Supplementary material

DotKnot: pseudoknot prediction using the probability dot plot under a refined energy model

Jana Sperschneider and Amitava Datta
janaspe@csse.uwa.edu.au

Algorithm 1: MWIS calculation for finding promising secondary structure elements, i.e. bulge loops, internal loops and multiloops with low free energy.

Input: A set of weighted stem intervals $D_s = \{s_1, s_2, \ldots, s_n\}$ where $s_i = [a_i : b_i]$. The sorted endpoints list $L = \{e_1, e_2, \ldots, e_{2n}\}$ for all stem intervals is given. The weight of a stem $s_i$ is its confidence indicator $c(s_i)$.

Output: A dictionary $D_L^s$ of stems interrupted by bulges and internal loops and a dictionary $D_M^s$ of multiloop structures. The weight of a stem $s_L^i \in D_L^s$ or $s_M^i \in D_M^s$ is its local free energy value $w(s_L^i)$.

1. $D_L^s \leftarrow \emptyset$
2. $D_M^s \leftarrow \emptyset$
3. for $j = 1$ to $2n$ do
4.   if $e_j$ is a right endpoint of corresponding stem interval $s_i$ then
5.     $S_i \leftarrow \{s_k \in D_s, s_L^i \in D_L^s | s_k, s_L^i \text{ inner stems for } s_i\}$ /* Flexible inner stem finding strategy, see Figure 1 */
6.     if $S_i$ is not empty then
7.       $S_{min}^i \leftarrow \text{MWIS}(S_i)$ /* Maximum weight independent set */
8.       $c(s_i) \leftarrow c(s_i) + \sum_{s_m \in S_{min}^i} c(s_m)$ /* Update stem weight */
9.       Add stem $s_L^i$ with internal stems $s_m \in S_{min}^i$ and updated weight $c(s_i)$ to $D_L^s$
10. end
11. end
12. forall $s_L^i \in D_L^s$ do
13.    $w(s_L^i) \leftarrow \text{local free energy evaluated by RNAeval}$
14.    if $w(s_L^i) \geq$ 0.0 kcal/mol then
15.      Delete $s_L^i$ from $D_L^s$
16.    end
17.    if $s_L^i$ is a multiloop structure then
18.      Delete $s_L^i$ from $D_L^s$
19.      Add stem $s_L^i$ to $D_M^s$
20. end
21. end

The flexible inner stem finding procedure covers the case where the outer stem and an inner stem compete for base pairs. Given an outer stem, a base pair overlap of at most two is allowed as long as this does not result in (inner or outer) stem sizes smaller than three base pairs (Figure 1). In other words, if one of the stems loses the shared base pairs, it still has to have length of at least three base pairs. There is no certainty about whether the outer or inner stem should lose some of its base pairs. If an overlap occurs, one of the stems is shortened in an arbitrary manner. Here, the inner stem is chosen as the one which loses its base pairs.
Figure 1: Flexible inner stem finding for an outer stem. In both cases, the outer and stem compete for two base pairs. (a) Both outer and inner stem have length longer than five base pairs, therefore an overlap of two base pairs is allowed. (b) The outer stem has length of four base pairs, therefore an overlap of two base pairs is not allowed. For inner and outer stem lengths of at least four base pairs an overlap of one base pair is allowed.

Example

This is an example for the 5SrRNA E.coli sequence. Note that not all possible stems are included for ease of presentation.

>5SColi
UGCCUGCCGGUCUUGGUGUGUGGUGUCCACUUGAGGGACUGGAAUGGUGAAGACUGCCUGCUUGGUAGUGUGGUGGUCCCAUGGGAGUGAAGGAACUGCCAGGCAU

Stem intervals displayed as intervals on the line.

```
1
20 0.78 31
10 0.16 22
48 0.37 57
46 0.31 55
61 0.66 77
60 0.19 69
88 0.37 108
89 0.37 107
91 0.41 107
95 0.91 104
```

Input

\[ D_s = \{ s_1, s_2, \ldots, s_{11} \} \]
\[ = \{ [1 : 119], [10 : 22], [20 : 31], [33 : 88], [46 : 55], [48 : 57], [60 : 69], [61 : 77], [89 : 108], [91 : 107], [95 : 104] \} \]

with corresponding confidence weights \( c(s_i) \) for each \( s_i = [a_i : b_i] \).

The sorted endpoints list \( L = \{ 1, 10, 20, 22, 31, 33, 46, 48, 55, 57, 61, 77, 88, 89, 91, 95, 104, 107, 108, 119 \} \) is given.
Calculation

| Endpoint | $s_i$ | $S_i$ | $S_{\text{min}}^i$ | $c(s_i)$ | $D_{L}^i$ |
|----------|------|------|-----------------|-----------|----------|
| 1        | $s_1 = [1 : 119]$ | - | - | - | - |
| 10       | $s_{10} = [10 : 22]$ | - | - | - | - |
| 20       | $s_9 = [20 : 31]$ | - | - | - | - |
| 22       | $s_2 = [10 : 22]$ | $\emptyset$ | - | - | - |
| 31       | $s_3 = [20 : 31]$ | $\emptyset$ | - | - | - |
| 33       | $s_4 = [33 : 88]$ | - | - | - | - |
| 46       | $s_5 = [46 : 55]$ | - | - | - | - |
| 48       | $s_6 = [48 : 57]$ | - | - | - | - |
| 55       | $s_5 = [46 : 55]$ | $\emptyset$ | - | - | - |
| 57       | $s_6 = [48 : 57]$ | $\emptyset$ | - | - | - |
| 60       | $s_7 = [60 : 69]$ | - | - | - | - |
| 61       | $s_8 = [61 : 77]$ | - | - | - | - |
| 69       | $s_7 = [60 : 69]$ | - | - | - | - |
| 77       | $s_8 = [61 : 77]$ | $\emptyset$ | - | - | - |
| 88       | $s_4 = [33 : 88]$ | $\{s_5, s_6, s_7, s_8\}$ | $\{s_6, s_8\}$ | 1.84 | $D_{L}^i = \{s_{4}^L\}$ |
| 89       | $s_9 = [89 : 108]$ | - | - | - | $D_{L}^i = \{s_9^L\}$ |
| 91       | $s_{10} = [91 : 107]$ | - | - | - | $D_{L}^i = \{s_{10}^L\}$ |
| 95       | $s_{11} = [95 : 104]$ | - | - | - | $D_{L}^i = \{s_{11}^L\}$ |
| 104      | $s_{12} = [95 : 104]$ | - | - | - | $D_{L}^i = \{s_{12}^L\}$ |
| 107      | $s_{13} = [91 : 107]$ | $\{s_{11}\}$ | $\{s_{11}\}$ | 1.32 | $D_{L}^i = \{s_{13}^L, s_{12}^L\}$ |
| 108      | $s_9 = [89 : 108]$ | $\{s_{11}\}$ | $\{s_{11}\}$ | 1.28 | $D_{L}^i = \{s_{10}^L, s_{11}^L\}$ |
| 119      | $s_1 = [1 : 119]$ | $\{s_2, \ldots, s_{11}, s_{4}^L, s_{10}^L, s_{9}^L\}$ | $\{s_3, \ldots, s_4^L, s_{10}^L\}$ | 4.86 | $D_{L}^i = \{s_{4}^L, s_{10}^L, s_{9}^L, s_{1}^L\}$ |

Output

As an output, the two dictionaries $D_{L}^s = \{s_{10}^L, s_9^L\}$ and $D_{M}^s = \{s_{4}^L, s_{1}^L\}$ are returned. In dot-bracket notation, the items in the two dictionaries look as follows:

$s_{10}^L$

> 91 107

CCCAUGCGAGAGUAGGG

((((.(((....)))))) ( -8.40)

$s_9^L$

> 89 108

UCCCCAUUGCGAGAUGGGA

(((((((.)))))))) ( -8.20)

$s_{11}^L$

> 33 88

GACCCCAUGCGAGAUGGGA

((((((((.)))))))) ( -20.90)

$s_{12}^L$

> 119

CCCAUGCGAGAUGGGA

((((((((.)))))))) ( -54.90)

Note that the structure $s_{12}^L$ corresponds to the minimum free energy structure for the 5SrRNA sequence as predicted by RNAfold.
Figure 2: During pseudoknot construction the two stems may compete for a base pair at the interhelix junction and three scenarios are possible: (a) loop $L_1$ can have a length of 0 nt, (b) loop $L_3$ can have a length of 1 nt or (c) loops $L_1$ and $L_3$ are both longer than the minimum lengths of 1 and 2 nt, respectively.

In case of an overlap at loop $L_2$, one of the stems is truncated according to certain rules. For pseudoknots with regular stems, an overlap of one or two base pairs is allowed. If the stems compete for a base pair at the junction, the conflict is resolved, if possible, using the stem shortening rules described in Algorithm 2. For an overlap of two base pairs, the rules are applied in an iterative fashion.

For pseudoknots with one interrupted and one regular stem, an overlap at interhelix loop $L_2$ of only one base pair is considered. If such an overlap occurs, the regular stem is shortened if this results in a stem length of at least 3 bp.

Note that in general a shortened stem length of 2 bp for either $S_1$ or $S_2$ is not allowed as a result.

**Algorithm 2**: Pseudoknot construction with regular stems: overlap of one base pair at $L_2$.

```plaintext
if stems $S_1$ and $S_2$ compete for one base pair at loop $L_2$ then
  if $|L_1| = 0$ nt $\land |S_1| > 3$ then
    Cut stem $S_1$ by one base pair
  end
  else if $|L_3| = 1$ nt $\land |S_2| > 3$ then
    Cut stem $S_2$ by one base pair
  end
  else if $|L_1| \geq 1$ nt $\land |L_3| \geq 2$ nt then
    if $|S_1| > 3$ then
      Cut stem $S_1$ by one base pair
    end
    else if $|S_2| > 3$ then
      Cut stem $S_2$ by one base pair
    end
  end
end

Recalculate the local energy weights for shortened stem $S_1$ or $S_2$
```
Algorithm 3: Recursive secondary structure formation in pseudoknot loops

Input: A set of pseudoknot candidates stored in the dictionary $D_p = \{p_1, \ldots, p_n\}$, the set of stems interrupted by bulge loops and internal loops $D_s^L$ and the set of multiloop structures $D_s^M$.

Output: A set of pseudoknot candidates stored in the dictionary $D_p = \{p_1, \ldots, p_n\}$ with secondary structure elements in each the three pseudoknot loops $L_1$, $L_2$, $L_3$. Each pseudoknot candidate $p_i$ has three (possible empty) lists $\text{list}_{i}^{\min}$, $\text{list}_{2}^{\min}$ and $\text{list}_{3}^{\min}$ which hold the set of secondary structures with best local free energy that fold in the three loops $L_1$, $L_2$ and $L_3$, respectively.

foreach $p_i \in D_p$ do
    list$_1$ ←− all stems $s_i \in D_s$, $s_i^L \in D_s^L$ and $s_i^M \in D_s^M$ contained in $L_1$
    list$_2$ ←− all stems $s_i \in D_s$, $s_i^L \in D_s^L$ and $s_k^M \in D_s^M$ contained in $L_2$
    list$_3$ ←− all stems $s_i \in D_s$, $s_i^L \in D_s^L$ and $s_k^M \in D_s^M$ contained in $L_3$
    if list$_1$ not empty then
        list$_{1i}^{\min}$ ←− MWIS(list$_1$) /* Maximum weight independent set */ /* with absolute local free energy weights */
    end
    if list$_2$ not empty then
        list$_{2i}^{\min}$ ←− MWIS(list$_2$) /* Maximum weight independent set */ /* with absolute local free energy weights */
    end
    if list$_3$ not empty then
        list$_{3i}^{\min}$ ←− MWIS(list$_3$) /* Maximum weight independent set */ /* with absolute local free energy weights */
    end
    Store recursive secondary structure elements list$_1^{\min}$, list$_2^{\min}$, list$_3^{\min}$ for $p_i$.
end