Les indices de diversité

25 août 2017

La fonction `intraHostDiversity` évalue l'indice de diversité entré en argument à chaque pas de temps.

1 Estimateur de richesse

L'estimateur de richesse S correspond au nombre de génotypes différents dans la population virale. Pour le calculer avec R, on utilise la fonction `countprop` qui calcule le nombre d'éléments non nuls dans un vecteur. On applique cette fonction au vecteur de proportions.

2 Index de Shannon

L'index de Shannon H est défini comme :

$$H = -\sum_{i=1}^{N} p_i \cdot ln(p_i),$$

avec p_i la proportion de variants de génotype i et N le nombre de génotypes différents dans la population virale. Pour le calculer avec R on utilise la fonction `diversity` du package *vegan* avec l'option "shannon". Cette fonction prend comme argument un vecteur de taille N dont chaque élément d'index i correspond à l'effectif du génotype i (vecteur de proportions multiplié par le nombre de virions à l'instant considéré).

3 Index de Simpson

L'index de Simpson λ est défini comme :

$$\lambda = 1 - \sum_{i=1}^{N} p_i^2,$$

avec p_i la proportion de variants avec la séquence i et N le nombre de séquences différentes dans la population virale. Pour le calculer avec R on utilise la même fonction que précédemment mais avec l'option "simpson".

4 Régularité de Pielou

La régularité de Pielou J_1 est définie comme :

$$J_1 = \frac{H}{ln(S)},$$

avec H l'index de Shannon et S l'estimateur de richesse. Pour calculer cet indice avec R on utilise les valeurs trouvées pour S et H.

5 Diversité nucléotidique ou p-distance

La diversité nucléotidique ou p-distance π pour un ensemble de N séquences est définie comme :

$$\pi = \frac{N}{(N-1)\cdot M}\sum_{i=1}^{N}\sum_{j=1}^{N}\hat{x}_i\cdot\hat{x}_j\cdot\delta_{ij},$$

avec $\hat{x}_i$ la fréquence de la séquence i dans la population virale et δ_{ij} le nombre de nucléotides différentes entre les séquences i et j. Pour calculer cette distance avec R on utilise la fonction `nuc.div` du package *pegas*. Cette fonction prend en arguments des séquences sous forme d'objets DNAbin. Pour ça dans la fonction `viralCompo` on récupère une matrice avec toutes les séquences (codées en 1234) et on passe cette matrice dans la fonction `toDNAbin` pour avoir l'objet qui nous convient (c'est-à-dire un ensemble de séquences DNAbin). La fonction `nuc.div` renvoie donc une valeur de p-distance sur l'ensemble des N séquences. Comme à partir d'un certain temps il y a beaucoup de séquences (environ 20 000), la distance est trop longue à calculer : on échantillonne donc un certain nombre de séquences choisi par l'utilisateur (avec un échantillon de 100 séquences les résultats sont corrects) et on calcule la p-distance à partir de ces séquences.

6 Distance de Jukes et Cantor

La distance de Jukes et Cantor entre deux séquences X et Y est définie comme :

$$d_{JC\{XY\}} = \frac{-3}{4}ln(1-\frac{4}{3}f_{XY}),$$

avec f_{XY} la dissimilarité entre X et Y (fraction de différences observées). Pour calculer cette distance avec R, on utilise la fonction `dist.dna` du package *ape*. Cette fonction prend aussi comme argument un ensemble de séquences DNAbin, on réalise donc une conversion comme précédemment. On échantillonne l'ensemble des séquences pour diminuer le temps de calcul. La fonction retourne une matrice de distance (décrit les distances entre les séquences 2 à 2) donc on fait une moyenne pour avoir une seule valeur.

7 Distance de Nei

La distance de Nei entre deux séquences X et Y de M nucléotides est définie comme :

$$d_{Nei} = -ln\left(\frac{\sum_{\ell=1}^{M}\sum_{i=\{A,C,G,T\}}x_{i\ell}y_{i\ell}}{\sqrt{\left(\sum_{\ell=1}^{M}\sum_{i=\{A,C,G,T\}}x_{i\ell}^2\right)\left(\sum_{\ell=1}^{M}\sum_{i=\{A,C,G,T\}}y_{i\ell}^2\right)}}\right),$$

avec $x_{i\ell}$ et $y_{i\ell}$ la fréquence de la nucléotide i au locus $\ell = \{1,...,M\}$. Pour calculer cette distance avec R on utilise la fonction `nei.dist` du package *poppr*. Cette fonction prend également comme argument un ensemble de séquences DNAbin, on réalise donc une conversion comme précédemment. On échantillonne l'ensemble des séquences pour diminuer le temps de calcul. La fonction retourne une matrice de distance (décrit les distances entre les séquences 2 à 2) donc on fait une moyenne pour avoir une seule valeur.