Exercices avec le logiciel

Épeuvre aMIG - Contrôle terminal - 26 juin 2007

J.R. Lobry, L. Duret, G. Perrière

11 mars 2008

Les trois problèmes sont complètement indépendants.

(Durée : 3 heures)
Documents autorisés
Envoyez (avant la fin de l'épreuve, heure de réception du mél faisant foi !)
votre compte-rendu au format PDF à :
`{lobry, duret, perriere}@biomserv.univ-lyon1.fr`.

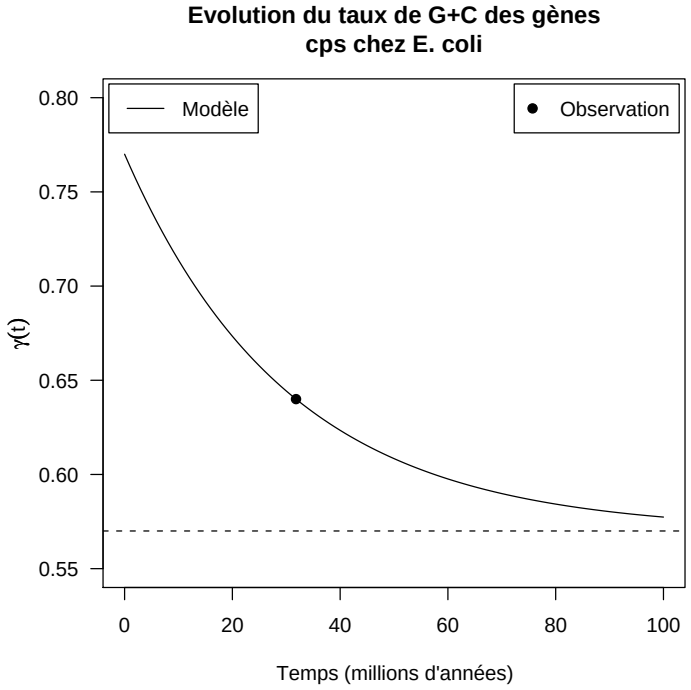
1 Problème 1 (J.R. Lobry)

L'évolution au cours du temps ($t \geq 0$) de la fréquence relative en bases C ou G, $\gamma(t)$, dans les génomes a été modélisée par Sueoka en 1962. Les hypothèses de ce modèle se traduisent sous la forme d'une équation différentielle,

$$\frac{d\gamma}{dt} = -(u+v)\gamma + v$$

où u et v sont des paramètres strictement positifs représentant les taux de mutation des bases C ou G vers A ou T, et des bases A ou T vers C ou G, respectivement.

1. Soit γ^* la valeur de $\gamma(t)$ telle que $\frac{d\gamma}{dt} = 0$. Exprimer γ^* en fonction de u et v.
2. Soit la condition initiale $\gamma_0 = \gamma(0)$. Donner l'expression de $\gamma(t)$ en fonction de t, γ_0, γ^*, u et v.
3. Etudier succintement le modèle obtenu en 2). Dans cette étude, on constatera en particulier l'existence de trois cas de figure, selon le sens de croissance de la fonction $\gamma(t)$.
4. Les gènes *cps* nécessaires pour la production d'acide colanique ont été acquis dans le génome de *Escherichia coli* par transfert horizontal en provenance de *Salmonella enterica*. Sachant que la fréquence relative en bases G ou C de l'ensemble des gènes est de 0.57 chez *E. coli*, de 0.77 chez *S. enterica*, et de 0.64 pour les gènes *cps* chez *E. coli*, avec $u + v = 3.3\ 10^{-8}\text{an}^{-1}$, en déduire depuis combien d'années les gènes *cps* sont présents dans le génome de *E. coli*.
5. Donner le code permettant de produire le graphique suivant :



2 Problème 2 (L. Duret)

La protéine MaProt (de souris) a été caractérisée en 1998. Cette protéine a été comparée à la banque de données SwissProt à l'aide du logiciel BLASTP. La liste des séquences similaires détectées à l'époque par BLASTP est indiquée ci-dessous :

```
################################################################################
BLASTP 2.0.5 [May-5-1998]

Query= MaProt

Database: SwissProt
           600,231 sequences; 186,808,058 total letters

                                                                 Score     E
Sequences producing significant alignments:                      (bits)  Value

Seq1    264 Human Seq1 protein.                                  1851  0
Seq2    193 Chicken  Seq2 protein                                 138  1e-55
Seq3    351 Zebrafish Seq3 protein.                               92  1e-03
Seq4    558 Drosophila Seq4 protein.                              31  0.7
Seq5    531 Caenorhabditis elegans Seq5 protein.                  42  0.9
################################################################################
```

En 2007, on a refait la recherche de similarité avec BLASTP sur la dernière version de la banque de données SwissProt. On a obtenu le résultat suivant :

```
################################################################################
BLASTP 2.1.8 [April-15-2007]

Query= MaProt

Database: SwissProt
           6,8102,836 sequences; 191,518,738,896 total letters

                                                                  Score     E
Sequences producing significant alignments:                       (bits)  Value

Seq1    264 Human Seq1 protein.                                   1851  0
Seq2    193 Chicken  Seq2 protein                                  138  1e-54
Seq3    351 Zebrafish Seq3 protein.                                92  0.01
Seq4    558 Drosophila Seq4 protein.                               31  7
Seq5    531 Caenorhabditis elegans Seq5 protein.                   42  9
################################################################################
```

1. Donnez la liste des homologues identifiés par BLAST en 2007 et ceux identifiés en 1998 (vos choix doivent être justifiés).
2. Donnez la définition de la "E-value".
3. Pourquoi la "E-value" est-elle différente en 2007 par rapport à 1998 ?
4. Quelle méthode de recherche de similarité pourrait-on utiliser pour gagner en sensibilité ?

3 Problème 3 (G. Perrière)

Soit l'arbre non raciné donné dans la figure 1, obtenu à partir d'une famille de gènes homologues présente chez les procaryotes (bactéries + archées).

1. Représentez cet arbre en utilisant le groupe d'organismes que vous jugez le plus pertinent si l'on considère une étude portant sur l'ensemble des bactéries.
2. En prenant la phylogénie de Calteau (2004) (*cf.* cours) comme référence, interprétez en termes de transferts horizontaux, de duplications et de pertes de gènes la phylogénie observée. Pour ce faire, vous pouvez représenter sur un deuxième arbre l'ensemble des évènements en question. Si plusieurs hypothèses sont possibles pour expliquer un regroupement atypique, ne représentez que celle étant la plus parcimonieuse (c'est-à-dire impliquant le moins d'évènements). On supposera qu'aucune erreur n'est intervenue dans la construction de l'arbre ci-dessus.
3. Par ailleurs, les deux gènes présents chez *Bacillus subtilis* sont-ils orthologues ou paralogues des deux gènes chez *Streptococcus pneumoniae* ? Expliquez votre réponse.

FIG. 1 – Arbre non raciné obtenu à partir d'une famille de gènes homologues présente chez les procaryotes