TP Phylogénie

On vous fournit deux jeux de données :

- **Dataset 1**: un alignement protéique de la **phosphoglycerate kinase (pgk)** chez 21 espèces bactériennes 21 du groupe des firmicutes (adapté de [1]) aux formats clustalw, fasta et phylip. Le phylum des firmicutes regroupe la plupart des bactéries monodermes (Gram positives), celles à GC% faible (mais pas les Actinobacteria à GC% élevé notamment). Des classiques de laboratoire : les genres Clostridium, Bacillus, Staphylococcus, Streptococcus, Enterococcus, Lactobacillus ... mais aussi les Mollicutes.
- **Dataset 2** : un alignement nucléique du gène de la **beta glucosidase gene (bglA)** chez 10 espèces bactériennes du genre *Staphylococcus* et *Listeria* (adapté de [2]) aux formats clustalw, fasta et phylip. *Le gène bglA de la betaglucosidase est un gène utilisé dans les études MLST, nous l'étudions ici chez 5 souches de listeria et 5 souches de staphylocoques.*

Exercice 1. Choisir le modèle d'évolution le plus adapté à chacun des deux jeux des données *(use modelgenerator.jar)*

Pour cela utiliser le programme « modelgenerator.jar ». Indiquer le modèle le lus adapté pour chaque jeu de données et pour chacun des trois critères (AIC1, AIC2 et BIC).

Modelgenerator : take as input fasta or phylip files, choose 4 categories of sites.

Solution dataset 1

> *modelgenerator.jar*

input file : pgk_firmicutes.phy

number of discrete gamma categories: 4

Output : modelgenerator.out (rename !)

Solution dataset 2

> *modelgenerator.jar*

input file : bglA_listeriaStpah.phy

number of discrete gamma categories: 4

Output : modelgenerator.out (rename !)

Exercice 2. Construire l'arbre de Neighbor-Joining des deux jeux de données

Pour cela utiliser les programmes *Seaview* ou *Phylip*. Utiliser l'outil *figtree pour enraciner l'arbre selon deux types de méthodes (barycentre ou un outgroup arbitraire, du type Myco_geni pour le dataset protéique et Staph_epider pour le dataset nucléique)*

Solution

- **SEAVIEW**
  - **open alignment then distance**

- **Method : bioNJ**
- **distance matrix : Observed/Poisson/kimura => Kimura**
- **Save unrooted/rooted tree (Newick format)**
- **PHYLIP: 2 steps**
- **dnadist/protdist (be careful of outfile names !)**
- **neighbor (PHYLIP) or BIONJ**

FIGTREE

- **figtree &**
- **open tree (Newick format)**
- **Left panel : Select 'Trees' / Root tree (User Selection or Midpoint)**

Exercice 3. Construire l'arbre de parsimonie des deux jeux de données

Pour cela utiliser les programmes *Seaview* ou *Phylip.*

Solution

SEAVIEW

- ***seaview* &**
- **open alignment, Choose parcimony**

PHYLIP

- **Or *dnapars/protpars* (PHYLIP)**
- **Interpret using *figtree***

Exercice 4. Construire un arbre de Maximum de Vraisemblance pour les deux jeux de données

Pour cela utiliser les programmes *seaview* ou *phyml.*

Seaview

Trees menu/PhyML : Model (GTR/WAG) Nucleotide equilibrium frequencies : optimized, 4 categories for rate variation (For WAG select also Optimized Invariable Sites)

Phyml

phyml -i bglA_listeriaStpah.phy -d nt --quiet -c 4 -a e -m GTR

results : bglA_listeriaStpah.phy_phyml*

phyml -i pgk_firmicutes.phy -d aa --quiet -c 4 -a e –v e -m WAG

results : pgk_firmicutes.phy_phyml*

Exercice 5. Construire un arbre par une méthode bayésienne pour les deux jeux de données

Build a MrBayes tree on the nucleic dataset.

(i) Read the nexus datafile

(ii) Set the evolutionary model

(iii) Run the analysis

(iv) Summarize the samples

SOLUTION

- ***Mr Bayes step by step:*** *mb*
- ***(I) Read the nexus datafile***
- *MrBayes > execute bglA_listeriaStpah.nex*
- ***(II) Set the evolutionary model***
- *MrBayes > lset nst=6 rates=invgamma*
- ***(III) Run the analysis***
- *MrBayes > mcmc ngen=20000 samplefreq=100 printfreq=100 diagnfreq=1000*
- ***Mr Bayes step by step***
- *During the run, Mr Bayes prints samples of substitution model parameters (*.p files) ant tree samples (*.t files)*
- ***(IV) End of the analysis ?***
- *If the standard deviation of split frequencies is below 0.05 after 10000 generations stop the run, otherwise keep adding generations*
- ***(V) Summarize the samples***
- ***Summarize the parameter values***
- *MrBayes > sump*
- ***Summarize the trees***
- *MrBayes > sumt*

Exercice 6. Calculer les valeurs bootstrap des arbres de NJ et parcimonie des deux jeux de données

Utiliser Seaview ou seqboot (programme Phylip).

Seaview (Parsimony/BioNJ) => choose 'Bootsrap with 100 replicates'

Or seqboot (PHYLIP program) + neighbor or dnapars (protpars)