

Projet de TP GENOM

Analyse globale de génomes procaryotes par distribution de mots

Les génomes procaryotes présentent en général des régularités de composition assez marquées, comme les proportions de dinucléotides (AA, AC, AT, AG, CA, CC, ...), de trinucléotides, et plus généralement de k-mères (les mots de k lettres). Cette propriété semble également s'étendre à la composition en acides aminés des protéines d'un même génome. Les raisons de ces régularités sont mal connues, mais les biais d'utilisation des codons (pour contrôler l'expression des gènes importants) ou l'adaptation à des milieux extrêmes (notamment chauds et très chauds) sont parfois invoquées. De nombreuses publications ont exploité ces signatures génomiques (ou profils de fréquence), notamment à des fins phylogénétiques (voir Bibliographie). Le but de ce TP consistera à produire des programmes qui calculent, représentent et utilisent ces signatures génomiques à diverses fins précisées ci-dessous.

Constitution du jeu de données

Nous nous intéresserons essentiellement aux génomes procaryotes, pour leur petite taille et pour leur faible proportion de séquences non codantes. La première étape consistera à télécharger les génomes complets et les séquences protéiques d'un certain nombre de bactéries et d'archées. La diversité génétique de ces organismes étant considérable, on prendra soin d'effectuer un échantillonnage aussi large que possible, par exemple en utilisant la liste d'organismes fournie en annexe. Les séquences sont téléchargeables à de nombreux endroits, et en particulier sur le site du NCBI (ftp://ftp.ncbi.nih.gov/genomes/archive/old_refseq/). Ce répertoire contient un répertoire *archaea* et un répertoire *bacteria*, qui eux même contiennent chacun un fichier [summary.txt](#). Ce fichier indique pour chaque répertoire quels génomes sont complets, et donc ceux qui sont intéressants. Les séquences protéiques sont repérées par une extension *faa* et les génomes par une extension *fna*. Il pourra être utile de développer un script pour télécharger ces génomes.

Calcul des profils

En utilisant le langage de votre choix, vous devrez mettre au point un programme qui calcule les proportions de k-mères (mots de k lettres, acides aminés ou nucléotides) dans un génome donné. Vu que l'on va essentiellement manipuler du texte, des langages comme Perl ou Python sont particulièrement recommandés. Attention, le nombre de k-mères augmente rapidement avec k (4^k pour les nucléotides, 20^k pour les acides aminés), il faudra donc mettre au point des structures de données appropriées, même si on se contentera de faibles valeurs de k (entre 2 et 8, par exemple)... On utilisera ensuite ce programme pour calculer la signature génomique de chaque génome, c'est-à-dire la proportion des différents mots de k lettres pour k fixé. Une représentation astucieuse de ces signatures (avec R, par exemple) serait un plus.

Exploitation des profils

Ce projet est un projet de recherche et l'exploitation des données, ainsi que leur traitement, est laissée à votre initiative. En particulier, l'utilisation de Google et des forums de bioinformatique et la

lecture attentive de la bibliographie sont vivement conseillées. Nous donnons quelques pistes d'exploitation intéressantes :

- 1) La signature génomique est-elle vraiment constante le long d'un génome, comme prétendu ? Comment quantifier cette variation ? Si c'est le cas, les transferts horizontaux récents dans un génome devraient avoir une composition très différente de la signature moyenne. Peut-on détecter des gènes candidats à un transfert récent par cette approche ?
- 2) Si les signatures génomiques des différents génomes sont très différentes, on peut utiliser cette propriété pour assigner un gène inconnu à un génome, en choisissant le génome dont la signature est la plus proche de celle du gène. Plusieurs méthodes statistiques permettent de faire ce genre d'analyse. La plus rudimentaire consiste à choisir le génome dont la signature est la plus proche. Une approche plus complexe (mais plus intéressante) serait d'employer des réseaux de neurones, pour prédire un génome à partir d'une composition.
- 3) A partir de la signature de deux génomes, on peut concevoir une distance qui quantifie à quel point ces signatures se ressemblent. Peut-on réaliser une classification hiérarchique des génomes à partir de ces distances entre paires de génomes ? L'arbre obtenu est-il très différent d'un arbre phylogénétique obtenu à partir de l'ARNr 16 des organismes correspondant ? Que peut-on en conclure ?

Organismes recommandés (mais vous êtes libres de choisir)

Archées

Aeropyrum pernix K1
 Archaeoglobus fulgidus DSM 4304
 Archaeoglobus profundus DSM 5631
 Caldivirga maquilingensis IC-167
 Candidatus Korarchaeum cryptofilum OPF8
 Candidatus Methanoregula boonei 6A8
 Candidatus Methanosphaerula palustris E1-9c
 Cenarchaeum symbiosum A
 Desulfurococcus kamchatkensis 1221n
 Haloarcula marismortui ATCC 43049
 Haloarcula marismortui ATCC 43049
 Halobacterium sp. NRC-1
 Halomicrobium mukohataei DSM 12286
 Haloquadratum walsbyi DSM 16790
 Halorhabdus utahensis DSM 12940
 Halorubrum lacusprofundi ATCC 49239
 Halorubrum lacusprofundi ATCC 49239
 Haloterrigena turkmenica DSM 5511
 Hyperthermus butylicus DSM 5456
 Ignicoccus hospitalis KIN4/I
 Metallosphaera sedula DSM 5348
 Methanobrevibacter ruminantium M1
 Methanocaldococcus fervens AG86
 Methanocella paludicola SANA E
 Methanococcoides burtonii DSM 6242
 Methanococcus aeolicus Nankai-3
 Methanococcus maripaludis C6
 Methanococcus vannielii SB
 Methanocorpusculum labreanum Z
 Methanoculleus marisnigri JR1
 Methanopyrus kandleri AV19
 Methanosarcina thermophila PT
 Methanosarcina acetivorans C2A
 Methanosarcina barkeri str. Fusaro
 Methanosarcina mazei Go1
 Methanosphaera stadtmanae DSM 3091
 Methanospirillum hungatei JF-1
 Nanoarchaeum equitans Kin4-M
 Natronomonas pharaonis DSM 2160
 Nitrosopumilus maritimus SCM1
 Picrophilus torridus DSM 9790

Pyrobaculum aerophilum str. IM2
 Pyrobaculum arsenaticum DSM 13514
 Pyrococcus abyssi GE5
 Pyrococcus furiosus DSM 3638
 Pyrococcus horikoshii OT3
 Staphylothermus marinus F1
 Sulfolobus acidocaldarius DSM 639
 Sulfolobus solfataricus P2
 Thermococcus gammatolerans EJ3
 Thermophilum pendens Hrk 5
 Thermoplasma acidophilum DSM 1728
 Thermoplasma volcanium GSS1
 Thermoproteus neutrophilus V24Sta

Bactéries

Acholeplasma laidlawii PG-8A
 Acidobacterium capsulatum ATCC 51196
 Akkermansia muciniphila ATCC BAA-835
 Alicyclobacillus acidocaldarius subsp. acidocaldarius DSM 446
 Aquifex aeolicus VF5
 Bacillus cereus Q1
 Bacillus pseudofirmus OF4
 Bacteroides fragilis YCH46
 Bdellovibrio bacteriovorus HD100
 Bordetella pertussis Tohama I
 Borrelia burgdorferi B31
 Campylobacter jejuni subsp. jejuni 81-176
 Candidatus Amoebophilus asiaticus 5a2
 Candidatus Cloacamonas acidaminovorans
 Candidatus Endomicrobium sp. Rs-D17
 Carboxydotherrmus hydrogenoformans Z-2901
 Chlamydia trachomatis 434/Bu
 Chlorobium chlorochromatii CaD3
 Chloroflexus aurantiacus J-10-fl
 Clostridium acetobutylicum ATCC 824
 Corynebacterium glutamicum ATCC 13032
 Coxiella burnetii RSA 493
 Cupriavidus taiwanensis
 Cupriavidus taiwanensis
 Cyanospora sp. ATCC 51142
 Cyanospora sp. ATCC 51142

Dehalococcoides ethenogenes 195
 Deinococcus radiodurans R1
 Deinococcus radiodurans R1
 Dictyoglomus thermophilum H-6-12
 Elusimicrobium minutum Pei191
 Fibrobacter succinogenes subsp. succinogenes S85
 Flavobacterium psychrophilum JIP02/86
 Fusobacterium nucleatum subsp. nucleatum ATCC 25586
 Gemmata obscuriglobus UQM 2246
 Gemmatimonas aurantiaca T-27
 Gloeobacter violaceus PCC 7421
 Leptospira interrogans serovar Lai str. 56601
 Leptospira interrogans serovar Lai str. 56601
 Magnetococcus sp. MC-1
 Methylophilum infernorum V4
 Mycoplasma genitalium G37
 Nostoc punctiforme PCC 73102
 Opitutus terrae PB90-1
 Pedobacter heparinus DSM 2366
 Pirellula staleyi DSM 6068
 Prochlorococcus marinus str. AS9601
 Psychrobacter arcticus 273-4
 Rhizobium leguminosarum bv. trifolii WSM1325
 Rhodopirellula baltica SH 1
 Rhodospirillum rubrum ATCC 11170
 Rickettsia rickettsii str. Iowa
 Shewanella putrefaciens CN-32
 Solibacter usitatus Ellin6076
 Synechococcus elongatus PCC 6301
 Thermanaerovibrio acidaminovorans DSM 6589
 Thermoanaerobacter tengcongensis MB4
 Thermobaculum terrenum ATCC BAA-798
 Thermobaculum terrenum ATCC BAA-798
 Thermodesulfobacterium yellowstonii DSM 11347
 Thermomicrobium roseum DSM 5159
 Thermotoga maritima MSB8
 Thermus thermophilus HB8

Bibliographie succincte

- Sims G. E., Jun S. R., Wu G. A. and Kim S. H. (2009). Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc Natl Acad Sci U S A* 106, 2677-2682.
- Deschavanne P., DuBow M. S. and Regeard C. (2010). The use of genomic signature distance between bacteriophages and their hosts displays evolutionary relationships and phage growth cycle determination. *Viol J* 7, 163.
- Deschavanne P. and Tuffery P. (2008). Exploring an alignment free approach for protein classification and structural class prediction. *Biochimie* 90, 615-625.
- Chapus C., Dufraigne C., Edwards S., Giron A., Fertil B. and Deschavanne P. (2005). Exploration of phylogenetic data using a global sequence analysis method. *BMC Evol Biol* 5, 63.
- Deschavanne P. J., Giron A., Vilain J., Fagot G. and Fertil B. (1999). Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol Biol Evol* 16, 1391-1399.