Abstract: Network properties of the ensemble of RNA structures

Peter Clote

Boston College, Chestnut Hill, MA 02467

Abstract

A neighbor of the RNA secondary structure s is obtained by removing, adding or shifting a base pair in s. Here, we describe the first efficient algorithm to compute the expected number of neighbors for the collection of all secondary structures of a given RNA sequence. This surprisingly complex algorithm permits a better understanding of kinetics of RNA folding when allowing defect diffusion, helix zippering, and related conformation transformations. Moreover, only when allowing shift moves does the network of secondary structures for certain RNAs satisfy the requirements of a small-world network.

RNA secondary structure kinetics is plays an essential role in certain biological processes, such as the *hok/sok* host-killing/suppression of killing (*hok/sok*) system that kills *E. coli* replicates if insufficient plasmids are transfered to the new daughter cell Nevertheless, RNA folding kinetics remains a difficult problem, since it is known that computation of optimal folding pathways is NP-complete [3].

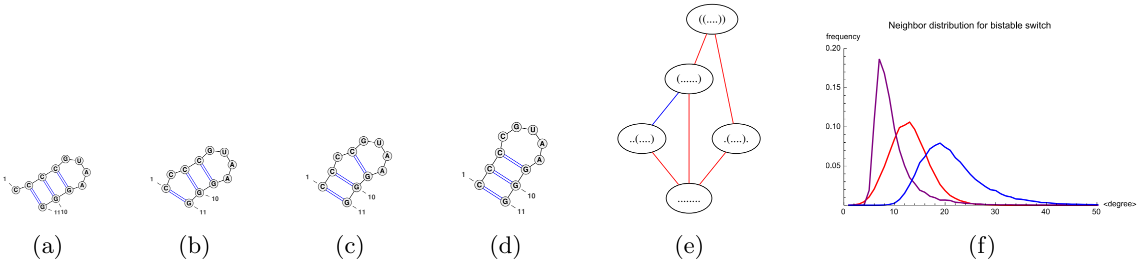
Due to the biological importance of RNA folding kinetics, users generally run a secondary structure kinetics program, such as `Kinfold`, `Kinefold`, `RNAKinetics`. However, repeated simulations must be performed, each requiring lengthy computation times – for instance, the population occupancy curve for yeast phe-tRNA required 3 months of CPU time on a 2.4 GHz Intel Pentium 4 running linux [4]). Coarse-grained approaches using spectral methods also exist, such as `Treekin`, basin hopping with `RNAlocmin`, and `Hermes`.

Shift moves allow a transition from secondary structure s to structure t in which the base pair (i, j) of s is modified while fixing one base; i.e. base pair transitions of the form $(i, j) \to (i, k)$ or $(i, j) \to (k, j)$. Panels (a)-(d) of Figure 1 depicts a particular type of shift move known as *defect diffusion*. Base pair addition, removal and shift moves constitute the default move set employed by the program `Kinfold` [2], with respect to which Wuchty [5] showed that the network of secondary structures of *E. coli* phe-tRNA (Sprinzl accession RF6280) is a *small-world network*.

The move set MS1 [resp. MS2] consists of base pair additions/removals [resp. additions/removals/shifts]. The network of the toy sequence ACGUACGU is illustrated in Figure 1(e), and the distribution of the number of neighbors of each structure for the 32 nt selenocysteine insertion sequence fruA is depicted in Figure 1(f). In this abstract, we describe an approach to efficiently compute the expected degree of an RNA network of secondary structures. Our work generalizes a recent paper [1], which describes a vastly simpler algorithm to compute the expected degree without consideration of shift moves. Since our algorithm is surprisingly complex, we state the recursions for the RNA *homopolymer* model and leave the extensions to the general Turner nearest neighbor energy model to the journal version of this paper.

We now sketch the approach taken. Let $\mathbf{a} = a_1, \ldots, a_n$ be an arbitrary but fixed RNA sequence. For any $1 \leq i \leq j \leq n$, let $a[i, j]$ denote the subsequence $a_i, \ldots, a_j$. A secondary structure on $a[i, j]$ is a set of non-crossing base pairs (x, y), for $i \leq x < y \leq j$, where (a, b) and (c, d) are crossing if $a < c < b < d$. A base pair (x, y) is *external* if there is no base pair (u, v) for which $u < x < y < v$; a position x is *visible* if there is no base pair (u, v) such that $u \leq x \leq v$. The set of all secondary structures on $a[i, j]$ is denoted by $\mathbb{SS}[i, j]$. Define $Q_{i,j} = \sum_{s \in \mathbb{SS}[i,j]} \exp(-E(s)/RT) \cdot N(s)$, where $N(s)$ is the number of secondary structures t of $a[i, j]$ obtained from the structure s by the addition, deletion or shift of a base pair. The partition function for $a[i, j]$ is defined by $Z_{i,j} = \sum_{s \in \mathbb{SS}[i,j]} \exp(-E(s)/RT)$. It follows that the expected number of neighbors (network degree) is $\frac{Q_{1,n}}{Z_{1,n}}$.

For simplicity, we state the recursions for $Q_{1,n}$ and $Z_{1,n}$ for the *homopolymer model*, in which any two positions $1 \leq i < j \leq n$ can form a base pair, provided only that $i + 1 < j$. For the homopolymer model, there is no RNA sequence $\mathbf{a} = a_1, \ldots, a_n$, but rather only the interval $[1, n] = \{1, \ldots, n\}$. Thus we speak of



(a) (b) (c) (d) (e) (f)

Figure 1: *(a-d):* Example of successive shift moves, corresponding to *defect diffusion*. *(e):* Network for the toy 8-mer ACGUACGU which has 5 nodes and 6 edges (hence 12 directed edges). The expected network degree is $\frac{12}{5} = 2.4$. Red edges indicate base pair addition or removal, while blue edges indicate shift moves. *(f):* Relative frequency for number of neighbors (degree) for the network of all secondary structures of the 25 nt bistable switch UGUACCGGAA GGUGCGAAUC UUCCG produced by exhaustive enumeration. The blue [resp. red resp. purple] curve corresponds to move set M2 [resp. (M2-M1) resp. M1]. Brute force analysis of the collection of all 83725 possible structures yields an expected network degree of 20.71 ± 6.91 [resp. 11.94 ± 3.93 resp. 8.77 ± 4.30] for move set MS2 [resp. MS2-MS1 resp. MS1].

a structure on $[i, j]$, rather than on $a[i, j]$. The energy of each structure in the homopolymer model is zero, so the probability of each structure s on $[i, j]$ equals one divided by the number of structures on $[i, j]$.

For $0 \leq n$, define Q_n to be the sum, taken over all structures s of $[1, n]$, of the number of base pair additions, removals or shifts of a base pair of s. Let Z_n denote the total number of homopolymer structures on $[1, n]$, where any two positions i, j can base-pair, as long as $j - i > 1$. Define $f(n, x)$ to be the number of secondary structures s for a length n homopolymer, such that s has x visible positions. Define $g(n, x)$ to be the number of secondary structures s for the length n homopolymer, such that s has x visible positions in the interval $[1, n - \theta - 1] = [1, n - 2]$, and position n is unpaired in s. Define the function E_n to be the number of *external base pairs* in all homopolymer structures on $[1, n]$. In the journal version of this paper, we give recursions for f, g, E and prove that

$$\begin{aligned} Q_n &= Q_{n-1} + 2 \sum_{k=0}^{n-\theta-2} Z_{k-1} \cdot Z_{n-k-1} + 2 \left(E_{n-1} + E_{n-2} + \sum_{r=1}^{n-4} E_r \cdot Z_{n-r-3} \right) + \\ & \sum_{x=2}^{n-\theta-1} x(x-1) \cdot g(n, x) + \sum_{k=1}^{n-\theta-1} (Z_{k-1} \cdot Q_{n-k-1}) + (Q_{k-1} \cdot Z_{n-k-1}) \end{aligned}$$

thus resulting in a cubic time algorithm to compute the expected network degree with respect to base pair additions, removals and shifts in the homopolymer case. The full paper also provides an even more complex extension to the Turner energy model.

Finally, we would like to thank the referees for valuable suggestions in a preliminary version of this abstract. This research was funded by the National Science Foundation grant DBI-1262439. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] P Clote. Expected degree for RNA secondary structure networks. *J Comp Chem*, 36(2):103–17, Jan 2015.

[2] C. Flamm, W. Fontana, I.L. Hofacker, and P. Schuster. RNA folding at elementary step resolution. *RNA*, 6:325–338, 2000.

[3] C. Thachuk, J. Maňuch, L. Stacho, and A. Condon. NP-completeness of the direct energy barrier height problem. *Natural Computing*, 10(1):391–405, 2011.

[4] M. Wolfinger, W.A. Svrcek-Seiler1, C. Flamm, and P.F. Stadler. Efficient computation of RNA folding dynamics. *J Phys. A: Math. Gen.*, 37:4731–4741, 2004.

[5] S. Wuchty. Small worlds in RNA structures. *Nucleic. Acids. Res.*, 31(3):1108–1117, February 2003.