Mori et al. BMC Genomics 2014, 15(Suppl 10):S6
http://www.biomedcentral.com/1471-2164/15/S10/S6

BMC Genomics

RESEARCH Open Access

Efficient calculation of exact probability distributions of integer features on RNA secondary structures

Ryota Mori[1*], Michiaki Hamada[2,3], Kiyoshi Asai[1,2]

From The 25th International Conference on Genome Informatics (GIW/ISCB-Asia)
Tokyo, Japan. 15-17 December 2014

Abstract

Background: Although the needs for analyses of secondary structures of RNAs are increasing, prediction of the secondary structures of RNAs are not always reliable. Because an RNA may have a complicated energy landscape, comprehensive representations of the whole ensemble of the secondary structures, such as the probability distributions of various features of RNA secondary structures are required.

Results: A general method to efficiently compute the distribution of any integer scalar/vector function on the secondary structure is proposed. We also show two concrete algorithms, for Hamming distance from a reference structure and for $5' - 3'$ distance, which can be constructed by following our general method. These practical applications of this method show the effectiveness of the proposed method.

Conclusions: The proposed method provides a clear and comprehensive procedure to construct algorithms for distributions of various integer features. In addition, distributions of integer vectors, that is a combination of different integer scores, can be also described by applying our 2D expanding technique.

Background

Recent investigations of coding and non-coding RNAs have proved that RNA molecules have more important roles in the regulation of living cells than those of our previous knowledge. It has also become clear that the structures of RNAs, especially the secondary structures, are one of the important features to identify the functions of RNAs. While the high-throughput methods to determine the secondary structures of RNAs are spreading, the importance of computational analyses of RNA sequences including prediction of secondary structures is increasing [1,2].

The free energy of each structure is connected to its existence probability. The existence probability of a secondary structure St of an RNA is given by the following canonical distribution:

$$P_{\mathrm{St}} = \frac{1}{Z} e^{-E_{\mathrm{St}}/(k_B T)} \tag{1}$$

$$Z = \sum_{\mathrm{St}} e^{-E_{\mathrm{St}}/(k_B T)}, \tag{2}$$

where P_{St} and E_{St} are respectively the existence probability and the free energy of the structure St, k_B is the Boltzmann constant and T is the temperature constant. Z is the normalizing factor known as the partition function, which is the summation of Boltzmann factor $e^{-E_{\mathrm{St}}/(k_B T)}$ among all the possible structures. The partition function of an RNA sequence and the free energy of each structure can be obtained by dynamic programming algorithms on the parameters determined experimentally [3,4].

The equation (1) shows that the structure with the highest existence probability is the structure of the minimum free energy. Therefore, it is natural to treat the secondary structure of the minimum free energy as the estimate of the secondary structure. The probability that an RNA folds into a particular structure is, however, generally

* Correspondence: mori@cb.k.u-tokyo.ac.jp
[1]Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa City, Chiba, Japan
Full list of author information is available at the end of the article

extremely low even if it is the structure of the minimum free energy, because of the combinatorial explosion [5]. For example, the probability of a particular secondary structure of some rRNAs are less than 10^{-22} no matter which structure is chosen. This means that the prediction of the secondary structure of an RNA and subsequent analyses based on the predicted secondary structure are not always reliable. It is therefore desirable to investigate properties of the probability distributions of the whole ensemble of the possible structures.

We propose in this paper a general method to efficiently compute the exact distribution of any integer quantity of the feature on each secondary structure. The proposed method has been motivated by the framework and its application to sequence alignments by [6]. Their framework is generally valid for integer functions on Boltzmann distributions whose partition function can be calculated by a linear dynamic programming. For the case of secondary structures of RNAs, however, the recursions in the dynamic programming of the partition function have more complicated forms including the products of combinations of DP matrix elements, which inhibits direct application of their framework. We have overcome the difficulty by expanding McCaskill algorithm, which is a well-known dynamic programming of the partition function of the secondary structures of RNAs using the energy parameters experimentally determined [7].

Naive implementations of our proposed method requires computational complexities of $O(n^3|S|^2\alpha)$ in time and $O(n^3 + \beta|S|)$ in space, where n is the length of the sequence, $|S|$ is the size of the integer score variation, which depends on the objective distribution while they never exceed n in the case of example problems in this paper, and α and β is the costs depending on the objective score. By adapting Discrete Fourier Transform, we can reduce those complexities to $O(n^3|S|\alpha)$ in time and $O(n^3 + \beta)$. The DFT in our method on RNA structures achieves an order-level improvement of the complexity, which could not achieved by the DFT on linear dynamic programmings in [6]. We can further reduce time complexity to $O(n^3|S|\alpha/U)$ by parallel computing using U computational units.

We demonstrate the effectiveness of the proposed method in several practical problems. The first example is the distribution of the Hamming distances from a reference structure. A practically equivalent algorithm and its acceleration have been implemented as RNAbor by [8] and [9], while we have reconstructed the algorithm by deducing from our general principle. The second example is the exact distributions of 5′ - 3′ distance. Conventional methods for analysing 5′ - 3′ distance only calculate mean length or assume over-simplified models. We propose here a novel algorithm to compute the whole distribution of 5′ - 3′ distance considering the thermodynamic properties of the RNAs. The final example is acceleration of RNA2Dfold, which is included in ViennaRNA package [10].In this example, the distribution of the Hamming distances from two specified reference structures are calculated. We show our method reduces computational complexity from $O(n^7)$ in time and $O(n^4)$ in re-space to less than $O(n^5/U)$ in time and $O(n^2U)$ in space, which is a similar idea proposed recently [11]. These examples indicate that our method offers a way to obtain a wide variety of distributions of integer quantities.

Methods

We first show the fundamental concepts of our proposed method in this section.

Definition of integer score distribution

Let us assume that s represents a mapping from $x \in U$ to an integer score $s(x) \in \mathbb{Z}$. In our case of RNA secondary structures, the U is the space of all the possible secondary structures for a given RNA sequence, and an integer score $s(x)$ represents a feature or a property assigned to each structure x. The integer score distribution is defined as the probability distribution $p(s)$ of $s(x)$ derived from the probability distribution $p(x)$ of x:

$$p(s) = \sum_{\{x|s=s(x)\}} p(x) \tag{3}$$

In this paper, we discuss on how to efficiently compute integer score distributions in general and in the specific cases for RNA secondary structures. Our proposed method for RNA secondary structures efficiently computes the exact distribution when $p(x)$ and $p(s)$ can be calculated by the dynamic programming algorithms sharing a same form.

A conventional model for integer score distribution

For a certain class of problems, including distributions of integer score of each sequence alignment, the partition function of the objective distribution can be calculated abstractly by Algorithm 1. Z is the partition function shown in equation (2). Z is a scalar array of length N representing the partition function of the problem size N, whose components for the dynamic programming are aligned in the computing order. $t(k|i)$ is a quantity proportional to the probability of the transition from state i to state k, which can be quite sparse in values.

Algorithm 1 An abstract form of calculating the partition function

1: $Z[0] = 1$
2: **for** $k = 1$ to N **do**
3: $\quad z[k] = \Sigma_{i=0}^{k-1} Z[i]t(k|i)$

4: **end for**
5: $Z = Z[N]$

[6] showed that if the partition function can be computed by Algorithm 1, integer score distributions are obtained by Algorithm 2, where $Z(x)$ is an array of polynomials of x, and $s(i, k)$ is the gain of the integer score in the transition from i to k.

Algorithm 2 A polynomial approach to integer score distributions proposed in [6]
1: $Z(x)[0] = 1$
2: **for** $k = 1$ to N **do**
3: $Z(x)[k] = \Sigma_{i=0}^{k-1} Z(x)[i]t(k|i)x^{s(i,k)}$
4: **end for**
5: Z = a sum of coefficients of polynomial $Z(x)[N]$

In Algorithm 2, $Z(x)[N]$ represents a polynomial in x whose factor z_S of x^i is proportional to the sum of the probabilities of obtaining score i among all the paths:

$$Z(x)[N] = \sum_{j=0}^{S_{max}} z_j x^j, \tag{4}$$

where S_{max} is the maximum score.

The p_S, the probability of obtaining score S, is finally calculated by the following equation:

$$p_S = \frac{z_S}{Z}, \tag{5}$$

A general model for integer score distribution of RNA secondary structure

In the case of RNA secondary structures, the dynamic programming for the partition function does not match to Algorithm 1. Therefore, we have to construct an algorithm different from Algorithm 2 for the calculation of integer score distributions on RNA secondary structures. As pseudo-code is shown in Algorithm 3, products of combinations between DP matrix elements and constant term c_k are required for the computation. The detailed description of this derivation is shown in the additional file 1 (Section S1).

Algorithm 3 A general polynomial approach to integer score distributions for the ensemble of RNA secondary structures
1: $Z(x)[0] = 1$
2: **for** $k = 1$ to N **do**
3: $Z(x)[k] = \Sigma_{i=0}^{k-1} Z(x)[i]t(k|i)x^{s(i,k)} + \Sigma_{i=0}^{k-2}\Sigma_{j=i+1}^{k-1} Z(x)[i]Z(x)[j]t(k|i,j)x^{s(i,j,k)} + c_k x^{s(k)}$
4: **end for**
5: Z = a sum of coefficients of polynomial $Z(x)[N]$

The partition function is dispersed according to the score of each secondary structure included in the whole ensemble. In other words, the coefficient of x^S in $Z(x)[N]$ represents proportional to the probability that the RNA structure has score S. After the calculation by Algorithm 3, p_S can be derived from equation (5).

Algorithm 3 requires computational complexities of $O(n^3 S_{max}^2 \alpha)$ in time and $O((n^2 + \beta)S_{max})$ in memory, where α and β is the complexities in time and in space respectively for the calculation of each integer score.

Adopting Discrete Fourier Transform (DFT)

Discrete Fourier Transform (DFT) is a Fourier Transform on a discrete sampling interval, which is employed in improving the efficiency of various computational problems as well as frequency analysis. According to [6], by applying DFT distributed processing is available for computing integer score distributions on sequence alignments. On RNA secondary structures, DFT reduces time complexity of computations in order-level as well as merely decentralize the procedure.

DFT $\mathcal{F}$ satisfies the following equation:

$$z = \mathcal{F}(\zeta), \tag{6}$$

where

$$z = (z_0, z_1, \cdots, z_{S_{max}}) \tag{7}$$

$$\zeta = (\zeta_0, \zeta_1, \cdots, \zeta_{S_{max}}) \tag{8}$$

$$\zeta_k = \frac{\Sigma_{j=0}^{S_{max}} z_j \left(\exp\left[2\pi i \frac{k}{S_{max}+1}\right]\right)^j}{S_{max}+1}. \tag{9}$$

In DFT approach, each x in the polynomials is replaced by a complex number on the unit circle to calculate $\boldsymbol{\zeta}$ instead of $\mathbf{z}$ directly. The relation of the two quantities are derived by comparing equations (4) and (9)):

$$\zeta_k = \frac{Z\left(\exp\left[2\pi i \frac{k}{S_{max}+1}\right]\right)[N]}{S_{max}+1}. \tag{10}$$

After $\boldsymbol{\zeta}$ is obtained, DFT extracts $\mathbf{z}$ from $\boldsymbol{\zeta}$ by $O(S_{max}^2)$ time.

Algorithm 4 shown below is the modification of our naive Algorithm 3 by adopting DFT approach.

Algorithm 3 suffers from heavy computations of $O(S_{max}^2)$ in time for products of polynomials if the degree S_{max} is large. In the recursions for $\boldsymbol{\zeta}$ in Algorithm 4, however, each computation for polynomial products is replaced to a computation of products of complex numbers, which requires only a constant time. While we still need to extract $\mathbf{z}$ from $\boldsymbol{\zeta}$ by $O(S_{max}^2\alpha)$ time, the total time complexity is reduced from $O(n^3 S_{max}^2 \alpha)$ to $O(n^3 S_{max}^2 \alpha)$. In addition, each ζ_k can be calculated

Algorithm 4 DFT-adopted approach for integer score distribution

1: /* DP phase (distributed processing is available) */
2: **for** $S = 0$ to S_{max} **do**
3: $x = \exp\left[2\pi i \frac{S}{S_{max}+1}\right]$
4: $Z[S][0] = 1$
5: **for** $k = 1$ to N **do**
6: $Z[S][k] = \Sigma_{p=0}^{k-1} Z[S][p]t(k|p)x^{s(p,k)} + \Sigma_{p=0}^{k-2}\Sigma_{q=p+1}^{k-1} Z[S][p]Z[S][q]t(k|p,q)x^{s(p,q,k)} + c_k x^{s(k)}$
7: **end for**
8: $\zeta_S = Z[S][N]$
9: **end for**
10: /* DFT phase*/
11: **for** $S = 0$ to S_{max} **do**
12: $z_S = \Sigma_{r=0}^{S_{max}} \zeta_r \exp\left[-2\pi i \frac{rS}{S_{max}+1}\right]/(1+S_{max})$
13: **end for**
14: $Z = \Sigma_{S=0}^{S_{max}} z_S$

individually so we can replace the computational cost to $O(n^3\alpha)$ time and $O((n^2+\beta)S_{max})$ space by adopting maximum parallelization, using ether multi-core units or cluster machines. Accordingly, the practical efficiency by utilizing DFT depends on parallelization environment strongly (Table 1).

McCaskill model

According to the above approach, we next construct and implement concrete formulas of computing a general integer score distribution for RNA secondary structures based on McCaskill model. McCaskill model is a standard procedure for computing partition function in equation (2) by a dynamic programming based on energy parameters. In this model, the partition function is obtained as $Z_{1,n}$ from the following recursive scheme of polynomial order:

Initialization ($1 \le i \le n$):

$$Z_{i,i} = 1.0 \tag{11}$$

$$Z^{\bullet}_{i,i} = Z^{m}_{i,i-1} = 0, \tag{12}$$

Recursion ($1 \le i \le j \le n$):

$$Z_{i,j} = 1.0 + \sum_{k=i}^{j-1} Z_{i,k} Z^{1}_{k+1,j} \tag{13}$$

Table 1 Required time and space

	Polynomial	DFT	DFT with U* units
Time	$O(n^3 S_{max}^2 \alpha)$	$O(n^3 S_{max}\alpha)$	$O(n^3 S_{max}\,\alpha/U)$
Space	$O((n^2+\beta)S_{max})$	$O(n^2+\beta)$	$O((n^2+\beta)U)$

*The number of parallelization units must be equal to S_{max} or less. If $U = S_{max}$, DFT with U units requires $O(n^3\alpha)$ in time and $O((n^2+\beta)S_{max})$ in space.

$$Z^{1}_{i,j} = \sum_{k=i+1}^{j} Z^{b}_{i,k} \tag{14}$$

$$Z^{b}_{i,j} = e^{f_1(i,j)} + \sum_{k=i+1}^{j-2}\sum_{l=k+1}^{j-1} Z^{b}_{k,l} e^{f_2(i,j,k,l)} + \sum_{k=i+2}^{j-1} Z^{m}_{i+1,k-1} Z^{m1}_{k,j-1} e^{f_3(i,j)} \tag{15}$$

$$Z^{m}_{i,j} = \sum_{k=i}^{j-1} \left(e^{f_4(k-i)} + Z^{m}_{i,k-1}\right) Z^{m1}_{k,j} \tag{16}$$

$$Z^{m1}_{i,j} = \sum_{k=i+1}^{j} Z^{b}_{i,k} e^{f_4(j-k)}, \tag{17}$$

where each $f_k(\cdot)$ ($k = 1 \cdots 4$) is the function corresponding to the energy contribution to each state, and the parameters of the functions are determined experimentally [3,4].

$$f_k(\cdot) = -\frac{\Delta E}{k_B T} \tag{18}$$

Although the second factor in the right hand side of the equation (15) indicates that this procedure requires $O(n^4)$ in time, it is usually reduced to $O(n^3)$ by assuming a reasonable threshold of the length of the internal loops.

Score accumulable McCaskill model

We modify McCaskill model recursions (equations (13)-(17)) to calculate integer score distribution under the concept described in the Approach section.

$$Z_{i,j} = x^{g_1(i,j)} + \sum_{k=i}^{j-1} Z_{i,k} Z^{1}_{k+1,j} x^{g_2(i,j,k)} \tag{19}$$

$$Z^{1}_{i,j} = \sum_{k=i+1}^{j} Z^{b}_{i,k} x^{g_3(i,j,k)} \tag{20}$$

$$Z^{b}_{i,j} = e^{f_1(i,j)} x^{g_4(i,j)} + \sum_{k=i+1}^{j-2}\sum_{l=k+1}^{j-1} Z^{b}_{k,l} e^{f_2(i,j,k,l)} x^{g_5(i,j,k,l)} + \sum_{k=i+2}^{j-1} Z^{m}_{i+1,k-1} Z^{m1}_{k,j-1} e^{f_3(i,j)} x^{g_6(i,j,k)} \tag{21}$$

$$Z^{m}_{i,j} = \sum_{k=i}^{j-1} \left(e^{f_4(k-i)} x^{g_7(i,j,k)} + Z^{m}_{i,k-1} x^{g_8(i,j,k)}\right) Z^{m1}_{k,j} \tag{22}$$

$$Z^{m1}_{i,j} = \sum_{k=i+1}^{j} Z^{b}_{i,k} e^{f_4(j-k)} x^{g_9(i,j,k)}, \tag{23}$$

Results

In this section, we show three examples to demonstrate how to construct algorithms for practical problems. The first and the second examples are the case to which our general model is directly applicable, where all we have to do is defining scoring functions. In the third example, we expand our model into two dimensions in order to describe a distribution of two dimensional integer vector.

We practically implemented and evaluated the concrete algorithms for those three examples with a distributed processing application by OpenMP on a dual quad-core Intel® Xeon® E5540 @2.53GHz with 17.6 GB RAM. The run time was measured as a mean of 10 random sequences by single or eight cores.

A distribution of the Hamming distance from a reference structure

Conventional RNA secondary structure estimation produces the most stable and possible structure or the representative structure such as a centroid in the whole ensemble. Those point estimations of the secondary structures, however, have a risk to neglect the information on the thermal fluctuations or significant suboptimal structures [12]. Some local structures might be relatively stable only at certain global structures, and some structures such as ribo-switches might have multiple distinct stable global structures besides the predicted structures [13]. RNAbor [8] is a software which exactly calculates the probability that RNA folds into the structures that have the same distance from a given structure. It informs us concentration of existence probability around a structure predicted by conventional software, which will help deeper understanding about biological behavior of RNA molecules. Our model is applicable for this problem since the distance between RNA secondary structures can be defined as an integer function. Here we reconstruct the algorithm from a viewpoint of our general principle described in the Approach section, motivated by the work by Newberg *et al*., while practically equivalent algorithm has been independently presented in [9].

Definition of distance

We employ the distance measure of RNA secondary structures used in RNAbor, which is defined as the Hamming distance between binary vectors representing the structures as described below.

$$S[i][j] = \begin{cases} 1 \text{ (if i - th and j - th bases make a pair)} \\ 0 \text{ (otherwise)} \end{cases}. \quad (24)$$

Let us call S a structure vector. The dimension of a structure vector is $\binom{n}{2} = n(n-1)/2$ for an RNA of length n.

Now we define the Hamming distance d of two structures by the Hamming distance of their structure vectors S_1 and S_2:

$$d = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} S_1[i][j] \oplus S_2[i][j]$$

$\oplus$: exclusive disjunction.

The Hamming distance between RNA structures never exceeds its sequence length n in spite of the high dimensions of structure vectors, we obtain $d_{max} \le n$ as the exact maximum of d by cubic time (See the Section S3 in the additional file 1).

Scoring functions

Recursions for calculating the distribution of d are easily derived by defining $gk(\cdot)$ $(k = 1 \cdots 9)$ in the equations (19)-(23) as appropriate integer functions. For instance:

$$g_6(i,j,k) = \sum_{p=k}^{j-1} S[p][j] + \sum_{q=i+1}^{j} S[i][q] + \sum_{p=i+1}^{k-1} \sum_{q=k}^{j} S[p][q] + 1 - 2S[i][j] \quad (26)$$

This $g_6(\cdot)$ returns an integer value that is newly accumulated as the gain of the Hamming distance from the reference structure by the corresponding transition (Figure 1). Although naive implementation for computation of $g_k(\cdot)$ requires quadratic order in time, a slight pre-calculation reduces this to constant time. We show full description of $g_k(\cdot)$ and $O(1)$ time calculation in the additional file 1 (Section S2). Accordingly, the total complexity using DFT is $O\left(n^3 d_{max}\right)$ in time and $O(n^2)$ in space, since $S_{max} = d_{max}$, $\alpha = 1$, and $\beta = O(n^2)$. It can be reduced to $O\left(n^3 d_{max}/U\right)$ in time and $O(n^2 U)$ in space if parallelization of U-units is available($U \le d_{max}$).

A distribution of RNA 5′- 3′distance

Recently, Yoffe *et al*. found that the distance of 5′ end and 3′ end of the RNA molecule tended to be short, largely independent of molecule lengths or sequence patterns [14]. They pointed out the relevance of these observations and biological interpretation especially about in viral RNA evolution. Clote *et al*. proposed a method for calculating an expected distance [15], but it might be helpful for RNA structure analysis to reveal the population of structures shorter than some threshold as well as mean length. A method for counting the 5′-3′ distances over all secondary structures has been proposed by [16], but their method assumes that all structures occur by the same probability and every base can make pairs with an arbitrary base except for pseudoknots. We propose the first algorithm for computing

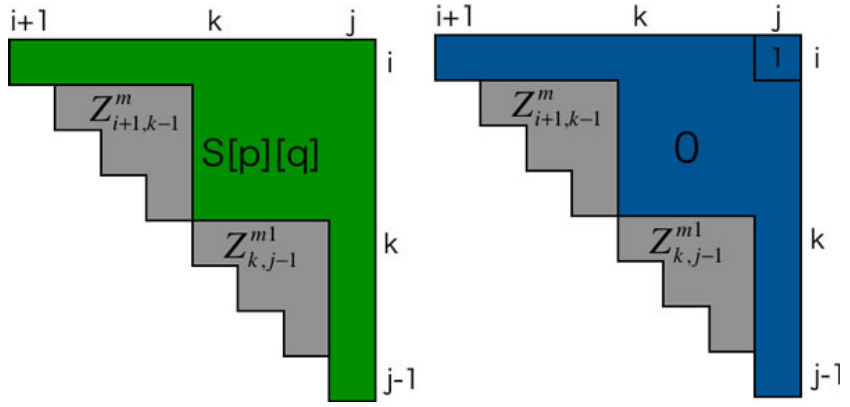

Figure 1 A simple concept illustration of the way to calculate the newly accumulated distance. Left and right pictures illustrate the vector of the reference structure and the transition which corresponds to the third term on right side of equation (21) respectively. $g_6(i, j, k)$ returns the Hamming distance between green and blue regions. Gray regions have been already considered in $Z^{m}_{i+1,k-1}$ and $Z^{m1}_{k,j-1}$.

the exact probability distribution of the 5′-3′ distances based on the McCaskill model.

Definition of 5′ - 3′ distance

We follow the work by Yoffe and colleagues as the definition of 5′ - 3′ distance $d_{5\prime-3\prime}$:

$$d_{5\prime-3\prime} = c_{ext} + h_{ext}, \tag{27}$$

where c_{ext} is the number of covalent bonds in the exterior loop and h_{ext} is the number of hydrogen bridges in the exterior loop (See Figure 2 for example).

Scoring functions

As with the case of the previous section, defining $gm(\cdot)$ ($m = 1 \cdots 9$) enables us to calculate the 5′-3′ distance distribution as following:

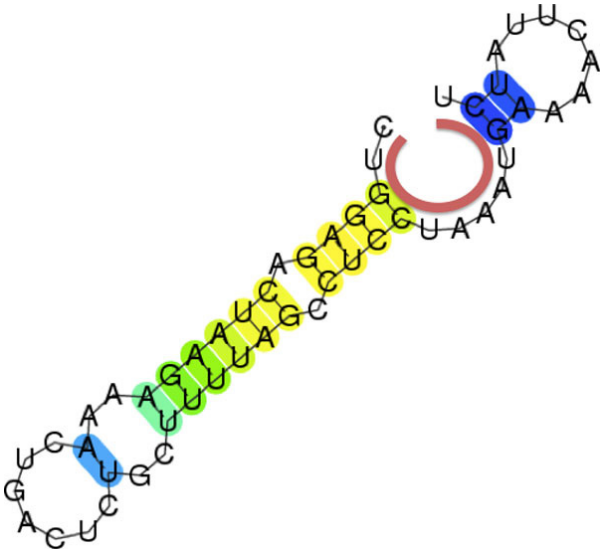

Figure 2 An example for introducing the definition of *d*5′ - 3′. The red arch represents the 5′ - 3′ distance, in this case, we have $d_{5\prime\text{ - }3\prime} = c_{ext} + h_{ext} = 3 + 8 = 11$.

$$g_1(i, j) = j - i \tag{28}$$

$$g_2(i, j, k) = 1 \tag{29}$$

$$g_3(i, j, k) = 1 + j - k \tag{30}$$

$$g_m(\cdot) = 0 \ (\text{for } m = 4, \ldots, 9) \tag{31}$$

The $g_1(i, j)$ is the 5′ - 3′ distance of the chain structure, which contains no base pairs. The $g_2(i, j, k)$ is the newly accumulated 5′ - 3′ distance, that is the junction of k-th and k+1-th bases. The $g_3(i, j, k)$ represents the sum of a hydrogen bridge by i-th and k-th bases and length of a chain structure from the k + 1-th base. Other functions $g_m(\cdot)$, $m = 4, \ldots, 9$) are irrelevant to 5′ - 3′ distance because their corresponding transitions for internal structures.

Total computational complexity using DFT with U parallel computing units, is $O(n^4/U)$ in time and $O(n^2U)$ in space ($S_{max} = n - 1$, $\alpha = \beta = 1$). In addition, since $Z^{b}_{i,j}$, $Z^{m}_{i,j}$, and $Z^{m1}_{i,j}$ do not contain variable x, therefore we can reduce the total amount of calculation by pre-computing them (See the Section S4 in the additional file 1).

A distribution of 2D RNA folding landscapes

RNA2Dfold is an application for 2D projections of RNA folding landscapes which are the two-dimensional probability distributions whose axis correspond to the Hamming distances from two independent given reference structures [10]. Such distributions provide profound information on the whole ensemble through the medium of landscapes depending on the given structures. The RNA2Dfold, however, has difficulty of computational cost; it requires $O(n^7)$ in time and $O(n^4)$ in space though the computational time can be drastically improved by utilizing sparse matrices. On the other

hand, extension of our proposed method reduces the complexity to less than $O(n^5)$ in time and $O(n^2)$ in space. Our method only calculate the distribution though RNA2Dfold also computes the minimum free energy structure of every combination of distances from the given structures. While a similar simplified algorithm has been proposed by [11], we construct an effective algorithm using DFT by expanding general principle described in previous sections.

Expanding the original model to two dimensions

The problem of computing the 2D folding landscape of an RNA, is defined as a natural expansion to two dimensions of the algorithm mentioned in the section. In this case, the objective distribution is defined on the two-dimensional discrete sample space which represents the Hamming distances from two given reference structures. Accordingly, we expand original model in equations (19)-(23) to two dimensions for the purpose of corresponding to two-dimensional score vectors. As shown in Algorithm S2, a vector variable $\mathbf{x} = (x_1, x_2)$ is employed to accumulate each component of a score vector $\mathbf{S} = (S_1, S_2)$ instead of applying a scalar variable x. The computational complexity of this model is $O(n^3 S_{1max} S_{2max} \alpha_1 \alpha_2 / U)$ in time and $O((n^2 + \beta_1 + \beta_2)U)$ in space, where U ($\leq S_{1max} S_{2max}$) is the number of parallel processing units, and α_k and β_k are time and space complexity for computing scoring function of k-th component.

Scoring functions

Now we can construct a model for the distribution of the Hamming distance from the two given structures by assigning S_1 and S_2 to the Hamming distance from the first and the second structures respectively. The total cost of this algorithm is $O(n^3 d_{1max} d_{2max}/U)$ in time and $O(n^2 U)$ in space. A concrete description is not shown here but in the Section S5 and S6 in the additional file 1.

Run time evaluation and sample outputs

Next we show the run time of the above three algorithms. We adopt the minimum free energy (MFE) structures as the reference structures for the algorithms in the section 4.1 and 4.3. The other reference structure for the algorithm in the section 4.3 is the open chain structure, that is a completely no base pairing structure. We measured the run time by single or 8 cores, though theoretically we can distribute the process up to S_{max} or $S_{1max} S_{2max}$.

As we can see in Figure 3, the curves of run time in each algorithms follow their theoretical orders, $O(n^3 dmax/U)$, $O(n^4/U)$, and $O(n^3 d_{1max} d_{2max}/U)$, where we consider d. to be proportional to RNA sequence length.

By way of example, we also illustrate outputs of our algorithms by using a sequence of tRNA. The secondary structure of tRNA is one of the most well-known

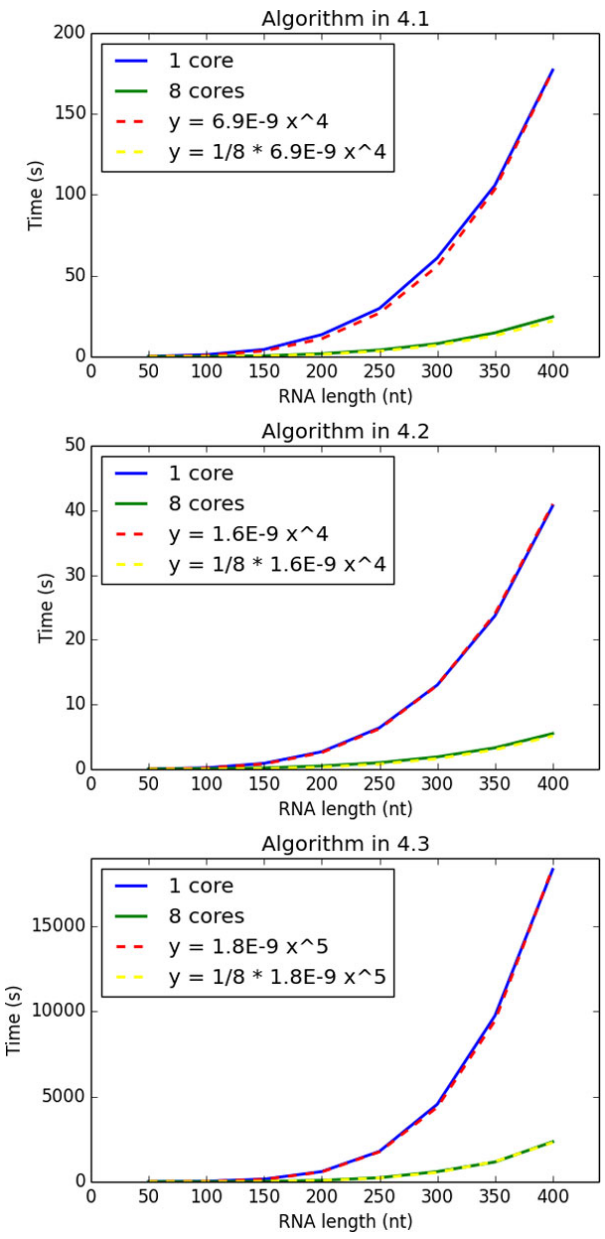

Figure 3 Run time along the sequence lengths compared with theoretical curves. Solid lines colored blue and green are run time measured as a mean of ten random sequences by single or eight cores respectively. Dashed lines are fitted curves theoretically expected.

structures called the cloverleaf structure (Figure 4(a)). However, prediction of the structure of a tRNA does not always have that shape. The CentroidFold [17], which is listed as one of the most accurate software tools in CompaRNA [18], predicts quite a different structure (Figure 4(b)). This disappointing example implies the limitation of RNA secondary structure predictions.

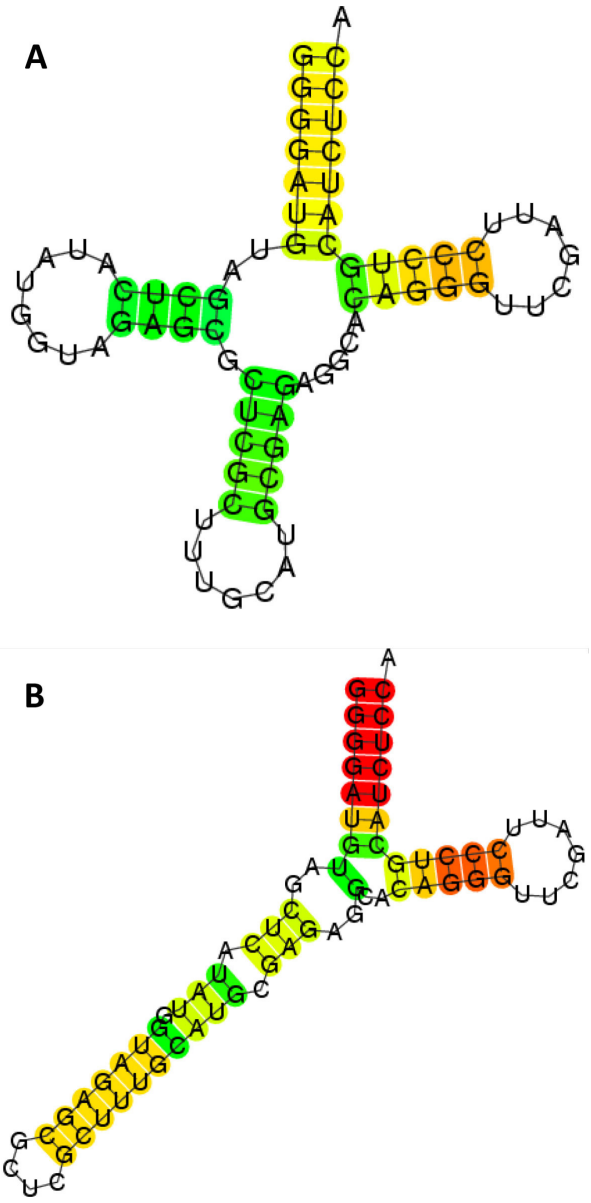

Figure 4 Estimated structures of a tRNA sequences. (a) A well-known cloverleaf structure, (b) a structure predicted by CentroidFold.

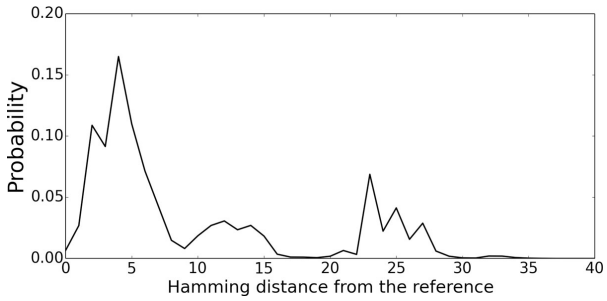

Figure 5 The landscape of the probability distribution of Hamming distance. X-axis represents the length of Hamming distance from the selected structure as a reference, and y-axis represents the probability that RNA folds into a structure which has each Hamming distance in the whole ensemble. We took a structure in Fig. 4(b) as the reference structure in this figure.

The probability distribution computed by the algorithm in the section 4.1 using the sequence and the reference structure illustrated in Figure 4(b) is shown in Figure 5. This probability landscape provides us an implication that this RNA might have sub-optimal structures around 25nt Hamming distance from the reference structure.

The peak around 25nt in Figure 5, however, may not form a concrete sub-optimal cluster, because the peak is just the sum of the probabilities of the structures that have the similar Hamming distances around 25nt. The number of such structures is very large and those structures may distributed widely in the structure space because of the combinatorial explosion of base pairs (See the Section S7 in the additional file 1). In order to illustrate the distribution more precisely, we show in Figure 6, the 2D distribution computed by the algorithm in the section 4.3 using the cloverleaf structure (Figure 4(a)) and the CentroidFold structure (Figure 4(b)) as the references. In Figure 6 there seems to exit quite a high potential barrier between the CentroidFold structure and the cloverleaf structure. Although the biological reason why there is such a large structure cluster other than the cloverleaf structures remains unclear, it might be related to tRNA base modification, which is known to contribute to structure stability [19,20].

We also draw a distribution of 5′- 3′ distance for the tRNA sequence, which is obtained by the algorithm in the section 4.2 (Figure 7). We can see almost all the structures (more than 99.7%) have the same 5′- 3′ distance although Figure 6 implies various structures are included in the ensemble. It indicates this tRNA is expected to fold into a certain compact structure near the 5′- 3′ ends.

Discussion and conclusions

Unreliable predictions of the RNA secondary structure have been prevented us from integrated analysis of RNA based on the estimated RNA structures. The energy model of the RNA secondary structure, however, offers rich information about the target RNA if we use appropriate algorithms to extract it. Such information is useful for analyzing not only the 3D structure prediction as a natural extension of secondary structure, but also the stabilities, the interactions with the other molecules, and so on.

In this paper, we proposed a general method to construct fast and accurate algorithms to compute the exact probability distributions of integer-valued features on the energy model of RNA secondary structures. We have shown that two useful algorithms, for Hamming

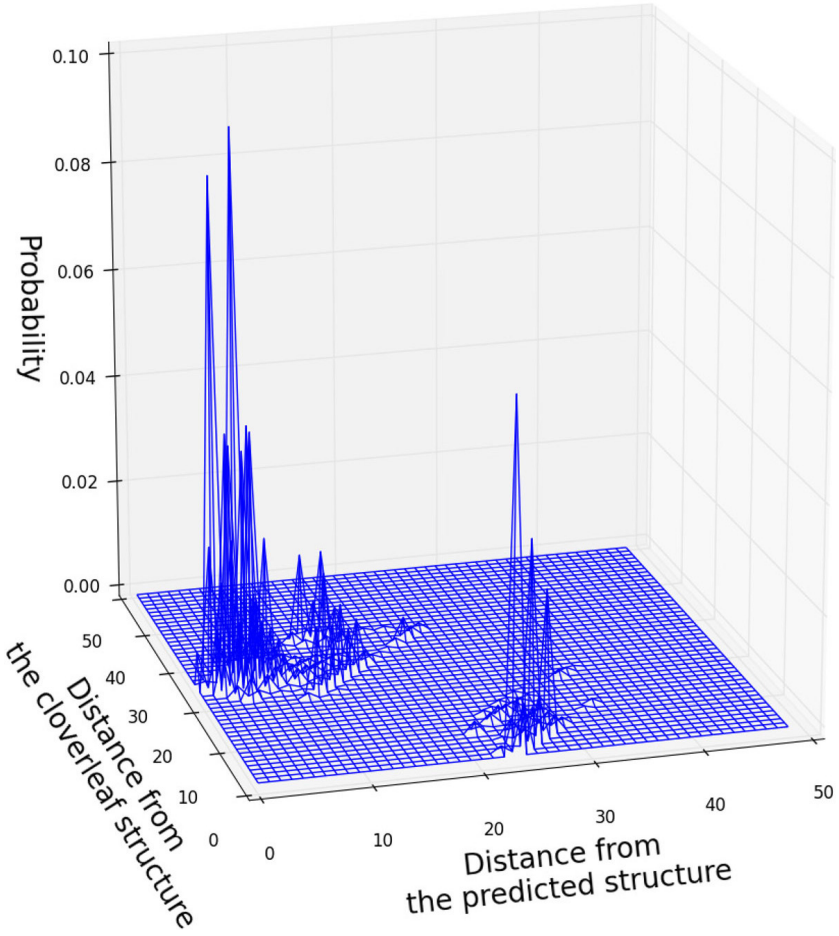

Figure 6 2D expansion of tRNA structure existence probability landscape. The landscape is drawn from the cloverleaf structure (Fig. 4(a)) and the predicted γ-centroid structure(Fig. 4(b)). We can see isolated population clusters around the both structures respectively.

distance from a reference structure and for 5′- 3′ distance, can be constructed by just assigning the score functions $gk(\cdot)$. We extended the general method of an integer score to the method of an integer vector (2D), for the distributions of Hamming distances from two reference structures. We also applied those algorithms to tRNA as an example, and demonstrated the usefulness of observing the landscapes of probability distributions of the features. Although in some applications there have been proposed practically equivalent algorithms, the proposed method offers a clear and comprehensive guideline to design algorithms for a wide variety of integer features. The web server for the distributions of the Hamming distances is available at http://rtools.cbrc.jp/cgi-bin/index.cgi. We don't show the precise implementations for the other applications, but the proposed method is applicable to the integer features such as number of base pairs, or frequency of specific structure motifs by a little modification of original McCaskill model. In addition, distributions of combination of different integer scores can be also visualized by applying the 2D expanding technique described in the previous section.

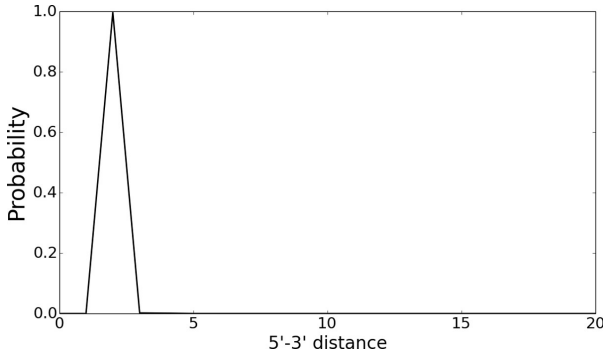

Figure 7 The 5′- 3′ distance distribution of the tRNA. Although Fig. 6 implies this RNA can be fold into various structures, almost all the elements have the common feature from the point of view of 5′-3′ distance.

Additional material

Additional file 1: Supplementary.pdf. We explained the detail of our algorithms or a little ingenuity in this file.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
Ryota Mori constructed and implemented general and concrete algorithms, analyzed and interpreted the data, and drafted the article. Michiaki Hamada helped and advised Ryota Mori throughout the study especially when Ryota Mori confronted technical difficulties. Kiyoshi Asai designed the constitution of the article, headed the critical revision of the content, and gave final approval of the article.

Acknowledgements
The authors thank to Toutai Mituyama and Yukiteru Ono for their help in integration of the software to the web page. The authors also thank to Hisanori Kiryu, Tomoshi Kameda, Junichi Iwakiri for useful discussions. The authors also thank to Ivo Hofacker *et al*. who developed Vienna RNA Package. This work was supported by JSPS KAKENHI Grant Numbers 13J06668, 24680031, 25240044, and MEXT KAKENHI Grant Number 221S0002.

Declarations
The publication charges for this work were funded by a Grant-in-Aid for Young Scientists (13J06668).
This article has been published as part of *BMC Genomics* Volume 15 Supplement 10, 2014: Proceedings of the 25th International Conference on Genome Informatics (GIW/ISCB-Asia): Genomics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/15/S10.

Authors' details
[1]Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa City, Chiba, Japan. [2]Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Koto Ward, Tokyo, Japan. [3]Faculty of Science and Engineering, Waseda University, Shinjuku Ward, Tokyo, Japan.

Published: 12 December 2014

References
1. Contrant M, Fender A, Chane-Woon-Ming B, Randrianjafy R, Vivet-Boudou V, Richer D, Pfeffer S: **Importance of the RNA secondary structure for the relative accumulation of clustered viral microRNAs.** *Nucleic Acids Research* 2014, **42(12)**:7981-7996.
2. Wan Y, Qu K, Zhang QC, Flynn RA, Manor O, Ouyang Z, Zhang J, Spitale RC, Snyder MP, Segal E, Chang HY: *Nature* 7485, 706-709.
3. Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH: **Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure.** *Proc Natl Acad Sci USA* 2004, **101(19)**:7287-7292.
4. Turner DH, Mathews DH: **NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure.** *Nucleic Acids Res* 2010, **38(Database)**:280-282.
5. Carvalho LE, Lawrence CE: **Centroid estimation in discrete high-dimensional spaces with applications in biology.** *Proc Natl Acad Sci USA* 2008, **105(9)**:3209-3214.
6. Newberg LA, Lawrence CE: **Exact calculation of distributions on integers, with application to sequence alignment.** *J Comput Biol* 2009, **16(1)**:1-18.
7. McCaskill JS: **The equilibrium partition function and base pair binding probabilities for RNA secondary structure.** *Biopolymers* 1990, **29(6-7)**:1105-1119.
8. Freyhult E, Moulton V, Clote P: **RNAbor: a web server for RNA structural neighbors.** *Nucleic Acids Res* 2007, **35(Web Server)**:305-309.
9. Senter E, Sheikh S, Dotu I, Ponty Y, Clote P: **Using the fast fourier transform to accelerate the computational search for RNA conformational switches.** *PLoS ONE* 2012, **7(12)**:50506.
10. Lorenz R, Flamm C, Hofacker IL: **2D Projections of RNA Folding Landscapes**.
11. Senter E, Dotu I, Clote P: **Efficiently computing the 2D energy landscape of RNA.** *Math Biol* 2014.
12. Ding Y, Chan CY, Lawrence CE: **RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble.** *RNA* 2005, **11(8)**:1157-1166.
13. Serganov A, Polonskaia A, Phan AT, Breaker RR, Patel DJ: **Structural basis for gene regulation by a thiamine pyrophosphate-sensing riboswitch.** *Nature* 2006, **441(7097)**:1167-1171.
14. Yoffe AM, Prinsen P, Gelbart WM, Ben-Shaul A: **The ends of a large RNA molecule are necessarily close.** *Nucleic Acids Res* 2011, **39(1)**:292-299.
15. Clote P, Ponty Y, Steyaert JM: **Expected distance between terminal nucleotides of RNA secondary structures.** *J Math Biol* 2012, **65(3)**:581-599.
16. Han HS, Reidys CM: **The 5'-3' distance of RNA secondary structures.** *J Comput Biol* 2012, **19(7)**:867-878.
17. Hamada M, Kiryu H, Sato K, Mituyama T, Asai K: **Prediction of RNA secondary structure using generalized centroid estimators.** *Bioinformatics* 2009, **25(4)**:465-473.
18. Puton T, Rother K, Kozlowski L, Tkalinska E, Bujnicki J: **A Server for Continuous Benchmarking of Automated Methods for RNA Structure Prediction.** 2011.
19. Copela LA, Chakshusmathi G, Sherrer RL, Wolin SL: **The La protein functions redundantly with tRNA modification enzymes to ensure tRNA structural stability.** *RNA* 2006, **12(4)**:644-654.
20. Tuorto F, Liebers R, Musch T, Schaefer M, Hofmann S, Kellner S, Frye M, Helm M, Stoecklin G, Lyko F: **RNA cytosine methylation by Dnmt2 and NSun2 promotes tRNA stability and protein synthesis.** *Nat Struct Mol Biol* 2012, **19(9)**:900-905.

doi:10.1186/1471-2164-15-S10-S6
Cite this article as: Mori *et al.*: Efficient calculation of exact probability distributions of integer features on RNA secondary structures. *BMC Genomics* 2014 **15**(Suppl 10):S6.