Exercices avec le logiciel

Épreuve aMIG - Contrôle terminal - 24 juin 2008

J.R. Lobry, L. Duret, A. Necşulea

27 juin 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, necsulea}@biomserv.univ-lyon1.fr`.

1 Taux de G+C et longueur des CDS (J.R. Lobry)

Un paramètre important pour la prédiction des séquences codantes des génomes bactériens est la longueur des CDS prédits : à partir de quand cette longueur est elle anormalement grande ar rapport à ce qui serait attendu au hasard en moyenne sous un modèle donné ? On s'intéresse ici à un modèle neutre très simple [5].

1.1 Probabilité des trois codons stops

On note θ le taux de G+C ($\theta \in [0, 1]$). Pour simplifier on considère que la deuxième règle de parité (PR2) est valide et que l'on a donc $P_C = P_G = \frac{\theta}{2}$ et $P_A = P_T = \frac{1-\theta}{2}$, où P_X représente la probabilité de tirer avec remise dans une urne la base X. Donner pour chacun des trois codons stops la probabilté de tirer ce codon en fonction de θ. Le code génétique standard est donné dans la table 1.

1.2 Probabilité d'un codon stop

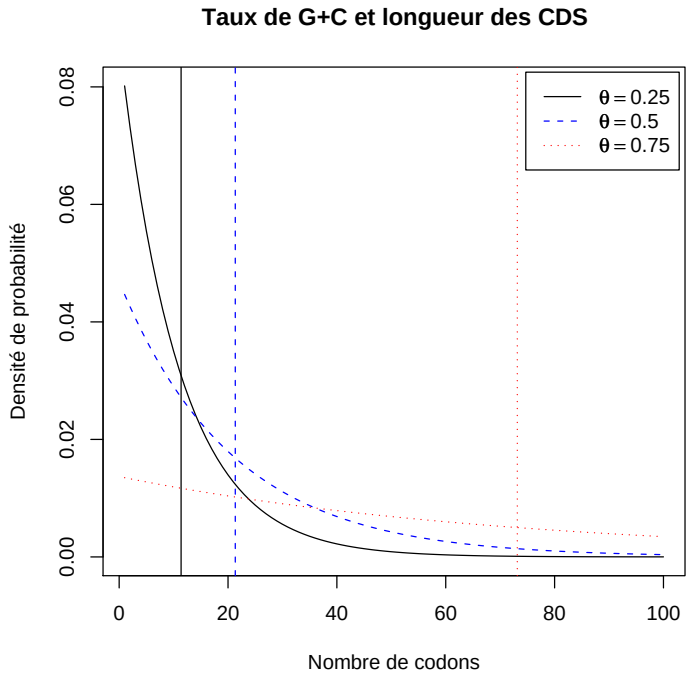
Donner la probabilité de tirer un codon stop en fonction de θ.

1.3 Fonction de densité de probabilité

On s'intéresse à la longueur des CDS c'est à dire au nombre de tirage consécutifs de codons non stop. Donner le code  permettant de représenter la fonction de densité de probabilité pour la longueur des CDS pour $\theta = 0.25$, $\theta = 0.5$ et $\theta = 0.75$. On veut de plus que l'espérance soit représentée par une ligne verticale pour avoir une représentation du type suivant :

TTT	Phe	TCT	Ser	TAT	Tyr	TGT	Cys
TTC	Phe	TCC	Ser	TAC	Tyr	TGC	Cys
TTA	Leu	TCA	Ser	TAA	Stp	TGA	Stp
TTG	Leu	TCG	Ser	TAG	Stp	TGG	Trp
CTT	Leu	CCT	Pro	CAT	His	CGT	Arg
CTC	Leu	CCC	Pro	CAC	His	CGC	Arg
CTA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CTG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
ATT	Ile	ACT	Thr	AAT	Asn	AGT	Ser
ATC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
ATA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
ATG	Met	ACG	Thr	AAG	Lys	AGG	Arg
GTT	Val	GCT	Ala	GAT	Asp	GGT	Gly
GTC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GTA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GTG	Val	GCG	Ala	GAG	Glu	GGG	Gly

TAB. 1 – Le code génétique standard. Les codons stop sont notés Stp.



1.4 Test du modèle neutre

Sachant que la longueurs des CDS dans les génomes bactériens est de l'ordre de 330 codons, que peut on dire du modèle neutre précédent ?

1.5 Prédiction des CDS

Est-il plus facile de prédire les CDS dans un génome à bas G+C ou bien dans un génome à haut G+C ? Pourquoi ?

1.6 Interprétation biologique

Supposons que nous soyons dans le cas où le modèle neutre précédent est rejeté. Comment interprétez vous ce résultat d'un point de vue biologique ?

1.7 Question subsidiaire

Passez aux questions suivantes avant d'essayer de répondre à cette question. S'il y a un effet du taux de G+C sur la longueur des CDS on devrait s'attendre à avoir des génomes plus denses en CDS pour les génomes bactériens à bas G+C. Les données de GOLD [2, 1, 4, 3] que vous avez utilisées en cours vous permettent-elles d'apporter des éléments de réponse à cette hypothèse ?

2 Problème 2 (L. Duret)

2.1 Génome de souris

Pour annoter une séquence génomique de souris, des chercheurs ont utilisé différentes approches de prédiction de gènes.

Les positions des exons identifiés par SIM4 et par GeneWise sont indiquées ci-dessous :

```
         SIM4     GeneWise
Exon 1  453..522 501..522
Exon 2 1422..1570 1422..1570
Exon 3 1931..2121 1931..2121
Exon 4 3156..3311 3156..3311
Exon 5 3854..3968 3854..3968
Exon 6 5789..5932 5789..5932
Exon 7 6841..7405 6841..6902
```

a Détaillez le protocole suivi pour obtenir ces résultats (expliquez l'objectif de chaque étape du protocole ; indiquez précisément les bases de données et logiciels utilisés)

b Quelle pourrait être l'explication des différences entre les résultats de SIM4 et ceux de GeneWise ? Comment procéderiez-vous pour le vérifier ?

2.2 Protéine de cheval

La protéine MaProt (de cheval) 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 2.1.8 [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  0.1
Seq4    558 Drosophila Seq4 protein.                             31  0.7
Seq5    531 Caenorhabditis elegans Seq5 protein.                 42  3
###########################################################################
```

a Donnez la liste des homologues identifiés par BLAST (justifiez votre choix).

b Donnez la définition de la "E-value".

c Quelle méthode de recherche de similarité pourrait-on utiliser pour gagner en sensibilité ?

3 Duplications de génome chez la levure (A. Necşulea)

La levure du boulanger *Saccharomyces cerevisiae* est un organisme polyploïde dégénéré ; cela signifie que son génome a subi un événèment de duplication complète, suivi de perte de gènes. Nous disposons actuellement de plusieurs génomes complètement séquencés appartenant au genre *Saccharomyces*, ainsi que d'autres génomes complets de champignons plus distants, comme par exemple celui de *Candida albicans*. On vous propose ici d'analyser des alignements de 5 familles de gènes orthologues, appartenant à *Saccharomyces cerevisiae*, *Saccharomyces kluyveri* et *Candida albicans*. Les alignements des séquences protéiques sont disponibles à l'adresse `http://biomserv.univ-lyon1.fr/~necsulea/tpphylo/examen/`, au format `mase` et `phylip`.

a Analysez les alignements donnés, d'abord avec une méthode de distance de type Poisson, et ensuite au maximum de vraisemblance, avec le modèle d'évolution JTT. Pouvez-vous dater la duplication de gènome ? A-t-elle eu lieu avant ou après la divergence des espèces *S. cerevisiae* et *S. kluyveri* ? Comment peut-on s'assurer de la fiabilité des résultats obtenus ? Détaillez tout votre raisonnement.

b Analysez les vitesses d'évolution des marqueurs, en particulier pour les deux copies des gènes présentes chez *S. cerevisiae*. Pouvez-vous proposer une explication pour l'incongruence entre les topologies soutenues par les différents marqueurs ?

c Essayez de proposer une explication pour les différences de vitesse d'évolution observées pour les deux copies de gènes de *S. cerevisiae*.

Références

[1] A. Bernal, U. Ear, and N. Kyrpides. Genomes online database (GOLD) : a monitor of genome projects world-wide. *Nucleic Acids Res.*, 29 :126–127, 2001.

[2] N.C. Kyrpides. Genomes online database (GOLD 1.0) : a monitor of complete and ongoing genome projects world-wide. *Bioinformatics*, 15 :773–774, 1999.

[3] K. Liolios, K. Mavrommatis, N. Tavernarakis, and N.C. Kyrpides. The genomes on line database (GOLD) in 2007 : status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research*, in press :D000–D000, 2008.

[4] K. Liolios, N. Tavernarakis, P. Hugenholtz, and N.C. Kyrpides. The genomes on line database (GOLD) v.2 : a monitor of genome projects worldwide. *Nucleic Acids Research*, 34 :D332–D334, 2006.

[5] J.L. Oliver and A. Marin. A relationship between GC content and coding-sequence length. *Journal of Molecular Evolution*, 43 :216–223, 1996.