Analyse de séquences sous R

Benoit Nabholz (benoit.nabholz@umontpellier.fr)

Partie 1 : Alignement de séquence

Objectif de l'exercice : Découvrir l'alignement de séquences en comparant les séquences de 9 mammifères placentaires et d'un groupe externe marsupial. L'alignement consiste à agencer les séquences les une par rapport aux autres de manière à obtenir les régions homologues en vis à vis, c'est-à-dire sur des colonnes identiques. L'idée étant que les nucléotides situés sur un même site (une même colonne) partagent une histoire phylogénétique commune (orthologie ou paralogie). Ici, nous allons réaliser une analyse sur la base de l'exon 11 du gène BRCA1 (BReast and ovarian CAncer susceptibility gene N°1).

Nous allons d'abord récupérer une partie des séquences sur la base de données GenBank de NCBI.

1. Extraction de données

- Connectez-vous à GenBank (http://www.ncbi.nlm.nih.gov/genbank/)

Pour récupérer les séquences, nous allons utiliser les numéros d'accession présentés dans le tableau ci-dessous. Ces numéros correspondent à des identifiants pour retrouver les séquences dans la base de données GenBank.

- **Téléchargez les séquences au format FASTA et renseigner les informations manquantes dans le tableau 1** (Table1.csv).

Pour cela, utilisez l'onglet «send:» en haut à droite. Stockez les séquences dans un **seul fichier texte**. Vous pouvez renommer les séquences avec un nom explicite (et sans espace!).

Table 1: Espèces et numéros d'accession

Nom_latin	Taxonomie	Accession
Vombatus ursinus	Marsupiaux	AF284031
Orycteropus afer	Afrothériens	AF284030
Elephas maximus	Afrothériens	AF284022
Dasypus novemcinctus	Xénarthres	AF283999
Bradypus tridactylus	Xénarthres	AF284002
Tupaia tana	Euarchontoglires	AF284006
Lepus capensis	Euarchontoglires	AF284005
Bos taurus	Laurasiathériens	AF284013
		AF284008

Exemple de code python pour renommer les séquences:

```
#! /usr/bin/env python
f = open("sequence.fasta")
for line in f:
    line = line.rstrip()
    if line.startswith(">"):
        genre = line.split(" ")[1]
        sp = line.split(" ")[2]
        print genre+"_"+sp
    else:
```

```
        print line
```

- Ouvrez le fichier FASTA avec le programme `seaview` (http://doua.prabi.fr/software/seaview , version linux 64 bit). `seaview` est normalement disponible directement en tapant `seaview` dans le terminal. **Que constatez-vous?**
- Lancez l'alignement dans seaview avec "Align" -> "Align all". **Que fait le logiciel d'alignement?**
- Nous allons maintenant calculer la taille des insertions/délétions (indels). -1 Ouvrir le logiciel R. -2 Charger la librairie ape:

```
library(ape)
```

-3 lire l'alignement à l'aide de la fonction `read.dna`:

```
align<-read.dna("BRCA1_aligned.fasta", format="fasta", as.character=TRUE)
```

L'alignement est stocké dans l'objet `align` sous forme d'une matrice avec les espèces en lignes et les sites en colonnes.

- Quelle est le résultat de la commande: `align[1,]` et `align[,1]`?
- A l'aide d'une boucle `for()` comptez la taille des délétions ("-") présentes dans la séquence de l'homme et de la souris. **Y a-t-il plus de délétions dans la séquences du hérisson ou de la vache?**

Partie 2 : Comparaison du nombre des insertions/délétions et des substitutions nucléotidiques

Objectif de l'exercice: Dans la partie 1, nous avons comptabilisé le nombre de "gap" entre espèces de mammifères. Ici, nous allons estimer le nombre de substitution nucléotidique et comparer ces deux quantités.

L'alignement de quatre gènes (dont BRCA1) est stocké dans le répertoire "sequence".

La fonction `dist.dna` du package `ape` permet de calculer le nombre de substitutions par sites entre les paires de séquences d'un alignement.

```
# lire l'alignement
BRCA1<-read.dna("BRCA1.fasta",format="fasta")
# calculer la divergence
dist.dna(BRCA1)
```

- A l'aide du code que vous avez déjà réalisé, calculez le nombre de délétions pour chacun des alignements et normalisez cette valeur par le nombre total de nucléotides de l'alignement (pour obtenir des délétions par paires de bases), **remplissez le tableau 2** (Table2.csv):

Table 2: Divergence et nombre de "gap" par paire de base pour quatres gènes

Genes	Div_Homo_Didelphis	Gap_Homo	Gap_Monodelphis
BRCA1	NA	NA	NA
PHF6	NA	NA	NA
RAP2C	NA	NA	NA
SEMA3F	NA	NA	NA

-Proposez un(des) hypothèse(s) pour expliquer les différences du nombre de substitutions par sites entre gènes.

-**Existe-t-il une relation entre la divergence moléculaire de chaque gène et le nombre "gap"" par paire de base? Si elle existe, comment interprétez cette relation?** Utilise la fonction `plot` dans R pour répondre à cette question.

Partie 3: Horloge Moléculaire

Objectif de l'exercice : Comparer la divergence génétique pour le gène NFYA entre l'homme et trois espèces de vertébrés: la vache (*Bos taurus*), l'opossum (*Monodelphis domestica*) et le Poulet (*Gallus gallus*). Cela nous permettra d'illustrer le fait que la divergence génétique s'accumule de manière régulière en fonction du temps de divergence ; c'est le concept d'**horloge moléculaire**.

- Calculez les divergences synonymes et non-synonymes toutes les espèces à l'aide de la fonction `kaks` pour les deux gènes. Divergence non-synonyme est noté ka ou d~n et la divergence synonyme ks ou d~s.

```
library(seqinr)
# Noté l'utilisation de la fonction read.alignement
NFYA<-read.alignment("sequences/NFYA.fasta",format="fasta")
kaks(NFYA)
```

- **Comment expliquez-vous qu'il y ait plus de substitutions synonymes entre homme et l'opossum qu'entre l'homme et la vache?**

Le registre fossile (voir l'article de Benton et al. 2009 disponible sur l'espace pédagogique), nous indique que la divergence homme/vache date de ~61,5 Millions d'années (MA) et homme/Poulet de ~312(MA).

- En appliquant le principe de l'horloge moléculaire ($ks = 2t\mu$), où μ est le taux de mutation et t le nombre de génération depuis l'ancêtre commun.
- Calculez le taux de mutation moyen entre l'homme et la vache puis entre les mammifères et les oiseaux.
- Déduisez le temps de divergence entre l'homme et l'opposum (*Monodelphis*) et comparez-le à la divergence proposée par le registre fossile (voir figure page 44 de Benton et al. 2009). Interpréter.