ILP formulation

- q is an upper bound of the total quantity of MDS (*lenght/2*)
- `=` is string equivalence
- `MIC[i,j]` (`MAC[i,j]`) is the substring starting at `i` and finishing at `j` (`i`,`j` being positions) of the MIC(MAC). Can be trivially defined using string concatenation and *MIC(i,c)* (*MAC(i,c)*).
- `reverse_complement(String)` is the Watson-Crick reverse complement of `String`
- Size of the Oxytricha Input genome: MIC is fragmented into ~750 000 MDSs, MAC into 300 000.
- Variables marked with $*$ are populated during the preprocessing phase.

objective function: $$min \sum_{i,j} MDS_{MACstart}(i,j)$$

Variables definitions

$$*Eq(i,j,h,l) = \begin{cases} 0 \\ 1, & \text{if } \texttt{MIC[i:j]} = \texttt{MAC[h:l]} \end{cases}$$

$$*cwc(i,j,h,l) = \begin{cases} 0 \\ 1, & \text{if } \texttt{MIC[i:j]} \text{ is the reverse complement of } \texttt{MAC[h:l]} \end{cases}$$

$$*Possible_{MDSMAC}(i,a,b) = \begin{cases} 0 \\ 1, & \text{if MDS } i \text{ can start at } a \text{ and finish at } b \text{ in the MAC} \end{cases}$$

$$*Possible_{MDSMIC}(i,a,b) = \begin{cases} 0 \\ 1, & \text{if MDS } i \text{ can start at } a \text{ and finish at } b \text{ in the MIC} \end{cases}$$

$$Possible_{assignment}(a,b,c,d) = Eq(a,b,c,d)$$

$$MDS_{MICstart}(i,j) = \begin{cases} 0 \\ 1, & \text{if MDS } i \text{ starts at position } j \text{ in the MIC} \end{cases}$$

$$MDS_{MICend}(i,j) = \begin{cases} 0 \\ 1, & \text{if MDS } i \text{ ends at position } j \text{ in the MIC} \end{cases}$$

$$MDS_{MACstart}(i,j) = \begin{cases} 0 \\ 1, & \text{if MDS } i \text{ starts at position } j \text{ in the MAC} \end{cases}$$

$$MDS_{MACend}(i,j) = \begin{cases} 0 \\ 1, & \text{if MDS } i \text{ ends at position } j \text{ in the MAC} \end{cases}$$

$$Inv(i) = \begin{cases} 0 \\ 1, & \text{if MDS } i \text{ is inverted in the MAC} \end{cases}$$

$$P_{start}(i,j) = \begin{cases} 0 \\ 1, & \text{if } MDS_{MACstart}(i,j) = 1, \text{ Pointer } i \text{ starts at position } j \text{ in the MAC} \end{cases}$$

$$P_{end}(i,j) = \begin{cases} 0 \\ 1, & \text{if } MDS_{MACend}(i-1,j) = 1 \text{, Pointer } i \text{ ends at position } j \text{ in the MAC} \end{cases}$$

$$*MAC(i,c) = \begin{cases} 0 \\ 1, & \text{if } c \text{ is the character at position } i \text{ in the MAC} \end{cases}$$

$$*MIC(i,c) = \begin{cases} 0 \\ 1, & \text{if } c \text{ is the character at position } i \text{ in the MIC} \end{cases}$$

$$Cov_{MIC}(i,j) = \begin{cases} 0 \\ 1, & \text{if MDS } i \text{ covers the position } j \text{ in the MIC} \end{cases}$$

$$Cov_{MAC}(i,j) = \begin{cases} 0 \\ 1, & \text{if MDS } i \text{ covers the position } j \text{ in the MAC} \end{cases}$$

Constraints

Internally Eliminated Sequences

$$IES(j) = \begin{cases} 0 \\ 1, & \text{if } i \text{ is part of an IES: } \sum_{0 \le i \le q} Cov_{MIC}(i,j) = 0 \end{cases}$$

MDSs must correspond to identical or reverse and complemented substrings of MIC and MAC. The following constraints enforce this fact:

$$MDS_{MICstart}(i,a) + MDS_{MICend}(i,b) + MDS_{MACstart}(i,c) + MDS_{MACend}(i,d) + Inv(i) - 5cwc(a,b,c,d) = 0$$

$$MDS_{MICstart}(i,a) + MDS_{MICend}(i,b) + MDS_{MACstart}(i,c) + MDS_{MACend}(i,d) - 4Eq(a,b,c,d) = Inv(i)$$

$$\sum_j MDS_{MICstart}(i,j) \le 1$$

$$\sum_j MDS_{MICend}(i,j) = \sum_j MDS_{MICstart}(i,j)$$

$$Cov_{MIC}(i,j) \ge MDS_{MICstart}(i,j)$$

$$Cov_{MAC}(i,j) \ge MDS_{MACstart}(i,j)$$

$$Cov_{MIC}(i,j) = 3 - (cov_{MIC}(i,j-1) + cov_{MAC}(i,j+1) + MDS_{MICstart}(i,j) + MDS_{MICend}(i,j))$$

$$Cov_{MAC}(i,j) = 3 - (cov_{MAC}(i,j-1) + cov_{MAC}(i,j+1) + MDS_{MACstart}(i,j) + MDS_{MACend}(i,j))$$

Validity Checks

$$\sum_{l \le j} MDS_{MICstart}(i,l) + \sum_{l \ge j} MDS_{MICend}(i,l) - Cov_{MIC}(i,j) = 2$$

$$\sum_{l \le j} MDS_{MACstart}(i,l) + \sum_{l \ge j} MDS_{MACend}(i,l) - Cov_{MAC}(i,j) = 2$$

$$Cov_{MIC}(i,-1) = 0$$

$$Cov_{MAC}(i,-1) = 0$$

$$Cov_{MIC}(i,j) = Cov_{MIC}(i,j-1) - MDS_{MICend}(i,j-1) + MDS_{MICstart}(i,j)$$

$$Cov_{MAC}(i,j) = Cov_{MAC}(i,j-1) - MDS_{MACend}(i,j-1) + MDS_{MACstart}(i,j)$$