

Alignement global d'une séquence d'acide nucléique guidé par une protéine

Eric Coissac

January 20, 2011

1 Cas d'un alignement avec une fonction de gap affine

1.1 Notation

On veut aligner une séquence nucléique N sur une séquence protéique P de manière à identifier les points de décalage du cadre de lecture (FrameShift F).

Soit $l_P = |P|$ et $l_N = |N|$.

Soit C une séquence nucléique inconnue correspondant à la traduction reverse de P , donc $l_C = 3 l_P$.

Les symboles de toutes ces séquences sont numérotés à partir de 1.

On définit ψ comme l'indice d'un nucléotide de la séquence P à l'intérieur de son codon. $\psi \in \{0, 1, 2\}$ et $\psi_i = (i - 1) \bmod 3$.

On note $\mathcal{C}(i, \psi)$ ou $\mathcal{C}_{i, \psi}$ le codon contenant le nucléotide N_i en sa position ψ . Le début du codon doit être potentiellement reconstruit par backtracking dans la matrice déjà calculée. La fin du codon est constituée des positions suivant éventuellement la position N_i .

1.2 Liste des opérations

On fait la différence entre un Gap qui correspond à une brèche de longueur multiple de 3 et un Frameshift qui correspond à une brèche de longueur non multiple de 3.

- **Pairing** (M_0)
- **OpenGap** (O_G)
- **ExtendGap** (E_G)
- **OpenFrameshift** (O_F)
- **ExtendFrameshift** (E_F)

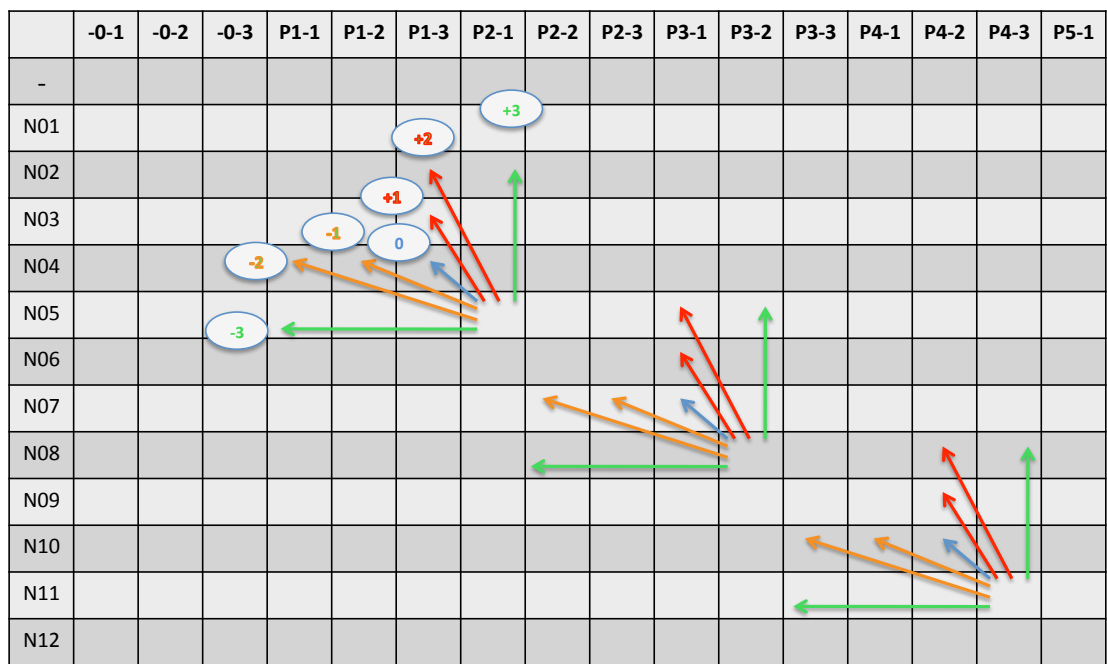


Figure 1: default

1.3 Fonction de reconstruction du codon

1.4 Fonction de match d'un codon et d'un acide aminé

$$S_{i,j} \mid \psi_j=0 = \max \left\{ \begin{array}{l} \textit{Match in phase} \\ S_{i-1,j-1} + \mathcal{S}(\mathcal{C}(N_i, N_{i+1}, N_{i+2}), \mathcal{C}(P_j, P_{j+1}, P_{j+2})/3 \\ \\ \textit{Frame shift by insertion} \\ S_{i-2,j-1} + \mathcal{S}(\mathcal{C}(N_i, N_{i+1}, N_{i+2}), \mathcal{C}(P_j, P_{j+1}, P_{j+2})/3 + F_{+1} \\ S_{i-3,j-1} + \mathcal{S}(\mathcal{C}(N_i, N_{i+1}, N_{i+2}), \mathcal{C}(P_j, P_{j+1}, P_{j+2})/3 + F_{+2} \\ \\ \textit{Frame shift by deletion} \end{array} \right. \quad (1)$$