Problème pratique de statistique n° pps065

Arbres génomiques

Anamaria Necşulea (M2 - EEB - 2005)

16 octobre 2016

La présence ou l'absence de 3307 familles de gènes protéiques homologues est compilée dans le génomes de 33 organismes dont 27 bactéries. Sur la marge espèce de ce tableau, on possède des mesures de dimensions et un arbre phylogénétique. Des données originales et des questions ouvertes proposées par Anamaria Necşulea (M2 - EEB - 2005).

Table des matières

1 Description du problème

Le principal objectif de la phylogénie moléculaire est de reconstituer la généalogie des espèces étudiées, en se basant sur les similarités observées entre les séquences (nucléiques ou protéiques) du marqueur moléculaire choisi. Toutefois, l'intervention des phénomènes comme le transfert horizontal, la duplication des gènes ou la présence des vitesses d'évolution différentes entre les séquences peut rendre cet objectif intangible. Le résultat est alors plutôt un "arbre des gènes" que l'arbre des espèces que l'on recherche.

En partant de cette observation, certains auteurs ont proposé des approches phylogénetiques basées sur la comparaison des génomes entiers ([1],[2]). A ce jour, plus de 200 génomes complètement sequencés et annotés ont été publiés, d'où la faisabilité de cette méthode.

Pour donner une distance entre deux génomes, il faut d'abord déterminer les groupes de gènes orthologues (c'est à dire les gènes qui sont dérivés d'un même ancêtre commun par spéciation) présents chez les organismes que l'on souhaite étudier. La recherche d'orthologues se fait sur la base de la similarité des séquences. Il existe plusieurs banques de données de gènes homologues (qui proviennent d'un même ancêtre commun), qui peuvent simplifier cette recherche (HOBACGEN [1], HOGENOM [2], COG [3]). Ensuite, une mesure de similarité entre deux espèces peut être définie à partir du nombre de gènes orthologues qu'elles ont en commun.

Il paraît vraisemblable que l'impact des événements de duplication ou des différences entre les taux dévolution des séquences sur la topologie de l'arbre phylogénétique soit diminué avec cette approche. Une fois les orthologues définis, la méthode n'utilise plus les similarités entre les séquences pour définir les distances entre les différents organismes. On peut toutefois se demander jusqu'à quel point on peut reconstituer l'histoire évolutive des organismes en analysant le contenu en gènes de leur génomes. Est-ce que cette approche permet de discriminer des organismes qui sont très proches, aussi bien que le font les méthodes classiques de phylogénie moléculaire ? Est-ce qu'elle est sensible à la taille des génomes ?

La construction d'arbres "génomiques" semble être une innovation importante dans la phylogénie moléculaire, mais la validité de ce type d'approche reste encore à être confirmée.

2 Données

La totalité des données se retrouvent dans le fichier **pps065.rda**.

1. http ://pbil.univ-lyon1.fr/databases/hobacgen.html
2. http ://pbil.univ-lyon1.fr/databases/hogenom.html
3. http ://www.ncbi.nlm.nih.gov/COG/

```
load(url("http://pbil.univ-lyon1.fr/R/donnees/pps065.rda"))
names(pps065)
```
```
[1] "COGs"  "descr" "orga"  "tree"
```

Les données ont été acquises à partir de la version initiale, datant de 2003, de la banque **COG** (Clusters of Orthologous Groups) [3]. Cette banque contient 3307 familles de gènes protéiques homologues.

La composante **descr** contient pour chacune de ces familles la description des protéines qu'elles encodent.

```
length(pps065$descr)
```
```
[1] 3307
```
```
pps065$descr[1:10]
```
```
 [1] "[H]_COG0001_Glutamate-1-semialdehyde_aminotransferase"
 [2] "[E]_COG0002_Acetylglutamate_semialdehyde_dehydrogenase"
 [3] "[D]_COG0003_Predicted_ATPase_involved_in_chromosome_partitioning"
 [4] "[P]_COG0004_Ammonia_permeases"
 [5] "[F]_COG0005_Purine_nucleoside_phosphorylase"
 [6] "[E]_COG0006_Xaa-Pro_aminopeptidase"
 [7] "[H]_COG0007_Uroporphyrinogen-III_methylase"
 [8] "[J]_COG0008_Glutamyl-_and_glutaminyl-tRNA_synthetases"
 [9] "[J]_COG0009_Putative_translation_factor_(SUA5)"
[10] "[E]_COG0010_Arginase/agmatinase/formimionoglutamate_hydrolase,_arginase_family"
```

Nous avons décidé d'étudier ici 33 organismes, dont 25 bactéries, 1 eucaryote et 7 archées. Pour chacun d'entre ces organismes et pour chaque famille de COG, nous avons noté la présence ou l'absence (codée en 0 et 1) des gènes appartenant à cette famille dans le génome étudié. Ici, nous avons ignoré le fait que plusieurs copies du même gène peuvent être présentes dans le même génome. Ces donnés sont présentées dans la composante **COGs**, sous la forme d'un tableau de 33 lignes et 3307 colonnes (avec des noms des lignes et des colonnes).

```
dim(pps065$COGs)
```
```
[1]   33 3307
```
```
pps065$COGs[1:10,1:8]
```
```
    COG0001 COG0002 COG0003 COG0004 COG0005 COG0006 COG0007 COG0008
Vch       1       1       0       0       0       1       1       1
Eco       1       1       0       1       1       1       1       1
Buc       0       1       0       0       0       0       1       1
Hin       0       0       0       0       0       1       0       1
Pae       1       1       0       1       1       1       1       1
Nme       1       1       0       1       0       1       1       1
Xfa       1       1       0       1       1       1       1       1
Rpr       0       0       0       0       0       1       0       1
Cje       1       1       0       0       0       1       0       1
Hpy       1       0       0       0       0       1       0       1
```

On dispose aussi dans la composante **orga** d'un tableau contenant, pour chaque organisme :

— l'abbréviation utilisée pour le désigner ;

— l'espèce ;

```
pps065$orga[,1:2]
```
```
   Abbr.                          Species
1    Vch                   Vibrio cholerae
2    Eco                 Escherichia coli
3    Buc                     Buchnera_sp
4    Hin             Haemophilus influenzae
5    Pae             Pseudomonas aeruginosa
6    Nme             Neisseria meningitidis
7    Xfa                 Xylella fastidiosa
8    Rpr              Rickettsia prowazekii
9    Cje               Campylobacter jejuni
10   Hpy                Helicobacter pylori
```

```
11   Bha                 Bacillus halodurans
12   Bsu                   Bacillus subtilis
13   Mtu           Mycobacterium tuberculosis
14   Mle                 Mycobacterium leprae
15   Bbu                 Borrelia burgdorferi
16   Tpa                   Treponema pallidum
17   Uur               Ureaplasma urealyticum
18   Mge                 Mycoplasma genitalium
19   Mpn                 Mycoplasma pneumoniae
20   Dra              Deinococcus radiodurans
21   Cpn             Chlamydophila pneumoniae
22   Ctr                Chlamydia trachomatis
23   Syn                    Synechocystis sp.
24   Aae                     Aquifex aeolicus
25   Tma                 Thermotoga maritima
26   Ape                    Aeropyrum pernix
27   Hbs                   Halobacterium sp.
28   Mth Methanobacterium thermoautotrophicum
29   Tac             Thermoplasma acidophilum
30   Afu                Archaeglobus fulgidus
31   Pab                    Pyrococcus abyssi
32   Pho                Pyrococcus horikoshii
33   Sce             Saccharomyces cerevisiae
```

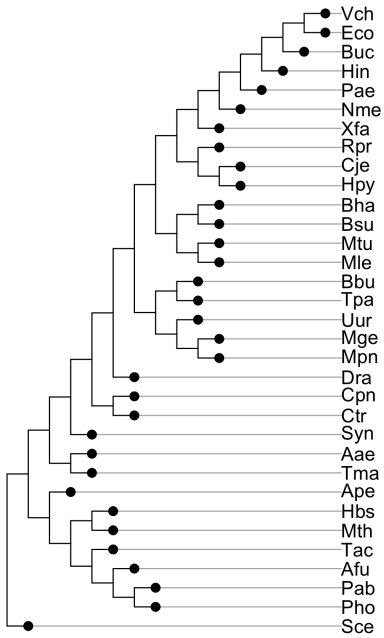
— le règne auquel il appartient (Bacteria,Archaea ou Eukaryota) ;
— la taille du génome, en millions de paires de bases ;
— le nombre total de gènes protéiques présents dans son génome ;
— le nombre de gènes protéiques répertoriés dans la base COG.

```
pps065$orga[,3:6]
```

```
    Kingdom Gen.size Nb.prot Nb.cogs
1   Bacteria     2.96    3835    2820
2   Bacteria     4.63    4275    3414
3   Bacteria     0.64     575     568
4   Bacteria     1.83    1714    1542
5   Bacteria     6.26    5567    4392
6   Bacteria     2.27    2080    1506
7   Bacteria     2.67    2831    1589
8   Bacteria     1.11     835     697
9   Bacteria     1.64    1634    1302
10  Bacteria     1.66    1576    1096
11  Bacteria     4.20    4066    2878
12  Bacteria     4.21    4118    2870
13  Bacteria     4.40    3927    2585
14  Bacteria     3.26    1605    1134
15  Bacteria     0.91    1637     696
16  Bacteria     1.13    1036     716
17  Bacteria     0.75     614     406
18  Bacteria     0.58     484     381
19  Bacteria     0.82     689     425
20  Bacteria     3.06    3187    2226
21  Bacteria     1.23    1054     648
22  Bacteria     1.04     895     631
23  Bacteria     3.57    3167    2159
24  Bacteria     1.55    1560    1329
25  Bacteria     1.86    1858    1527
26   Archaea     1.67    1841    1178
27   Archaea     2.01    2605    1701
28   Archaea     1.75    1873    1330
29   Archaea     1.56    1482    1230
30   Archaea     2.18    2420    1872
31   Archaea     1.76    1768    1456
32   Archaea     1.74    1800    1378
33 Eukaryota    14.21    5955    2290
```

On dispose enfin, pour ces 33 espèces, d'un arbre phylogénétique construit avec la méthode de la parcimonie sur des séquences d'ARN de la petite sous-unité ribosomique. Cet arbre est donné (au format Newick) dans la composante **tree**.

```
library(ade4)
plot(newick2phylog(pps065$tree))
```



3 Questions

La principale question qu'on se pose est : est-ce que l'analyse statistique de ces données permet de retrouver les relations de parenté qui s'établissent entre les organismes ? Est-ce que ce jeu de données peut nous dire plus que "les organismes qui ont de génomes de taille importante ont beaucoup de gènes en commun" ?

On pourra aussi se demander, lorsque la discrimination entre les espéces est possible, quels sont les gènes COG qui permettent de faire cette discrimination.

En 1999, Tekaia *et al.,* ([1]) ont utilisé l'analyse des correspondances et des méthodes de classification pour construire l'arbre des espèces, en partant d'un jeu de données similaire (mais beaucoup plus complet). Le jeu de données proposé ici permettra peut-être de valider cette méthodologie.

Les conclusions que l'on tire doivent être rapportées à la qualité du jeu de de données dont on dispose, et ne peuvent donc pas atteindre le problème de la validité de la construction de "arbres génomiques". Si l'on accepte que la comparaison entre espèces sur la base de leur contenu en gènes est appropriée, l'analyse de ce jeu de données pourra donner une idée de l'exhaustivité de la base COG.

4 Références

1. Tekaia,F.,Lazcano,A.,Dujon,B. **The genomic tree as revealed from whole proteome comparisons.** *Genome Research* 1999, **9**, 550-557 .

2. Snel,B., Bork,P., Huynen, M.A. **Genome phylogeny based on gene content.** *Nature Genetics* 1999, **21**, 108-110 .

3. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M. *et al.* **The COG database : an updated version includes eukaryotes.** *BMC Bioinformatics* 2003,**4** :41 .

5 Liens

1. http ://www.ncbi.nlm.nih.gov/COG/old/
2. http ://pbil.univ-lyon1.fr/databases/hobacgen.html
3. http ://pbil.univ-lyon1.fr/databases/hogenom.html