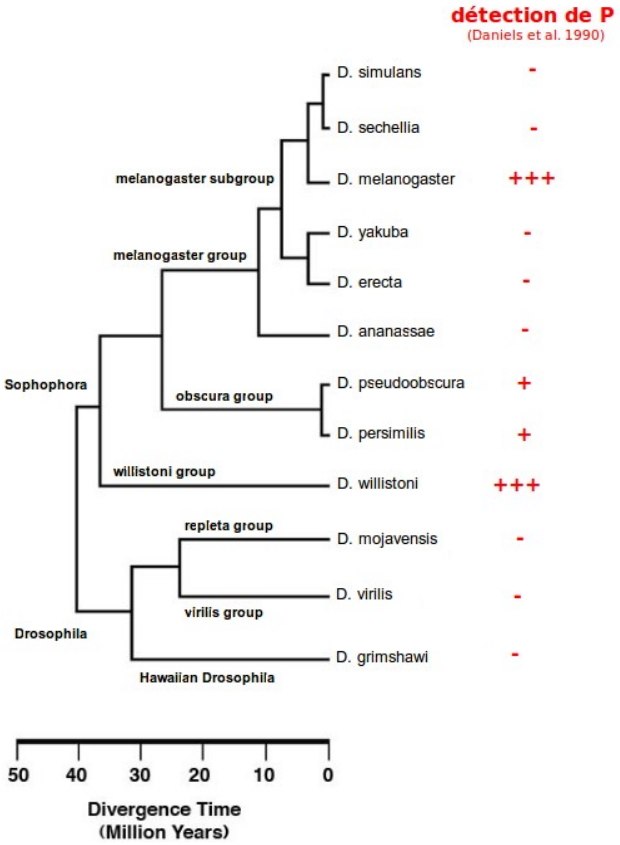
Détection de transfert horizontal d'élément transposable

marie.fablet@univ-lyon1.fr ; camille.mayeux@univ-lyon1.fr

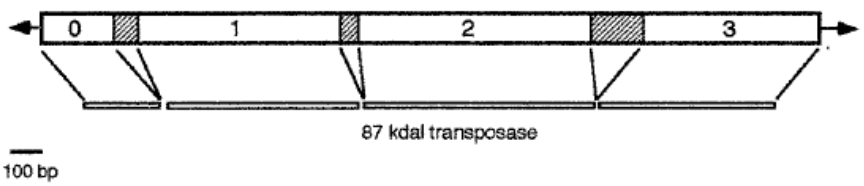
Les fichiers nécessaires à ce TP se trouvent à l'adresse suivante :
`ftp://pbil.univ-lyon1.fr/pub/cours/fablet/GGE/TP_HT-ET`

Introduction

P est un transposon de drosophile mis en évidence récemment dans le génome des souches naturelles de *Drosophila melanogaster*, les vieilles souches de laboratoire en étant dépourvues. La distribution discontinue de l'élément P dans les espèces du sous-groupe *melanogaster* a, entre autres, suggéré l'hypothèse de l'apparition de ce transposon dans le génome de *D. melanogaster* par transfert horizontal. Au cours de ce TP, vous allez éprouver cette hypothèse de transfert horizontal par différentes méthodes.



Phylogénie de quelques espèces de drosophiles, incluses dans le programme de séquençage de génomes complets. Distribution de l'élément P dans ces espèces (d'après Daniels et al., 1990).



Structure de l'élément P de référence chez *D. melanogaster* : 4 exons (numérotés de 0 à 3) constituent la séquence codante d'une transposase et sont encadrés de répétitions inversées (symbolisées par des flèches). D'après Clark & Kidwell, 1997, PNAS.

Flybase est une ressource disponible en ligne (`http://flybase.org`) pour la génétique et la génomique des drosophiles. Vous allez y récupérer l'identifiant de la séquence de référence pour l'élément P de *Drosophila melanogaster*. Pour cela, dans l'encadré Quick Search, dans le menu déroulant Data class choisissez " natural transposon ", et entrez la requête " P-element ". Téléchargez le fichier `transposon_sequence_set.embl.txt.gz` que vous trouvez dans la rubrique Sequences & Components. Vous pouvez alors obtenir l'identifiant correspondant au transposon P (cherchez " P-element " dans le texte).

Ce fichier contient également l'annotation de cette séquence de référence de l'élément P. Récupérez les positions de P dans la séquence de référence ainsi que les positions de l'exon 2, sur lequel vous réaliserez la suite des analyses.

Sur GenBank (`http://www.ncbi.nlm.nih.gov/genbank/`), consultez la page correspondant à cette séquence (en renseignant l'identifiant dans la rubrique Nucleotide), et récupérez la séquence correspondant à l'exon 2, que vous enregistrerez dans un fichier texte au format fasta.

Pour la suite de ce TP, vous ne travaillerez que sur l'exon 2 de P. Vous allez constituer un jeu de séquences provenant de différentes espèces de drosophiles.

Grâce à l'identifiant de la séquence de référence, vous allez interroger la banque de séquences GenBank par BLAST (`http://www.ncbi.nlm.nih.gov/BLAST/`). Choisissez Basic BLAST, nucleotide BLAST, puis " nucleotide collection (nr/nt) " dans le menu déroulant Database. Collez l'identifiant de P dans la zone Enter Query Sequence. Que constatez-vous ?

Nous utiliserons les séquences : « Drosophila willistoni P transposable element P-Dwill_Ilha10 transposase gene, complete cds », « Drosophila mediopunctata P element transposase gene, complete cds » et « Drosophila sturtevanti P transposable element P_Dst_Apazapan10S10 truncated transposase gene, complete cds ». Pour chacun de ces éléments P, récupérez la séquence de l'exon 2, que vous enregistrez au format fasta dans le fichier que vous avez précédemment créé.

Vous utiliserez également deux autres séquences de P provenant respectivement de *Drosophila pallidosa* et de *Drosophila bifasciata* (fichier `P_pal-bif.txt`). Ajoutez ces séquences à votre fichier de séquences et modifiez les noms des autres séquences de façon à les rendre courts et explicites pour la suite des analyses.

1. Construction de phylogénies

La méthode la plus communément utilisée pour mettre en évidence un événement de transfert horizontal est la comparaison de la phylogénie des espèces à la phylogénie de l'élément dans les espèces. Les éventuelles incohérences entre les deux peuvent alors être expliquées par des transferts horizontaux.

Commencez par établir la phylogénie des espèces étudiées dans ce TP. Pour connaître le groupe d'espèces auquel appartient une espèce non indiquée sur cette figure, consultez Flybase (onglet Species, puis Phylogeny).

Pour réaliser la phylogénie des éléments P dans ces espèces, vous allez utiliser l'éditeur d'alignements Seaview (`http://pbil.univ-lyon1.fr/softwares/seaview.html`). La version 4 de Seaview permet de construire des arbres phylogénétiques. Enregistrez l'alignement au format fasta.

À partir de cet alignement, construisez un arbre phylogénétique. Que concluez-vous ?

2. Analyse de l'usage du code

Le code génétique est dit "dégénéré", ce qui signifie que plusieurs codons peuvent coder le même acide aminé. Tous les codons correspondant à un même acide aminé ne sont pas nécessairement utilisés dans les mêmes proportions par un organisme donné ; la façon dont un organisme traite ces différents codons est appelée "usage du code". Dans la mesure où l'usage du code peut beaucoup différer entre espèces, certains auteurs proposent que l'analyse de l'usage du code peut permettre de mettre en évidence un événement de transfert horizontal récent. Quel est le résultat attendu sous l'hypothèse de transfert horizontal ?

Commencez par déterminer la phase codante des séquences que vous analysez avec l'outil en ligne ORF Finder (`http://www.ncbi.nlm.nih.gov/orffinder/`).

Vous allez analyser l'usage du code à l'aide du logiciel R. R permet d'utiliser de très nombreux "packages", adaptés à différents types d'analyses statistiques. Un package R contient un ensemble de fonctions, dont on peut afficher la documentation en tapant le nom de la fonction précédé d'un point d'interrogation. Vous utiliserez ici les packages seqinr (pour l'analyse des séquences) et ade4 (pour l'analyse multivariée).

Chargez les packages avec la commande :

```
library(nom_du_package)
```

Quelques exemples de fonctionnalités de seqinr :

```
tablecode()        ## ouvre une fenêtre avec le code génétique
SEQINR.UTIL        ## cf notice jointe
```

Le package seqinr permet de lire des fichiers de séquences au format fasta.

Chargez les séquences à considérer dans un objet nommé `seqP` :

```
seqP=read.fasta("nom_du_fichier")
names(seqP)
```

Déterminez l'usage du code pour ces différentes séquences. Pour cela, dans seqinr, il existe la fonction `uco` (consultez la documentation avec la commande `?uco`).

```
calc=function(x)  {return(as.vector(uco(x)))}
tabuco=function(x)      {return (data.frame(lapply(x,calc),
                        row.names=SEQINR.UTIL$CODON.AA$CODON))}
tabuco.P=tabuco(seqP)
```

Vous allez maintenant étudier ces données d'usage du code grâce à une analyse factorielle des correspondances (AFC). L'AFC est une représentation graphique des données qui permet de visualiser les "correspondances" entre modalités des facteurs. Le package ade4 permet la réalisation de cette AFC avec la fonction `dudi.coa`.

```
P.coa=dudi.coa(tabuco.P)
scatter(P.coa,clab.col=1,clab.row=0)
```

Que concluez-vous ?

Comparez ces résultats à ceux correspondant au reste du génome de ces organismes. Vous allez pour cela utiliser les séquences du gène de l'alcool déshydrogénase (adh), qui sont contenues dans le fichier `adh.txt`. Malheureusement, nous ne disposons pas de cette séquence pour toutes les espèces de l'étude.

Élargissez également l'analyse à d'autres éléments transposables, à l'aide des séquences contenues dans le fichier `seqET.txt`.

Comparez l'usage du code pour les trois types de séquences.

```
tabuco.total=cbind(tabuco.P,tabuco.adh,tabuco.ET)

plot(total.coa$co,type="n")
points(total.coa$co[1:6,],pch=20)                    ## P
points(total.coa$co[7:11,],pch=21,col="blue")        ## adh
points(total.coa$co[12:25,],pch=22,col="red")        ## ET
legend("topright",c("P","adh","ET"),col=c("black","blue","red"),pch=c(20,21,22))
```

Que concluez-vous ?

3. Analyse de la divergence

Pour éprouver une hypothèse de transfert horizontal, il est également possible d'analyser les taux de divergence des séquences et les pressions de sélection auxquelles elles sont soumises. Si les éléments sont plus divergents que les gènes des hôtes, on peut penser que les éléments étaient déjà dans les génomes hôtes au moment de la divergence des espèces ; en revanche, si les éléments sont moins divergents que les gènes des hôtes, cela peut indiquer que les éléments ont récemment été introduits dans les génomes hôtes, ce qui est en faveur de l'hypothèse de transfert horizontal. La divergence entres les séquences est estimée par Ks.

Calculez les valeurs de Ks pour les séquences étudiées ici avec la fonction `kaks` de `seqinr`. Il faut pour cela travailler sur des séquences préalablement alignées. Il faut également que les longueurs des séquences soient un multiple de 3 et que la première position soit une première position de codon.

```
seqP.align=read.alignment("Pseq2_phase_3_align.txt", format="fasta")
kaks.seqP=kaks(seqP.align)
```

Comparez avec les valeurs obtenues pour le gène adh.

Que concluez-vous ?