Problème pratique de statistique n° pps067

L'usage du code génétique chez *Borrelia garinii*

Anamaria Necşulea (M2 - EEB - 2005)

16 octobre 2016

832 gènes de *Borrelia garinii* contre 64 codons pour 20 acides aminés : Anamaria Necşulea introduit à un problème difficile. Quelles variables permettent une prédiction de l'usage du code ? Pour un premier accès à `seqinr` ? Pour sortir des sentiers battus de l'AFC ?

Table des matières

1 Introduction

1.1 L'analyse de l'usage du code

La nature dégénerée du code génétique fait que la plupart des acides aminés peuvent être codés par plusieurs codons (triplets de nucleotides), que l'on appelle alors synonymes. Tous les codons synonymes ne sont pas utilisés à la même fréquence, et chaque espèce a son propre pattern d'usage du code [4].

Pour certaines espèces, l'usage du code peut s'expliquer par une approche sélectionniste : le choix des codons synonymes est fait de façon à optimiser l'efficacité de la traduction. En effet, pour traduire les ARNm en protéines il faut passer par l'intermédiaire des ARN de transfert, qui ont des abondances différentes à l'intérieur de la cellule. Ainsi, un codon reconnu par un ARNt très abondant sera préféré (dans les gènes fortement exprimés) car il augmente la vitesse de traduction. Cette hypothèse a été validée par l'observation d'une corrélation entre l'expressivité des gènes et le biais d'usage du code [6].

Dans d'autres cas, la pression de mutation peut fournir une explication vraisemblable de l'usage du code. La composition en nucléotides A,C,G,T d'un génome sera corrélée avec les préférences entre les différents codons synonymes [1].

L'usage global (synonyme et non-synonyme) des codons inclut un effet au niveau des acides aminés. Par exemple, selon la nature des protéines et leur localisation sub-cellulaire, il y aura une préférence pour des acides aminés hydrophiles ou hydrophobes.

Dans certaines situations particulières (comme par exemple pour les bactéries du genre *Chlamydia* ou pour *Borrelia burgdorferi*), un facteur important qui déterminera le pattern d'usage du code sera la localisation des gènes sur le chromosome, relative à l'origine de réplication. Ainsi, les gènes qui se trouvent sur le brin **leading** n'utiliseront pas les mêmes codons que ceux du brin **lagging** [8].

1.2 L'organisme étudié

Les bactéries du genre *Borrelia*, classe des Spirochètes, sont des parasites qui provoquent chez l'homme la maladie de Lyme (ou borréliose). Le génome de *Borrelia garinii*, qui est répandue en Europe, a été récemment séquencé [3]. Il est composé d'un chromosome linéaire de 904246 paires de bases, ainsi que de plusieurs plasmides. Du point de vue de la topologie du chromosome, les *Borrelia* représentent une exception au sein des bactéries, qui possèdent généralement des chromosomes circulaires.

2 Problématique

Ce jeu de données devrait permettre de répondre à plusieurs questions.

Tout d'abord, on doit se demander quels sont les facteurs qui régissent l'usage du code chez cet organisme. Est-ce que la pression de séléction (évaluée par la corrélation entre le biais d'usage du code et le niveau d'expression des gènes) prime sur la pression de mutation (détérminée par l'effet de la composition en bases), ou le contraire ? Est-ce que la hydrophobicité des acides aminés a un effet au niveau de l'usage des codons ?

En même temps, on peut s'interroger sur la validité de l'indice d'adaptation des codons dans cette situation. Bien que nous l'ayons défini à partir des gènes fortement exprimés (codant pour des protéines ribosomiques), il peut y avoir des effets de confusion, suite auxquels le CAI ne serait pas bien aproprié pour décrire l'expressivité des gènes.

Et finalement, on pourra s'intéresser à l'effet du positionnement des gènes sur les deux brins du chromosome sur l'usage des codons.

3 Description des données

Les séquences codantes de *Borrelia garinii* ont été récupérées à partir de la banque de données GenBank, grâce au système d'intérrogation des banques *WWW-Query* [7]. Seulement les séquences codantes complètes et de longueur supérieure à 150 paires de bases ont été conservées.

Le jeu de données final comprend 832 gènes protéiques. Les fréquences des codons ont été calculées pour chaque gène, ainsi que les contenus en G+C et les indices de hydropathie de Kyte et Doolittle [5].

L'origine de réplication du chromosome a été localisée grâce à Oriloc [2], pour ensuite déterminer la position de chaque gène sur les brins **leading** ou **lagging**.

Faute de données décrivant l'abondance des ARNt et les niveaux d'expression des gènes chez *Borrelia garinii*, on a utilisé l'indice d'adaptations des codons ou CAI [9]. Cet indice utilise un ensemble de gènes pour lesquels on suppose un fort niveau d'expression (ici les protéines ribosomiques) pour évaluer l'apport de chaque codon à l'expressivité. Pour chaque gène on calcule ensuite un score en fonction des fréquences des codons. Les gènes qui ont un CAI fort devraient avoir aussi un fort niveau d'expression, et inversement pour les CAI faibles.

4 Lecture des données

Le tout est disponible dans l'objet **pps067**.

```
load(url("http://pbil.univ-lyon1.fr/R/donnees/pps067.rda"))
names(pps067)
```

```
[1] "codons" "info"   "rib"
```

La composante **codons** est un tableau (832 x 64) contient pour chacun des 832 gènes les fréquences des 64 codons.

```
dim(pps067$codons)
```

```
[1] 832  64
```

```
names(pps067$codons)
```

```
 [1] "aaa" "aac" "aag" "aat" "aca" "acc" "acg" "act" "aga" "agc" "agg" "agt" "ata"
[14] "atc" "atg" "att" "caa" "cac" "cag" "cat" "cca" "ccc" "ccg" "cct" "cga" "cgc"
[27] "cgg" "cgt" "cta" "ctc" "ctg" "ctt" "gaa" "gac" "gag" "gat" "gca" "gcc" "gcg"
[40] "gct" "gga" "ggc" "ggg" "ggt" "gta" "gtc" "gtg" "gtt" "taa" "tac" "tag" "tat"
[53] "tca" "tcc" "tcg" "tct" "tga" "tgc" "tgg" "tgt" "tta" "ttc" "ttg" "ttt"
```

Le code génétique n'est pas inclu dans les données, mais il peut être obtenu d'une façon très élégante en utilisant la librairie **seqinr** de R :

```
library(seqinr)
tablecode()
# words(3)
# translate(apply(words(3),1,s2c))
```

Genetic code 1 : standard

T T T	Phe	T C T	Ser	T A T	Tyr	T G T	Cys
T T C	Phe	T C C	Ser	T A C	Tyr	T G C	Cys
T T A	Leu	T C A	Ser	T A A	Stp	T G A	Stp
T T G	Leu	T C G	Ser	T A G	Stp	T G G	Trp
C T T	Leu	C C T	Pro	C A T	His	C G T	Arg
C T C	Leu	C C C	Pro	C A C	His	C G C	Arg
C T A	Leu	C C A	Pro	C A A	Gln	C G A	Arg
C T G	Leu	C C G	Pro	C A G	Gln	C G G	Arg
A T T	Ile	A C T	Thr	A A T	Asn	A G T	Ser
A T C	Ile	A C C	Thr	A A C	Asn	A G C	Ser
A T A	Ile	A C A	Thr	A A A	Lys	A G A	Arg
A T G	Met	A C G	Thr	A A G	Lys	A G G	Arg
G T T	Val	G C T	Ala	G A T	Asp	G G T	Gly
G T C	Val	G C C	Ala	G A C	Asp	G G C	Gly
G T A	Val	G C A	Ala	G A A	Glu	G G A	Gly
G T G	Val	G C G	Ala	G A G	Glu	G G G	Gly

Pour passer des codons à l'acide aminé codé faire simplement :

```
w <- names(pps067$codons)
w <- translate(sapply(w,s2c))
w
```

```
 [1] "K" "N" "K" "N" "T" "T" "T" "T" "R" "S" "R" "S" "I" "I" "M" "I" "Q" "H" "Q" "H"
[21] "P" "P" "P" "P" "R" "R" "R" "R" "L" "L" "L" "L" "E" "D" "E" "D" "A" "A" "A" "A"
[41] "G" "G" "G" "G" "V" "V" "V" "V" "*" "Y" "*" "Y" "S" "S" "S" "S" "*" "C" "W" "C"
[61] "L" "F" "L" "F"
```

seqinr contient tout le nécessaire pour travailler sur ce type de données. Par exemple, pour passer du code IUAPC à un caractère au noms des acides aminés, utiliser

```
   data(SEQINR.UTIL)
   SEQINR.UTIL$CODON.AA[c(1:5,60:64),]
   CODON  AA L
1    aaa Lys K
2    aac Asn N
3    aag Lys K
4    aat Asn N
5    aca Thr T
60   tgt Cys C
61   tta Leu L
62   ttc Phe F
63   ttg Leu L
64   ttt Phe F
   aaa(w)
 [1] "Lys" "Asn" "Lys" "Asn" "Thr" "Thr" "Thr" "Thr" "Arg" "Ser" "Arg" "Ser" "Ile"
[14] "Ile" "Met" "Ile" "Gln" "His" "Gln" "His" "Pro" "Pro" "Pro" "Pro" "Arg" "Arg"
[27] "Arg" "Arg" "Leu" "Leu" "Leu" "Leu" "Glu" "Asp" "Glu" "Asp" "Ala" "Ala" "Ala"
[40] "Ala" "Gly" "Gly" "Gly" "Gly" "Val" "Val" "Val" "Val" "Stp" "Tyr" "Stp" "Tyr"
[53] "Ser" "Ser" "Ser" "Ser" "Stp" "Cys" "Trp" "Cys" "Leu" "Phe" "Leu" "Phe"
```

La composante **info** est un tableau (832 x 6) contient pour chaque gène :

- ⋆ le numéro d'accession GenBank (**$Acc**) ;
- ⋆ le contenu en guanine et cytosine (**$GC**) ;
- ⋆ le contenu en guanine et cytosine en troisième position des codons(**$GC3**) ;
- ⋆ l'indice de hydropathie de Kyte et Doolittle (**$K.D**) ;
- ⋆ l'indice d'adaptation des codons (CAI) (**$CAI**) ;
- ⋆ la localisation sur le chromosome (brins leading ou lagging) (**$strand**).

```
   dim(pps067$info)
[1] 832   6
   names(pps067$info)
[1] "Acc"    "GC"     "GC3"    "K.D"    "CAI"    "strand"
```

Enfin, la composante **rib** est un vecteur de caractères à 53 composantes contenant la liste des numéros d'accession GenBank pour les protéines ribosomales utilisées pour calculer le CAI.

```
   length(pps067$rib)
[1] 53
   pps067$rib[1:8]
[1] "CP000013.RPLI" "CP000013.RPSR" "CP000013.RPSF" "CP000013.RPSB" "CP000013.RPSA"
[6] "CP000013.RPLT" "CP000013.RPMI" "CP000013.RPME"
```

Références

[1] M. Bulmer. The selection-mutation-drift theory of synonymous codon usage. *Genetics*, 129 :897–907, 1991.

[2] A.C. Frank and J.R. Lobry. Oriloc : prediction of replication boundaries in unannotated bacterial chromosome. *Bioinformatics*, 16 :560–561, 2000.

[3] G. Glockner, R. Lehmann, A. Romualdi, S. Pradella, U. Schulte-Spechtel, M. Schilhabel, B. Wilske, J. Suhnel, and M. PLatzner. Comparative analysis of the *Borrelia garinii* genome. *Nucleic Acids Research*, 32(20) :6038–6046, 2004.

[4] R. Grantham, C. Gautier, and M. Gouy. Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. *Nucleic Acids Research*, 8 :1892–1912, 1980.

[5] J. Kyte and R.F. Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157(1) :105–132, 1982.

[6] M.Gouy and C. Gautier. Codon usage in Bacteria : correlation with gene expressivity. *Nucleic Acids Research*, 10(22) :7055–7074, 1982.

[7] G. Perriere and M. Gouy. WWW-query : an on-line retrieval system for biological sequence banks. *Biochimie*, 78(5) :364–369, 1996.

[8] H. Romero, A. Zavala, and H. Musto. Codon usage in *Chlamydia trachomatis* is the result of strand-specific mutational biases and a complex pattern of selective forces. *Journal of Molecular Biology*, 157(1) :105–132, 1982.

[9] P.M. Sharp and W.H. Li. The codon adaptation index – a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research*, 15(3) :1281–1295, 1987.