Projet Clustal

Introduction

Le programme Clustal a été développé en 1988 par Desmond G.Higgins et Paul M.Sharp. Les auteurs souhaitaient implémenter un programme permettant de réaliser un alignement de séquence multiple afin de comparer les évolutions des séquences, mais aussi leurs différences et similarités structurelle, ainsi que les régions qu'elles ont en communs. La complexité de ce projet était de fournir un programme pouvant fonctionner sur des ordinateurs ayant une faible puissance de calcul. Pour cela, les auteurs ont utilisé un alignement progressif introduit par Feng, Doolittle et Willie Taylor.

Matériels et Méthodes

Les séquences à étudier sont récupérées sur la base de données de séquences protéiques Uniprot.

Pour cette étude, un jeu de donnée de 5 séquences est utilisé. Elle concerne des séquences protéiques du gène CIROP provenant de différents organismes.

Pour effectuer leur comparaison, l'algorithme de Needleman et Wunch est utilisé. C'est un algorithme de programmation dynamique qui permet d'effectuer un alignement globale maximale entre deux chaînes de caractères. Il crée une matrice de longueur n * m (où n correspond à la longueur de la première séquence et m celle de la seconde longueur) ou tout d'abord la première colonne et ligne sont rempli par les pénalités de gap.

Par la suite, la matrice est remplie en prenant le score le plus élevée pour chaque acide aminé. Elle permet donc de fournir le score d'alignement maximum pour les deux séquences, mais également de fournir l'alignement optimal. Plus le score d'alignement est élevé, plus les séquences sont similaires.

L'algorithme prend en entrée les séquences à comparer ainsi que la valeur de pénalité du gap (par default -5) et la matrice de similarité blossum 62.

La matrice réunissant tous les scores d'alignements maximums pour chaque séquence est appelée matrice de score. Cette matrice de score est convertie en matrice de distance afin d'effectuer un clustering hiérarchique. Contrairement à la matrice de score, les plus petites valeurs correspondent aux séquences les plus proches.

Le clustering hiérarchique permet à chaque itération de fusionner deux classes étant définies comme plus proche.

Dans l'objectif d'utiliser des bonnes pratiques de programmation, le programme a été codé selon la référence pep8.

Résultats et discussions

Les 5 séquences ont été comparés avec une pénalité de gap définie à -5.

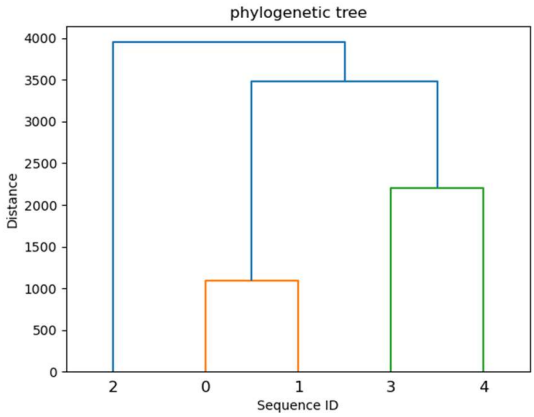

Figure 1 : Arbre phylogénétique

Le script fournit l'arbre phylogénétique représenté sur la figure 1. Sur celui-ci, il est possible de voir que la première et seconde séquence forme un groupe et les séquences 4 et 5 en forment un également. Ces groupes sont ensuite reliés. La séquence la plus éloignée est donc la troisième séquence.

Afin de savoir si le script fournit un arbre phylogénétique correct, il est comparé à celui du programme clustal omega.

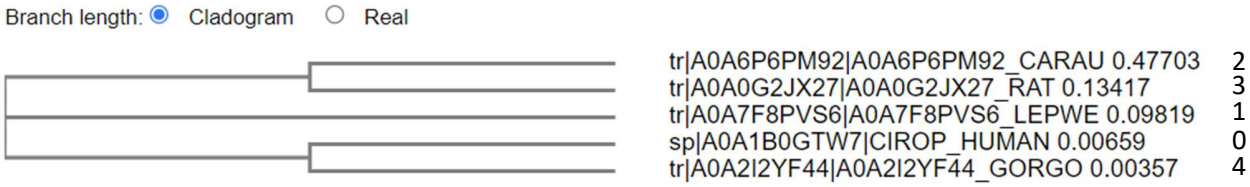

Figure 2 : Arbre phylogénétique selon le programme CLUSTAL OMEGA

Clustal omega utilise l'algorithme UPGMA pour construire l'arbre phylogénétique. En comparant les deux arbres, il est possible de voir qu'il y a le même nombre de groupes. Cependant, ce ne sont pas les mêmes séquences qui appartiennent à chacun des groupes. Il semble donc qu'il y ait une divergence entre les algorithmes. Il faudrait donc par la suite au lieu d'utiliser un package dans le script qui réalise automatiquement un clustering hiérarchique, mettre en place une fonction UPGMA afin de construire l'arbre. L'algorithme Neighbor-joining pourrait permettre également de construire un arbre phylogénétique. Cependant il n'est pas optimisé pour prendre un grand nombre de séquence.

Une fois que l'arbre est donné, l'alignement multiples des séquences peut être réalisé. Cependant mon programme ne permet pas de le faire. Il peut seulement comparer des séquences par pair.

Les séquences 1 et 2 ont été comparés car elles sont relativement proche.

```
alignement1='LLLLPPLVLRVAASRCLHDETQKSVSLLRPPFSQLPSK-SRSSSLTLPSSRDPQPLRIQSCYLGDHISDGAWDPEGEGMRG-GS-RALAAVREATQRIQAVLAVQGP
LLLSRDPAQYCHAVWGDPDSPNYHRCSLLNPGYKGESCLGAKIPDTHLRGYALWPEQGPPQLVQPDGPGVQNTDFLLYVRVAHTSKCHQETVSLCCPGWSTAAQSQLTAALTSWAQRRGF
VMLPRLCLKLLGSSNLPTLASQSIRITGPSVIAYAACCQLDSEDRPLAGTIVYCAQHLTSPSLSHSDIVMATLHELLHALGFSGQLFKKWRDCPSGFSVRENCSTRQLVTRQDEWGQLLL
TTPAVSLSLAKHLGVSGASLGVPLEEEEGLLSSHWEARLLQGSLMTATFDGAQRTRLDPITLAAFKDSGWYQVNHSAAEELLWGQGSGPEFGLVTTCGTGSSDFFCTGSGLGCHYLHLDK
GSCSSDPMLEGCRMYKPLANGSECWKKENGFPAGVDNPHGEIYHPQSRCFFANLTSQLLPGDKPRHPSLTPHLKEAELMGRCYLHQCTGRGAYKVQVEGSPWVPCLPGKVIQIPGYYGLL
FCPRGRLCQTNEDINAVTSPPVSLSTPDP-LFQLSLELAGPPGHSLGKEQQEGLAEAVLEALASKGGTGRCYFHGPSITTSLVFTVHMWKSPGCQGPSVATLHKALTLTLQKKPLEVYHG
GANFTTQPSKLLVTSDHNPSMTHLRLSMGLCLMLLILVGVMGTTAYQKRATLPVRPSASYHSPELHSTRVPVRGIRE'
alignement2='MFLLPLLLLGTAASRCLHEETQKSVTLLRPHLS-QPAPNFRSSALTLPGSRDPQPLRIRTCYIRDPVSDGAWDPEGAGMRGGPAALALAAARQAAQQLQGIFAVQGP
LLLSRDPAQYCHAVWGDPDTPNYHRCSLLNPGYKGESCLGAKIPDTHLRGYALWPEQGAPQLVQPDGPGVQDTDFLLYVWVAHTSKC----------------------------
---------H---------G------EPSVMAYAACCHLDSEDRPLAGTIVYCAQHLTSPSLSHSDIVTVTLHELLHALGFSGQLFKKWRDCPSGPSAREDCSTRQQVTRRDEWGQLLL
TTPTVSHSLARHLGVPGALPGVP-LEEEGPSSSHWEARLLQGSIMTATFDGARRTRLDPITLAAFKDSGWYRVNHSAAEGLLWGRGSGLEFGLVTTCGAGSSDFFCTGSGLGCHYLHLDK
GSCSSDPVLEGCRMYKPLANGSECWKKENGFPPGAENPHGEIYHPQSRCFLANLTSQLLREDKTGHPSEIPHPKEAELTGRCYLHQCTERGAYKVQAQGSPWAPCLPGKAIQIPGYSGLL
FCPRGRLCQTNEVTNAVTSPPVS-LPPQDLSFQLSFQLAGPPGHSVGKEELDGLTEAVLQALVSRGSPSRCYFHSPSTTTSLVFTVYMGKSPGCQGPSVGTLHRALTLTLQKKPLEVHFG
GASFTTGHTKLLVTLDRNPFVTHLALSTGLCLTLLILVGALGTVAYQKRATLRVAPSAPHHSPELQDTRGPAGGIRE'
```

Figure 3 : Alignement optimal des séquences 1 et 2

La quantité de gap est relativement faible, d'après le programme, il y a 7 % de gap au niveau de l'alignement 2.

```
CIROP_HUMAN       1 MLLLLLLLLLLLPPLVLRVAASRCLHDETQKSVSLLRPPFSQLPSKSRSSS    50
                        :.|||.|:|..|||||||:||||||:||||..||.....|||:
A0A7F8PVS6_LE     1 -------MFLLPLLLLGTAASRCLHEETQKSVTLLRPHLSQPAPNFRSSA    43

CIROP_HUMAN      51 LTLPSSRDPQPLRIQSCYLGDHISDGAWDPEGEGMRGG--SRALAAVREA    98
                    ||||.||||||||::||:.|.:|||||||||.|||||  :.||||.|:|
A0A7F8PVS6_LE     44 LTLPGSRDPQPLRIRTCYIRDPVSDGAWDPEGAGMRGGPAALALAAARQA    93

CIROP_HUMAN       99 TQRIQAVLAVQGPLLLSRDPAQYCHAVWGDPDSPNYHRCSLLNPGYKGES   148
                    .|::|.:.:|||||||||||||||||||||:|||||||||||||||||
A0A7F8PVS6_LE     94 AQQLQGIFAVQGPLLLSRDPAQYCHAVWGDPDTPNYHRCSLLNPGYKGES   143

CIROP_HUMAN      149 CLGAKIPDTHLRGYALWPEQGPPQLVQPDGPGVQNTDFLLYVRVAHTSKC   198
                    ||||||||||||||||||||.||||||||||:|||||||.|||||||
A0A7F8PVS6_LE    144 CLGAKIPDTHLRGYALWPEQGAPQLVQPDGPGVQDTDFLLYVWVAHTSKC   193

CIROP_HUMAN      199 HQETVSLCCPGWSTAAQSQLTAALTSWAQRRGFVMLPRLCLKLLGSSNLP   248
                    |.|
A0A7F8PVS6_LE    194 HGE-----------------------------------------------   196

CIROP_HUMAN      249 TLASQSIRITGPSVIAYAACCQLDSEDRPLAGTIVYCAQHLTSPSLSHSD   298
                         |||:||||||.|||||||||||||||||||||||||||||||||
A0A7F8PVS6_LE    197 -----------PSVMAYAACCHLDSEDRPLAGTIVYCAQHLTSPSLSHSD   235

CIROP_HUMAN      299 IVMATLHELLHALGFSGQLFKKWRDCPSGFSVRENCSTRQLVTRQDEWGQ   348
                    ||..|||||||||||||||||||||||.|.||:|||||.|||:|||||
A0A7F8PVS6_LE    236 IVTVTLHELLHALGFSGQLFKKWRDCPSGPSAREDCSTRQQVTRRDEWGQ   285

CIROP_HUMAN      349 LLLTTPAVSLSLAKHLGVSGASLGVPLEEEEGLLSSHWEARLLQGSLMTA   398
                    ||||||.||.|||:||||.||..|||| ||||..||||||||||||:|||
A0A7F8PVS6_LE    286 LLLTTPTVSHSLARHLGVPGALPGVPL-EEEGPSSSHWEARLLQGSIMTA   334

CIROP_HUMAN      399 TFDGAQRTRLDPITLAAFKDSGWYQVNHSAAEELLWGQGSGPEFGLVTTC   448
                    |||||:|||||||||||||||:|||||||.||||:|||.||||||||
A0A7F8PVS6_LE    335 TFDGARRTRLDPITLAAFKDSGWYRVNHSAAEGLLWGRGSGLEFGLVTTC   384
```

```
#=======================================
#
# Aligned_sequences: 2
# 1: CIROP_HUMAN
# 2: A0A7F8PVS6_LEPWE
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 790
# Identity:     589/790 (74.6%)
# Similarity:   630/790 (79.7%)
# Gaps:          68/790 ( 8.6%)
# Score: 3103.0
#
#
```

Figure 4 : Debut de l'alignement par paire fournit par le logiciel EMBOSS

En comparant les résultats obtenus avec l'alignement fournit par le logiciel EMBOSS, une région non conservée est observée de façon similaire entre les deux programmes. Le pourcentage de gap observé par le logiciel EMBOSS est de 8,6 %, il est légèrement plus élevé que celui obtenu par le programme développé.

Cela peut s'expliquer par les paramètres donnés aux différents programmes qui peuvent par exemple favoriser l'apparition de gap ou non.

Conclusion et Perspectives

Le programme nécessite des améliorations afin de produire un arbre étant plus précis et de fournir un alignement multiple. Pour pouvoir implémenter l'alignement multiple, il serait nécessaire de parcourir l'arbre phylogénétique, afin de récupérer les groupes de séquences les plus proches et de comparer chacune de ses séquences par un algorithme itératif.