mori[3] and Kiyoshi Asai[1,2*]

From Proceedings of the 28th International Conference on Genome Informatics: bioinformatics
Seoul, Korea. 31 October - 3 November 2017

Abstract

Background: It is known that functional RNAs often switch their functions by forming different secondary structures. Popular tools for RNA secondary structures prediction, however, predict the single 'best' structures, and do not produce alternative structures. There are bioinformatics tools to predict suboptimal structures, but it is difficult to detect which alternative secondary structures are essential.

Results: We proposed a new computational method to detect essential alternative secondary structures from RNA sequences by decomposing the base-pairing probability matrix. The decomposition is calculated by a newly implemented software tool, RintW, which efficiently computes the base-pairing probability distributions over the Hamming distance from arbitrary reference secondary structures. The proposed approach has been demonstrated on ROSE element RNA thermometer sequence and Lysine RNA ribo-switch, showing that the proposed approach captures conformational changes in secondary structures.

Conclusions: We have shown that alternative secondary structures are captured by decomposing base-paring probabilities over Hamming distance. Source code is available from http://www.ncRNA.org/RintW.

Keywords: RNA secondary structure, Dynamic programming, Base-pairing probability, Partition function

Background

Secondary structures of RNA are known to be thermodynamically fluctuated and the number of possible secondary structures of RNA are huge. We can predict the secondary structure by software tools, but on Boltzmann distribution of the secondary structure the probability of the 'best' secondary structure predicted is usually very small [1]. For example, the probability of the canonical 'clover leaf' secondary structure of a tRNAs is often less than one percent. One the other hand, the marginal probabilities on each base of the secondary structural contexts, such as base-pairs, loops, bulges, multi loops, are not necessarily very small and carry important structural information [2]. Among them, the **base-pairing probabilities (BPPs)** [3], which are often greater than eighty percent, are convenient to observe the local stability of the secondary structures. Furthermore, the estimators based on maximum expected accuracy, such as the MEA (Maximum Expected Accuracy) estimator [4] and the γ-centroid estimator [5], can be calculated by posterior decoding of BPPs without using further information of the RNA sequence [6].

Several functional RNAs, such as RNA thermometers and ribo-switches, change their functions by forming different secondary structures. It is difficult, however, to detect such structural changes. There are bioinformatics tools to predict suboptimal structures [7], but it is difficult to detect which alternative secondary structures are essential.

*Correspondence: asai@k.u-tokyo.ac.jp
[1] Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, University of Tokyo, 5-1-5 Kashiwanoha, 277-8561 Kashiwa, Japan
[2] Artificial Intelligence Research Center (AIRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-3-26 Aomi, Koto-ku, 136-0064 Tokyo, Japan
Full list of author information is available at the end of the article

When the structural change occurs, the energy landscape, equivalently the probability distribution of the secondary structures, also changes, but we may expect that the information of the alternative secondary structures is included in the original probability distribution. For example, the two alternative structures of an aptamer verified by structure-specific RNase experiments were supported by two competing potential stems in the BPPs [8]. In order to characterize the alternative secondary structures more clearly, the other types of marginal probabilities, the distributions over Hamming distance are useful. The general idea has been proposed by [9] as an algorithm for exact calculation of distributions on integers applied to sequence alignment. For RNA secondary structures, algorithms and tools that calculate the probability distribution of the secondary structures over the Hamming distance from the reference structure has been proposed (RNAbor [10, 11], RintD [12]). If the distribution is concentrated near to the reference structure, the structure is regarded as reliable (i.e. low credibility limit). If there are multiple clusters of secondary structures, the distribution should have multiple peaks (See "Results" section). If the second reference structure is appropriately selected, we can obtain informative 2D distribution over Hamming distances from the two reference structures using RNA2Dfold [7] or RintD [12] (See "Results"section).

On the secondary structure over Hamming distance, a method to calculate approximate MEA estimators over Hamming distance has been introduced [13]. We propose a method to calculate exact decomposition of BPPs over each range of the Hamming distance and the exact MEA-based estimators (MEA estimator and γ-centroid estimator) by posterior decoding of decomposed BPPs as the representative secondary structure of the range of the Hamming distance. The existence of the cluster is visualized by the distribution over the Hamming distances from the two reference structures.

In order to establish the method to decompose the BPPs of a whole RNA sequence, we have developed RintW, a new computational tool that efficiently compute the exact base-pairing probability distribution over the Hamming distance from the reference structure. RintW is an extension of the calculation of partition function over Hamming distance in RintD [12]. RintD efficiently computes the secondary structure probability distribution over the Hamming distance from an arbitrary reference structure, by applying polynomial approach and Discrete Fourier Transform (DFT) to McCaskill's algorithm [3]. In order to apply their approach to the base-pairing probabilities over Hamming distance, we derived outside algorithm for outside partition function from McCaskill's algorithm of base-pairing probabilities [3]. RintW runs computational complexities of $O\left(L^4 H_{\max}\right)$ in time and $O\left(L^3\right)$ in space, where L is the sequence length and $H_{\max}$ $(< L)$ is the maximum values of the Hamming distance from an arbitrary structure.

Methods

In this paper, we present a new computational method to detect the essential alternative structures from RNA sequences. Figure 1 shows an overview of the method. Firstly, we calculate an estimate of the secondary structure of the given RNA sequence, as the reference secondary structure. Secondly, the probability distribution of the secondary structure over the Hamming distance from the reference secondary structure is calculated by RintD [12] (Fig. 1 left). If there are multiple peaks in the distribution, which does not guarantee but implies that there are multiple clusters of secondary structures, we detect the ranges of Hamming distance for each potential cluster (Fig. 1 left, A and B). Thirdly, by newly implemented RintW, the base-paring probability matrix over each range of Hamming distance is calculated (Fig. 1 middle). Finally, the representative secondary structure for each cluster, which will be used as one of the reference structures for 2D analysis over Hamming distances, is estimated by posterior decoding of the corresponding base-pairing matrix (Fig. 1 right).

1D analysis over Hamming distance

Reference secondary structure selection

In the proposed method, the reference secondary structure is the structure estimated by CentroidFold [5]. Alternatively, of course, any reliable prediction by another tool, such as the minimum free energy structure by Mfold [14] can be used.

DP for secondary structure probability over Hamming distance

The distribution of secondary structures over Hamming distance from the reference strucure is calculated by RintD [12], while there are several tools for similar calculation (RNAbor [10, 11], RNA2Dfold [7]).

The Hamming distance between two secondary structures of an RNA sequence of the length L is defined as the Hamming distance of the upper triangle binary matrices $\sigma_{i,j}$ as follows:
for $1 \leq i < j \leq L$,

$$\sigma_{i,j} = \left[\begin{array}{l} 1;\ \text{a base-pair } (i,j) \text{ exists in the secondary structure} \\ 0;\ \text{otherwise} \end{array} \right. \tag{1}$$

The probability distribution over Hamming distance, the marginal probability that the Hamming distance of the secondary structure is d from the reference structure σ, is written as follows:

$$p(d, \sigma) = \frac{Z_{1,L}(d, \sigma)}{Z_{1,L}}, \tag{2}$$

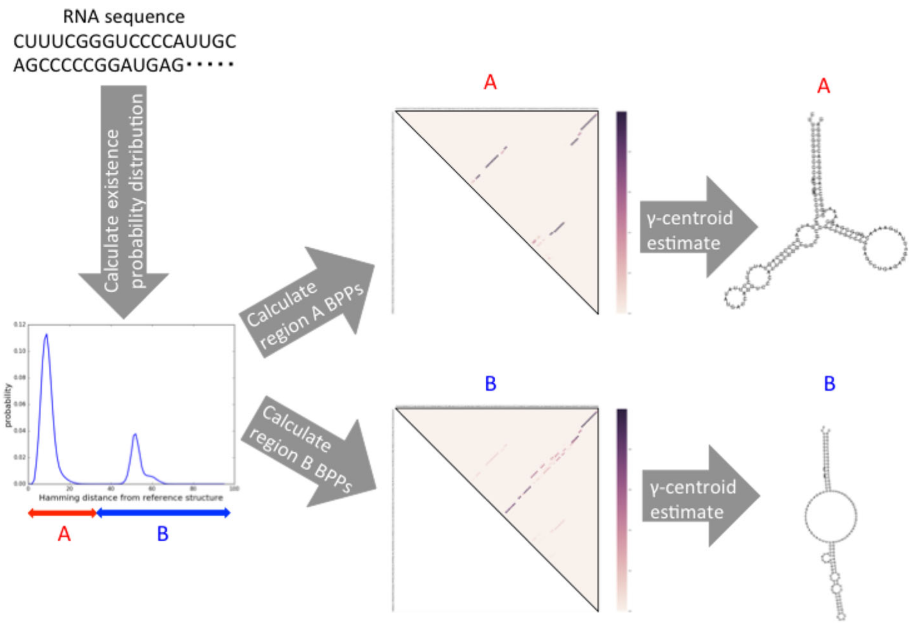

Fig. 1 An overview of capturing the alternative secondary structures. The distribution of the secondary structures over the Hamming distance from the reference structure is calculated (left). The base-pairing probability matrix is calculated for each peak of the distribution (middle). The γ-centroid estimater of the secondary structure is calculated from each base-pairing probability matrix (right)

where $Z_{1,L}$ is the partition function of all the secondary structures, and $Z_{1,L}(d,\sigma)$ is the partition function over Hamming distance d from the reference structure σ, which is the sum of all the Boltzmann factors of secondary structures whose Hamming distance from the reference structure σ is d. The partition function $Z_{1,L}$ is calculated by McCaskill's algorithm [3]. The partition function over Hamming distance, $Z_{1,L}(d,\sigma)$, can be calculated by a dynamic programming (DP) adding Hamming distance as an additional dimension of DP matrices to McCaskill's algorithm. For practical implementation, however, the computational cost of this DP is too high. In the efficient calculation of RintD, this DP was mapped to a DP on polynomials and was converted to Discrete Fourier Transformation (DFT).

The DP on polynomials in RintD [12] is shown in Algorithm 1. In this algorithm, DP matrices, $Z_{i,j}$, $Z^1_{i,j}$, $Z^b_{i,j}$, $Z^m_{i,j}$, $Z^{m1}_{i,j}$, are polynomials, whose terms store the sum of Boltzmann factors of subsequence $[i,j]$ as the coefficients and the Hamming distance from the reference structure σ as the power of the dummy variable x. $Z_{i,j}$ corresponds to the general case, $Z^1_{i,j}$ has exactly one outmost base pair whose 5′ base is i, $Z^b_{i,j}$ is the partition function conditioned by (i,j) base-pairing, $Z^m_{i,j}$ and $Z^{m1}_{i,j}$ correspond to the multi loops. The functions f_1 to f_5 are the free energy of unpaired bases depending on the structural contexts and the lengths of the subsequences.

The functions g^Z_1 to g^Z_8 are the gains of the Hamming distance for the transitions from the term in the right

Algorithm 1 DP on polynomials of the partition function over Hamming distance [12]

The DP variables, Zs, are polynomials of x.
Zs and the gain functions g^Z_1 to g^Z_8 are the functions of the reference structure σ.

Initialization:

for $1 \le i \le L$

$$Z_{i,i} = 1.0$$

$$Z^1_{i,i} = Z^b_{i,i} = Z^m_{i,i} = Z^m_{i,i-1} = Z^{m1}_{i,i} = 0.0$$

Recursion:

for $1 \le i < j \le L$

$$\begin{aligned}
Z_{i,j} &= x^{g^Z_1(i,j)} + \sum_{h=i}^{j-1} Z_{i,h} Z^1_{h+1,j} x^{g^Z_2(i,j,h)} \\
Z^1_{i,j} &= \sum_{h=i+1}^{j} Z^b_{i,h} x^{g^Z_3(i,j,h)} \\
Z^b_{i,j} &= e^{-f_1(i,j)/kT} x^{g^Z_4(i,j)} \\
&\quad + \sum_{h=i+1}^{j-2} \sum_{\ell=h}^{j-1} Z^b_{h,\ell} e^{-f_2(i,h,\ell,j)/kT} x^{g^Z_5(i,h,\ell,j)} \\
&\quad + \sum_{h=i+2}^{j-1} Z^m_{i+1,h-1} Z^{m1}_{h,j-1} e^{-f_3(i,j)/kT} x^{g^Z_6(i,j,h)} \\
Z^m_{i,j} &= \sum_{h=i}^{j-1} \left[\begin{array}{l} e^{-f_4(i,h-1)/kT} x^{g^Z_7(i,j,h)} \\ +Z^m_{i,h-1} x^{g^Z_8(i,j,h)} \end{array} \right] Z^{m1}_{h,j} \\
Z^{m1}_{i,j} &= \sum_{h=i+1}^{j} Z^b_{i,h} e^{-f_4(h+1,j)/kT} x^{g^Z_3(i,j,h)}
\end{aligned} \tag{3}$$

hand side of the equation to the left hand side Z's, which mostly correspond to the number of the base-pairs in the reference structure where base-pair should not exist in the specific structural contexts. Using the following definitions

$$g_0(i,j,h) = \sum_{p=i}^{h} \sum_{q=h+1}^{j} \sigma_{p,q}, \qquad g_0^Z(i,j) = \sum_{p=i}^{j-1} \sum_{q=p+1}^{j} \sigma_{p,q}, \tag{4}$$

the gain functions are described as follows:

$$\begin{aligned} g_1^Z(i,j) &= g_0^Z(i,j), \\ g_2^Z(i,j,h) &= g_0(i,j,h), \\ g_3^Z(i,j,h) &= g_0(i,j,h+1) + g_0^Z(h+1,j), \\ g_4^Z(i,j) &= g_0^Z(i,j) + 1 - 2\sigma_{p,q}, \\ g_5^Z(i,j,h,\ell) &= g_0^Z(i,j) - g_0^Z(h,\ell) + 1 - 2\sigma_{p,q}, \end{aligned}$$

$$\begin{aligned} g_6^Z(i,j,h) &= g_0^Z(i,j) - g_0^Z(i+1,h-1) \\ &\quad - g_0^Z(h,j-1) + 1 - 2\sigma_{p,q}, \\ g_7^Z(i,j,h) &= g_0^Z(i,h-1) + g_0(i,j,h), \\ g_8^Z(i,j,h) &= g_0(i,j,h-1). \end{aligned}$$

$Z(d,\sigma)$, the partition function over the Hamming distance, is obtained as the coefficients of the polynomials as follows:

$$Z_{1,L} = \sum_{d=d_{\min}}^{d_{\max}} Z(d,\sigma) x^d, \tag{5}$$

where $d_{\max}$ is the maximum Hamming distance from the reference structure, which is no greater than the length of the sequence, and $d_{\min}$ is the minimum, which is usually 0.

Accelerated calculation by discrete fourier transformation

Algorithm 1 includes the multiplication of the polynomials, which is computationally expensive. According to [9], distributed processing by DFT is available for dynamic programing of distribution on integer function. This acceleration has been implemented in RintD [12] for the partition function.

The DP on polynomials in Algorithm 1 can be converted to DP on complex numbers by substituting

$$x = \exp\left[2\pi i \frac{r}{\Delta}\right], \tag{6}$$

where $\Delta = d_{\max} - d_{\min} + 1$, then the DP matrices of the partition function over Hamming distance becomes matrices of complex numbers:

$$\tilde{Z}_{i,j}(\sigma) = \sum_{d=d_{\min}}^{d_{\max}} Z_{i,j}(d,\sigma) \exp\left[2\pi i \frac{rd}{\Delta}\right]. \tag{7}$$

In the DFT approach, the partition function over Hamming distance is rewritten as follows:

$$\begin{aligned} Z(d,\sigma) &= \sum_{s=d_{\min}}^{d_{\max}} Z(s,\sigma)\delta_{sd} \\ &= \sum_{s=d_{\min}}^{d_{\max}} Z(s,\sigma) \sum_{r=d_{\min}}^{d_{\max}} \frac{\exp\left[2\pi i \frac{r(s-d)}{\Delta}\right]}{\Delta} \\ &= \frac{1}{\Delta} \sum_{r=d_{\min}}^{d_{\max}} \exp\left[2\pi i \frac{-rd}{\Delta}\right] \sum_{s=d_{\min}}^{d_{\max}} Z(s,\sigma) \exp\left[2\pi i \frac{rs}{\Delta}\right] \\ &= \frac{1}{\Delta} \sum_{r=d_{\min}}^{d_{\max}} \exp\left[2\pi i \frac{-rd}{\Delta}\right] \tilde{Z}_{1,L}(\sigma). \end{aligned} \tag{8}$$

In (8), $\tilde{Z}_{1,L}(\sigma)$ can be computed as a complex number by Algorithm 1, substituting $x = \exp\left[2\pi i(r/(\Delta))\right]$ as shown in (7), and the entire (8) can be computed by DFT. Note that the above DP and DFT are computable in parallel.

Range detection

If the distribution over Hamming distance is concentrated around the reference structure, the structure is reliable (i.e. low credibility limit). If the distribution of Hamming distance has multiple peaks (as in Fig. 1 left), we detect the range of Hamming distance for each peak. It should be noted that a peak is generally not a guaranteed structural cluster but a potential candidate of a cluster, because the structures within a range of the Hamming distance from the reference structure may have mutual Hamming distance up to the double of the maximum Hamming distance of the range. On the contrary, a cluster whose members have small mutual Hamming distances is always observed as a peak in the distribution over Hamming distance. The first peak, however, tends to be a real cluster if the maximum Hamming distance of the range is reasonably small.

Decomposition of base-pairing probability

DP for base-paring probabilities over Hamming distance

The base-pairing probability matrix over Hamming distance, $P^b_{ij}(d,\sigma)$, which is the marginal probability that the secondary structure has Hamming distance d from the reference structure σ and that the i-th base and the j-th base form a base-pair in the secondary structure, can be calculated by a DP algorithm adding Hamming distance as an additional dimension to the DP algorithm for base-pairing probability [3]. Because direct calculation of this DP is computationally impractical, we have developed a polynomial-DFT approach similar to RintD.

The original DP algorithm for base-pairing probabilities [3], however, included divisions of DP matrices, which is inappropriate for the polynomial-DFT approach. In order to avoid this problem, we rewrote the base-pairing probability over Hamming distance, as follows

$$P^b_{ij}(d,\sigma) = \sum_t \frac{Z^b_{ij}(d-t)\,W^b_{ij}(t)}{Z(d,\sigma)}. \tag{9}$$

$Z^b_{ij}(d-t)$ and $Z(d,\sigma)$ are obtained by Algorithm 1. The $Z^b_{ij}(d-t)$ is the inside partition function over Hamming distance, which is defined as the sum of all the Boltzmann factors of the structures of the subsequence $[i,j]$ when (i,j) is a base-pair and the Hamming distance of subsequence $[i,j]$ is $d-t$. $Z(d,\sigma)$ is the partition function of the whole RNA sequence over Hamming distance, which is defined as the sum of all the Boltzmann factors of the structures of the whole sequence having Hamming distance d. $W^b_{ij}(t,\sigma)$ is the outside partition function over Hamming distance, which is defined as the sum of all the Boltzmann factors of the structures outside of the (i,j) base-pair when the Hamming distance outside of $\left[i,j\right]$ is t. The algorithm for outside partition function is given in the next subsection.

The base-pairing probability matrix $P^b\left[r_{\min},r_{\max}\right]$ on the range $[r_{\min},r_{\max}]$ of Hamming distance, is obtained by integrating each Boltzmann factor in (9) over the range of Hamming distance as follows:

$$P^b_{ij}\left[r_{\min},r_{\max}\right] = \frac{\sum_{d\in[r_{\min},r_{\max}]}\sum_t Z^b_{ij}(d-t)\,W^b_{ij}(t)}{\sum_{d\in[r_{\min},r_{\max}]} Z(d,\sigma)}. \tag{10}$$

Outside partition function over Hamming distance

In order to calculate the base-pairing probability over Hamming distance by (9), we apply the polynomial-DFT approach to the outside partition function over Hamming distance. The dynamic programming on polynomials of the outside partition function is described in Algorithm 2. This algorithm corresponds to the outside algorithm of the stochastic context free grammar (SCFG), and iterations are computed from long (i,j) to short (i,j), while McCaskill's algorithm of the partition function corresponds to the inside algorithm of SCFG and runs from short (i,j) to long (i,j).

Algorithm 2 DP on polynomials of the outside partition function over Hamming distance

The DP variables, Zs and Ws are polynomials of x.
Zs, Ws, and the gain functions g_1 to g_5 are the functions of the reference structure σ.

Initialization:

$$W^b_{1,L} = \left[\begin{array}{ll} 1.0 & \text{if } (1,L) \text{ can form a base-pair} \\ 0.0 & \text{otherwise} \end{array}\right.$$

Recursion:
for $1 \le i < j \le L$

$$W^b_{i,j} = Z_{1,i-1}Z_{j+1,L}\,x^{g_1^{\mathrm{W}}(i,j)} \tag{11}$$

$$+\sum_{\substack{h<i\\ \ell>j}} W^b_{h,\ell}\,e^{-f_2(h,i,j,\ell)/kT}\,x^{g_2^{\mathrm{W}}(h,i,j,\ell)} \tag{12}$$

$$+\sum_{\substack{h<i\\ \ell>j}} W^b_{h,\ell}\,e^{-f_3/kT}\left[\begin{array}{l} Z^m_{h+1,i-1}\,e^{-f_4(j,\ell)/kT}\,x^{g_3^{\mathrm{W}}(h,i,j,\ell)} \\ +Z^m_{j+1,\ell-1}\,e^{-f_4(h,i)/kT}\,x^{g_4^{\mathrm{W}}(h,i,j,\ell)} \\ +Z^m_{h+1,i-1}Z^m_{j+1,\ell-1}\,x^{g_5^{\mathrm{W}}(h,i,j,\ell)} \end{array}\right]. \tag{13}$$

In Algorithm 2, (11) represents the case that (i,j) base-pair is not included in any other base-pair, (12) represents the case of (i,j) base-pair being included in another base-pair (h,ℓ), while no base-pair in the subsequence (h,i) or (j,ℓ). In (13), (i,j) base-pair is included in another base-pair (h,ℓ), while at least one base-pair only in the subsequence (h,i) in the first line, at least one base-pair only in the subsequence (j,ℓ) in the second line, and at least one base-pair in both of the subsequences (h,i) and (j,ℓ) in the third line. This outside algorithm requires the partition functions Z and Z^m of Algorithm 1. $Z_{i,j}$ is the partition function of subsequence $\left[i,j\right]$, and $Z^m_{i,j}$ is the partition function of subsequence $\left[i,j\right]$ that is a part of multi loop and that includes at least one base-pair. The functions f_2 and f_4 are the same as in Algorithm 1, the free energy of unpaired bases depending on the structural contexts.

The functions g_1^{W} to g_5^{W} are the gain function of Hamming distance, for the transitions from the term in the right hand side of the equation to the left hand side W's, which mostly correspond to the number of the paired bases in the reference structure where no base pair may exist in the specific structural contexts. The functions g_1^W to g_5^W are defined using g_0^Z in (4) as follows, and illustrated in Fig. 2.

$$\begin{aligned}
g_1^{\mathrm{W}}(i,j) &= g_0^{\mathrm{Z}}(1,L) - g_0^{\mathrm{Z}}(i,j) - g_0^{\mathrm{Z}}(1,i-1) \\
&\quad - g_0^{\mathrm{Z}}(j+1,L), \\
g_2^{\mathrm{W}}(h,i,j,\ell) &= g_0^{\mathrm{Z}}(h,\ell) - g_0^{\mathrm{Z}}(i,j) + 1 - 2\sigma_{h,\ell}, \\
g_3^{\mathrm{W}}(h,i,j,\ell) &= g_2^{\mathrm{W}}(h,i,j,\ell) - g_0^{\mathrm{Z}}(h+1,i-1), \\
g_4^{\mathrm{W}}(h,i,j,\ell) &= g_2^{\mathrm{W}}(h,i,j,\ell) - g_0^{\mathrm{Z}}(j+1,\ell-1), \\
g_5^{\mathrm{W}}(h,i,j,\ell) &= g_2^{\mathrm{W}}(h,i,j,\ell) - g_0^{\mathrm{Z}}(h+1,i-1) \\
&\quad - g_0^{\mathrm{Z}}(j+1,\ell-1).
\end{aligned}$$

$W_{ij}^b(d,\sigma)$, the outside partition function over Hamming distance, which is the sum of all the Boltzmann factors outside of the (i,j) base-pair, whose Hamming distance outside of (i,j) from the reference structure σ is d, is obtained as the coefficients of the polynomials as

$$W_{i,j}^b(\sigma) = \sum_d W_{ij}^b(d,\sigma)x^d. \tag{14}$$

Accelerated calculation by discrete fourier transformation

Algorithm 2 also includes the multiplication of the polynomials, in the third term of (13) . We applied DFT approach to the outside partition function over Hamming distance. Substituting same complex number to x as in (6), we obtain

$$\tilde{W}_{i,j}^b(\sigma) = \sum_{d=d_{\min}}^{d_{\max}} W_{i,j}^b(d,\sigma)\exp\left[2\pi i\frac{rd}{\Delta}\right]. \tag{15}$$

The outside partition function over Hamming distance is rewritten as

$$\begin{aligned}
W^b(d) &= \sum_{s=d_{\min}}^{d_{\max}} W^b(s)\delta_{sd} \\
&= \sum_{s=d_{\min}}^{d_{\max}} W^b(s) \sum_{r=d_{\min}}^{d_{\max}} \frac{\exp\left[2\pi i\frac{r(s-d)}{\Delta}\right]}{\Delta} \\
&= \frac{1}{\Delta}\sum_{r=d_{\min}}^{d_{\max}} \exp\left[2\pi i\frac{-rd}{\Delta}\right] \sum_{s=d_{\min}}^{d_{\max}} W^b(s)\exp\left[2\pi i\frac{rs}{\Delta}\right] \\
&= \frac{1}{\Delta}\sum_{r=d_{\min}}^{d_{\max}} \exp\left[2\pi i\frac{-rd}{\Delta}\right]\tilde{W}_{1,L}(\sigma).
\end{aligned} \tag{16}$$

In (16), $\tilde{W}_{1,L}(\sigma)$ can be computed by Algorithm 2, substituting $x = \exp[2\pi i(r/(\Delta))]$. The entire (16) can be computed by DFT. Note that the above dynamic programming and DFT are computable in parallel.

Secondary structure predictions using decomposed base-pairing probabilities

Once we obtain the partition functions over Hamming distance, Z^b and W^b, we can calculate the base-pairing probabilities (BPPs) $P^b(d)$ on each Hamming distance d by Eq. (9), and then BPPs, $P_{ij}^b[r_{\min}, r_{\max}]$ for each peak by Eq. (10). As the representative secondary structures of each peaks, the γ-centroid estimator is computed from each corresponding BPP by the posterior decoding [4].

Results

Application to Lysine riboswitch

We applied our method to an RNA called the Lysine riboswitch. The Lysine riboswitch RNA is 5'-UTR region of *lysC* and is known to be regulated by concentration of lysine [15]. The sequence was taken from *lysC* of *B. Subtilis* (J03294.1:2297–2537). The secondary structure predicted by CentroidFold (with $\gamma = 1$) was chosen as

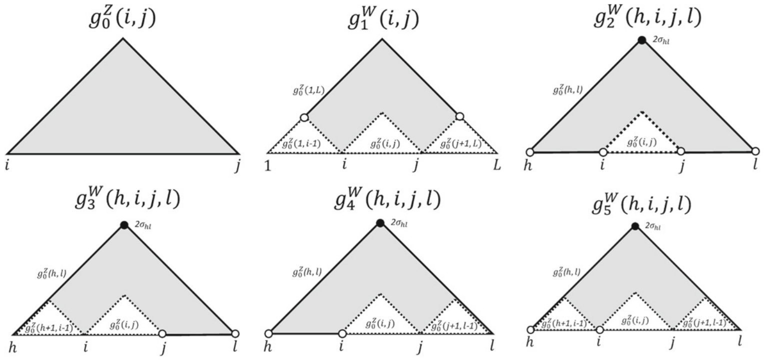

Fig. 2 Schematic illustration of gain functions for the powers of polynomials in Algorithm 2

a reference structure. First we analyzed the distribution of the structures over Hamming distance from the reference structure σ, represented as $P(d, \sigma)$ in (2). Figure 3a represents the plot of $P(d)$, where multiple peaks were observed in the distribution. We split the range of the Hamming distance at $d = d'$ such that $P(d' - 1) > P(d')$ and $P(d') < P(d' + 1)$. Furthermore, only the region $P(d) > \exp(Q/RT)$ was considered, where $Q = -10$ kcal/mol, R is the gas constant, and $T = 310$ K (37 °C). As a result, ranges of $[r_{\min}, r_{\max}]$ were determined to be [9, 56], [56, 82], and [82, 95]. The base-pairing probability matrices were then calculated according to the procedures described in "Methods" section. The γ-centroid estimator ($\gamma = 1$) were then reconstructed by posterior decoding of the base-pairing probability matrices as the representative structures (alternative structures hereafter).

Figure 3b and c represent the secondary structure of the reference structure, and those of the alternative structures, respectively. As expected from the large Hamming distance change, the alternative structure obtained from the Hamming distance range $[82, 95]$ had a considerable structural change from the reference structure (hereafter we call this structure *alt3*). Furthermore, it can be seen that two experimentally important sequences (red and skyblue circles) of the RNA, form an antiterminator stem. Experimentally, it has been considered that 3' pair of the antiterminator stem (colored skyblue in Fig. 3b also forms a terminator hairpin (skyblue and blue circles) and its transition is modulated by the concentration of lysine. Disrupting either of the antiterminator stem, or the terminator hairpin formation by mutations leads to the loss of the riboswitch function [15]. We further applied the RintD [12] algorithm to both the wild-type

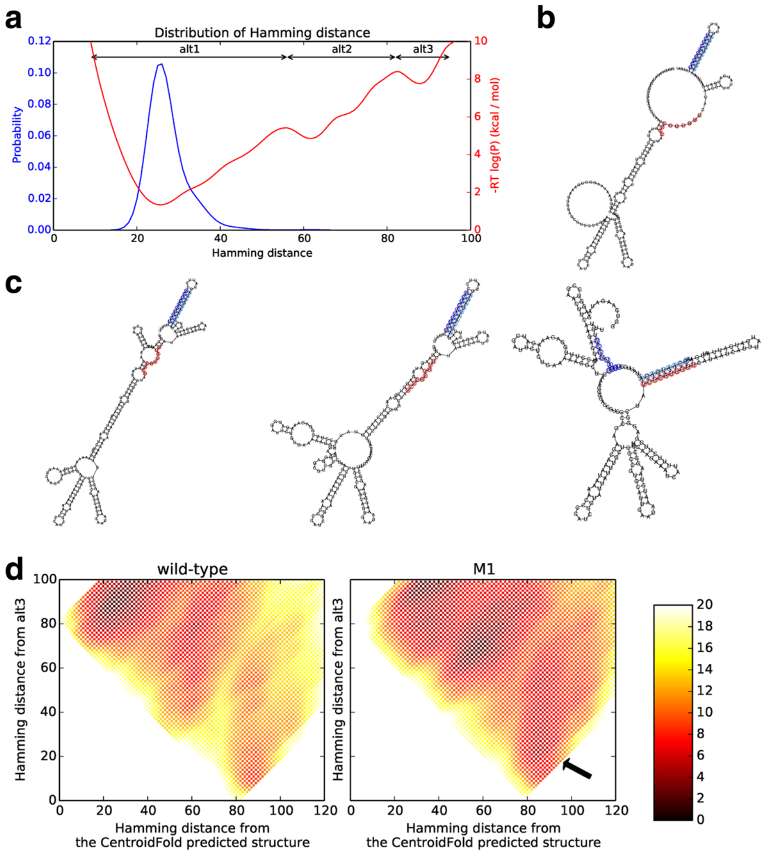

Fig. 3 a The distribution of Lysine riboswitch secondary structures projected to the Hamming distance from the reference structure. Both the probabilities and their logarithm are plotted. **b** The secondary structure representations of alternative structures. **c** The alternative structures of the the Hamming distance range $[9, 56]$ (left), $[56, 82]$ (middle) and $[82, 95]$ (right, "alt3"). Terminator and antiterminator sequences are marked with circles (see the main text.) Figures were generated by RNAplot in the ViennaRNA package [7]. **d** The 2D distribution of Lysine riboswitch secondary structures projected to the Hamming distances from the structure predicted by CentroidFold and the alt3 structures, calculated by using RintD [12]. The probabilities were converted to the free energy (i.e. $-RT \log P$) and were plotted. (left) The case with the wild-type RNA sequence. (right) The case with the M1 mutant. Secondary structures that have low Hamming distances to the alt3 structure are more stable with the M1 mutant than those with the wild-type (marked with an arrow). Free energies greater than 20 kcal/mol are plotted in white. Note the checkered pattern appears due to difficulties in achieving some Hamming distance constraints and a mutant *M1* that has modified sequence at the 5'-pair of the terminator hairpin, with mutations U234A and G235C. These mutations have been known to disrupt the pairing of the terminator hairpin. We use the structure predicted by CentroidFold, and alt3 as the reference structures for the RintD-2D [12]. Figure 3d shows the mutations to the RNA sequence increase the ratio of the alternative structure found in the wild-type sequence. It is thus inferred that a structural transition between the structure predicted by CentroidFold and alt3 reflects the functionally relevant structural change of the Lysine riboswitch.

We note that in the case of Fig. 3 the probability that a randomly sampled sequence falls into the Hamming-distance range of $[82, 95]$ is 1.43×10^{-5}. These structures have a low total probability when RNA is not interacting to other molecules, but this weak peak may imply a potential structural change under the interaction with the ligand. We also note that such a low-probability structural cluster is hard to find using existing methods, such as the random sampling of the secondary structures.

Application to ROSE element thermometer

The ROSE element thermometer is a functional RNA encoded in 5'-UTR of a mRNA, which changes its structure according to the temperature to regulate the translation of the downstream mRNA. The ROSE element thermometer prohibits the binding of ribosomes to the Shine-Dalgano (SD) sequence in 30 °C, but the structure change in 45 °C enables ribosomes to bind SD sequence to promote the translation of the mRNA.

Figure 4 shows the secondary structure probability distribution over Hamming distance of the ROSE element thermometer, where three major peaks for potential structural clusters are observed. We classified the three potential structural clusters A, B and C by Hamming distances, $[0, 10]$, $[11, 34]$ and $[35, 40]$. The abundance of the clusters in their probabilities along the changing temperature is shown in Fig. 5.

We then calculate the base-pairing probabilities over Hamming distance by RintW, and γ-centroid estimators ($\gamma = 1$) for the three clusters. The esimated γ-centroid structures are shown in Fig. 6. Using the γ-centroid estimators of the first and the third peak as the reference structures, the secondary structure probability distribution over Hamming distances (2D) were calculated under three different temperatures (Fig. 7), which shows the change of the probability landscape depending on the temperature. The change of probabilities of the three clusters depending on the temperature is shown in Fig. 5. It can be observed that cluster A (Hamming distance $d \in [0, 10]$) is dominant in low temperature and that cluster C ($d \in [35, 40]$)become stronger in high temperature.

Discussion

It is known that the functions of RNAs are closely correlated to their secondary structures, but limited reliability of secondary structure predictions have been preventing effective functional analyses. There are many tools for secondary structure predictions, but any point estimate of secondary structure has very small probability.

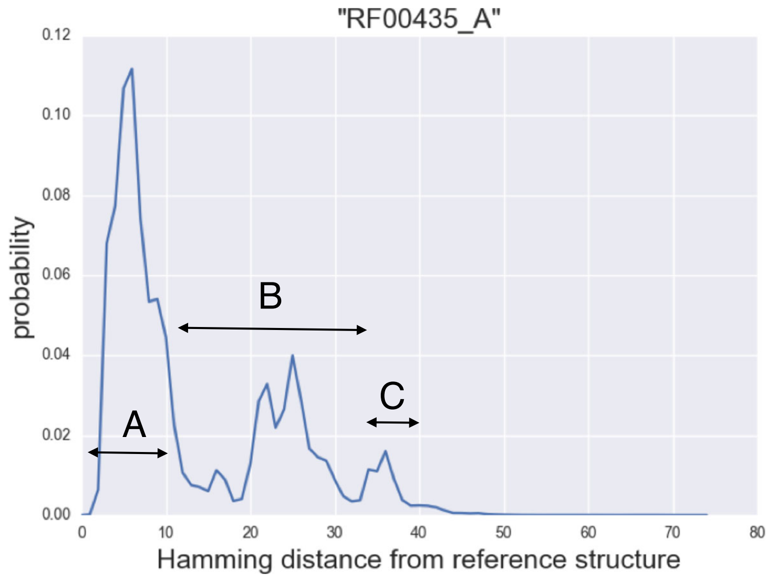

Fig. 4 Distribution over Hamming distance (1D) of ROSE element thermometer. The reference structure was taken from the γ-centroid estimator ($\gamma = 1$) with the temperature of 30 °C

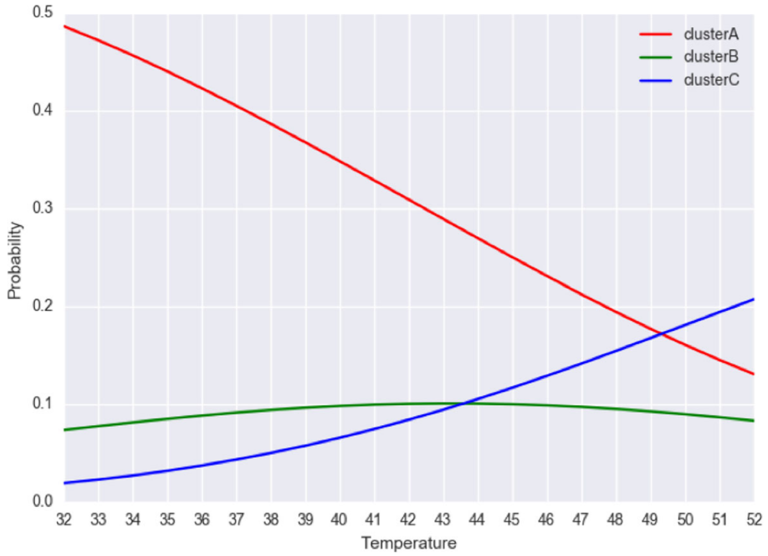

Fig. 5 Probabilities of three structural clusters of ROSE element thermometer on different temperatures

If the probabilities are concentrated in a cluster, it is possible to estimate an appropriate representative structure in the cluster. The Boltzmann distribution, however, often has multiple clusters of the concentration. Such concentrations may reflect of the essential alternative structures associated to switching functions, which are observed in several functional RNAs such as ribo-switches and thermometers.

There is no convenient tool to capture such essential alternative structures, though some tools can predict sub-optimal structures [7]. In this paper, we demonstrated that potential clusters of essential alternative secondary structures can be found by 1D analysis over Hamming distance. After detecting the candidates of clusters and defining the range of Hamming distance for each, the decomposed base-pairing probability matrix is computed by RintW. By using the representative structure calculated by posterior decoding of the base-pairing probability matrix, the 2D probability distributions (Figs. 7 and 3 bottom) are obtained by RintD [12]. If necessary, we can re-define the structural cluster based on the 2D distribution to re-calculate the decomposition of base-pairing probabilities.

In case of the ROSE element thermometer, the detection of the essential structural cluster from Fig. 4 was straightforward. The detection of the clusters, however, is not always easy. In the case of the Lysine riboswitch (Fig. 3), it was difficult to detect the peaks from simple 1D probability distribution, and log probability was informative. Such low probability region may have been ignored previously to overlook functionally relevant structures. It will be our future work to develop a method to find functionally important structural clusters by detailed analysis of the distribution, and also by better distance measure of the secondary structures than the Hamming distance.

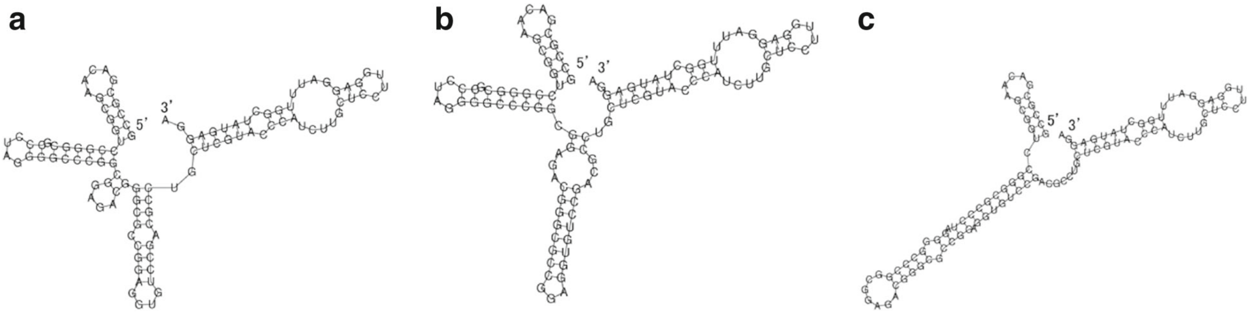

Fig. 6 γ-centroid structures ($\gamma = 1$) of three clusters of ROSE element thermometer. Structures **a**, **b** and **c** correspond to those obtained from the Hamming distance ranges of [0, 10], [11, 34] and [35, 40]

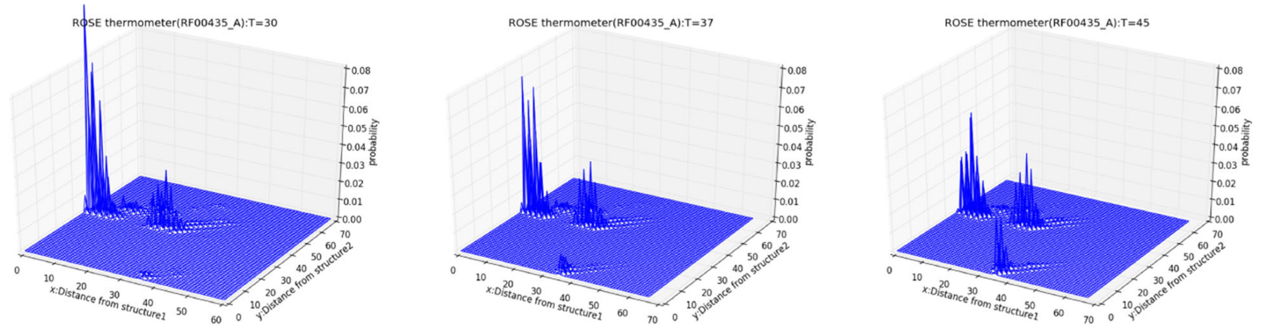

Fig. 7 Distribution over Hamming distance (2D) of ROSE element thermometer, under the temperatures, 30 °C (left), (b) 37 °C (middle) and 45 °C (right). The first reference structure was obtained using the γ-centroid estimator ($\gamma = 1$) with the Hamming distance [0, 10] (Structure A, in Fig. 6). The second reference structure is obtained similarly from the Hamming distance [35, 40] (Structure C, in Fig. 6)

Conclusion

In this paper, we presented a new method to detect the essential alternative secondary structures from RNA sequences by decomposing the base-pairing probability matrix. In order to calculate the decomposition, we have developed RintW, which efficiently calculates the inside/outside partition functions over Hamming distance and the base-pairing probabilities. Those calculations utilized dynamic programming mapped to polynomials and application of discrete Fourier transformation. By applying the method to the Lysine riboswitch and ROSE element RNA thermometer, potential alternative structural clusters, which may reflect their change in conformation, were observed. In the case of the ROSE element RNA thermometer, it was shown that changing temperature affected abundance of the clusters in their probabilities. Those results have shown that our method have a strong potential to analyze functional RNAs which have essential alternative structures.

Acknowledgements

The authors greatly thank to the contributors of Vienna RNA package, whose library was necessary in the developed software tool, RintW. The authors thank to Risa Kawaguchi, who suggested possibility of dynamic programing of base-pairing probabilities over Hamming distance. The authors also thank to the members of Artificial Intelligence Research Center, AIST, and the members of Hisanori Kiryu's laboratory for useful discussions.

Funding

This work was supported in part by JSPS KAKENHI Grant Numbers JP25240044 and JP16H06279 to K.A., JP16K17778 to S.S., and JP16K16143 to J.I. The funding for the publication of this article was provided by KAKENHI Grant Number 16H02484.

Availability of data and materials

Source code is available from http://www.ncRNA.org/RintW.

About this supplement

This article has been published as part of *BMC Bioinformatics* Volume 19 Supplement 1, 2018: Proceedings of the 28th International Conference on Genome Informatics: bioinformatics. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-19-supplement-1.

Authors' contributions

TH implemented outside partition function and analyzed RNAs in computational experiments. SS evaluated computational experiment of Lysine Riboswich and wrote the manuscript of this part. JI evaluated and prepared the RNA sequences to be analyzed. RM implemented inside partition function and helped TH for the implementation of outside partition function. KA organized the research, derived the equations of outside partition functions, and wrote the manuscript. All the authors have read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, University of Tokyo, 5-1-5 Kashiwanoha, 277-8561 Kashiwa, Japan. [2]Artificial Intelligence Research Center (AIRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-3-26 Aomi, Koto-ku, 136-0064 Tokyo, Japan. [3]Unique Co. Ltd., 6-6-20 Daita, Setagaya-ku, 155-0033 Tokyo, Japan.

Published: 19 February 2018

References

1. Ding Y, Chan CY, Lawrence CE. RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. RNA. 2005;11(8):1157–66.
2. Fukunaga T, Ozaki H, Terai G, Asai K, Iwasaki W, Kiryu H. CapR: revealing structural specificities of RNA-binding protein target recognition using CLIP-seq data. Genome Biol. 2014;15(1):16.
3. McCaskill JS. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers. 1990;29(6-7):1105–19.
4. Do CB, Woods DA, Batzoglou S. CONTRAfold: RNA secondary structure prediction without physics-based models. Bioinformatics. 2006;22(14):90–8.
5. Hamada M, Kiryu H, Sato K, Mituyama T, Asai K. Prediction of RNA secondary structure using generalized centroid estimators. Bioinformatics. 2009;25(4):465–73.
6. Hamada M, Kiryu H, Iwasaki W, Asai K. Generalized centroid estimators in bioinformatics. PLoS ONE. 2011;6(2):16450.

7. Lorenz R, Bernhart SH, Honer Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. ViennaRNA Package 2.0. Algorithms Mol Biol. 2011;6:26.
8. Adachi H, Ishiguro A, Hamada M, Sakota E, Asai K, Nakamura Y. Antagonistic RNA aptamer specific to a heterodimeric form of human interleukin-17A/F. Biochimie. 2011;93(7):1081–8.
9. Newberg LA, Lawrence CE. Exact calculation of distributions on integers, with application to sequence alignment. J Comput Biol. 2009;16(1):1–18.
10. Freyhult E, Moulton V, Clote P. Boltzmann probability of RNA structural neighbors and riboswitch detection. Bioinformatics. 2007;23(16):2054–62.
11. Freyhult E, Moulton V, Clote P. RNAbor: a web server for RNA structural neighbors. Nucleic Acids Res. 2007;35(Web Server issue):305–9.
12. Mori R, Hamada M, Asai K. Efficient calculation of exact probability distributions of integer features on RNA secondary structures. BMC Genomics. 2014;15 Suppl 10:6.
13. Clote P, Lou F, Lorenz WA. Maximum expected accuracy structural neighbors of an RNA secondary structure. BMC Bioinformatics. 2012;13 Suppl 5:6.
14. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 2003;31(13):3406–15.
15. Sudarsan N, Wickiser JK, Nakamura S, Ebert MS, Breaker RR. An mRNA structure in bacteria that controls gene expression by binding lysine. Genes Dev. 2003;17(21):2688–97.