SCIENCES SORBONNE UNIVERSITÉ

Lence Alex

Projet de comparaison de paires de génomes : recherche de synténie par dotplot

(Programmation avancée en python et algorithmes d'analyse de séquences)

Année : 2020

Introduction

L'informatique et la biologie sont aujourd'hui deux disciplines en interaction continue, que ce soit dans l'emprunt de concepts biomimétiques appliqués à l'informatique ou à l'inverse l'utilisation de l'informatique pour le traitement et la modélisation de données.

Durant mon projet j'ai eu l'occasion d'utiliser l'informatique à des but de filtrage de l'information. Les dotplots sont une représentation graphique qui permettent de visualiser rapidement les homologies entre séquences ou entre génomes. Dans notre cas nous nous sommes entièrement penchés sur la représentation graphique de l'homologie entre génome. Le but était de représenter sur un axe gradué les gènes homologues, dont l'homologie dépendrait de variable de filtration comme un seuil de e-value, un pourcentage d'identité, un pourcentage de recouvrement ou même des fonctions similaires.
Il fallait également que nous nous penchions sur une interface graphique, et par conséquent qu'on travaille sur l'ergonomie de notre programme pour qu'il puisse être utilisé par n'importe qui et de préférence sans que cette personne n'ai à lire les instructions d'utilisation. L'interface graphique nécessitait un soin tout particulier puisqu'elle était surtout destinée à des personnes non formées à l'informatique.

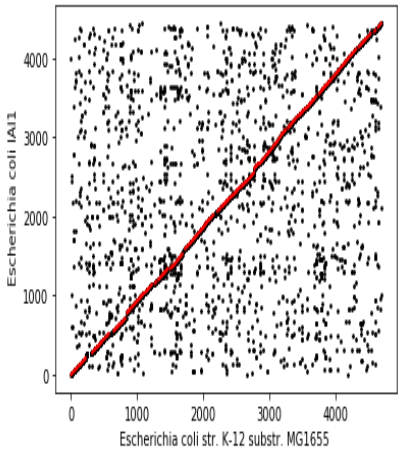
L'homologie entre deux séquences indique que ces deux séquences sont issues d'un ancêtre commun. Deux gènes homologues sont soit issus d'un évènement de spéciation soit de duplication. Cependant l'homologie est un concept biologique assez particulier puisqu'il ne s'appuie pas véritablement sur des faits mais plutôt sur des supposions, puisque deux gènes sont considérés homologues s'ils partagent un niveau de ressemblance supérieur à un seuil fixé et selon le temps d'évolution deux gènes dont l'homologie est réel peuvent ne pas être décelé comme tel du à un éloignement évolutif.
Pour diminuer cette erreur on va essayer de trouver des blocs de synténie. Les blocs de synténie sont des séquences d'ADN conservé entre deux espèces, plus ces blocs sont long, plus la distance évolutive entre 2 espèces est petite. De façon générale les blocs de synténie renseignent sur la distance évolutive et les différentes mutations entre les génomes comme on pourra le voir dans la suite du rapport.

Pour les résultats suivants j'ai choisi d'utiliser une taille minimale pour les blocs de synténie de 10 ce qui me semblait apporter les résultats les plus visibles. J'ai également choisi une distance maximale entre 2 gènes d'un même bloc de synténie de 30 qui pour les mêmes raisons semblait correspondre à des résultats plus probant.

Résultats

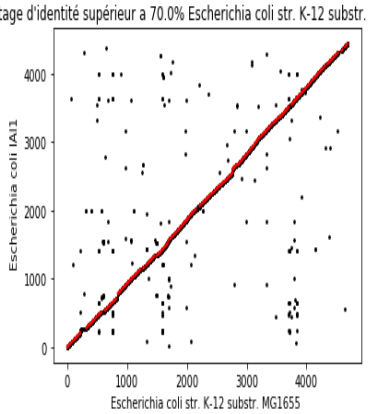
Escherichia coli

Dotplot pour une e-value maximum de 1e-50 Escherichia coli str. K-12 substr. MG1655 VS Escherichia coli IAI1



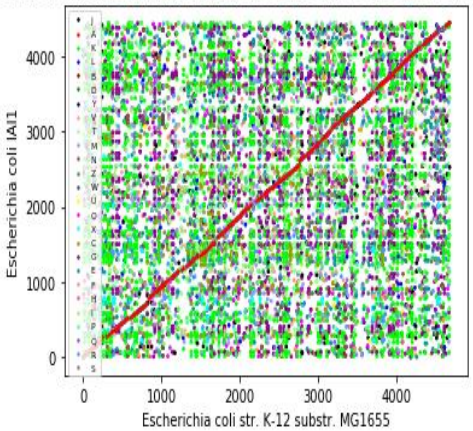
Dans un premier temps intéressons-nous à un filtrage selon la e-value pour un fichier blast entre *Escherichia coli str. K-12 substr. MG1655* et *Escherichia coli IAI1*. On observe pour une e-value maximale de 10^{-50} une assez bonne identité entre ces deux organismes. On remarque quelques insertions de gènes dans chacune des séquences en témoigne les « trous/décalage » dans les blocs de synténie. Cependant on ne voit pas apparaitre d'inversion dans les séquences. Il semble donc que ces deux organismes soient assez proches en termes de phylogénie. De plus même en changeant le seuil de e-value les blocs de synténie ne changent pas ce qui montre une très bonne conservation des séquences entre ces deux organisme.

Dotplot pour un poucentage d'identité supérieur a 70.0% Escherichia coli str. K-12 substr. MG1655 VS Escherichia coli IAI1



Dans un second temps lorsqu'on s'intéresse au pourcentage d'identité entre les gènes, on observe également une conservation de séquence entre ces deux organismes, ce qui concorde avec les résultats obtenus avec la e-value. Ici j'ai pris volontairement un pourcentage d'identité très élevé pour montrer la conservation entre ces deux organismes en termes de constitution de séquence.

Dotplot fonctionnel Escherichia coli str. K-12 substr. MG1655 VS Escherichia coli IAI1



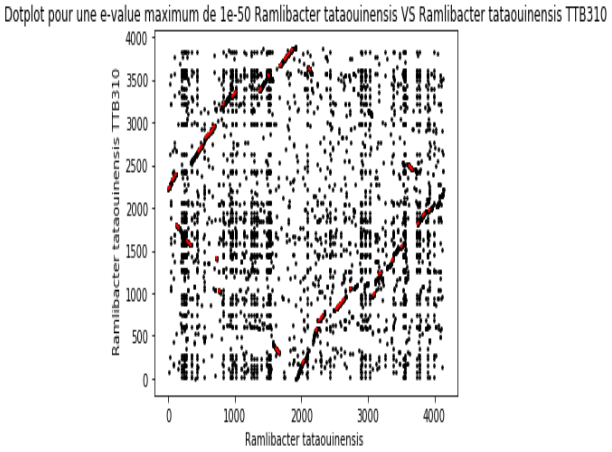
Enfin les résultats sont tout aussi visible en observant les fonctions des différents gènes de ces deux organismes. En effet sur ce plot on considère que deux gènes sont homologues si ils partagent la même fonction. Or ici on observe les mêmes blocs de synténie ce qui indique que les fonctions dans les gènomes des ces deux organismes sont extremement bien conservés.

En conclusion il apparait que ces deux organismes sont relativement proche d'un point de vue génomique puisque partagent sur une grande partie de leur génome les mêmes séquences et les mêmes fonctions de gènes. Il semble cependant exister une distance assez importante entre ces deux organismes visibles sur l'arbre phylogénétique d'*Escherichia coli* (Chaudhuri and Henderson, 2012).

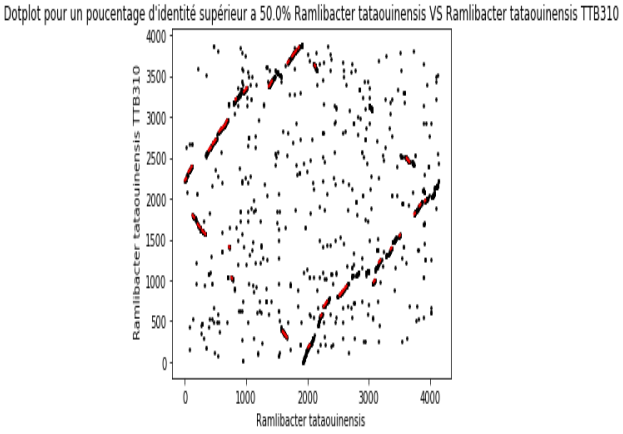
Ramlibacter tataouinensis

On va maintenant observer via les représentations dotplot les homologies de séquences entre *Ramlibacter tataouinensis* et *Ramlibacter tataouinensis TTB310*.

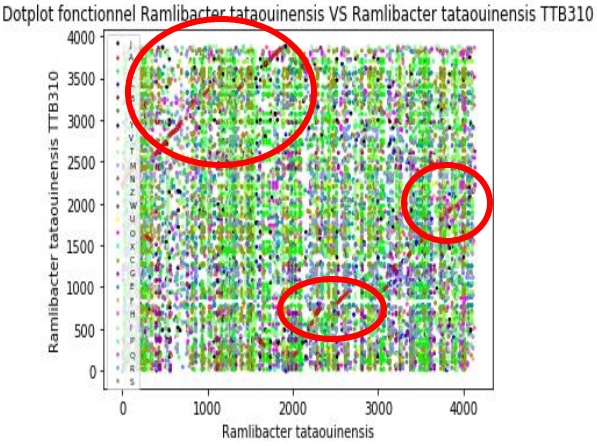
Si on regarde le dotplot pour une value inférieure a 10^{-50} on remarque contrairement à Escherichia un plus fort réarengement au sein des génomes. En effet on remarque deux régions sur le graphique comprenant des blocs de synténie, la région en haut à gauche et la région en bas à droite. On remarque également certains blocs qui semblent faire le lien entre ces deux régions. Les régions en haut à droite et en bas à gauche semblent indiquer que chez l'ancêtre commun de ces deux organismes une partie du génome « terminal » a été déplacé en début de son génome ce qui à conduit à un évènement de spéciation distinguant *Ramlibacter tataouinensis* et *Ramlibacter tataouinensis TTB310*. Les blocs qui « font le lien » témoignent quant à eux d'inversion de séquence entre ces deux organismes.



Lorsqu'on se penche sur le pourcentage d'identité entre ces deux génomes on observe plus ou moins les mêmes résultats pour un pourcentage d'identité minimal de 50%. A noter que pour un pourcentage d'identité plus important on perd certain bloc de synténie ce qui indique une plus faible homologie entre les génomes. Cependant 50% est une valeur d'identité tout à fait convenable pour parler d'homologie. Et il semble que cette disposition en « carré » est assez bien conservée.



En regardant un dotplot qui indique les fonctions similaires entre ces deux organismes on observe le même nombre de blocs de synténie (42 pour les fonctions 44 pour un pourcentage d'identité de 50%). Il semble donc que ces fonctions soient bien des fonctions et plus généralement des gènes homologues entre ces deux organismes.

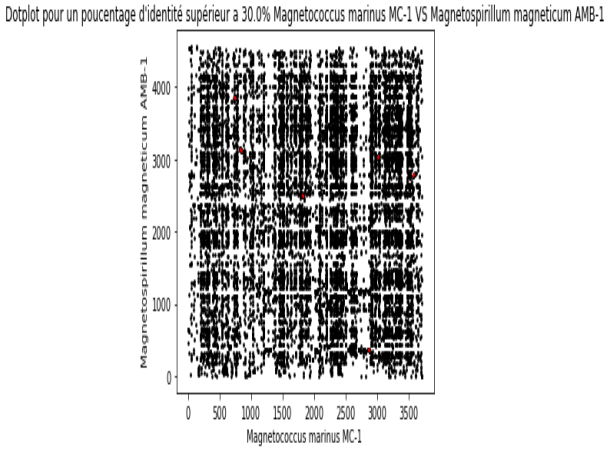


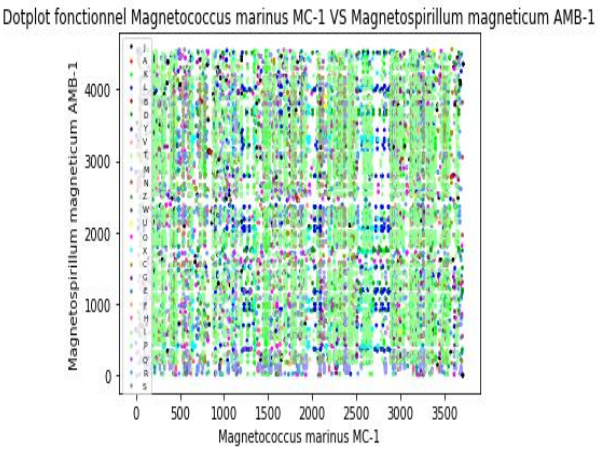
Il apparait à travers l'analyse de l'homologie entre ces deux organismes une plus grande variation au sein de leur génome. Cependant on a pu observer des régions qui semblent assez bien conservée entre ces deux organismes. Il apparait cependant sur l'arbre phylogénétique une parenté très proche entre ces deux organismes (Lee et al., 2014) .

Magnetococcus

Dans cette partie nous allons nous intéresser aux organismes *Magnetococcus marinus MC-1* VS *Magnetospirillum magneticum AMB-1*.

Dans un premier temps cette fois nous allons regarder un dotplot pour un pourcentage d'identité de 30%, soit le minimum pour considérer deux gènes comme homologues. On observe rapidement que très peu de blocs de synténie sont retrouvés (5 trouvés pour un pourcentage d'identité de 30%). Il semble donc que ces deux organismes soient assez éloignés sur l'arbre phylogénétique.





De même que pour l'identité on trouve très peu de bloc de synténie, mais il est intéressant de noter des régions entières uniquement représentées par une certaine fonction, ce que l'on n'observait pas chez les autres organismes, notamment la région entre 2500 et 3000 de *Magnetococcus marinus MC-1* ou on ne retrouve quasiment exclusivement que des gènes impliqués dans la réplication (J : bleu), ce qui indique une multiplication de cette fonction chez *Magnetospirillum magneticum AMB-1*.

En conclusion, l'homologie entre ces deux organismes ne semble pas vérifiée. De plus même en changeant la taille minimale des blocs de synténie ou la distance entre deux gènes d'un même bloc de synténie, il ne semble pas qu'on puisse améliorer ces résultats. D'après les résultats en phylogénie il semble bel et bien que la distance phylogénétique entre ces deux organismes est assez importante (Zhang et al., 2017).

Conclusion

En conclusion on voit assez rapidement que l'outil de dotplot apporte des résultats facilement compréhensibles et qui donne une idée globale de la proximité évolutive des organismes. Cependant on ne peut pas utiliser cet outils pour définir un arbre phylogénétique, cet outils n'a pour but que de donner un résultat observable, mais n'est pas en mesure de véritablement opérer de test statistique pour conclure sur quoique ce soit.
En effet lorsqu'on regarde les arbres phylogénétiques de chaque *espèce* on observe parfois une homologie certaine qui n'est pas présente sur l'arbre phylogénétique comme pour *Escherichia Coli*. A l'inverse on peut observer une homologie incertaine alors que la proximité sur l'arbre phylogénétique est proche comme pour *Ramlibacter tataouinensis*. Cela montre bien que le dotplot n'a pas de prétention à révéler de vraies proximités phylogénétiques mais de simplement comparer les homologies de séquence entre deux organismes.

Les améliorations possibles à ce programme pourraient être de visualiser l'emplacement des gènes dans les génomes et représenter le déplacement de ceci au sein des génomes rendant ainsi mieux compte de l'évolution des génomes entre deux organismes.

Références

Chaudhuri, R.R., and Henderson, I.R. (2012). The evolution of the Escherichia coli phylogeny. Infection, Genetics and Evolution *12*, 214–226.

Lee, H.J., Lee, S.H., Lee, S.-S., Lee, J.S., Kim, Y., Kim, S.-C., and Jeon, C.O. (2014). Ramlibacter solisilvae sp. nov., isolated from forest soil, and emended description of the genus Ramlibacter. International Journal of Systematic and Evolutionary Microbiology, *64*, 1317–1322.

Zhang, H., Menguy, N., Wang, F., Benzerara, K., Leroy, E., Liu, P., Liu, W., Wang, C., Pan, Y., Chen, Z., et al. (2017). Magnetotactic Coccus Strain SHHC-1 Affiliated to Alphaproteobacteria Forms Octahedral Magnetite Magnetosomes. Front. Microbiol. *8*, 969.