Fusion des projets de programmation AR1 et AR2 PI4 2021/2022 : **Alignement de séquences et PhyloApp on Planet 622**

Proposé par Alice Rogier

alice.rogier@inserm.fr

1 Contexte

Nous sommes toujours en 24762. Les utilisateurs sont très contents de l'interface que l'équipe AR2 a développée. Ils aimeraient cependant mieux comprendre comment se mesure la distance génétique entre les espèces. Fort heureusement, une autre équipe de chercheurs en exobiologie, l'équipe AR1, a développé une superbe application permettant de visualiser un alignement et d'obtenir un score de similarité entre deux séquences d'ADN. Pourquoi ne pas lier vos efforts ?

2 Indications

Pour créer une belle interface, il est impératif de comprendre les objectifs des deux projets. Lisez-les donc attentivement.
L'équipe AR1 (équipe "amont") a développé une interface qui permet de calculer un score de similarité entre deux courtes séquences d'ADN. L'équipe AR2 (équipe "avale") a développé une interface qui permet de visualiser un arbre phylogénétique à partir d'une matrice de distances.

Pour le moment, l'équipe avale calcule la matrice de distances de départ à partir d'entiers modélisant des gènes. Des courtes séquences d'ADN vont donc remplacer ces entiers. Des scores de similarité seront calculés entre les séquences deux à deux. L'utilisateur doit pouvoir visualiser les alignements de séquences donnant ces scores, comme il le peut avec l'interface développée par l'équipe amont. À partir des similarités calculées avec l'alignement des séquences, il faut remplir la matrice des distances intitiale de l'algorithme UPGMA. Or la similarité entre deux séquences a et b est un entier relatif ($s_{(a,b)} \in \mathbb{Z}$), et plus elle est élevée, plus les séquences sont proches et leur distance devrait être faible.

Il faut donc trouver un moyen de convertir les similarités en distances. En voici un (ni le seul, ni le meilleur) :

Soit X un vecteur de similarités X=(s_1,s_2, ... s_n).
On va d'abord normaliser les similarités s_i entre 0 et 1 à l'aide de la formule suivante :

$$Z_{s_i} = \frac{s_i - min(X)}{max(X) - min(X)} \tag{1}$$

Soit $Z_{s_{(a,b)}}$ et $d_{(a,b)}$ respectivement la similarité normalisée et la distance entre a et b. On définit $d_{(a,b)}$ de la façon suivante :

$$d_{(a,b)} = 1 - Z_{s_{(a,b)}} \tag{2}$$

3 Exemple

Sur la planète 622, on a séquencé un gène homologue à 5 espèces :

Espèce	Séquence du gène
Bronteroc (b)	GATTACA
Totoro (t)	GATTACAAA
Marsupilami (m)	GTCGACGCA
Hippogriffe (h)	GATTACT
Chewbacca (c)	CTCTTCCCGCAAA

En prenant un score de match de 1, de mismach de -1 et de gap de -2, on a obtenu les similarités suivantes :

Similarités
$s(b,t) = 3$
$s(b,m) =$ -3
$s(b,h) = 5$
$s(b,c) =$ -11
$s(t,m) =$ -1
$s(t,h) = 1$
$s(t,c) =$ -5
$s(m,h) =$ -5
$s(m,c) =$ -5
$s(h,c) =$ -13

Ici, on a:
X=$(3, -3, 5, -11, -1, 1, -5, -5, -5, -13)$
$min(X) = s(h,c) =$ -13 et $max(X) = s(b,h) = 5$

Voici un tableau résumant tous les calculs pour convertir les similarités de ces gènes en distances.

Similarité	Similarité normalisée	Distance
$s(b,t) = 3$	$Z_{s(b,t)} = 0.89$	$d(b,t) = 0.11$
$s(b,m) =$ -3	$Z_{s(b,m)} = 0.56$	$d(b,m) = 0.44$
$s(b,h) = 5$	$Z_{s(b,h)} = 1$	$d(b,h) = 1$
$s(b,c) =$ -11	$Z_{s(b,c)} = 0.11$	$d(b,c) = 0.89$
$s(t,m) =$ -1	$Z_{s(t,m)} = 0.67$	$d(t,m) = 0.33$
$s(t,h) = 1$	$Z_{s(t,h)} = 0.78$	$d(t,h) = 0.22$
$s(t,c) =$ -5	$Z_{s(t,c)} = 0.44$	$d(t,c) = 0.56$
$s(m,h) =$ -5	$Z_{s(m,h)} = 0.44$	$d(m,h) = 0.56$
$s(m,c) =$ -5	$Z_{s(m,c)} = 0.44$	$d(m,c) = 0.56$
$s(h,c) =$ -13	$Z_{s(h,c)} = 0$	$d(h,c) = 1$

À titre d'exemple, $Z_{s(b,t)} = (3+13)/5+13 = 16/18 = 0.89$, et $d(b,t) = 1$ - $0.89 = 0.11$.