Reconstruction phylogénétique

Nous allons étudier la phylogénie du gène de la thiorédoxine pour les 10 organismes suivants :

1. Helicobacter Pylori (Accession Number : `P66928`)
2. Bacillus subtilis (Accession Number : `P14949`)
3. Homo sapiens (Accession Number : `P10599`)
4. Penicillium chrysogenum (Accession Number : `P34723`)
5. Listeria monocytogenes (Accession Number : `P0A4L4`)
6. Escherichia coli (Accession Number : `P0AA26`)
7. Gallus gallus (Accession Number : `P08629`)
8. Mus musculus (Accession Number : `P10639`)
9. Neurospora crassa (Accession Number : `P42115`)
10. Drosophila melanogaster (Accession Number : `P47938`)

Pour chacun de ces organismes est donné entre parenthèse le numéro d'accession permettant d'identifier la protéine codant pour la thiorédoxine.

Acquisition des données

Question 1

Allez chercher les séquences protéiques de la thiorédoxine de ces dix organismes via la section Protein du NCBI (vous vérifierez que vous avez bien 10 résultats).

```
P66928 P14949 P10599 P34723 P0A4L4 P0AA26 P08629 P10639 P42115 P47938
```

Question 2

Pour chaque organisme, à l'aide de Entrez ou de vos connaissances, identifiez sa taxonomie.

A partir de cette information, en déduire une classification plausible des 10 organismes.

Question 3

Pourquoi avoir choisi la thiorédoxine ?

Pour faire une phylogénie, il faut commencer par trouver un critère de classification commun à toutes les espèces concernées. La thiorédoxine est une protéine que l'on trouve chez tous les organismes vivants, car elle intervient de dans de nombreux processus cellulaires. C'est donc un bon point de départ.

Préparation des données

Pour pouvoir comparer ces séquences afin de construire une phylogénie, nous devons construire un alignement multiple.

> Question 4
>
> Réalisez un alignement multiple des 10 séquences avec Clustal. Il est **avant** conseillé de modifier l'entête des fichiers fasta en plus compréhensible et plus court - par exemple, remplacer
>
> `>gi|135767|sp|P08629.2|THIO_CHICK RecName: Full=Thioredoxin; Short=Trx`
>
> par
>
> `>THIO_CHICKEN.`
>
> Choisissez le format PHYLIP comme format de sortie.
>
> Conservez le résultat de l'alignement.

Avant de construire une phylogénie, il faut s'assurer de la correction de l'alignement multiple. Pour cela, il est possible de tirer parti de connaissances extérieures sur la fonction des séquences étudiées, en vérifiant par exemple que les domaines connus soient bien alignés. Pour un ensemble de protéines ayant la même fonction, on doit retrouver un même domaine dans leurs séquences protéiques. Celui-ci doit être visible au niveau de l'alignement.

Pour identifier les domaines des protéines liées à la thiérodoxine, nous allons interroger la banque de données InterPro.

> Question 5
>
> Lancez la recherche de domaines connus sur la protéine de l'humain avec Interpro.
>
> Repérez les positions des domaines conservés (sur la protéine humaine).
>
> Identifiez ces positions sur l'alignement multiple des dix séquences protéiques.
>
> Vérifiez que le domaine est présent et bien aligné dans toutes les séquences.

Le bon alignement du domaine sur toutes les séquences nous donne une indication sur la pertinence de notre alignement : notre alignement est au moins correct au niveau de ces domaines.

Une petite remarque : en général, il est également souhaitable de "nettoyer" l'alignement multiple, en supprimant les régions non informatives, celles qui sont mal conservées. Sur cet exemple, comme les

séquences sont relativement bien conservées, cela n'est pas nécessaire.

Matrice de distances

Il existe plusieurs techniques pour reconstruire un arbre phylogénétique à partir de données moléculaires. En cours, vous avez vu ou vous verrez les méthodes de parcimonie et les méthodes de distance (comme UPGMA). Pour ce TP, vous allez appliquer une méthode de distance, appelée *Neighbor Joining*. *Neighbor Joining* est conçue dans le même esprit que *UPGMA*. Elle regroupe les séquences deux par deux progressivement à partir de la matrice de distances. Mais, contrairement à *UPGMA* l'hypothèse d'une *distance ultramétrique* n'est plus faite, ce qui évite de nombreuses erreurs.

Nous allons transformer cet alignement en matrice de distance. A partir de l'alignement obtenu sur Clustal, nous calculons pour chaque couple d'espèces leur distance évolutive. Le fonctionnement est similaire au calcul du score de l'alignement.

> Question 6
>
> À partir de la page contenant l'alignement multiple obtenu avec *Clustal*, téléchargez le fichier de résultat sur votre ordinateur (vous pouvez aussi le copier-coller dans un éditeur de texte et l'enregistrer).
>
> Vous utilisez le résultat de l'alignement multiple avec le logiciel *ProtDist* qui calcule une matrice de distance. Sauvegardez le résultat.

Vous devrez obtenir une matrice de distances

[Résultats]

> Question 7
>
> A première vue, dans cette matrice, quels sont les organismes les plus proches, les plus éloignés ?

Construction de l'arbre

Avec la matrice de distance, nous allons calculer un arbre phylogénétique **non enraciné** en utilisant la méthode de *Neighbor Joining*.

> Question 8
>
> Pour cela, nous allons utiliser le site T-Rex
>
> Choisissez *matrice de distance* comme type de données et collez-y la matrice de distance obtenue précédemment.
>
> Dans les résultats, le format Newick décrit l'arbre sous forme textuelle. Ce format n'est pas destiné à être lu par un humain mais à être fourni à un autre programme (par exemple pour

dessiner l'arbre).

Ici on nous propose déjà de visualiser l'arbre, il suffit de cliquer sur le lien qui propose la visualisation.

Le formulaire de lancement du programme proposait différentes méthodes de reconstruction. Essayez d'autres méthodes et comparez les arbres obtenus. Sont-ils identiques ? Sinon qu'est-ce qui les différencient ?

Evaluation de la robustesse de l'arbre

L'arbre que vous venez d'obtenir semble globalement réaliste. Mais il se peut que localement, ou concernant les longueurs des branches, celui-ci ne soit pas correct (la longueur des branches est indicative de l'évolution). Il est possible de tester la robustesse d'un arbre phylogénétique avec des techniques de *bootstrap*. L'idée du *bootstrap* est que si l'on effectue des petits changements sur les données on doit être capable de retrouver le même arbre. Concrètement, à partir de l'alignement multiple initial obtenu avec *Clustal*, on construit de nouveaux alignements qui sont obtenus en échangeant des colonnes. C'est légitime car les méthodes de reconstruction phylogénétique supposent que les sites évoluent de manière indépendante. Pour chaque nouvel alignement, on construit une matrice des distance, puis l'arbre correspondant.

Question 9

Retournez sur la page de *T-Rex*, choisissez cette fois *Sequences* comme type de données. Il faut lui fournir les séquences dans un format particulier, le format Phylip. Vous pouvez avoir un exemple de ce format en cliquant sur le lien *Data example*. Fournissez au logiciel les séquences au format voulu. Dans l'option Bootstrap, rentrez 50.

Que pensez-vous de l'arbre **strict** obtenu ?

De l'origine des cichlidés

Cet exemple est du à Xavier Vekemans, d'après l'article *"Similar morphologies of cichlid fish in Lakes Tanganyika and Malawi are due to convergence."* [PMID: 8025722].

Nous nous intéressons à l'évolution des *Cichlidés*, une famille de poissons d'eau douce. Nous nous intéressons à leur évolution dans deux grands lacs Africains : le lac Malawi (A) et le lac Tanganyika (B).

View Larger Map

Parmi les espèces de ces deux lacs, nous porterons notre intérêt sur 12 d'entre elles : *Petrochromis sp, Bathybates ferox, lobochilotes labiatus, Tropheus brichardi, Cyphotilapia frontosa* et *Julidochromis ornatus* pour le lac Tanganyika ; *Petrotilapia sp., Rhamphochromis sp., Placidochromis milomo, Pseudotropheus microstoma, Cyrtocara moori* et *Melanochromis auratus* pour le lac Malawi.

Ces 12 espèces (6 par lacs) se ressemblent deux à deux d'un point de vue morphologique. Pour une espèce du lac Tanganyika (T) on peut lui associer une espèce du lac Malawi (M). Voici les correspondances :

Tanganyika	Malawi	Ressemblances
Petrochromis sp	*Petrotilapia sp.*	Herbivore très efficace, adapté au raclage des algues, même habitat
Bathybates ferox	*Rhamphochromis sp.*	Gros prédateurs pélagiques, même morphologie hydrodynamique
lobochilotes labiatus	*Placidochromis milomo*	Prédateurs pétricoles, grosses lêvres charnues et molles
Tropheus brichardi	*Pseudotropheus microstoma*	Herbivore, bouche infère et dents bicuspides, petite taille

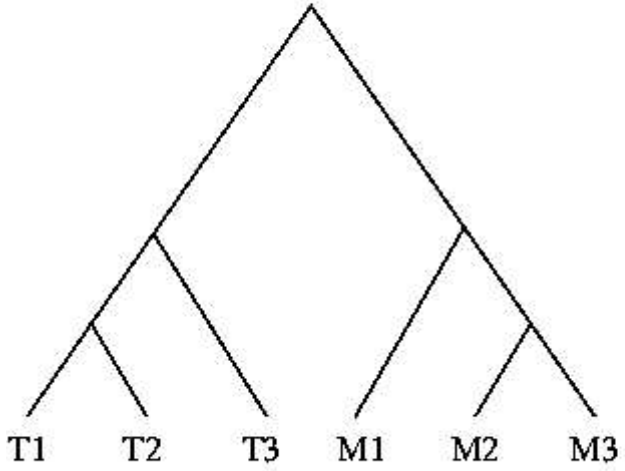
Cyphotilapia frontosa	*Cyrtocara moori*	Les mâles portent une bosse frontale
Julidochromis sp.	*Melanochromis auratus*	Rayures horizontales sur le corps

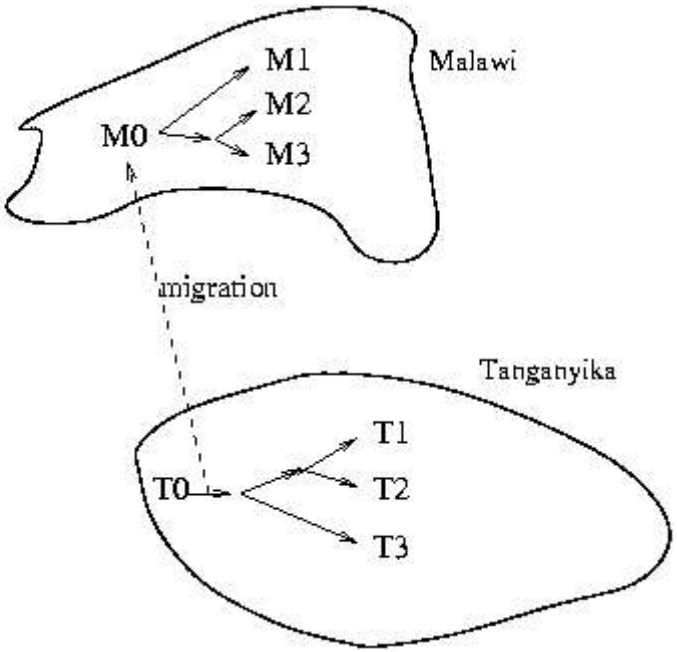
On peut éventuellement expliquer ces similitudes par deux types d'évolution

Homoplasie (convergence des caractères) :

Il y a eu évolution indépendante des poissons dans les deux lacs. les caractères similaires sont dus au milieu similaire ainsi qu'aux mêmes pressions évolutives.

Un des deux lacs (T) est plus ancien que l'autre, on peut supposer qu'il y a d'abord eu une migration de T vers M puis évolution indépendante dans les deux lacs.

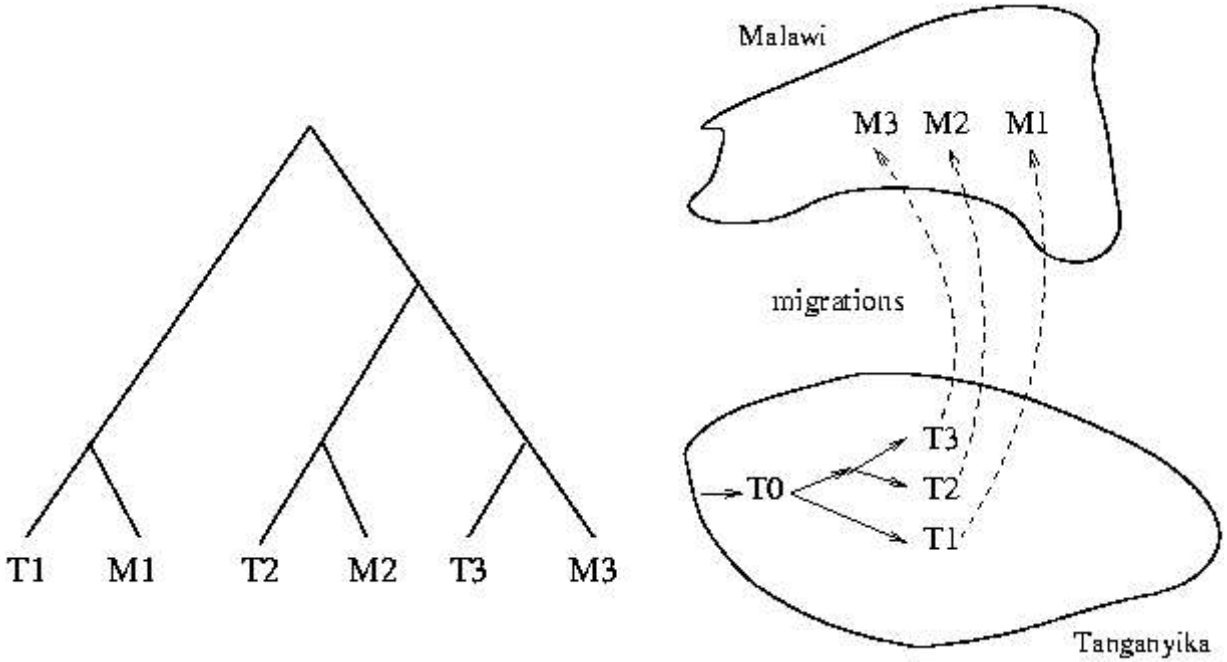




Homologie :

On considère ici que toutes les ressemblances sont dues à l'existence d'un ancêtre commun pour chacune des 6 espèces possédant les mêmes caratères que les 2 espèces respectives des deux lacs.

Il y a donc eu plusieurs évenements de migration de T vers M (un pour chaque espèce).



Le choix de l'hypothèse

Votre but est de déterminer quelle hypothèse est la plus vraissemblance en vous basant sur une analyse phylogénétique. Pour éviter de baser votre phylogénie sur des caractères adaptatifs (morphologiques) vous allez utiliser des séquences non codantes et donc non soumise à la pression évolutive.

Voici les 12 séquences : all.fas.

Question 10

A vous de faire l'analyse phylogénétique selon les différentes méthodes vues en cours :

- Faites une analyse par *UPGMA* et *Neighboor Joining*, d'abord sans, puis avec *bootstrap*. Que constatez vous?
- Faites également une phylogénie *parsimonieuse*, et par *maximum de vraisemblance*. Les arbres non enracinés sont-ils topologiquement identiques?

Aide

Logiciels

Vous pouvez vous aider des logiciels disponibles sur *Mobyle*, en particulier :

1. **Alignement :** *ClustalW*.
2. **Calcul des distances :***DNAdist* calcule une matrice de distances à partir de l'alignement multiple de *ClustalW*.
3. **Calcul de l'arbre non enraciné à l'aide d'une méthode *UPGMA* ou *Neighboor Joining* :** à partir de la matrice de distances, vous pouvez utiliser *Neighbor* qui implémente les deux algorithmes.
4. **Calcul de l'arbre non enraciné avec la parcimonie ou le maximum de vraisemblance :** *DNAPars* et *DNAML* construisent cet arbre en utilisant directement l'alignement multiple de *ClustalW*.
5. **Générer des boostraps :** chacun des logiciels précédents donne accès à un paramètre de bootstrap. Vous utiliserez le *Consensus* strict ou majoritaire quant cela sera nécéssaire (Consense).
6. **Dessiner les arbres :** vous utiliserez les logiciel *DrawTree* (arbre non enraciné), et *DrawGram* (arbre enraciné).

L'enracinement de l'arbre

On vous propose deux jeux de séquences de la famille des *Cichlidés*, l'un provenant de l'espèce *Geophagus brasiliensis* d'Amérique du Sud (newworld.fas), l'autre de l'espèce *Cyprichromis Leptosoma* de l'île de Malasa située dans le lac Tanganyika. (malasaisland.fas).

Question 11

Lequel des deux jeux est inutile pour l'enracinement? Pour quelle raison?

Construire l'arbre phylogénétique de l'ensemble, et trouver l'enracinement de l'arbre des 12 espèces

L'évolution des Cichlidés

Question 12

Depuis environ quelle période les espèces de *Cichlidés* d'Amérique du Sud ont divergé de leurs cousins éloignés d'Afrique ? (aide)

Pour conclure, le site Tree of Life permet de voyager dans l'arbre phylogénétique des espèces. Localisez y le super-odre des *Cichlidés* (aide).

Page réalisée par les membres de l'équipe bonsai, mise à jour en novembre 2011