Université de Nantes

UFR Sciences et Techniques

Master informatique parcours
"optimisation en recherche opérationnelle (ORO)"
Année académique 2019-2020

Distanciel
-
Algorithmics in genomics

Auteur :
Corentin Pelhâtre
Adrien Cassaigne

Référent :
Irena Rusu

29 novembre 2019

Table des matières

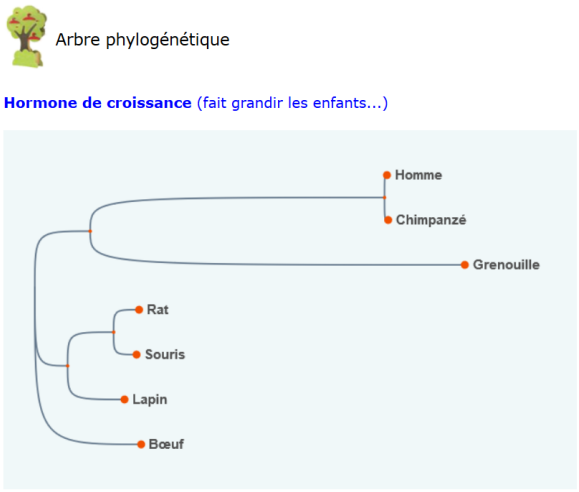
1 Introduction :

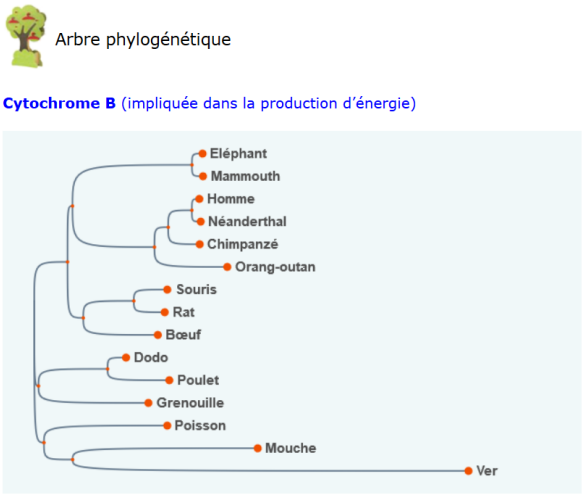
Dans le cadre du module X3IO060 : "Algorithmics in Genomics" nous avons a réaliser un travail en distanciel. Ce document est le rapport de notre travail. Il comportera tous les éléments demandés à l'exceptions des 36 arbres générés que nous avons placé dans des documents ".pdf" dans le dossier "./arbres".

Bien que cela ne soit pas demandé, nous avons fourni tous nos codes lors de la réalisation du projet. Vous trouverez, "src/main/main.jl" le fichier utile à la génération des fichiers.txt de donnée pour le site "nbci", et, "src/main/main2.jl" pour tout ce qui concerne la génération de la matrice de distance.

2 Serious game :

Nous avons commencé cette introduction à la construction d'arbre de phylogénie en étudiant les résultats obtenus par le choix des deux protéines suivantes : l'hormone de croissance et la Cytochrome B. Après avoir suivi les quelques manipulations demandés, nous avons construits les deux arbres suivant :





Afin de comparer les deux arbres obtenus, nous allons nous concentrer sur l'étude des espèces en commun, soit les six suivantes : l'homme, le chimpanzé, la grenouille, la souris, le rat et le boeuf.

On remarque ainsi que des espèces très proches comme l'homme et le chimpanzé ou encore la souris et le rats le reste lors de la construction de l'arbre que se soit pour la première ou la seconde protéines. On peut également noter que la distance qui sépare ces deux sous-groupes reste à peu près équivalente entre les deux arbres de phylogénie. Les différences se retrouvent donc avec les deux dernières espèces. Tout d'abord la grenouille, qui est plutôt proche de l'homme en s'appuyant sur l'hormone de croissance, et qui lorsque l'on observe l'arbre issu de la Cytochome B est assez éloigné. Elle se retrouve aussi loin de l'homme que le rats ou la souris, en ayant très peu de caractéristiques en commun. Dans le sens inverse, le boeuf qui partage uniquement un ancêtre commun (base de l'arbre de phylogénie) avec toutes les autres espèces, se retrouve proche du rat et de la souris lorsque l'on étudie la protéine de la Cytochrome B.

La construction de ces arbres phylogéniques se base sur l'alignement des séquences (en acides aminés) de la protéine sélectionnée. Les séquences d'acides aminés qui servent donc à l'élaboration de ces arbres diffères. De plus, on tient compte de la ressemblance des acides aminés pour établir une "proximité" entre certaines espèces, mais nous n'avons aucune information sur la différence qu'engendre deux acides aminés différents à une même position. On ne nous donne pas non plus d'informations nous permettant de savoir si par exemple l'acide aminé **S** est plus proche de **I** où de **Q**. Cette comparaison nous permet toutefois d'observer que malgré toutes les différences de ces espèces on retrouve un nombre assez important d'acide aminé en commun pour une même protéine.

Hormone de croissance (fait grandir les enfants...)

```
☑ Souris        -MATDSRTSWLLTVSLLCLLWPQEASAFPAMPLSSLFSNAVLRAQHLHQLAADTYKEFER AYIPEG
☑ Rat           -MAADSQTPWLLTFSLLCLLWPQEAGAFPAMPLSSLFANAVLRAQHLHQLAADTYKEFER AYIPEG
☑ Lapin         -MAAGSWTAGLLAFALLCLPWPQEASAFPAMPLSSLFANAVLRAQHLHQLAADTYKEFER AYIPEG
☑ Bœuf          MMAAGPRTSLLLAFALLCLPWTQVVGAFPAMSLSGLFANAVLRAQHLHQLAADTFKEFER TYIPEG
☑ Homme         -MATGSRTSLLLAFGLLCLPWLQEGSAFPTIPLSRLFDNAMLRAHRLHQLAFDTYQEFEE AYIPKE
☑ Chimpanzé     -MAPGSRTSLLLAFGLLCLPWLQEGSAFPTIPLSRLFDNAMLRAHRLHQLAFDTYQEFEE AYIPKE
☑ Grenouille    -MATGFCSSFGLLVVLL-LKNVADVGAFPSVPLFSLFTNAVSRAQYIHMLAADTYRDYER TYITDE
```

Alignement des séquences (en acides aminés) de la protéine que vous avez sélectionnée, dans différentes espèces.
les colonnes colorées mettent en évidence les positions conservées entre les séquences des différentes espèces:
acides aminés identiques.
acides aminés avec les mêmes propriétés biochimiques.
acides aminés peu similaires du point de vue biochimique.
acides aminés non similaires du point de vue biochimique.

Cytochrome B (impliquée dans la production d'énergie)

```
☐ Mais          --MTIRNQRFSLLKQPIYSTLNQHLIDYPTPSNLSYWWGFGCLAGICLVIQIVTGVFLAMHYTPHVDLAF
☐ Riz           --MTIRNQRFSLLKQPIYSTLNQHLIDYPTPSNLSYWWGFGSLAGICLVIQIVTGVFLAMHYTPHVDLAF
☐ Levure        --MAFRKS------NVYLSLVNSYIIDSPQPSSINYWWNMGSLLGLCLVIQIVTGIFMAMHYSSNIELAF
☑ Eléphant      -MTHIRKS------HPLLKIINKSFIDLPTPSNISTWWNFGSLLGACLITQILTGLFLAMHYTPDTMTAF
☑ Mammouth      -MTHIRKS------HPLLKILNKSFIDLPTPSNISTWWNFGSLLGACLITQILTGLFLAMHYTPDTMTAF
☑ Homme         -MTPMRKI------NPLMKLINHSFIDLPTPSNISAWWNFGSLLGACLILQITTGLFLAMHYSPDASTAF
☑ Néanderthal   -MTPMRKI------NPLMKLINHSFIDLPTPSNISAWWNFGSLLGACLILQITTGLFLAMHYSPDASTAF
☑ Chimpanzé     -MTPTRKI------NPLMKLINHSFIDLPTPSNISAWWNFGSLLGACLILQITTGLFLAMHYSPDASTAF
☑ Orang-outan   -MTPMRKT------NPLMKLINHSLIDLPTPSNISAWWNFGSLLGACLIIQTITGLFLAMHYSPDATTAF
☑ Souris        -MTNMRKT------HPLFKIINHSFIDLPAPSNISSWWNFGSLLGVCLMVQIITGLFLAMHYTSDTMTAF
☑ Rat           -MTNIRKS------HPLFKIINHSFIDLPAPSNISSWWNFGSLLGVCLMVQILTGLFLAMHYTSDTMTAF
☑ Bœuf          -MTNIRKS------HPLMKIVNNAFIDLPAPSNISSWWNFGSLLGICLILQILTGLFLAMHYTSDTTTAF
☑ Dodo          ------------------------------------WWNFGSLLGICLMTQILTGLLLAAHYTADTTLAF
☑ Poulet        MAPNIRKS------HPLLKMINNSLIDLPAPSNISAWWNFGSLLAVCLMTQILTGLLLAMHYTADTSLAF
☑ Grenouille    MAPNIRKS------HPLIKIINNSFIDLPTPSNISSLWNFGSLLGVCLIAQIITGLFLAMHYTADTSMAF
☑ Poisson       -MTSLRKT------HPVLKIANDALVDLPTPLNISAWWNFGSLLGLCLITQILTGLFLAMHYTSDISTAF
☑ Mouche        MNKPLRNS------HPLFKIANNALVDLPAPINISSWWNFGSLLGLCLIIQILTGLFLAMHYTADINLAF
☑ Ver           --MKINNS--------LLNFVNGMLVTLPSSKTLTLSWNFGSMLGMVLIFQILTGTFLAFYYTPDSLMAF
```

Alignement des séquences (en acides aminés) de la protéine que vous avez sélectionnée, dans différentes espèces.
les colonnes colorées mettent en évidence les positions conservées entre les séquences des différentes espèces:
acides aminés identiques.
acides aminés avec les mêmes propriétés biochimiques.
acides aminés peu similaires du point de vue biochimique.
acides aminés non similaires du point de vue biochimique.

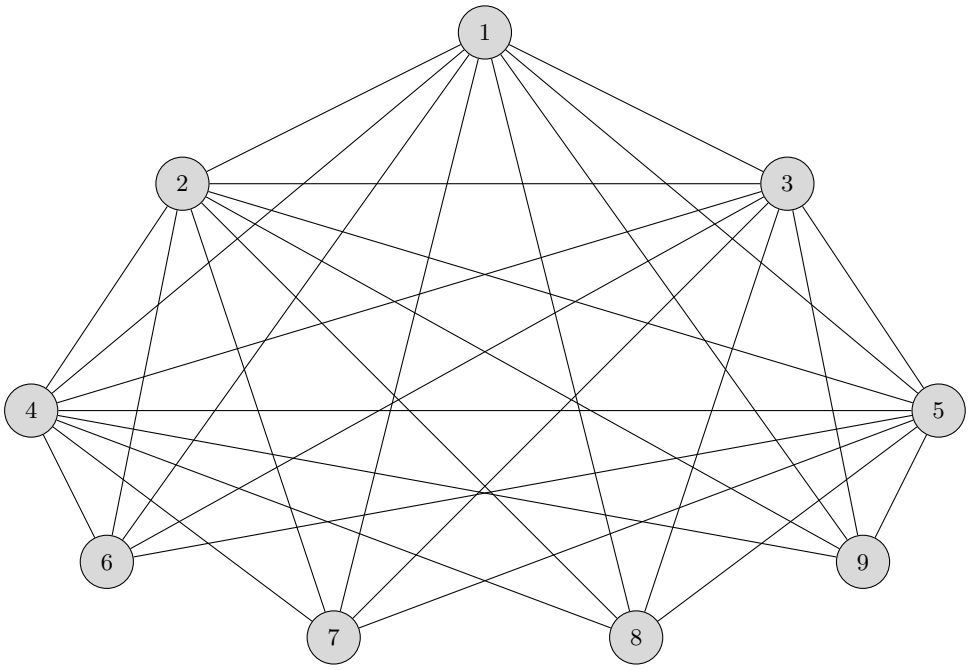
3 Construire une phylogénie

3.1 Introduction

Dans le cadre de ce distanciel, nous avons choisi de nous intéresser aux six protéines suivante : BMP4, Distalless, FibronogenB, Insuline, Myoglobine, Sonic . La base de données de NBCI sur ces six protéines nous a permis d'étudier les espèces suivante : Homme, Loup, Macaque, Souris, Rat, Coq, Chimpanzé, Bovins, Poisson. Cette comparaison est possible par l'intermédiaire des séquences d'acides aminés disponible sur le site internet. Nous allons générer un certains nombre d'arbres de phylogénie en fonction de l'étude des espèces et des protéines. Cette étude va nous permettre de récolter un grand nombre d'information à ce sujet. L'objectif est donc de synthétiser les résultats obtenus afin de réaliser un unique arbre de phylogénie. L'enjeux est ainsi de définir une proximité entre les différentes espèces malgré les différences que l'on peut observer lorsqu'on les compare sur l'ensemble des protéines.

3.2 Construire un premier graphe

Afin de formuler notre problème sous forme de graphe nous avons pris pour sommet les espèces et nous avons défini les arêtes de la façon suivante : (i, j) est une arête du graphe lorsqu'on a comparé les deux séquences d'acides aminés des espèces i et j à l'aide de Blast. Voici le résultat obtenu :



avec :
— 1 : Bovins
— 2 : Coq
— 3 : Homme
— 4 : Loup
— 5 : Souris
— 6 : Chimpanzé
— 7 : Macaque
— 8 : Rat
— 9 : Poisson

3.3 Amélioration du graphe

Après définition de notre graphe, nous lui avant ajouté de l'information par l'intermédiaire des arêtes. Dans cette optique là, nous avons cherché à ajouter un poids. Nous avons alors utilisé les données récoltées lors de la construction des arbres de phylogénie pour calculer un score à associer à chaque arête. Nous avons pu obtenir un score entre espèces par l'intermédiaire de NBCI en fonction de la protéine étudiée. Ce score reflète la similarité entre deux espèces, plus il est élevé, plus leur séquence d'acides aminés sont proches. Afin d'obtenir un score unique pour chaque arête de notre graphe, voici la démarche que nous avons suivie :

Soit deux espèces : e_i, e_j et p_k une protéine avec $(i, j) \in$ {Bovin, Coq, Homme, Loup, Souris, Chimpanzé, Poisson zèbre, Macaque, Rat, Grenouille et $k \in$ {BMP4, Distal-less, FibronogenB, Insuline, Myoglobine, Sonic}. Notons $s^k_{i,j}$ le score entre les espèces i et j pour la protéine k

Une fois le score extrait des arbres pour chaque protéine, nous l'avons normé afin d'obtenir une certaine équivalence entre les protéines et ainsi éviter le biais possible dû aux différentes longueurs des séquences comparées. Nous avons par la suite symétrisé la matrice de distance afin de supprimer la différence du à la taille d'une espèce par rapport à l'autre. Le score entre deux espèce est donc ensuite calculé par la formule suivante :

$$score_{i,j} = \frac{\sum_k s^k_{i,j}}{6} \forall (i, j)$$

On obtient donc la matrice des poids suivant :

0.0	0.294	0.587	0.595	0.559	0.586	0.583	0.553	0.002
0.294	0.0	0.445	0.365	0.432	0.35	0.347	0.341	0.065
0.587	0.445	0.0	0.659	0.628	0.814	0.772	0.624	0.05
0.595	0.365	0.659	0.0	0.62	0.651	0.66	0.641	0.063
0.559	0.432	0.628	0.62	0.0	0.62	0.621	0.725	0.058
0.586	0.35	0.814	0.651	0.62	0.0	1.0	1.0	1.0
0.583	0.347	0.772	0.66	0.621	1.0	0.0	1.0	1.0
0.553	0.341	0.624	0.641	0.725	1.0	1.0	0.0	1.0
0.002	0.065	0.05	0.063	0.058	1.0	1.0	1.0	0.0

Nous avons par ailleurs fixé à 1.0 les arcs non présents (distance maximal suite à la normalisation). Cette technique c'est avéré non efficace (non valide pour la condition des 4 points) nous avons donc choisi de travailler seulement sur les 5 espèces formant un graphe complet :

— 1 : Bovins
— 2 : Coq
— 3 : Homme
— 4 : Loup
— 5 : Souris

et ainsi obtenir la matrice suivante :

0.0	0.294	0.587	0.595	0.559
0.294	0.0	0.445	0.365	0.432
0.587	0.445	0.0	0.659	0.628
0.595	0.365	0.659	0.0	0.62
0.559	0.432	0.628	0.62	0.0

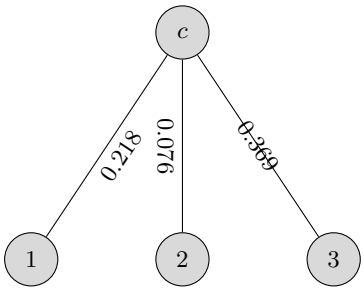
3.4 Récolte d'informations

Notre première intuition fut de sélectionner les arêtes avec un score maximum afin d'expliquer au mieux notre phylogénie et donc d'avoir des branches dans notre arbre en ayant des espèces "proche" entre deux noeuds. Avec cette notion de distance entre espèces, nous nous sommes ensuite penché sur les méthodes de résolution vu en cours. C'est ainsi que nous avons choisi de suivre le déroulé de l'algorithme "AdditivePhylogenya". Dans cette optique il faut d'abord vérifier que notre matrice de score obtenue vérifié la condition des quatre points détaillé en cours. Nous avons rencontré quelques difficultés à comprendre cette contrainte et à rendre cette condition satisfaite et donc à adapter notre matrice pour pouvoir poursuivre. Nous avons alors décidé de nous intéresser au sous-problème contenant les cinqs espèces qui nous ont servi de base à la comparaison des séquences.

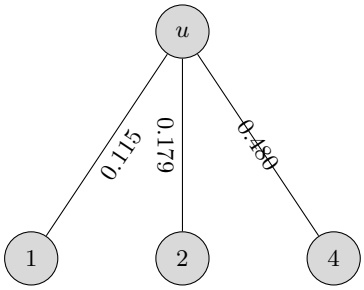
3.5 Construction de l'arbre de phylogénie

Nous allons donc détailler l'exécution de cet algorithme afin de réaliser notre arbre de phylogénie pour résumer l'information obtenu à partir des arbres générés.

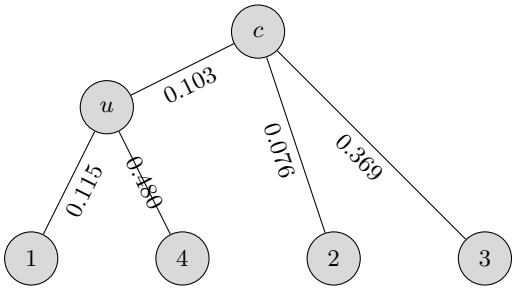
Lors de la première itération nous définissons notre arbre de base avec les trois premières espèces.



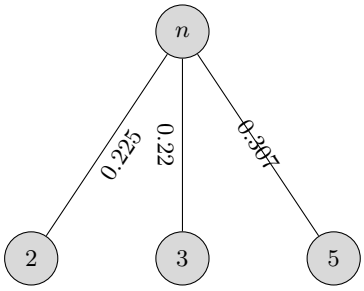
Une fois construit nous insérons la quatrième espèce en résolvant le système à 3 équations, 3 inconnues afin d'obtenir un nouveau noeud.



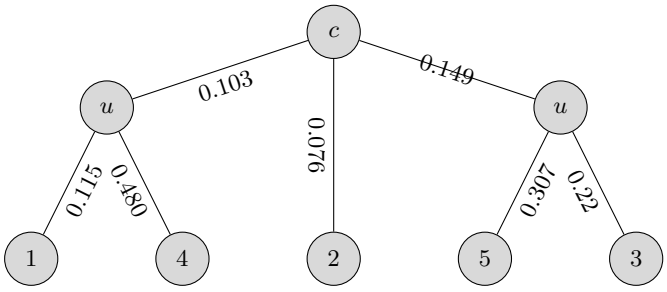
Ce noeud est réalisable et peut être introduit dans notre arbre sur la branche menant à la première espèce.



De la même façon nous insérons la cinquième espèce en branchant sur celle menant à la troisième espèce. Nous obtenons ainsi un arbre de phylogénie pour les cinq espèces étudiées.



Ce qui nous donne l'arbre final :



4 Conclusion :

Pour conclure, nous avons rencontré un grand nombre de difficultés dans le parsage et la récupération des données du site "nbci" dans des formats utilisables. Cela correspond avec ce qui nous a été dis sur la difficulté de collaboration biologistes/informaticiens. Nous avons pu observer à quel point le problème de trouver une phylogénie peut être complexe au vue des différences et ressemblances possibles entre les espèces, ce qui vient confirmer l'idée que l'on avait déjà pu avoir lorsque l'on s'était intéressé à la phylogénie basée sur les caractères lors de l'étude des articles.

Par ailleurs, nous avons pu en apprendre beaucoup plus sur les arbres phylogénétiques grâce à ce distanciel. Cela nous sera d'une grande aide pour la suite du module.