

Notredame-Higgins-Heringa (T-Coffee)

Unit Tests

Hint: Many test values are taken from project Algorithms for Bioninformatics of Alexander Mattheis or the lectures.

Test 1 (Hint: Notation from original Feng-Doolittle paper used!)

Input

Sequence a: ACGT
Sequence b: AT
Sequence c: GCT

Gap opening: 0 (easifies later visual proofment)
Enlargement: -2
Match: 1 (and 0 for placeholder #)
Mismatch: -1

Output (Computation: Global Primary Library)

	Alignment- Length	Gaps	Gap- starts	Score
(a,b)	4	2	1	-2
(a,c)	4	1	1	-1
(b,c)	3	1	1	-2

a: ACGT
 * *

b: A__T

a: ACGT
 |* *

c: GC_T

b: _AT
 |*

c: GCT

Hint: More alignments exists, but only one is computed!

Output (Computation: Weight Primary Library)

1. Conversion (not used in implementation – only for visualization)

für (a,b):

{(1,1),(4,2)}

für (a,c):

{(1,1),(2,2),(4,3)}

für (b,c):

{(1,2),(2,3)}

$$Pool = \left\{ \begin{array}{l} \{(1,1), (4,2)\} \\ \{(1,1), (2,2), (4,3)\} \\ \{(1,2), (2,3)\} \end{array} \right\}$$

2. Weight computation

$$L_{i,j}^{s_1, s_2} = 0$$

$$L_{i,j}^{s_1, s_2} = L_{i,j}^{s_1, s_2} + \text{weight}(A(a, b)) \text{ where } \text{weight}(A(a, b)) = \text{seq}_{ID}(A(a, b))$$

$$PrimLib = \left\{ \begin{array}{l} \{L_{1,1}^{a,b}, L_{4,2}^{a,b}\} \\ \{L_{1,1}^{a,c}, L_{2,2}^{a,c}, L_{4,3}^{a,c}\} \\ \{L_{1,2}^{b,c}, L_{2,3}^{b,c}\} \end{array} \right\} = \left\{ \begin{array}{l} \left\{ \frac{2}{2} \cdot 100, \frac{2}{2} \cdot 100 \right\} \\ \left\{ \frac{2}{3} \cdot 100, \frac{2}{3} \cdot 100, \frac{2}{3} \cdot 100 \right\} \\ \left\{ \frac{1}{2} \cdot 100, \frac{1}{2} \cdot 100 \right\} \end{array} \right\} = \left\{ \begin{array}{l} \{100, 100\} \\ \left\{ \frac{200}{3}, \frac{200}{3}, \frac{200}{3} \right\} \\ \{50, 50\} \end{array} \right\}$$

Output (Computation: Extended Primary Library)

Hint: $L_{i,j}^{a,b} = L_{j,i}^{b,a}$ or it would not make sense

(a triple like **three T in a column** should be recognized irrelevant of order)
else:

$$ExtendedLib = \left\{ \begin{array}{l} \{EL_{1,1}^{a,b}, EL_{4,2}^{a,b}\} \\ \{EL_{1,1}^{a,c}, EL_{2,2}^{a,c}, EL_{4,3}^{a,c}\} \\ \{EL_{1,2}^{b,c}, EL_{2,3}^{b,c}\} \end{array} \right\} = \left\{ \begin{array}{l} \{100, 100\} \\ \left\{ \frac{200}{3}, \frac{200}{3}, \frac{350}{3} \right\} \\ \{50, 50\} \end{array} \right\}$$

correct:

$$ExtendedLib = \left\{ \begin{array}{l} \{EL_{1,1}^{a,b}, EL_{4,2}^{a,b}\} \\ \{EL_{1,1}^{a,c}, EL_{2,2}^{a,c}, EL_{4,3}^{a,c}\} \\ \{EL_{1,2}^{b,c}, EL_{2,3}^{b,c}\} \end{array} \right\} = \left\{ \begin{array}{l} \{100, 150\} \\ \left\{ \frac{200}{3}, \frac{200}{3}, \frac{350}{3} \right\} \\ \left\{ 50, \frac{350}{3} \right\} \end{array} \right\} \quad \text{Hint: because of Triple-Match } \frac{350}{3}$$

$$\begin{aligned} EL_{1,1}^{a,b} &= L_{1,1}^{a,b} + \sum_{x \in \mathcal{S} \setminus \{a,b\}} \sum_{k \in Pos(x_{gapfree})} \min(L_{1,k}^{a,x}, L_{k,1}^{x,b}) \\ &= L_{1,1}^{a,b} + \sum_{x \in \{c\}} \sum_{k \in \{1,2,3\}} \min(L_{1,k}^{a,x}, L_{k,1}^{x,b}) \\ &= 100 + \min(L_{1,1}^{a,c}, L_{1,1}^{c,b}) + \min(L_{1,2}^{a,c}, L_{2,1}^{c,b}) + \min(L_{1,3}^{a,c}, L_{3,1}^{c,b}) \\ &= 100 + \min\left(\frac{200}{3}, 0\right) + 0 + 0 = 100 \end{aligned}$$

$$\begin{aligned} EL_{4,2}^{a,b} &= 100 + \min(L_{4,1}^{a,c}, L_{1,2}^{c,b}) + \min(L_{4,2}^{a,c}, L_{2,2}^{c,b}) + \min(L_{4,3}^{a,c}, L_{3,2}^{c,b}) \\ &= 100 + 0 + 0 + \min\left(\frac{200}{3}, 50\right) = 150 \end{aligned}$$

$$\begin{aligned}
EL_{1,1}^{a,c} &= L_{1,1}^{a,c} + \sum_{x \in \{b\}} \sum_{k \in \{1,2\}} \min(L_{1,k}^{a,x}, L_{k,1}^{x,b}) \\
&= \frac{200}{3} + \min(L_{1,1}^{a,b}, L_{1,1}^{b,c}) + \min(L_{1,2}^{a,b}, L_{2,1}^{b,c}) \\
&= \frac{200}{3} + \min(100, 0) + \min(0, 50)
\end{aligned}$$

$$\begin{aligned}
EL_{2,2}^{a,c} &= L_{2,2}^{a,c} + \sum_{x \in \{b\}} \sum_{k \in \{1,2\}} \min(L_{2,k}^{a,x}, L_{k,2}^{x,b}) \\
&= \frac{200}{3} + \min(L_{2,1}^{a,b}, L_{1,2}^{b,c}) + \min(L_{2,2}^{a,b}, L_{2,2}^{b,c}) \\
&= \frac{200}{3} + \min(0, 50) + 0
\end{aligned}$$

$$\begin{aligned}
EL_{4,3}^{a,c} &= L_{4,3}^{a,c} + \sum_{x \in \{b\}} \sum_{k \in \{1,2\}} \min(L_{4,k}^{a,x}, L_{k,3}^{x,c}) \\
&= \frac{200}{3} + \min(L_{4,1}^{a,b}, L_{1,3}^{b,c}) + \min(L_{4,2}^{a,b}, L_{2,3}^{b,c}) \\
&= \frac{200}{3} + 0 + \min(100, 50) \\
&= \frac{350}{3} \approx 116.67
\end{aligned}$$

$$\begin{aligned}
EL_{1,2}^{b,c} &= L_{1,2}^{b,c} + \sum_{x \in \{a\}} \sum_{k \in \{1,2,3,4\}} \min(L_{1,k}^{b,x}, L_{k,2}^{x,c}) \\
&= 50 + \min(L_{1,1}^{b,a}, L_{1,2}^{a,c}) + \min(L_{1,2}^{b,a}, L_{2,2}^{a,c}) + \min(L_{1,3}^{b,a}, L_{3,2}^{a,c}) + \min(L_{1,4}^{b,a}, L_{4,2}^{a,c}) \\
&= 50 + \min(100, 0) + \min\left(0, \frac{100}{3}\right) + 0 + 0
\end{aligned}$$

$$\begin{aligned}
EL_{2,3}^{b,c} &= L_{2,3}^{b,c} + \sum_{x \in \{a\}} \sum_{k \in \{1,2,3,4\}} \min(L_{2,k}^{b,x}, L_{k,3}^{x,c}) \\
&= 50 + \min(L_{2,1}^{b,a}, L_{1,3}^{a,c}) + \min(L_{2,2}^{b,a}, L_{2,3}^{a,c}) + \min(L_{2,3}^{b,a}, L_{3,3}^{a,c}) + \min(L_{2,4}^{b,a}, L_{4,3}^{a,c}) \\
&= 50 + 0 + 0 + 0 + \min\left(100, \frac{200}{3}\right) \\
&= \frac{350}{3}
\end{aligned}$$

Output (Distances)

$$D(a, b) = -\ln S_{a,b}^{eff} \approx 0.9808 \text{ (look into Feng-Doolittle Unit-Tests)}$$

$$\begin{aligned} S_{a,c}^{rand} &= \frac{1}{4} \left(\begin{array}{l} s(A_a, G_b) \cdot N_A(a) \cdot N_G(b) + s(A_a, C_b) \cdot N_A(a) \cdot N_C(b) + s(A_a, T_b) \cdot N_A(a) \cdot N_T(b) \\ + s(C_a, G_b) \cdot N_C(a) \cdot N_G(b) + s(C_a, C_b) \cdot N_C(a) \cdot N_C(b) + s(C_a, T_b) \cdot N_C(a) \cdot N_T(b) \\ + s(G_a, G_b) \cdot N_G(a) \cdot N_G(b) + s(G_a, C_b) \cdot N_G(a) \cdot N_C(b) + s(G_a, T_b) \cdot N_G(a) \cdot N_T(b) \\ + s(T_a, G_b) \cdot N_T(a) \cdot N_G(b) + s(T_a, C_b) \cdot N_T(a) \cdot N_C(b) + s(T_a, T_b) \cdot N_T(a) \cdot N_T(b) \end{array} \right) \\ &+ 1 \cdot enlarge \\ &= \frac{1}{4} \left(\begin{array}{l} (-1) + (-1) + (-1) \\ + (-1) + 1 + (-1) \\ + 1 + (-1) + (-1) \\ + (-1) + (-1) + 1 \end{array} \right) + 1 \cdot (-2) = \frac{-6}{4} - 2 = -3.5 \end{aligned}$$

$$S_{a,c}^{max} = \frac{4+3}{2} = 3.5$$

$$S_{a,c}^{eff} = \frac{S(a, c) - S_{a,c}^{rand}}{S_{a,c}^{max} - S_{a,c}^{rand}} = \frac{-1 - (-3.5)}{3.5 - (-3.5)} = \frac{2.5}{7}$$

$$D(a, c) = -\ln(S_{a,c}^{eff}) \approx 1.0296$$

$$\begin{aligned} S_{b,c}^{rand} &= \frac{1}{3} \cdot \left(\begin{array}{l} s(A_a, G_b) \cdot N_A(a) \cdot N_G(b) + s(A_a, C_b) \cdot N_A(a) \cdot N_C(b) + s(A_a, T_b) \cdot N_A(a) \cdot N_T(b) \\ + s(T_a, G_b) \cdot N_T(a) \cdot N_G(b) + s(T_a, C_b) \cdot N_T(a) \cdot N_C(b) + s(T_a, T_b) \cdot N_T(a) \cdot N_T(b) \end{array} \right) \\ &+ 1 \cdot enlarge \\ &= \frac{1}{3} \cdot \left(\begin{array}{l} (-1) + (-1) + (-1) \\ + (-1) + (-1) + 1 \end{array} \right) - 2 = \frac{-4}{3} - 2 = -\frac{10}{3} \end{aligned}$$

$$S_{b,c}^{max} = \frac{2+3}{2} = 2.5$$

$$S_{b,c}^{eff} = \frac{S(b, c) - S_{b,c}^{rand}}{S_{b,c}^{max} - S_{b,c}^{rand}} = \frac{-2 - \left(-\frac{10}{3}\right)}{2.5 - \left(-\frac{10}{3}\right)} = \frac{\frac{4}{3}}{\frac{35}{6}} = \frac{8}{35}$$

$$D(b, c) = -\ln(S_{b,c}^{eff}) \approx 1.4759$$

Output (Phylogenetic Tree : look into Feng-Doolittle Unit-Tests)

1.

$$d_{min} = 1$$

	a	b	c
a	0	1	1
b		0	1.5
c			0

2.

$$\mathcal{C} = ((\mathcal{C} - \{a\}) - \{b\}) \cup \{d\}$$

	a	b	c	d
a	0	1	1	
b		0	1.5	
c			0	1.25
d				0

3.

$$\text{dist}(d, a) = \text{dist}(d, b) = \frac{1}{2} = 0.5$$

4.

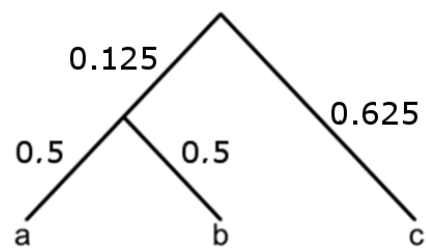
$$\text{dist}(c, d = \{a, b\}) = \frac{|a| \cdot \text{dist}(c, a) + |b| \cdot \text{dist}(c, b)}{|a| + |b|} = \frac{1 \cdot 1 + 1 \cdot 1.5}{1 + 1} = 1.25$$

1) $d_{min} = 1.25$

2) $\mathcal{C} = ((\mathcal{C} - \{c\}) - \{d\}) \cup \{e\}$

3) $\text{dist}(e, c) = \text{dist}(e, d) = \frac{d_{min}}{2} = 0.625$

	a	b	c	d	e
a	0	1	1		
b		0	1.5		
c			0	1.25	
d				0	
e					0



Output (Joinment)

Gap opening: 0 → Gotoh not needed for calculation
 Enlargement: 0

$$ExtendedLib = \left\{ \begin{array}{l} \{EL_{1,1}^{a,b}, EL_{4,2}^{a,b}\} \\ \{EL_{1,1}^{a,c}, EL_{2,2}^{a,c}, EL_{4,3}^{a,c}\} \\ \{EL_{1,2}^{b,c}, EL_{2,3}^{b,c}\} \end{array} \right\} = \left\{ \begin{array}{l} \{100, 150\} \\ \{\frac{200}{3}, \frac{200}{3}, \frac{350}{3}\} \\ \{50, \frac{350}{3}\} \end{array} \right\}$$

1. a~b:

		A ₁	T ₂
	0	0	0
A ₁	0	100	100
C ₂	0	100	100
G ₃	0	100	100
T ₄	0	100	250

ACGT
 A##T
 Score 250

⋮

2.

A C G T and G C T
 A # # T

ab~c: (every char with every other char, so A_a with G_c and A_b with G_c → and then average)

			G ₁	C ₂	T ₃
		0	0	0	0
A ₁	A ₁	0	100/3	100/3	100/3
C ₂	#	0	100/3	200/3	200/3
G ₃	#	0	100/3	200/3	200/3
T ₄	T ₂	0	100/3	200/3	550/3

$$\frac{EL_{1,1}^{a,c} + EL_{1,1}^{b,c}}{2} = \frac{\frac{200}{3} + 0}{2} = \frac{100}{3}$$

$$\frac{EL_{1,2}^{a,c} + EL_{1,2}^{b,c}}{2} = \frac{0 + 50}{2} = 25$$

$$\frac{EL_{1,3}^{a,c} + EL_{1,3}^{b,c}}{2} = \frac{0 + 0}{2} = 0$$

$$\frac{EL_{2,1}^{a,c} + 0}{2} = \frac{0 + 0}{2} = 0$$

$$\frac{EL_{2,2}^{a,c} + 0}{2} = \frac{\frac{200}{3} + 0}{2} = 100/3$$

$$\frac{EL_{2,3}^{a,c} + 0}{2} = \frac{0 + 0}{2} = 0$$

$$\frac{EL_{3,1}^{a,c} + 0}{2} = \frac{0 + 0}{2} = 0$$

$$\frac{EL_{3,2}^{a,c} + 0}{2} = \frac{0 + 0}{2} = 0$$

$$\frac{EL_{3,3}^{a,c} + 0}{2} = \frac{0 + 0}{2} = 0$$

$$\frac{EL_{4,1}^{a,c} + EL_{2,1}^{b,c}}{2} = \frac{0 + 0}{2} = 0$$

$$\frac{EL_{4,2}^{a,c} + EL_{2,2}^{b,c}}{2} = \frac{0 + 0}{2} = 0$$

$$\frac{EL_{4,3}^{a,c} + EL_{2,3}^{b,c}}{2} = \frac{\frac{350}{3} + \frac{350}{3}}{2} = \frac{350}{3}$$

Output (Final)

ACGT

A__T

GC_T

SoP-Score -5