Algorithme de micro assemblage entre les lectures pair-end

The Author

2 mai 2012

1 La fonction de score

1.1 notation

P_X : probabilité de la base lue sur la séquence X
P_Y : probabilité de la base lue sur la séquence Y
P_X et P_Y sont liés au score de la base donnée par le séquenceur. Approximativement par la relation suivante (varie en fonction du schéma de score).

$$P_{\text{base lue}} = 1 - e^{-\frac{Q}{10}} \tag{1}$$

Pour les Solexa, le type d'erreur prise en compte est une substitution. Nous ne considérons que ce séquenceur et posons que l'erreur à pu se produite de manière équivalente vers toutes les autres bases.
S_{match} : score d'un match plein entre deux nucléotides ($S_{match} > 0$)
$S_{mismatch}$: score d'un mismatch plein entre deux nucléotides ($S_{match} < 0$)

1.2 Score d'un match dans l'alignement

Le score d'un match (ou d'un mismatch) est en fait la somme d'une fraction d'un match plein et d'une fraction d'un mismatch.

	A	C	G	T
A	$\mathbf{P_X.P_Y}$	$P_X.(1-P_Y)/3$	$P_X.(1-P_Y)/3$	$P_X.(1-P_Y)/3$
C	$P_Y.(1-P_X)/3$	$\mathbf{(1-P_X)/3.(1-P_Y)/3}$	$(1-P_X)/3.(1-P_Y)/3$	$(1-P_X)/3.(1-P_Y)/3$
G	$P_Y.(1-P_X)/3$	$(1-P_X)/3.(1-P_Y)/3$	$\mathbf{(1-P_X)/3.(1-P_Y)/3}$	$(1-P_X)/3.(1-P_Y)/3$
T	$P_Y.(1-P_X)/3$	$(1-P_X)/3.(1-P_Y)/3$	$(1-P_X)/3.(1-P_Y)/3$	$\mathbf{(1-P_X)/3.(1-P_Y)/3}$

TABLE 1 – cas d'un match AA apparent

$$\begin{aligned} S'_{(X=Y)} &= S_{match}\{ & P_X P_Y \\ & & + \ 3\tfrac{(1-P_X)}{3}\tfrac{(1-P_Y)}{3}\} \\ & +S_{mismatch}\{ & 3P_Y\tfrac{(1-P_X)}{3} \\ & & + \ 3P_X\tfrac{(1-P_Y)}{3} \\ & & + \ 6\tfrac{(1-P_X)}{3}\tfrac{(1-P_Y)}{3}\} \end{aligned} \tag{2}$$

$$\begin{aligned} S'_{(X=Y)} &= S_{match}\{ & P_X P_Y + \tfrac{(1-P_X)(1-P_Y)}{3}\} \\ & +S_{mismatch}\{ & P_Y(1-P_X) \\ & & + \ P_X(1-P_Y) \\ & & + \ \tfrac{2}{3}(1-P_X)(1-P_Y)\} \end{aligned} \tag{3}$$

$$\begin{aligned} S'_{(X=Y)} &= S_{match}\{ & \tfrac{4P_XP_Y-P_Y-P_X+1}{3}\} \\ & +S_{mismatch}\{ & -\tfrac{4P_XP_Y-P_Y-P_X-2}{3}\} \end{aligned} \tag{4}$$

$$S'_{(X=Y)} = \frac{1}{3}((4P_XP_Y - P_X - P_Y)(S_{match} - S_{mismatch}) + (S_{match} + 2S_{mismatch})) \tag{5}$$

$$S'_{(X=Y)} = \frac{1}{3}((4\frac{1}{4}P_Y - \frac{1}{4} - P_Y)(S_{match} - S_{mismatch}) + (S_{match} + 2S_{mismatch})) \tag{6}$$

$$S'_{(X=Y)} = \frac{S_{match} + S_{mismatch}}{4} \tag{7}$$

	A	C	G	T
A	$\mathbf{P_X}.\mathbf{(1-P_Y)/3}$	$P_X.P_Y$	$P_X.(1-P_Y)/3$	$P_X.(1-P_Y)/3$
C	$(1-P_X)/3.(1-P_Y)/3$	$\mathbf{P_Y}.\mathbf{(1-P_X)/3}$	$(1-P_X)/3.(1-P_Y)/3$	$(1-P_X)/3.(1-P_Y)/3$
G	$(1-P_X)/3.(1-P_Y)/3$	$P_Y.(1-P_X)/3$	$\mathbf{(1-P_X)/3}.\mathbf{(1-P_Y)/3}$	$(1-P_X)/3.(1-P_Y)/3$
T	$(1-P_X)/3.(1-P_Y)/3$	$P_Y.(1-P_X)/3$	$(1-P_X)/3.(1-P_Y)/3$	$\mathbf{(1-P_X)/3}.\mathbf{(1-P_Y)/3}$

TABLE 2 – cas d'un mismatch AC apparent

$$\begin{aligned} S'_{(X\neq Y)} &= S_{match}\{ & P_Y\tfrac{(1-P_X)}{3} \\ & & + \ P_X\tfrac{(1-P_Y)}{3} \\ & & + \ 2\tfrac{(1-P_X)}{3}\tfrac{(1-P_Y)}{3}\} \\ & +S_{mismatch}\{ & P_XP_Y \\ & & + \ 2P_Y\tfrac{(1-P_X)}{3} \\ & & + \ 2P_X\tfrac{(1-P_Y)}{3} \\ & & + \ 7\tfrac{(1-P_X)}{3}\tfrac{(1-P_Y)}{3}\} \end{aligned} \tag{8}$$

$$S'_{(X\neq Y)} = \frac{1}{9}((P_Y + P_X - 4P_XP_Y)(S_{match} - S_{mismatch}) + 2S_{match} + 7S_{mismatch}) \quad (9)$$

$$S'_{(X\neq Y)} = \frac{1}{9}((S_{match} - S_{mismatch})(P_Y + \frac{1}{4} - 4\frac{1}{4}P_Y) + 7S_{mismatch} + 2S_{match}) \quad (10)$$

$$S'_{(X\neq Y)} = \frac{1}{4}(S_{match} + 3S_{mismatch}) \quad (11)$$

$$S'_{(X=Y)} = \frac{1}{3}((4P_XP_Y - P_X - P_Y)(S_{match} - S_{mismatch}) + (S_{match} + 2S_{mismatch})) \quad (12)$$

$$S'_{(X\neq Y)} = \frac{1}{9}((P_Y + P_X - 4P_XP_Y)(S_{match} - S_{mismatch}) + 2S_{match} + 7S_{mismatch}) \quad (13)$$