Supplementary Text for Rfold: An exact algorithm for computing local base pairing probabilities

Hisanori Kiryu, Taishin Kin, and Kiyoshi Asai

November 26, 2007

1 The Rfold Model

The state transition rules of the Rfold model that incorporates the full energy model is defined by,

$$\begin{aligned}
\text{Outer} &\longrightarrow \epsilon|\text{Outer}\cdot a|\text{Outer}\cdot\text{Stem}\\
\text{Stem} &\longrightarrow b_< \cdot \text{Stem}\cdot b_>|b_<\cdot\text{StemEnd}\cdot b_>\\
\text{StemEnd} &\longrightarrow s_n|s_m\cdot\text{Stem}\cdot s_n(m+n>0)|\text{Multi}\\
\text{Multi} &\longrightarrow a\cdot\text{Multi}|\text{MultiBif}\\
\text{MultiBif} &\longrightarrow \text{Multi1}\cdot\text{Multi2}\\
\text{Multi1} &\longrightarrow \text{MultiBif}|\text{Multi2}\\
\text{Multi2} &\longrightarrow \text{Multi2}\cdot a|\text{Stem}
\end{aligned}$$

where ϵ represents the null terminal symbol; a, an unpaired nucleotide character; s_k, an unpaired base string of length k; and ($b_<$, $b_>$), a base pair. There are 7 non terminal symbols Outer, Stem, StemEnd, Multi, MultiBif, Multi1, and Multi2 in the grammar. Outer emits outer bases as in the main text. Stem emits all the base pairs. StemEnd represents the end of each stem from which either a hairpin loop (StemEnd $\rightarrow$ s_n), an interior/bulge loop (StemEnd $\rightarrow$ $s_m\cdot\text{Stem}\cdot s_n$), or a multiloop (StemEnd $\rightarrow$ Multi) is emitted. Multi represents a complete multiloop. Multi2, Multi1, and MultiBif represent parts of a multiloop structure that contain exactly one, one or more, and two or more base pairs in the loop, respectively.

We have described the unambiguity of the grammar in parsing outer regions in the main text. As shown in Figure1, it is easy to see unambiguity also holds for the structures that do not contain multiloops.

For multiloops, we illustrate the unambiguity of the grammar by an example shown in Figure2. As shown in the figure, the 5'-most segment of unpaired bases (denoted by (a)) is emitted by successive Multi $\rightarrow$ Multi transitions after

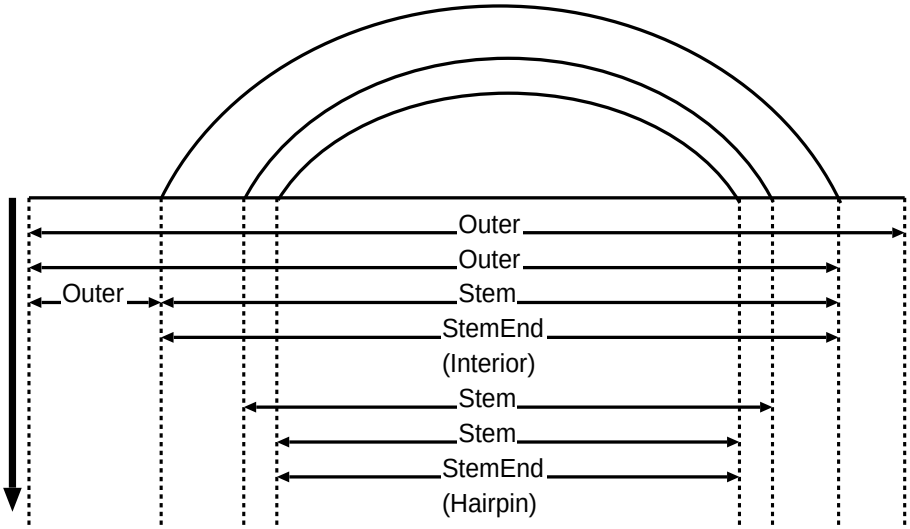

Figure 1: Derivation of a secondary structure with no multiloops. The base pairs are represented by arcs. For each state, the sequence segment that is parsed by the descendants of the state in the parse tree is shown by horizontal arrow. State transitions occur from top to bottom (thick arrow).

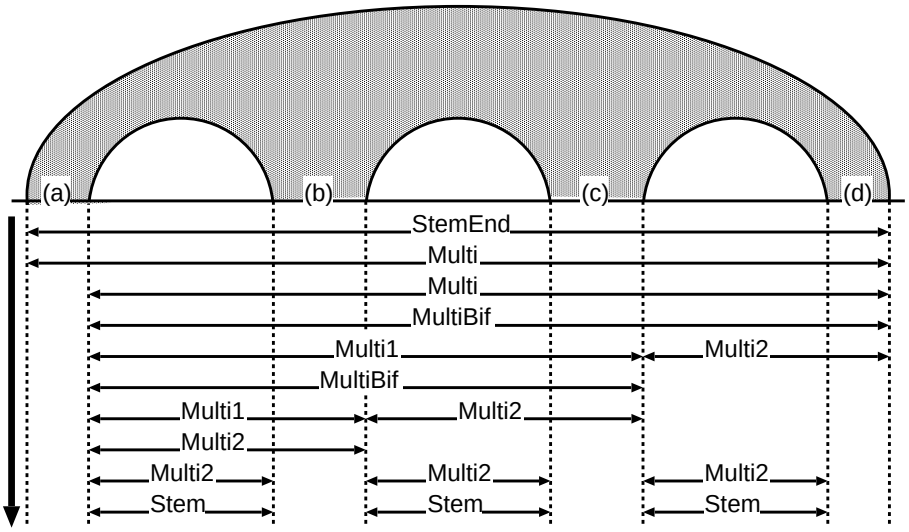

Figure 2: Derivation of a multiloop structure. The base pairs are represented by the arcs. and the multiloop is shown as the shaded region.

a StemEnd → Multi transition. All the other unpaired segments ((b),(c), and (d)) are emitted by successive Multi2 → Multi2 transitions. In general, an arbitrary multiloop is unambiguously parsed in the same manner because of the following properties of the grammar:

- Multi is the only state that emits unpaired bases on the left hand side.
- Multi2 is the only state that emits unpaired bases on the right hand side.
- StemEnd is the unique direct parent of Multi other than itself
- The 5'-most bases of MultiBif, Multi1, and Multi2 are always the left bases of some base pairs.

2 The Inside and Outside Algorithms

The inside algorithm for the grammar in the previous section is as follows;

$$\alpha_{\mathrm{Stem}}(i,j) = \sum \begin{cases} \alpha_{\mathrm{Stem}}(i+1,j-1) \cdot t(\mathrm{Stem} \to \mathrm{Stem}) \\ \alpha_{\mathrm{StemEnd}}(i+1,j-1) \cdot t(\mathrm{Stem} \to \mathrm{StemEnd}) \end{cases}$$

$$\alpha_{\mathrm{MultiBif}}(i,j) = \sum \begin{cases} \alpha_{\mathrm{Multi1}}(i,k) \cdot \alpha_{\mathrm{Multi2}}(k,j) \cdot t(\mathrm{MultiBif} \to \mathrm{Multi1} \cdot \mathrm{Multi2}) \\ \text{for } i < k < j \end{cases}$$

$$\alpha_{\mathrm{Multi2}}(i,j) = \sum \begin{cases} \alpha_{\mathrm{Stem}}(i,j) \cdot t(\mathrm{Multi2} \to \mathrm{Stem}) \\ \alpha_{\mathrm{Multi2}}(i,j-1) \cdot t(\mathrm{Multi2} \to \mathrm{Multi2}) \end{cases}$$

$$\alpha_{\mathrm{Multi1}}(i,j) = \sum \begin{cases} \alpha_{\mathrm{Multi2}}(i,j) \cdot t(\mathrm{Multi1} \to \mathrm{Multi2}) \\ \alpha_{\mathrm{MultiBif}}(i,j) \cdot t(\mathrm{Multi1} \to \mathrm{MultiBif}) \end{cases}$$

$$\alpha_{\mathrm{Multi}}(i,j) = \sum \begin{cases} \alpha_{\mathrm{Multi}}(i+1,j) \cdot t(\mathrm{Multi} \to \mathrm{Multi}) \\ \alpha_{\mathrm{MultiBif}}(i,j) \cdot t(\mathrm{Multi} \to \mathrm{MultiBif}) \end{cases}$$

$$\alpha_{\mathrm{StemEnd}}(i,j) = \sum \begin{cases} t(\mathrm{StemEnd} \to (\mathrm{Hairpin})) \\ \alpha_{\mathrm{Stem}}(i',j') \cdot t(\mathrm{StemEnd} \to (\mathrm{Interior}) \to \mathrm{Stem}) \\ \text{for } i \le i' < j' \le j,\ 0 < (j-j') + (i'-i) \le C \\ \alpha_{\mathrm{Multi}}(i,j) \cdot t(\mathrm{StemEnd} \to \mathrm{Multi}) \end{cases}$$

$$\alpha_{\mathrm{Outer}}(j) = \sum \begin{cases} 1 \text{ if } j = 0 \\ \alpha_{\mathrm{Outer}}(j-1) \cdot t(\mathrm{Outer} \to \mathrm{Outer}) \\ \alpha_{\mathrm{Outer}}(k) \cdot \alpha_{\mathrm{Stem}}(k,j) \cdot t(\mathrm{Outer} \to \mathrm{Outer} \cdot \mathrm{Stem}) \\ \text{for } (j-W) < k < j \end{cases}$$

where C denotes the maximum length of interior/bulge loops. $\alpha_s(\dots)$ represent the inside variables for the state s. $t(s \to s')$ represent the transition scores associated with the state transition $s \to s'$, which are functions of the energy

parameters. We have not explicitly written the position (i, j, etc.) dependencies of $t(s \to s')$ for simplicity of notation.

The corresponding outside algorithm is given by,

$$\beta_{\text{Outer}}(j) = \sum \begin{cases} 1 \text{ if } j = N+1 \\ \beta_{\text{Outer}}(j, j+1) \cdot t'(\text{Outer} \to \text{Outer}) \\ \alpha_{\text{Stem}}(j,k) \cdot \beta_{\text{Outer}}(k) \cdot t'(\text{Outer} \to \text{Outer} \cdot \text{Stem}) \\ \text{ for } j < k < (j+W) \end{cases}$$

$$\beta_{\text{StemEnd}}(i,j) = \beta_{\text{Stem}}(i-1, j+1) \cdot t'(\text{Stem} \to \text{StemEnd})$$

$$\beta_{\text{Multi}}(i,j) = \sum \begin{cases} \beta_{\text{StemEnd}}(i,j) \cdot t'(\text{StemEnd} \to \text{Multi}) \\ \beta_{\text{Multi}}(i-1,j) \cdot t'(\text{Multi} \to \text{Multi}) \end{cases}$$

$$\beta_{\text{Multi1}}(i,j) = \sum \begin{cases} \beta_{\text{MultiBif}}(i,k) \cdot \alpha_{\text{Multi2}}(j,k) \cdot t'(\text{MultiBif} \to \text{Multi1} \cdot \text{Multi2}) \\ \text{ for } j < k < (i+W) \end{cases}$$

$$\beta_{\text{Multi2}}(i,j) = \sum \begin{cases} \beta_{\text{Multi2}}(i,j+1) \cdot t'(\text{Multi2} \to \text{Multi2}) \\ \beta_{\text{Multi1}}(i,j) \cdot t'(\text{Multi1} \to \text{Multi2}) \\ \beta_{\text{MultiBif}}(k,i) \cdot \alpha_{\text{Multi1}}(k,j) \cdot t'(\text{MultiBif} \to \text{Multi1} \cdot \text{Multi2}) \\ \text{ for } (j-W) < k < i \end{cases}$$

$$\beta_{\text{MultiBif}}(i,j) = \sum \begin{cases} \beta_{\text{Multi1}}(i,j) \cdot t'(\text{Multi1} \to \text{MultiBif}) \\ \beta_{\text{Multi}}(i,j) \cdot t'(\text{Multi} \to \text{MultiBif}) \end{cases}$$

$$\beta_{\text{Stem}}(i,j) = \sum \begin{cases} \alpha_{\text{Outer}}(i) \cdot \beta_{\text{Outer}}(j) \cdot t'(\text{Outer} \to \text{Outer} \cdot \text{Stem}) \\ \beta_{\text{StemEnd}}(i',j') \cdot t'(\text{StemEnd} \to (\text{Interior}) \to \text{Stem}) \\ \text{ for } i' \le i < j \le j',\ 0 < (i-i') + (j'-j) \le C \\ \beta_{\text{Multi2}}(i,j) \cdot t'(\text{Multi2} \to \text{Stem}) \\ \beta_{\text{Stem}}(i-1, j+1) \cdot t'(\text{Stem} \to \text{Stem}) \end{cases}$$

where $\beta_s(\dots)$ represent the outside variables for state s. $t'(s \to s')$ represent the transition scores associated with the outside algorithm that are determined such that $\alpha_s(X) \cdot \beta_s(X)/Z$ ($X = (i,j)$ or i) are the sums of the posterior probabilities of all the parse trees that pass through the cell X as state s. Here, $Z = \alpha_{\text{Outer}}(N+1)$ denotes the partition function.