Repeat Detector Algorithm

V. Dion & T. Schuepbach

March 7, 2022

The present document gathers pieces of informations for the understanding of repeat detecter tool. The algorithm is presented here in details.

Contents

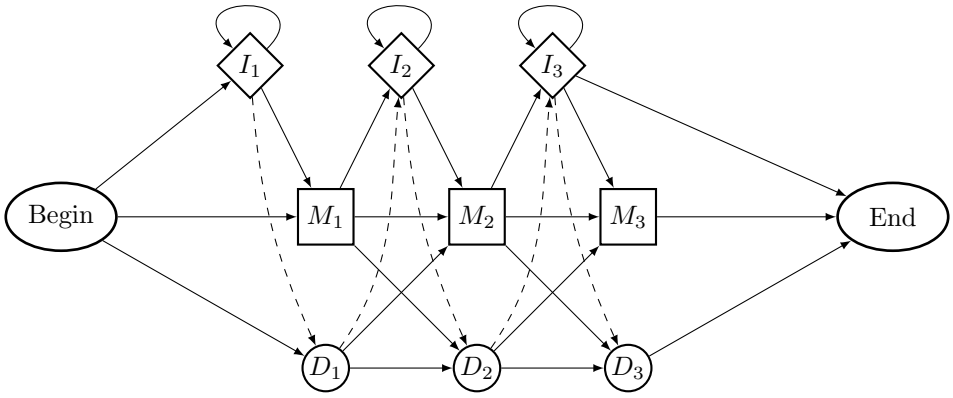

Figure 1: Simplified PfTools algorithm profile schema for a 3 sequence length profile. Simplifications arise from not showing all connections from Begin to nodes [I_i,M_i,D_i] as well as nodes [I_i,M_i,D_i] to End. It is worth noting the dashed that should not be used even though available. Indeed, mismatches should replace Insertion-Deletion or Deletion-Insertion paths thanks to extreme score loss.

Definitions

We define a sequence string of length N_s as $S = (s_1, s_2, \ldots, s_N)$ with s_i, $i \in [1, N_s]$, belonging to the alphabet $\mathcal{A}$. The cardinality of the set $\mathcal{A}$ is further on referred to as N_α. Given that an alignment fills in a matrix $(N_s + 1) \times (N_p + 1)$, where N_p denotes the length of the profile. Each element of the matrix, denoted as a cell, $C_{i,j}$, $i \in [0, N_s]$, $j \in [0, N_p]$, holds pieces of informations of the current 4 possible states: *Match*, *Insertion*, *Deletion* and *Score*. That is $C^M_{i,j}$, $C^I_{i,j}$, $C^D_{i,j}$ and $C^S_{i,j}$ respectively. It is worth noting that $C^S_{i,j}$ is useless when $i = 0$ or $j = 0$.

Generalized profiles provide tables of scores to be used in the computation of $C^\alpha_{i,j}$, $\alpha \in \{M, I, D, S\}$. Those are

- the match/mismatch score table $M \in \mathbb{Z}^{N_p \times N_\alpha}$,
- the insertion score table $I \in \mathbb{Z}^{N_p - 1 \times N_\alpha}$,
- the deletion score vector $D \in \mathbb{Z}^{N_p}$,
- the initial input score vectors $F^\delta \in \mathbb{Z}^{N_p+1}$, $\delta \in \{M, I, D\}$,
- the final output score vectors $L^\delta \in \mathbb{Z}^{N_p}$, $\delta \in \{M, I, D\}$,
- the state transition score vectors $T^{\beta \to \gamma} \in \mathbb{Z}^{1+N_p}$, β, $\gamma \in \{M, I, D, B, E\}$.

1 Standard PfTools algorithm

Let us define the indexing function $f(s) : \mathcal{A} \mapsto [1, N_\alpha]$ which translate the character s of the alphabet $\mathcal{A}$ into its corresponding index within the score tables M, I. The recurrence relations for $C^\alpha_{i,j}$, $\alpha \in \{M, I, D, S\}$ pictured in figure 2 are as follow:

$$C^M_{i,j} = \begin{cases} F^M_0 , & i = 0,\ j = 0 \\ \max \begin{cases} C^D_{0,j-1} + D_j + T^{D \to M}_j \\ F^M_j \end{cases} , & i = 0,\ j \in [1, N_p] \\ \max \begin{cases} C^I_{i-1,0} + I_{j,f(s_i)} + T^{I \to M}_0 \\ T^{B \to M}_0 \end{cases} , & i \in [1, N_s - 1],\ j = 0 \\ \max \begin{cases} C^M_{i-1,j-1} + M_{j,f(s_i)} + T^{M \to M}_j \\ C^I_{i-1,j} + I_{j,f(s_i)} + T^{I \to M}_j \\ C^D_{i,j-1} + D_j + T^{D \to M}_j \\ T^{B \to M}_j \end{cases} , & i \in [1, N_s - 1],\ j \in [1, N_p] \end{cases} \quad (1)$$

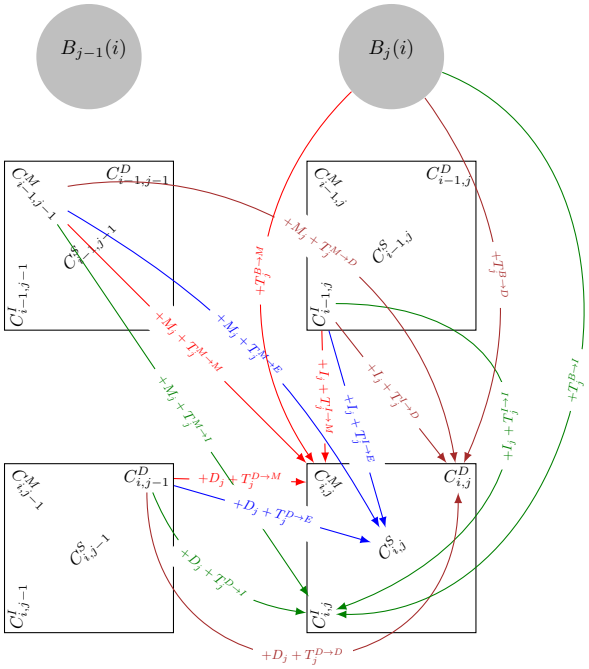

Figure 2: Standard PfTools algorithm schema

$$C^I_{i,j} = \begin{cases} F^I_0 \ , & i=0,\ j=0 \\ \max \begin{cases} C^D_{0,j-1} + D_j + T^{D\to I}_j \\ F^I_j \end{cases} , & i=0,\ j\in[1,N_p] \\ \max \begin{cases} C^I_{i-1,0} + I_{f(s_i),0} + T^{I\to I}_0 \\ T^{B\to I}_0 \end{cases} , & i\in[1,N_s-1],\ j=0 \\ \max \begin{cases} C^M_{i-1,j-1} + M_{j,f(s_i)} + T^{M\to I}_j \\ C^I_{i-1,j} + I_{j,f(s_i)} + T^{I\to I}_j \\ C^D_{i,j-1} + D_j + T^{D\to I}_j \\ T^{B\to I}_j \end{cases} , & i\in[1,N_s-1],\ j\in[1,N_p] \end{cases} \tag{2}$$

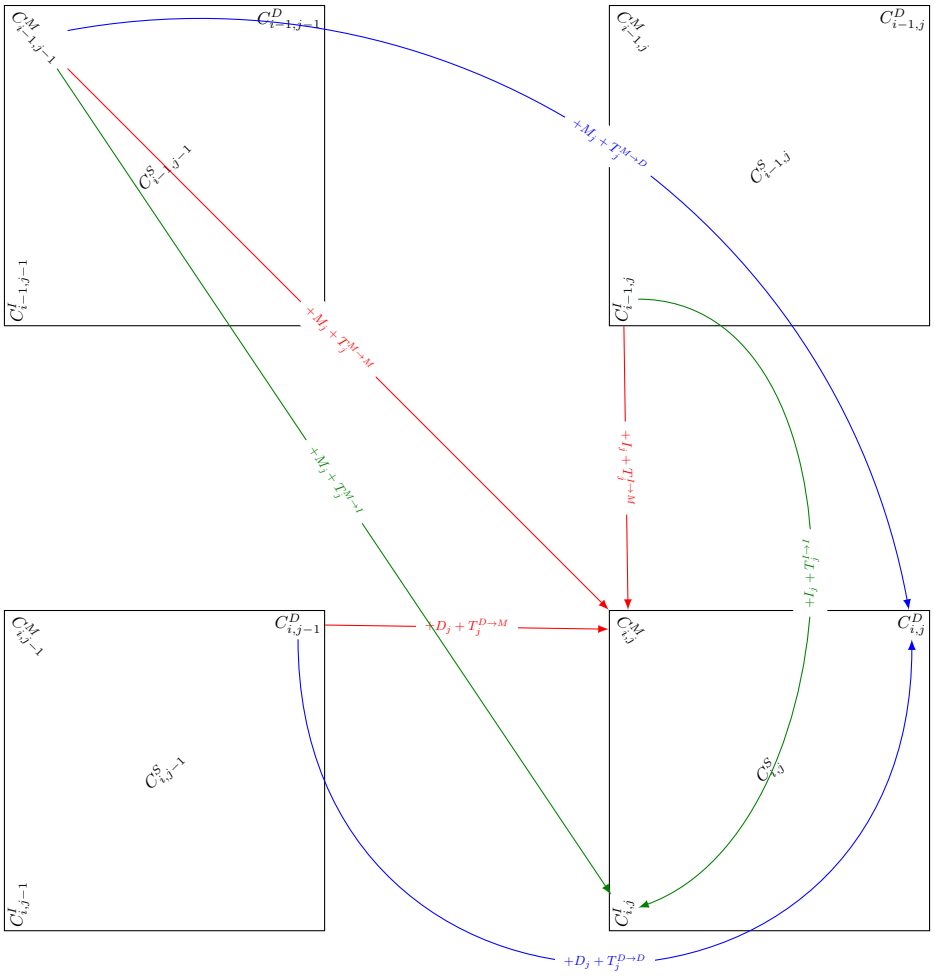

Figure 3: Standard PfTools algorithm schema

$$C^D_{i,j} = \begin{cases} F^D_0 \ , & i=0,\ j=0 \\ \max \begin{cases} C^D_{i,j-1} + D_j + T^{D\to D}_j \\ F^M_j \end{cases} , & i=0,\ j\in[1,N_p] \\ \max \begin{cases} C^I_{i-1,0} + I_{j,f(s_i)} + T^{I\to D}_j \\ T^{B\to M}_0 \end{cases} , & i\in[1,N_s-1],\ j=0 \\ \max \begin{cases} C^M_{i-1,j-1} + M_{j,f(s_i)} + T^{M\to D}_j \\ C^I_{i-1,j} + I_{j,f(s_i)} + T^{I\to D}_j \\ C^D_{i,j-1} + D_j + T^{D\to D}_j \\ T^{B\to D}_j \end{cases} , & i\in[1,N_s-1],\ j\in[1,N_p] \end{cases} \tag{3}$$

$$C^S_{i,j} = \begin{cases} \max \begin{cases} C^M_{i-1,j-1} + M_{j,f(s_i)} + T^{M\to E}_j \\ C^I_{i-1,j} + I_{j,f(s_i)} + T^{I\to E}_j \\ C^D_{i,j-1} + D_j + T^{D\to E}_j \end{cases} , & i\in[1,N_s-1],\ j\in[1,N_p] \\ \max \begin{cases} C^M_{i-1,j-1} + M_{j,f(s_i)} + L^M_j \\ C^I_{i-1,j} + I_{j,f(s_i)} + L^I_j \\ C^D_{i,j-1} + D_j + L^D_j \end{cases} , & i=N_s,\ j\in[1,N_p] \\ \text{useless} , & i=0 \text{ or } j=0 \end{cases} \tag{4}$$

Running PfSearch, either version 2 or 3, first computes the cell matrix $C^\alpha_{i,j}$, $\alpha\in\{M,I,D,S\}$, then analyzes the latter looking for the best match i.e. the highest score $C^S_{k,l}$ where the range of l is in agreement with the type of alignment sought. For a *global* alignment, one enforces the match to end on the last profile position, hence $l=N_p$. On the other hand, a *local* alignment allows $l\in[0,N_p]$. It is worth mentionning that any type of alignment is also subject to the profile declaration and therefore it may well happen that the backward tracing encounters an entry point before reaching the profile's length. As an example, one can try to elucidate the path given in figure **??**.

2 Repeat Decoder

The Repeat detecter algorithm is a bit more complex than the standard PfTools one as it incorporates feedback loop to accomodate potential replication of the profile.

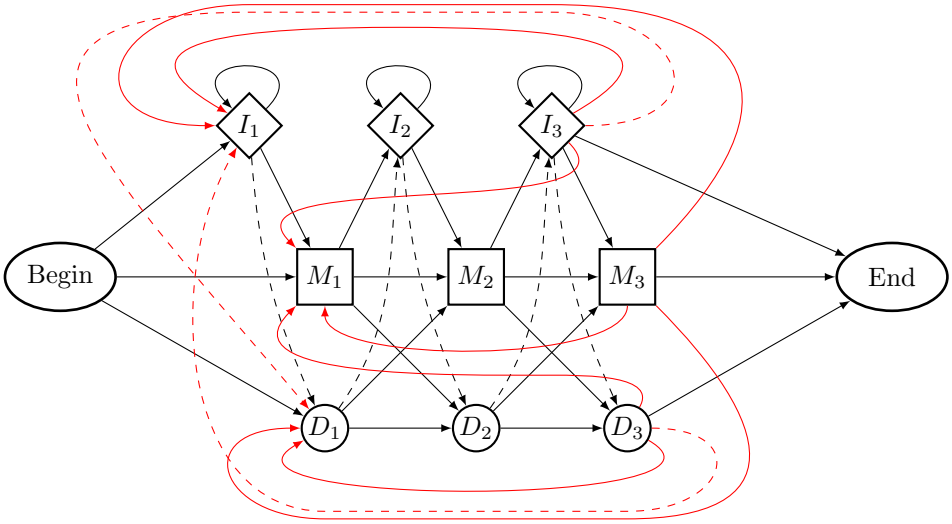

Figure 4: Simplified PfTools algorithm profile schema for a 3 sequence length profile. Simplifications arise from not showing all connections from Begin to nodes $[I_i,M_i,D_i]$ as well as nodes $[I_i,M_i,D_i]$ to End. It is worth noting the dashed lines that should not be used even though available. Indeed, mismatches should replace Insertion-Deletion or Deletion-Insertion paths thanks to extreme score loss.