Formation Alignement/ Phylogénie/Motifs

Hélène Chiapello, Géraldine Pascal, Claire Hoede

INRA
Unités GenPhyse et MIAT, Plateforme Genotoul-bioinfo

24 Février 2017

Contexte de la formation

- Le point de départ : une demande d'une équipe de recherche du LISBP (E. Laville *et al*.)

  “ ***Pouvoir caractériser fonctionnellement (fonction, substrat) des familles complexes d'enzymes de synthèse ou de dégradation de sucres****” (CAZy)*

- C'est une question de recherche ;-)

Tour de table

Qui êtes vous ?
Quelle expérience avez-vous en bioinfo ?
Quelles attentes avez-vous pour cette formation ?

Objectifs de la formation

- Acquérir les **concepts de base en bioinformatique** pour
  - -Réaliser des alignements de séquences multiples
  - -Construire et manipuler des arbres phylogénétiques
  - -Rechercher des motifs conservés dans les séquences
- Acquérir une **autonomie de pratique des outils bioinformatique** sans utiliser la ligne de commande
- Vous proposer une **démarche bioinformatique générique adaptée à votre question de recherche**

Un point de départ

AMERICAN SOCIETY FOR MICROBIOLOGY | Applied and Environmental Microbiology

Dividing the Large Glycoside Hydrolase Family 43 into Subfamilies: a Motivation for Detailed Enzyme Characterization

Keith Mewis,[a] Nicolas Lenfant,[b,c] Vincent Lombard,[b,c] Bernard Henrissat[b,c,d]

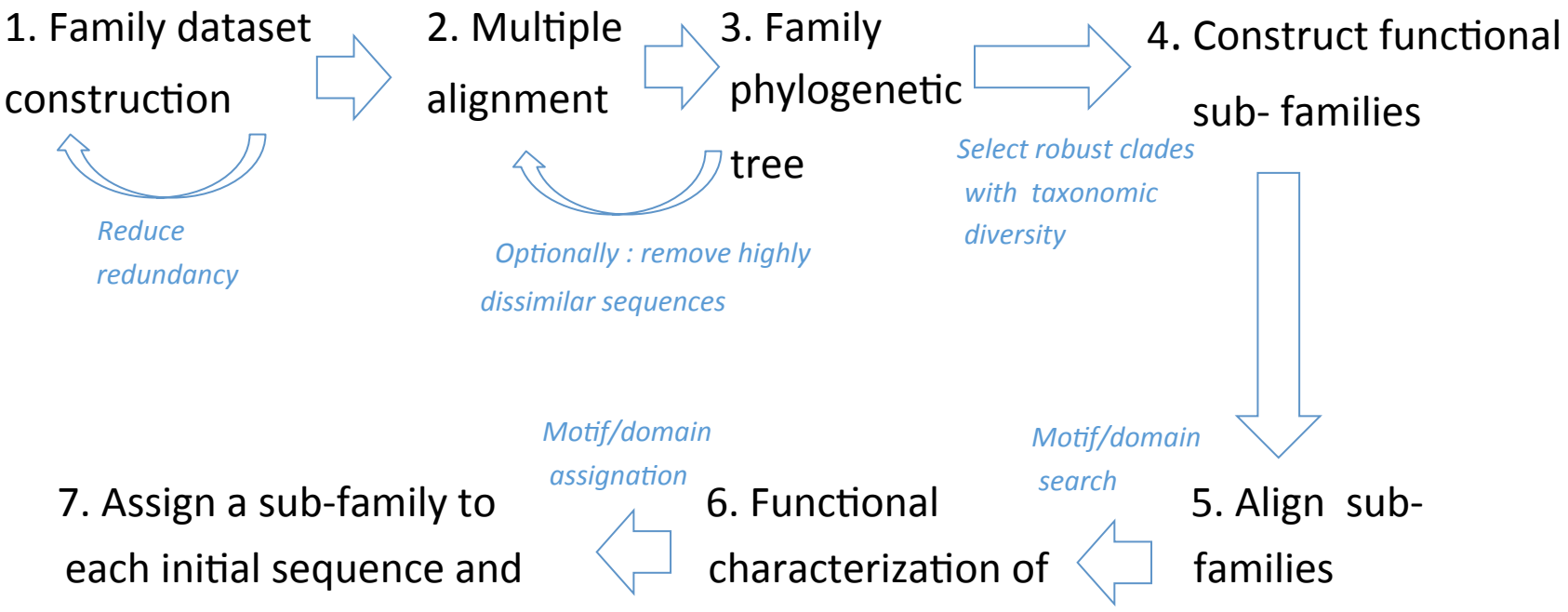
Genome Science and Technology Program, University of British Columbia, Vancouver, BC, Canada[a]; Architecture et Fonction des Macromolécules Biologiques, CNRS, Aix-Marseille Université, Marseille, France[b]; INRA, USC 1408 AFMB, Marseille, France[c]; Department of Biological Sciences, King Abdulaziz University, Jeddah, Saudi Arabia[d]

The rapid rise in DNA sequencing has led to an expansion in the number of glycoside hydrolase (GH) families. The GH43 family currently contains α-L-arabinofuranosidase, β-D-xylosidase, α-L-arabinanase, and β-D-galactosidase enzymes for the debranching and degradation of hemicellulose and pectin polymers. Many studies have revealed finer details about members of GH43 that necessitate the division of GH43 into subfamilies, as was done previously for the GH5 and GH13 families. The work presented here is a robust subfamily classification that assigns over 91% of all complete GH43 domains into 37 subfamilies that correlate with conserved sequence residues and results of biochemical assays and structural studies. Furthermore, cooccurrence analysis of these subfamilies and other functional modules revealed strong associations between some GH43 subfamilies and CBM6 and CBM13 domains. Cooccurrence analysis also revealed the presence of proteins containing up to three GH43 domains and belonging to different subfamilies, suggesting significant functional differences for each subfamily. Overall, the subfamily analysis suggests that the GH43 enzymes probably display a hitherto underestimated variety of subtle specificity features that are not apparent when the enzymes are assayed with simple synthetic substrates, such as pNP-glycosides.

GH43 family :

- Need to be divided into subfamilies
- A strategy proposed to identify functional differences between subfamilies

Une démarche bioinfo générique



Adapted from Mewis et al. App. And Env. Microb. 2016

Déroulé de la formation

3 parties :

- Partie 1 : Alignement de séquences - Géraldine Pascal
- Partie 2 : Construction d'arbres phylogénétiques - Hélène Chiapello
- Partie 3 : Recherche de motifs fonctionnels - Claire Hoede

Et de nombreux travaux pratiques sur un dataset de 62 séquences de CAZY GH130 (E. Laville)

Partie 1 - la stratégie utilisée :

Material and Methods - Mewis et al. App. And Env. Microb. 2016

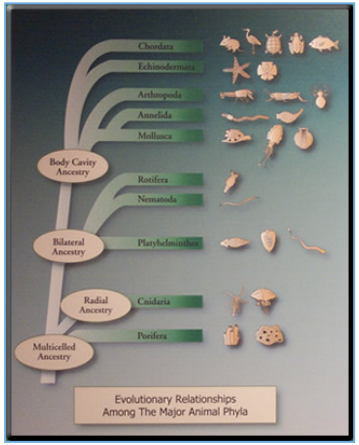
- *All complete GH43 domain sequences were taken from the CAZy database -> 62 sequences from Elisabeth*
- *To reduce redundancy and improve the processing time, sequences were clustered at 95% similarity by using CD-Hit*
- *In order to generate high-quality and relevant alignments, MAFFT was used to iteratively remove highly dissimilar sequences*
- *With these sequences, was used to generate a phylogenetic tree based on the midpoint root method -> NJ*

Alignement de séquence, quelques points de théorie

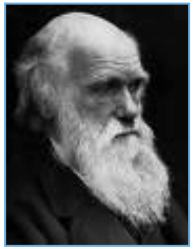
La théorie de l'évolution

Les espèces se modifient au cours du temps et donnent naissance à de nouvelles espèces.

Selon Jean-Baptiste Lamarck (1744-1829), les espèces évoluent en adoptant des caractères acquis par les individus au cours de leur vie.



Charles Darwin (1809-1882) émet l'hypothèse de la sélection du plus apte (ou sélection naturelle) parmi des individus naturellement variant.



La déduction par homologie, ou le «dogme central» de la bioinformatique

Si la bioinformatique «marche», c'est parce que l'évolution des gènes laisse une trace parfaitement visible lorsque l'on compare leur séquence

Évolution des gènes = mutations, insertions, délétion.

Les gènes des organismes modernes sont issus de remaniement de gènes ancestraux

On peut donc déduire la fonction de la plupart des gènes par comparaison avec les gènes «homologues» d'autres espèces.
(homologue = qui a un ancêtre commun)

La déduction par homologie, ou le «dogme central» de la bioinformatique

Les régions fonctionnelles des gènes (sites catalytique, de fixation, etc.) sont soumises à sélection. Elles sont relativement préservées par l'évolution car des mutations trop radicales sont désavantageuses.

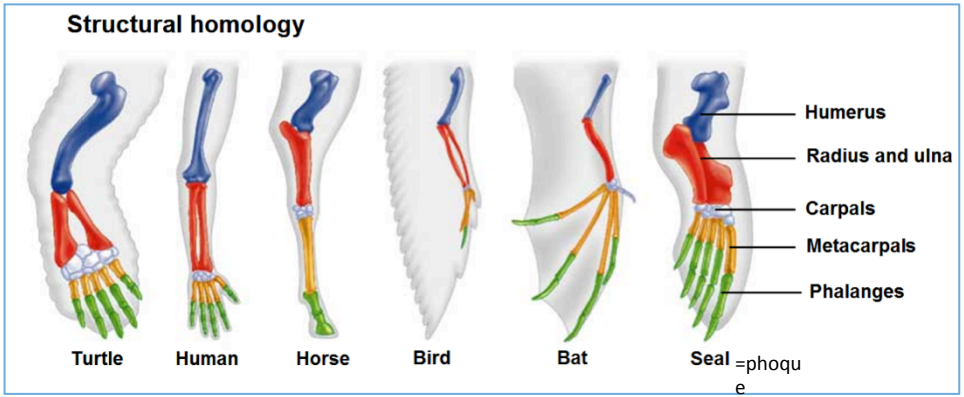
Les régions non fonctionnelles subissent peu de pressions de sélection et divergent rapidement au fur et à mesure que s'accumulent les mutations.

L'homologie de séquence

En bioinformatique aussi Homologie = parenté = ancêtre commun

L'aile de l'aigle est homologue à l'aile du perroquet (oiseaux) et à la patte de la tortue (reptile)

L'aile de la chauve souris est homologue à la patte du cheval et au bras de l'homme (chiroptère, ongulé et primate - mammifères).



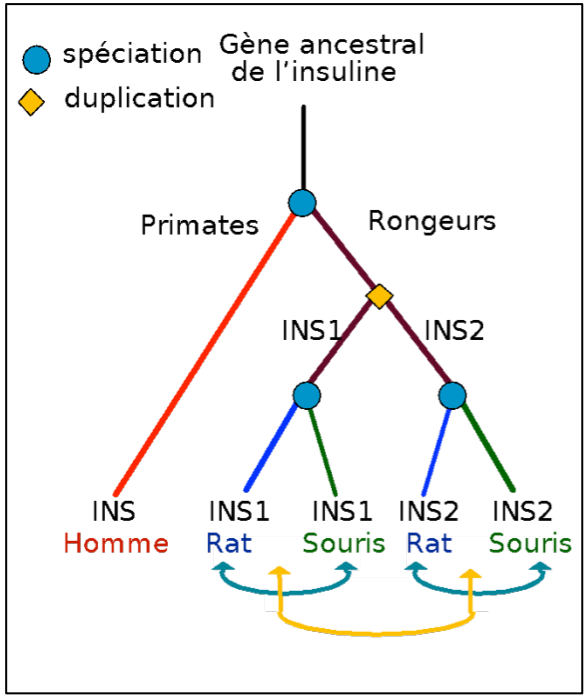
L'homologie de séquence

On est homologue ou on ne l'est pas.

Donc on ne dit pas: "~~très homologue~~", "~~faible homologie~~", «~~22% d'homologie~~», etc.

Pour une notion quantitative, on parle de similitude ("très similaire", etc.) ou d'identité (28% d'identité)

Orthologie et paralogie



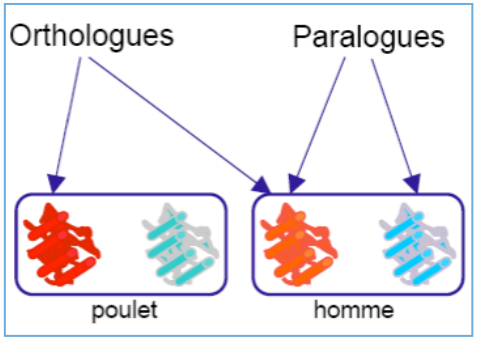
Homologie : 2 gènes sont homologues s'ils ont un ancêtre commun

Orthologie: 2 gènes sont orthologues s'ils ont divergé à la suite d'un événement de spéciation
ex: INS1 et INS2

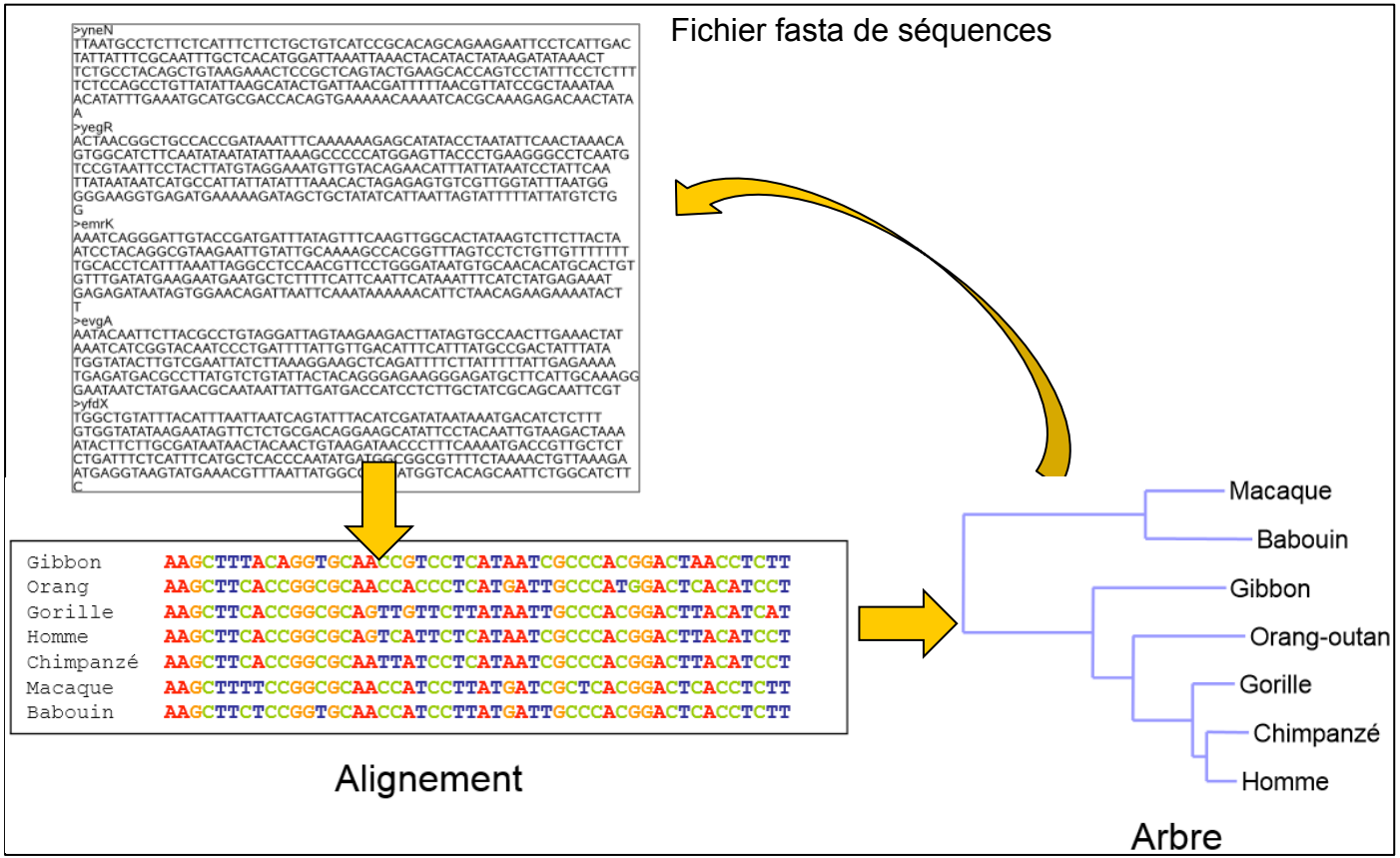
Paralogie: 2 gènes sont paralogues s'ils ont divergé après un événement de duplication
ex: INS1 de rat et INS1 de souris

Fonction et homologie

- Homologie n'implique pas même fonction: par exemple l'aile de l'oiseau et le bras humain n'ont pas la même fonction
- Des orthologues rapprochés (p. ex. homme/souris) ont le plus souvent la même fonction dans l'organisme.
- Des orthologues distants (p. ex. homme/mouche) ont plus rarement le même rôle phénotypique, mais peuvent exercer le même rôle dans une voie donnée.
- Les paralogues acquièrent rapidement des fonctions différentes



Les différentes étapes de la reconstruction d'arbre phylogénétique



Les différentes étapes de la reconstruction d'arbre phylogénétique

<u>Point de départ :</u>

Un ensemble de séquences *homologues* alignées.

Avoir sa séquence et rechercher des séquences *similaires* dans les banques de données par blast par exemple

OU

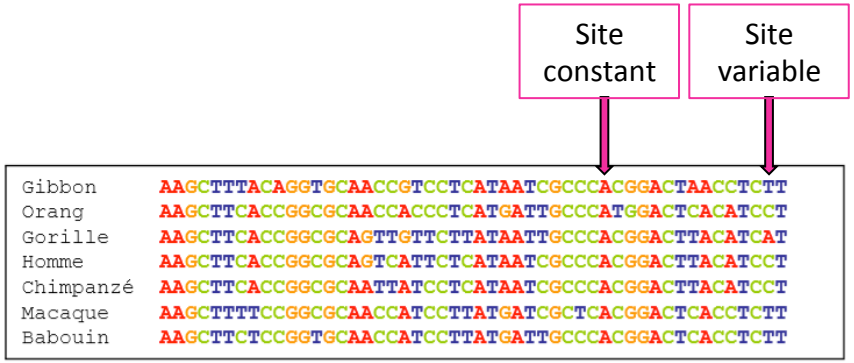
Avoir *sa propre liste* de séquences déduite d'expériences biologiques

Séquences protéiques ou nucléiques

Format standard fasta ou phylip

<u>Alignement multiple</u>

Chaque position dans l'alignement constitue un site.



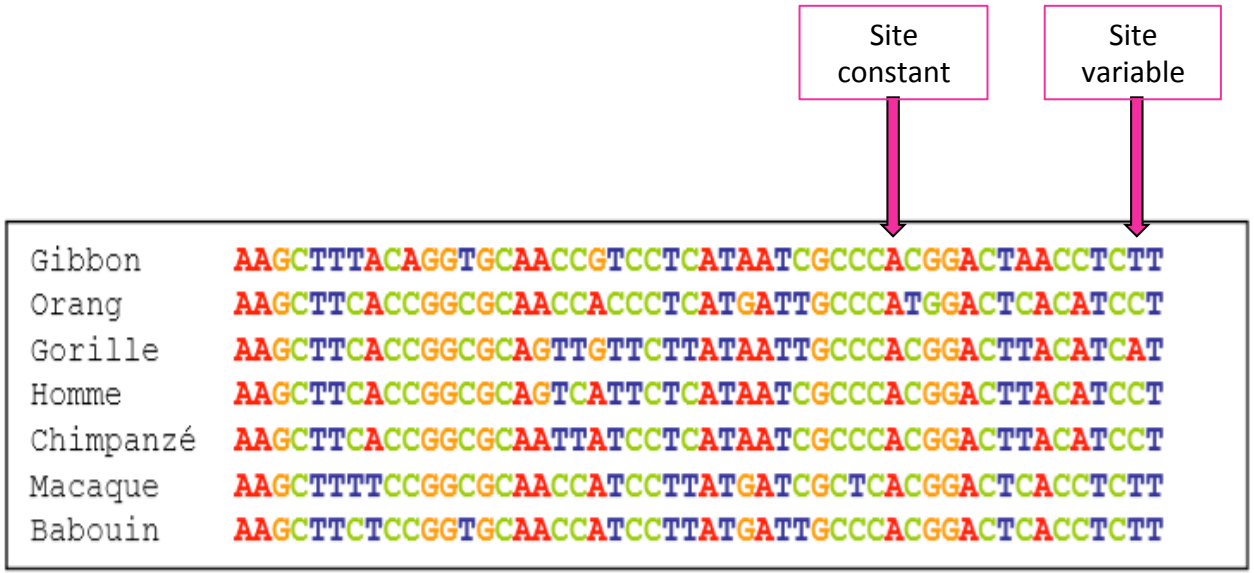
<u>Résultat obtenu :</u>

Un arbre décrivant les relations évolutives entre les sé
i.e. un arbre phylogénétique.

L’alignement multiple

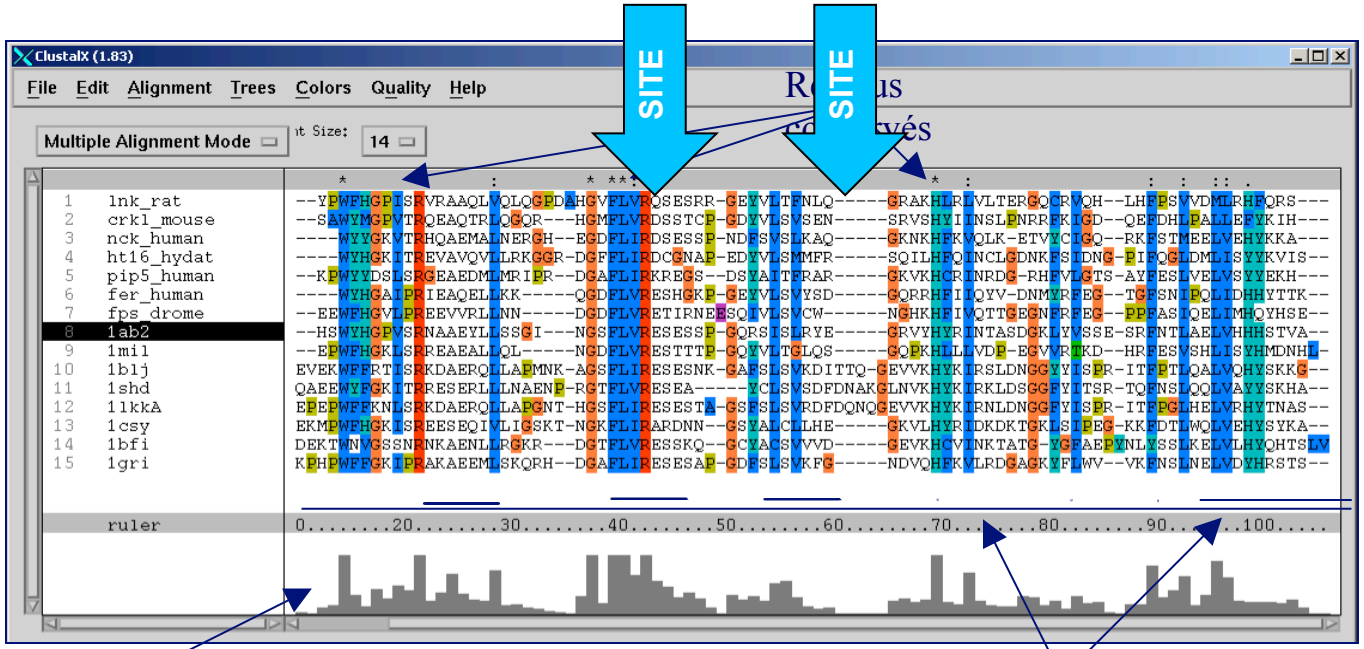
Qu'est-ce qu'un alignement multiple ?

Une représentation d'un ensemble de séquences, dans lesquelles les résidus équivalents (d'un point de vue fonctionnel ou structural) sont alignés en colonnes (un site).



Qu’est-ce qu’un alignement multiple ?

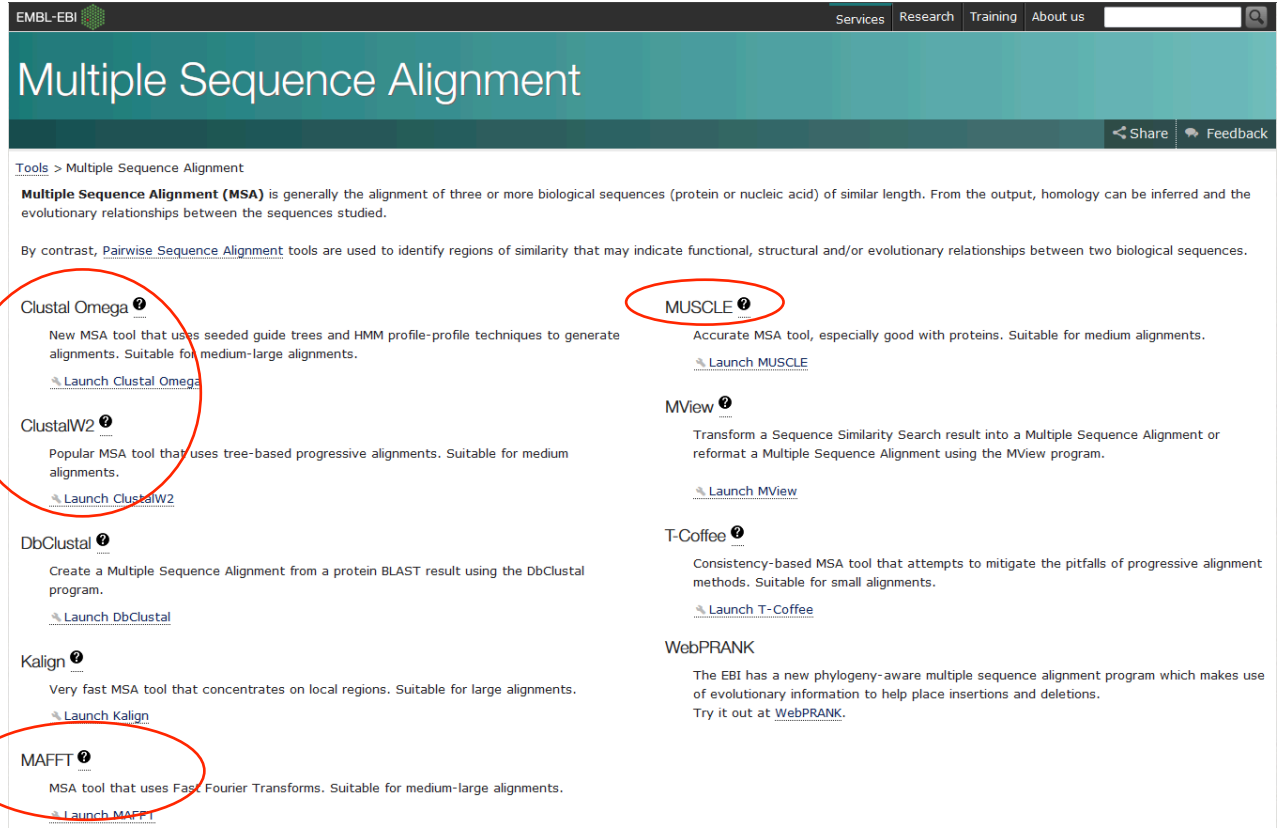
Vue d’un alignement dans un éditeur d’alignement



L’alignement multiple au format fasta

```
>danio
MKATSGPNGLRLCVLSCLFVLHIFTRVAGNMASPSHFGVAST----EVHRNRHPKRRNPH
FRPPVESRTDDDGMRLMLSLYRIAADADGRPKQHKIFGSNTIRLLQASTTEKHFPPTSSD
LQYTYTVKYELNDL-LLDKLVKASFMYLRSPMSSRL----------------PYICEASV
TSLQNPLEGDRITMGPRSRWTETDVTDHV--------SESKDGHVSFFARYWCTKPEHKR
SV----AHRKRF---------PPQHHLRAPLLLLFLEENKHPVEWG--------------
-KSFPPLSRPRTRR
>oryzias
MRATRGTRSLLRTCALTCFLFLLCSTCA-GAATRRNVASHPSTSKRSRRSHQSAKHMGAH
HRPLTDEQKADQNLQFMLSLYRSAAEPDGRPKQHRKFGSNTVRLLRPTASSVHYRPTSGD
LCYTFTVQYKLDTLP-SEQLVRASFIHLRTSSLTLSTSQT----------VEPPQCRAQI
ASF---GGKSLAVLEPHEQWTETDITAHVSAHILQGRDINEEGHLTLTAQYWCTEPGKPD
-----VDERKRW---------NSEPHLEAPSLLLYLEEERENST--LELKDSFLDALN--
-SPTSSS-------
>takifugu
----------------------------MDDRKLSPSRRSSLRRSRARLPDGGAH----
HRPLTDEQKADQNLRFMLNLYQSAAEPDGRPKQHRKFGSNTVRLLKPSASSVRYLPAP--
-----TVQYHMDTLP-TEQLVRASFVHLRSSSATSSSLNAT-------QGAKPPRCEARI
TSL---GQESLVTLEPHEQWMETDITAHVR----QDKNQSPGKFLTLTGQYWCAAEDALV
HSEEDVGLKWWWTRLQERGRRSGEHHLEVPSLLLYLEEQREEQR----------------
---------RHRR-
>tetraodon
------------------------------------------------------------
-RPLTDEQKADQNLQFMMSLYRSAAEPDGRPKQHRKFGSNTVRLLKPSASSVSYLPASPD
HQYHFSVQYHLDTLP-SEQLVRASFVHLRSAP-APFNSSQG---------APPPRCRARV
APL---GRESLVVLEPHQRWTETDITAHVR----QRQDQSPGGALTLTAQYRCTAPMAAQ
GGG---GLPRPW------AQRRGGQHLEVPSLLLYLEEERDGNVWMGDLLG---------
-----PEQRRRRRR
>gasterosteus
------------------------------------------------------------
-----------QNLQFMLSLYRSAAEPDGRPKQHRKFGSNTVRLLRPSAASVRHLPASPD
HRYSFTVQYNLDTVP-SEQLIRASFIHLRSAPSSSSARP-----------LRPPRCRAQI
---PSLGKASLVTLEPHERWTETDITAHVR----RGRSRGPGGHLTLTAQYWCTAWGGGF
--------------------PSTGGSPTSRSRLFFYTWRRS-----------------
-GRTPPQRTPRTRR
```

Les outils les plus populaires



Construire un alignement multiple

Approches traditionnelles

A. alignement multiple optimal
B. alignement multiple progressif
C. alignement multiple itératif

Approches traditionnelles

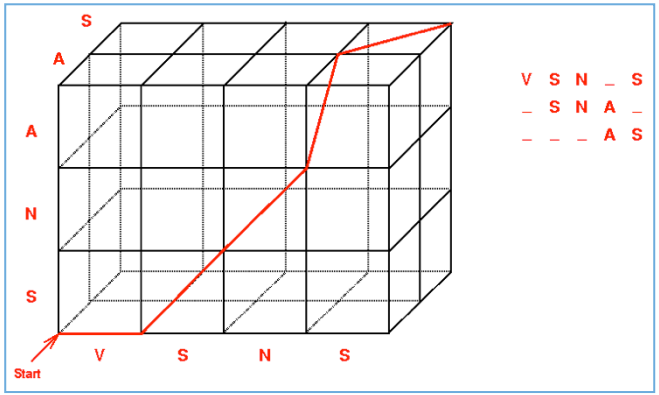
A. alignement multiple optimal
B. alignement multiple progressif
C. alignement multiple itératif

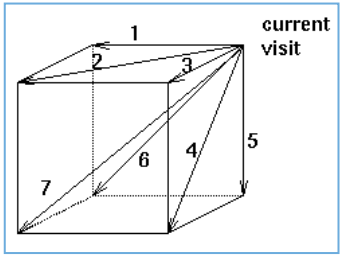
A. Alignement multiple optimal

Extension directe des programmes dynamiques des alignements de séquences par paires à N dimensions (Sankoff, 1975).

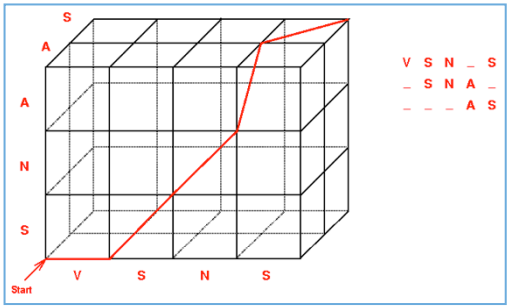
Examine l'ensemble des alignements possibles afin de trouver l'alignement optimal

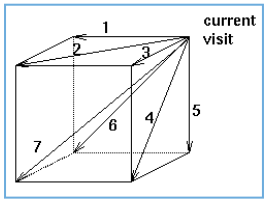
Exemple: alignement de 3 séquences





A. Alignement multiple optimal





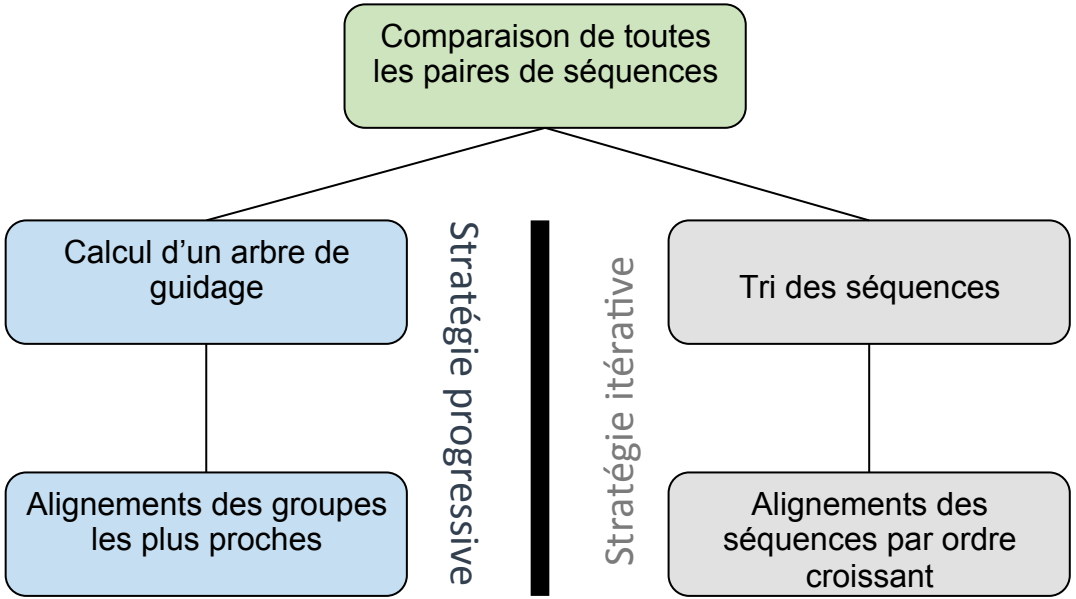
Problème

L'alignement mathématique optimisé n'est pas nécessairement l'alignement biologique optimal.

Le temps CPU (temps de calcul sur ordinateur) et la mémoire requise sont prohibitifs pour un usage classique (temps requis est proportionnel à N^k avec k séquences de longueur N).

En pratique, moins de 10 séquences peuvent être alignées.

Stratégies d'alignements non-optimales



Approches traditionnelles

A. alignement multiple optimal
B. alignement multiple progressif
C. alignement multiple itératif

B. Alignement multiple progressif

Évite le calcul de l'ensemble des alignements possibles

Non garantie d'obtenir l'alignement optimal

Principe :

Les séquences (ou groupe de séquences) sont alignées progressivement par paires

-> exemple Clustal, MUSCLE

B. Alignement multiple progressif

Une approche alternative pour aligner des séquences multiples est de réaliser un alignement progressif.
L'algorithme procède en plusieurs étapes:

- Calculer une matrice de distances, qui indique la distance entre chaque paire de séquences.
- Construire un arbre guide qui regroupe en premier lieu les séquences les plus proches, et remonte en regroupant progressivement les séquences les plus éloignées.
- Utiliser cet arbre pour aligner progressivement les séquences.

Il s'agit d'une approche heuristique. Cette approche est possible pour un grand nombre de séquences, mais ne peut pas garantir de retourner l'alignement optimal.

B. Alignement multiple progressif

1 ère étape: construction de la matrice de distance On effectue un alignement par paires entre chaque paire de protéines

A partir de chaque alignement par paire, calculer la distance entre les deux séquences.

B. Alignement multiple progressif

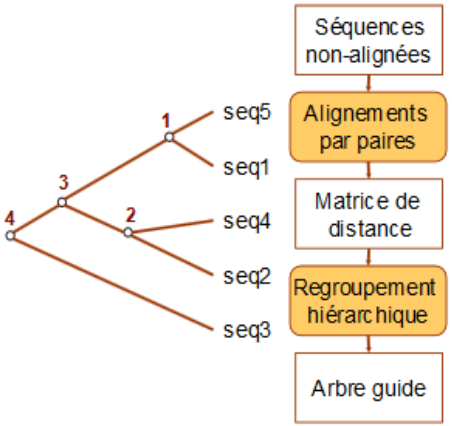
2ème étape : construire l'arbre guide

On peut calculer un arbre à partir d'une matrice de distance par regroupement hiérarchique.

On commence par regrouper les deux séquences les plus proches (groupe 1). On regroupe ensuite les groupes les plus proches. Les deux séquences les plus proches (groupe 2). Un groupe avec un groupe (groupe 3). Une séquence avec un groupe précédent (groupe 4).

Cet arbre sera ensuite utilisé comme guide pour déterminer l'ordre d'incorporation des séquences dans l'alignement multiple.

Attention ! Cet arbre ne doit pas être interprété comme un arbre phylogénétique. Il sert uniquement à identifier les similarités les plus fortes entre séquences pour construire l'alignement multiple.

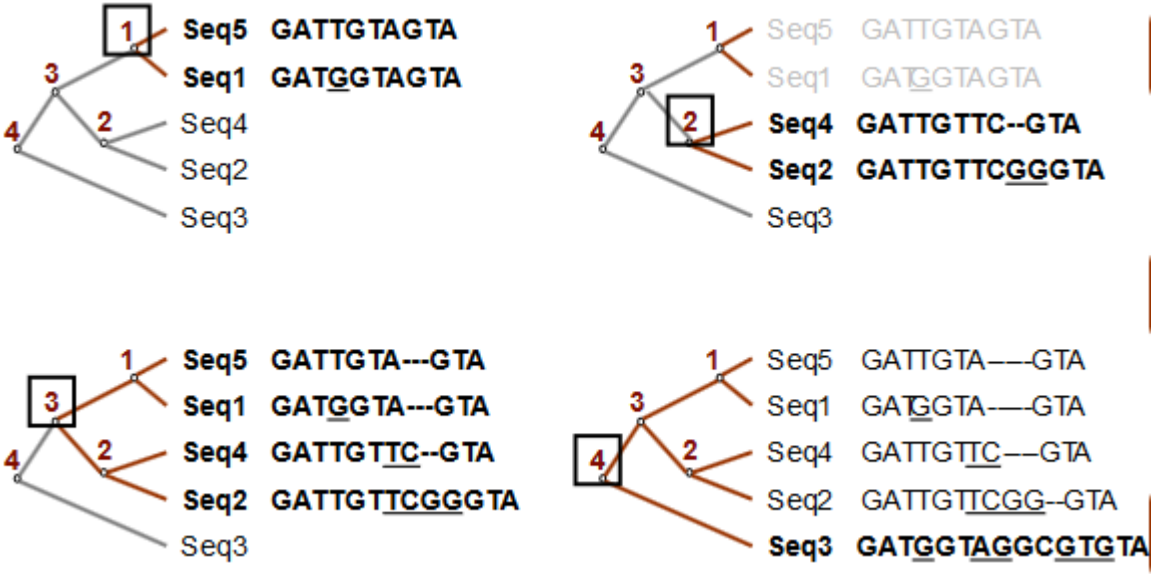


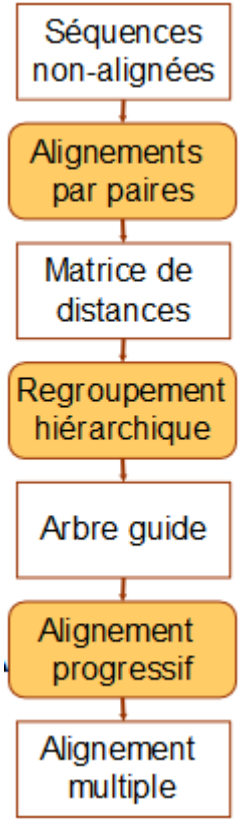
B. Alignement multiple progressif

3 ème étape: alignement multiple
On construit un alignement multiple en incorporant progressivement les séquences selon leur ordre de branchement dans l'arbre guide, en remontant des plus proches aux plus éloignées.

B. Alignement multiple progressif

3 ème étape: alignement multiple





Approches traditionnelles

A. alignement multiple optimal
B. alignement multiple progressif
C. alignement multiple itératif

C. Alignement multiple itératif

Les méthodes itératives calculent une solution sous-optimale et continuent de la modifier intelligemment en utilisant la programmation dynamique ou d'autres méthodes jusqu'à ce que la solution converge. Contrairement aux méthodes progressives, les méthodes itératives peuvent corriger de façon dynamique les erreurs d'alignement.

MAFFT: méthode itérative (Katoh et al, 2002) d'alignement multiple

MAFFT = programmes de nouvelle génération

MAFFT a été écrit dans le but explicite d'accélérer considérablement le processus d'alignement multiple,

permettant ainsi d'aligner un grand nombre de séquences sans pour autant sacrifier à la qualité de l'alignement.

MAFFT: méthode itérative (Katoh et al, 2002) d'alignement multiple

3 grandes étapes :

1ère ETAPE

- Chaque acide aminé est décrit par sa polarité et son volume, les séquences sont réécrites dans ce système. Par exemple, les acides aminés hydrophobes I, L, M et V forment un seul groupe, de même que D, E,N et Q, etc.

suite de lettres (chaque séquence) → une suite de valeurs numériques.

- Les nucléotides sont recodées en utilisant les fréquences locales des quatre bases.
- Puis les segments de similarité entre chaque paire de séquences sont repérés au moyen d'un algorithme de calcul appelé transformée de Fourier rapide, ou Fast Fourier Transform (FFT) en anglais.
- Les paires de séquences sont ensuite alignées sur la base de ces segments de similarité. Sauf pour des séquences très divergentes, ce procédé permet d'aligner toutes les paires de séquences environ 10 fois plus vite que ClustalW (approche progressive).

MAFFT: Alignement multiple

2ème ETAPE

Un arbre de guidage est ensuite calculé à partir des alignements précédents.

La distance entre 2 séquences est estimée à partir du nombre de mots de 6 lettres (*k-mers*) que ces séquences partagent dans ce nouvel alphabet

MAFFT: Alignement multiple

3ème ETAPE

Les séquences sont ensuite alignées progressivement en suivant l'ordre indiqué par l'arbre de guidage.

MAFFT: Alignement multiple

Plusieurs programmes sont proposés sur la page de garde du site Web de MAFFT.

Ce que nous venons de décrire correspond à l'option FFT-NS-1.

Ce nom un peu barbare signifie Fast Fourier Transform-New Scoring matrix-1 step.

MAFFT: Alignement multiple

MAFFT peut occasionnellement procéder à un deuxième passage (approche itérative).

Dans ce dernier, l'alignement réalisé précédemment sert à re-calculer la distance entre chaque paire de séquences, un nouvel arbre de guidage et un nouvelle alignement multiple.

Cette option s'appelle FFT-NS-2.

MAFFT: Alignement multiple

MAFFT peut procéder à un raffinement de l'alignement.

Dans ce cas, l'arbre de guidage est scindé en deux, puis les deux moitiés sont réalignées.

On recommence ainsi tant que le score d'alignement s'améliore.

On procède alors à un nombre *i* d'itérations *i* étant inconnu *a priori*.

Cette option porte le nom de FFT-NS-i.

On peut, sur la page de garde de MAFFT, limiter à deux le nombre d'itérations (« two cycles only »).

MAFFT: Alignement multiple

Il faut par ailleurs noter que le site de MAFFT propose des programmes fondés non pas sur la transformées de Fourier rapide, mais sur l'algorithme de programmation dynamique.

Ainsi le programme nommé G-INS-i aligne les paires de séquences suivant l'algorithme global de Needlman-Wunsch, calcule un arbre de guidage, aligne toutes les séquences suivant cet arbre et procède enfin à un raffinement de l'alignement comme décrit ci-avant.

Les programmes L-INS-i et E-INS-i procèdent de la même façon, mais avec l'algorithme d'alignement local de Smith-Waterman.

Bien entendu, ces programmes, nettement plus lents, ne conviennent pas pour un grand nombre de séquences.

Enfin, le programme Q-INS-i est spécifiquement dédié à l'alignement de séquences d'ARN.

MAFFT - bilan des paramètres

1. Mode basique, rapide — juste progressif
   a) FFTNS1 (fftns --retree 1)
   b) FFTNS2 (fftns) (same as mafft --retree 2)

OK jusqu'à 1 000 séquences facilement alignables

2. Mode intermédiaire — progressif + itérations
   a) FFTNSI (fftnsi) default two cycles, or e.g. fftnsi --maxiterate 1000
   b) NWNSI (nwnsi) same as FFTNSI, but no FFT, Needleman-Wunsch only.

OK entre 100 et 500 séquences

3. Mode avancé — progressif + itérations + consistance (cf. T-Coffee)
   a) EINSI (einsi) Smith-Waterman (plusieurs régions similaires même ordre)
   b) LINSI (linsi) Smith-Waterman stricte (1 région similaires)
   c) GINSI (linsi) global Needleman-Wunsch

Publication MAFFT

- https://academic.oup.com/mbe/article/30/4/772/1073398/MAFFT-Multiple-Sequence-Alignment-Software-Version

Visitons le site de MAFFT

- http://mafft.cbrc.jp/alignment/server/

MAFFT version 7

Multiple alignment program for amino acid or nucleotide sequences

Download version
- Mac OS X
- Windows
- Linux
- Source

Online version
- **Alignment**
- mafft --add
- Merge
- Phylogeny
- Rough tree

Merits / limitations
Algorithms
Tips
Benchmarks
Feedback
Follow

For a large number of short sequences, try an experimental service (2016/Jul).

Multiple sequence alignment and NJ / UPGMA phylogeny

Input:
Paste protein or DNA sequences in fasta format. Example

Au lieu d'utiliser les séquence brutes fournies par Elisabeth nous allons les clusteriser car vous aurez beaucoup de séquences à soumettre à MAFFT en réalité.

or upload a **plain text** file: Choisissez un fichier Aucun fichier choisi

Use structural alignment(s)
Allow unusual symbols (Selenocysteine "U", Inosine "i", non-alphabetical characters, *etc.*) Help

UPPERCASE / lowercase:
- Same as input
- Amino acid → UPPERCASE / Nucleotide → lowercase

Direction of nucleotide sequences: Help
- Same as input
- Adjust direction according to the first sequence (accurate enough for most cases)
- Adjust direction according to the first sequence (only for highly divergent data; **extremely slow**)

Output order:
- Same as input
- Aligned

Notify when finished (optional; recommended when submitting large data):
Email address:

Submit Reset

Visitons le site de CD-HIT

- http://weizhong-lab.ucsd.edu/cdhit_suite/cgi-bin/index.cgi?cmd=Server%20home

CD-Hit permet de regrouper les séquences très proche (ici 95%), limite la redondance et donc le nombre de séquences. Basiquement, CD-HIT est un algorithme progressif glouton qui débute avec la séquence la plus longue et sera donc la séquence représentative du premier cluster, puis traite les séquences restantes de la plus longue à la plus court pour classer chaque séquence comme une séquence redondante ou représentative basée sur ses similitudes avec la séquence existante.

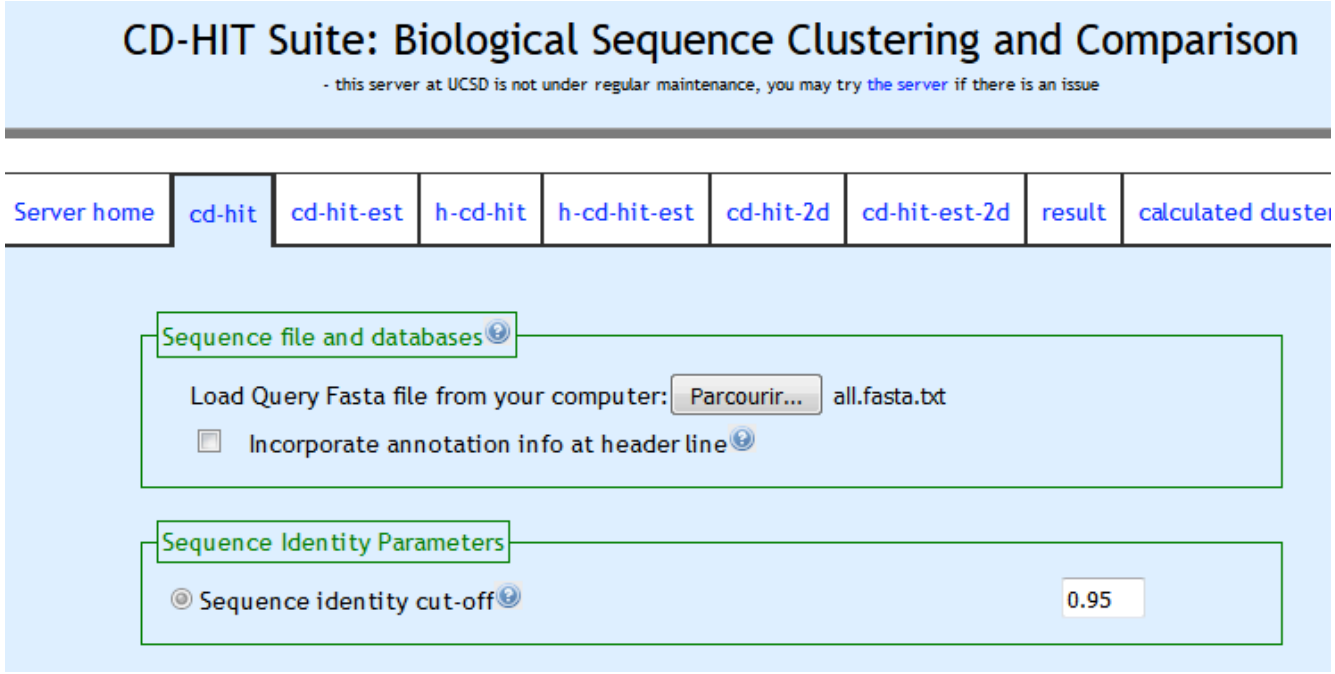
Visitons le site de CD-HIT

- http://weizhong-lab.ucsd.edu/cdhit_suite/cgi-bin/index.cgi?cmd=Server%20home

CD-Hit -> Très utile dans le cadre de très grand jeux de données.

Input data :
http://genoweb.toulouse.inra.fr/~formation/AlignClusteringLISBP/data

le site de CD-HIT



Résultats de CD-Hit

You job 1487863706 is finished.
Program you ran: cd-hit

You input file is all.fasta.txt and we named it as 1487863706.fas.0
Summary information for 1487863706.fas.0 included in 1487863706.fas.0.stat
You required 1 runs for sequence clustering

1. Fasta file for representative sequences at 95% identity is 1487863706.fas.1
   Summary information for 1487863706.fas.1 included in 1487863706.fas.1.stat
   Corresponding cluster file is1487863706.fas.1.clstr
   Sorted cluster file by size is 1487863706.fas.1.clstr.sorted

---

Generated shell script is run-1487863706.sh

```
faa_stat.pl 1487863706.fas.0
cd-hit -i 1487863706.fas.0 -d 0 -o 1487863706.fas.1 -c 0.95 -n 5 -G 1 -g 1 -b 20 -s 0.0 -aL 0.0 -aS 0.0
faa_stat.pl 1487863706.fas.1
clstr_sort_by.pl no < 1487863706.fas.1.clstr > 1487863706.fas.1.clstr.sorted
clstr_list.pl 1487863706.fas.1.clstr 1487863706.clstr.dump
gnuplot1.pl < 1487863706.fas.1.clstr > 1487863706.fas.1.clstr.1; gnuplot2.pl 1487863706.fas.1.clstr.1 1487863706.fas.1.clstr.1.png
clstr_list_sort.pl 1487863706.clstr.dump 1487863706.clstr_no.dump
clstr_list_sort.pl 1487863706.clstr.dump 1487863706.clstr_len.dump len
clstr_list_sort.pl 1487863706.clstr.dump 1487863706.clstr_des.dump des
```

57 clusters de séquences

```
>BAC_FRA_CC_emb_CAH06518.1 cluster21 [Bacteroides fragilis NCTC 9343]
MSLFNDKVAKLLAGHEALLMRKNEPVEEGNGVITRYRYPVLTAAHTPVFWRYDLNEETNPFLMERIGMNATLNAGAIKWDGKYLMLVRVEGAD
>BAC_OVA_CC_gb_EDO10988.1 cluster21 [Bacteroides ovatus ATCC 8483]
MNNMKSTFLFLLTTTMMTCTAYGQSSNHKENKLPDWAFGGFERPKNVNPVISPIENTKFYCPLTKDSIAWESNDTFNPAATLYNGEIVVLYRA
>BAC_THE_CC_gb_AAO76140.1 cluster46 [Bacteroides thetaiotaomicron VPI-5482]
MNKIQIPWEERPVGCTDVMWRYSQNPVIGRYHIPSSNSIFNSAVVPFKDGFAGVFRCDNKAVQMNIFTGFSKDGIHWDISHEPIQFKAGNTEM
>BAC_THE_CC_gb_AAO78885.1 cluster26 [Bacteroides thetaiotaomicron VPI-5482]
MKSTFLFLVTTTMMTCTALGQPSNDKKNVLPDWAFGGFERPQGANPVISPIENTKFYCPMTQDYVAWESNDTFNPAATLHDGKIVVLYRAEDK
>CAL_POL_CC_ref_WP_026485574.1 cluster1 [Caldanaerobius polysaccharolyticus]
MSKQLIVGEALPNIPWQDRPAGCNDVVWRYQQNPVIPRDLIPSSNSIFNSAVVPFNGEFAGVFRCDTKSRQMEIHSGRSKDGLNWEIDHERIK
>CEL_MIX_CC_gb_AAS19693.1 cluster16 [Cellvibrio mixtus]
MSSFKEKAKALLQQHETLITKKNAVKRDGNGIYDCYENPILTAEHAPVFWRYDLNEKTNPHLMERQGINAAFNSGAMYWNGKYILAVRVEGVD
>DTA_FER_CC_gb_ACT94389.1 cluster27 [Dyadobacter fermentans DSM 18053]
MKNFIAGFAICTSILTGQAFAQTENKLPDWAFGGFKRPAGVNPVVSPDSTTRFFCPMNKREVDWESNDTFNPAATMKNGKIVVLYRAEDKAGK
>LIS_INN_CC_emb_CAC96089.1 cluster31 [Listeria innocua Clip11262]
MNIYRYEENPLITPLDVKPIHEGFEVIGAFNGGVAEYNGEVLLLLRVAEKPVSEDPEIVLAPVYNAKNKELELQSFRLDDENYDFEDPRMIRS
>RUM_ALB_CC_gb_ADU21379.1 cluster24 [Ruminococcus albus 7 = DSM 20455]
MIHEKYTEMRNEQEALLSRKNTKTSFYNGIYDRYEHPVLTREHIPLHWRYDLNKETNPFFQERLGINAVFNAGAIKLNDRYCLVARVEGNDRK
>RUM_ALB_CC_gb_ADU20661.1 cluster38 [Ruminococcus albus 7 = DSM 20455]
MKTQIINGVSLPNIPWQDKPADCKDVIWRYDANPIIPRDQLPTSNSIFNSAVVPYESEKGKFAGVFRVDDKCRNMELHAGFSKDGIHWDINPD
>THE_SP_CC_gb_ABY93074.1 cluster51 [Thermoanaerobacter sp. X514]
MFRLTRLSNKPILSPIKEHEWEKEAVFNAAVIYEGNKFHLFYRASNNKFVLNTEKPEEKYKFVSSIGYAVSEDGINFERFDKPVLVGEIPQEA
>THE_SP_CC_gb_ABY93073.1 cluster54 [Thermoanaerobacter sp. X514]
MIKLKRLSDKPVLMPKAENEWERAAVFNTAAIYDNGLFHLIYRATDIGPHAKYGKYISRLGYAVSKDGINFMRLDKPVMSNETEQELRGLEDP
>UNC_ORG_CC_gb_ADD61463.1 cluster3 [uncultured organism]
MSMSSKVIIPWEERPAGCKDVLWRSVANPIIPRDLLPTSNSIFNSAVVPFGDGFAGVFRCDDTSRRMRLHVGFSKDAINWNIKEEPLKFQCDD
>BAC_XYL_emb_CBK68185.1 cluster28 [Bacteroides xylanisolvens XB1A]
MKRKLQNIAYLLMAAAFVASCSEKKQISEFPDWAWTDFQRPEGVNPIISPDTTTLFYCPMRQDSIAWEASDTFNPAATIYDGKVVVLYRAEDN
>CLO_SP_gb_AEY67872.1 cluster37 [Clostridium sp. BNL1100]
MSIIGKTLKNIPWQDKPLGCSSVIWRHEGNPIIGWNPTPRTARIYNSSVVPWNSGFAGIFRADHKDGKARIHVGFSSDGVNWNVEDAPIVWHD
>PAE_POL_gb_ADO59098.1 cluster32 [Paenibacillus polymyxa SC2]
MSKLNVDLDVGNVRIVGDSLSKMPWQDKPEGSEAPVWRHTENPVIGRNPVPGIARIFNSAVAPYEGRFVGIFRAETINGRPHLHLGWSDDGLA
>PRE_DEN_gb_AGB28392.1 cluster39 [Prevotella dentalis DSM 3688]
MNTLKIQGPALPGMPWEDRAENDKHVMWRSHRNPIIPRDLLPTSNSIFNSAVVPFGDGYAGVFRCDDTNRRMALHAGFSTNGIDWHINEAPLR
>LAC_PHY_gb_ABX42090.1 cluster47 [Lachnoclostridium phytofermentans ISDg]
MKNIPWEPRPVDCEDVVWRYSKNPIIHRNEIKRSNSIFNSAVVPFKDGYAGVFRCDDKRREMLLHAGFSVDGVKWNINPEPIEFQSEVEDSEP
```

Sauver le fichier fasta en 57clusters.fasta

MAFFT version 7

Multiple alignment program for amino acid or nucleotide sequences

Download version
- Mac OS X
- Windows
- Linux
- Source

Online version
- **Alignment**
- mafft --add
- Merge
- Phylogeny
- Rough tree

Merits / limitations
Algorithms
Tips
Benchmarks
Feedback

Follow

For a large number of short sequences, try an experimental service (2016/Jul).

Multiple sequence alignment and NJ / UPGMA phylogeny

Input:
Paste protein or DNA sequences in fasta format. Example

input = données issus de CD-HIT
57clusters.fasta

or upload a **plain text** file: Choisissez un fichier Aucun fichier choisi

Use structural alignment(s)
Allow unusual symbols (Selenocysteine "U", Inosine "i", non-alphabetical characters, *etc.*) Help

UPPERCASE / lowercase:
- Same as input
- Amino acid → UPPERCASE / Nucleotide → lowercase

Direction of nucleotide sequences: Help
- Same as input
- Adjust direction according to the first sequence (accurate enough for most cases)
- Adjust direction according to the first sequence (only for highly divergent data; **extremely slow**)

Output order:
- Same as input
- Aligned

Notify when finished (optional; recommended when submitting large data):
Email address:

Submit Reset

Advanced settings

Strategy:

- (•) Auto (FFT-NS-1, FFT-NS-2, FFT-NS-i or L-INS-i; depends on data size) ***Updated***

Progressive methods

- ( ) FFT-NS-1 (Very fast; recommended for >2,000 sequences; progressive method)
- ( ) FFT-NS-2 (Fast; progressive method)
- ( ) G-INS-1 (Slow; progressive method with an accurate guide tree)

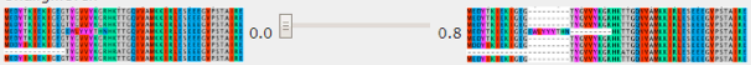
Iterative refinement methods

- ( ) FFT-NS-i (Slow; iterative refinement method)
- ( ) E-INS-i (Very slow; recommended for <200 sequences with multiple conserved domains and long gaps) Help ***Updated*** (2015/Jun)
- ( ) L-INS-i (Very slow; recommended for <200 sequences with one conserved domain and long gaps) Help
- ( ) G-INS-i (Very slow; recommended for <200 sequences with global homology) Help
- ( ) Q-INS-i (Extremely slow; secondary structure of RNA is considered; recommended for a global alignment of highly divergent ncRNAs with <200 sequences × <1,000 nucleotides; the number of iterative cycles is restricted to two, 2016/May) Help

Align unrelated segments, too? *in Alpha Testing* (2014/Mar)

If the input data is expected to be globally conserved but locally contaminated by unrelated segments, try 'Unalignlevel>0' and possibly 'Leave gappy regions'.

Unalignlevel:



↑ Default

This feature is available only when G-INS-1 or G-INS-i is selected in the **Strategy** section above.

- (•) Try to align gappy regions anyway
- ( ) Leave gappy regions (Not recommended for >~1,000 sequences)

Parameters:
Scoring matrix for amino acid sequences: BLOSUM62 ▾
Scoring matrix for nucleotide sequences: 200PAM / κ=2 ▾
↑ Switch it to '1PAM / κ=2' when aligning closely related DNA sequences.
Gap opening penalty: 1.53 (1.0 – 5.0)
Offset value: 0.0 (0.0 – 1.0)

Score of N in nucleotide data: Example
↓ Long stretches of Ns tend to be gapped (excluded from the alignment).
(nzero) N has no effect on the alignment score.
(nwildcard) N is treated like a wildcard. ***Experimental option*** (2016/Apr/26)
↑ Try this if Ns should be aligned with usual letters.

Mafft-homologs (Collects homologs from SwissProt by BLAST and performs profile-based alignments; Protein only): Help
On
Show homologs (if any)
Number of homologs: 50 (5 – 200)
Threshold: *E* = 1e-10 (1e-5 – 1e-40)

Plot LAST hits (DNA only):
The top sequence vs the others The longest sequence vs the others
Plot and alignment Plot only Alignment only
Threshold: score=39 (E=8.4e-11) ▾

Submit Reset

Alignement -> score -> matrice de substitution

- **Distance d'édition**
  - Selon ce concept, le bon alignement est celui qui minimise les opérations à réaliser pour passer d'une séquence à l'autre.
  - Opérations: conservation, remplacement/mutation, délétion, insertion. Une pénalité peut être affectée à chaque opération, par exemple c=0, m=1, d=2, i=2. La distance finale entre les deux séquences (distance d'édition) est la somme de ces pénalités.

	Seq 1	CAGTGGT-GC	
	Seq 2	CA-TCGTAGC	c=0, m=1, d=2, i=2.
Ou, variante:	distance	ccicmccdcc = 0+0+2+0+1+0+0+2+0+0 = 5	
	ressemblance	ccicmccdcc = 2+2-1+2-1+2+2-1+2+2 = 11	c=2, m=-1, d=-1, i=-1.

- **Une délétion à l'intérieur d'une séquence est considéré comme une insertion dans la séquence lui faisant face.**

Matrices de Substitution

- Matrice 4X4 (nucléotides) ou 20x20 (acides aminés) décrivant la distance ou la similitude entre résidus.
  - Estiment le coût ou le taux de remplacement d'un résidu par un autre (distance).
  - Le choix d'une matrice affecte fortement le résultat de l'analyse. Chaque matrice de score représente implicitement une théorie évolutive donnée

Matrices DNA

	A	C	G	T
A	2	-1	-1	-1
C	-1	2	-1	-1
G	-1	-1	2	-1
T	-1	-1	-1	2

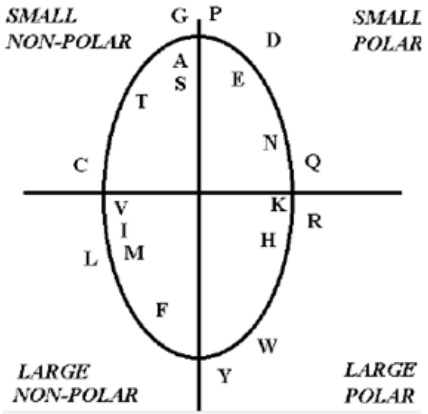
Matrice identité

	A	C	G	T
A	3	-1	1	-1
C	-1	3	-1	1
G	1	-1	3	-1
T	-1	1	-1	3

Matrice transition/transversion

Matrices de Substitution

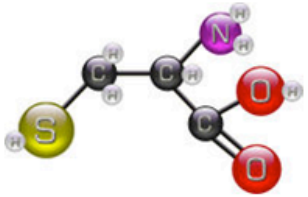
- Matrices fondées sur le code génétique
  - Les scores sont déterminés en fonction du nombre commun de nucléotides présents dans les codons des acides aminés, ce qui revient à considérer le minimum de changements nécessaires en bases pour convertir un acide aminé en un autre.
- Matrices fondées sur les propriétés physicochimiques
  - Les plus courantes sont celles basées sur le caractère hydrophile ou hydrophobe des protéines. Ces matrices sont peu utilisées.



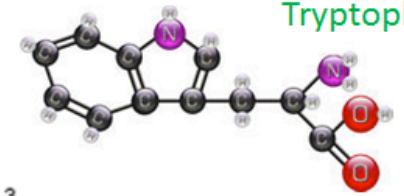
Une représentation bidimensionnelle des propriétés des aa calculée d'après la matrice de Dayhoff par G. Vriend, Centre for Molecular and Biomolecular Informatics, University of Nijmegen

Matrice de Dayhoff (1979)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X
A	2																						
R	-2	6																					
N	0	0	2																				
D	0	-1	2	4																			
C	-2	-4	-4	-5	12																		
Q	0	1	1	2	-5	4																	
E	0	-1	1	3	-5	2	4																
G	1	-3	0	1	-3	-1	0	5															
H	-1	2	2	1	-3	3	1	-2	6														
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5													
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6												
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5											
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6										
F	-4	-4	-4	-6	-4	-5	-5	-5	-2	1	2	-5	0	9									
P	1	0	-1	-1	-3	0	-1	-1	0	-2	-3	-1	-2	-5	6								
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2							
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3						
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	-0	-6	-2	-5	17					
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10				
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4			
B	0	-1	2	3	-4	1	2	0	1	-2	-3	1	-2	-5	-1	0	0	-5	-3	-2	2		
Z	0	0	1	3	-5	3	3	-1	2	-2	-3	0	-2	-5	0	0	-1	-6	-4	-2	2	3	
X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0



Cystéine



Tryptophane

Matrice de Blosum 62

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

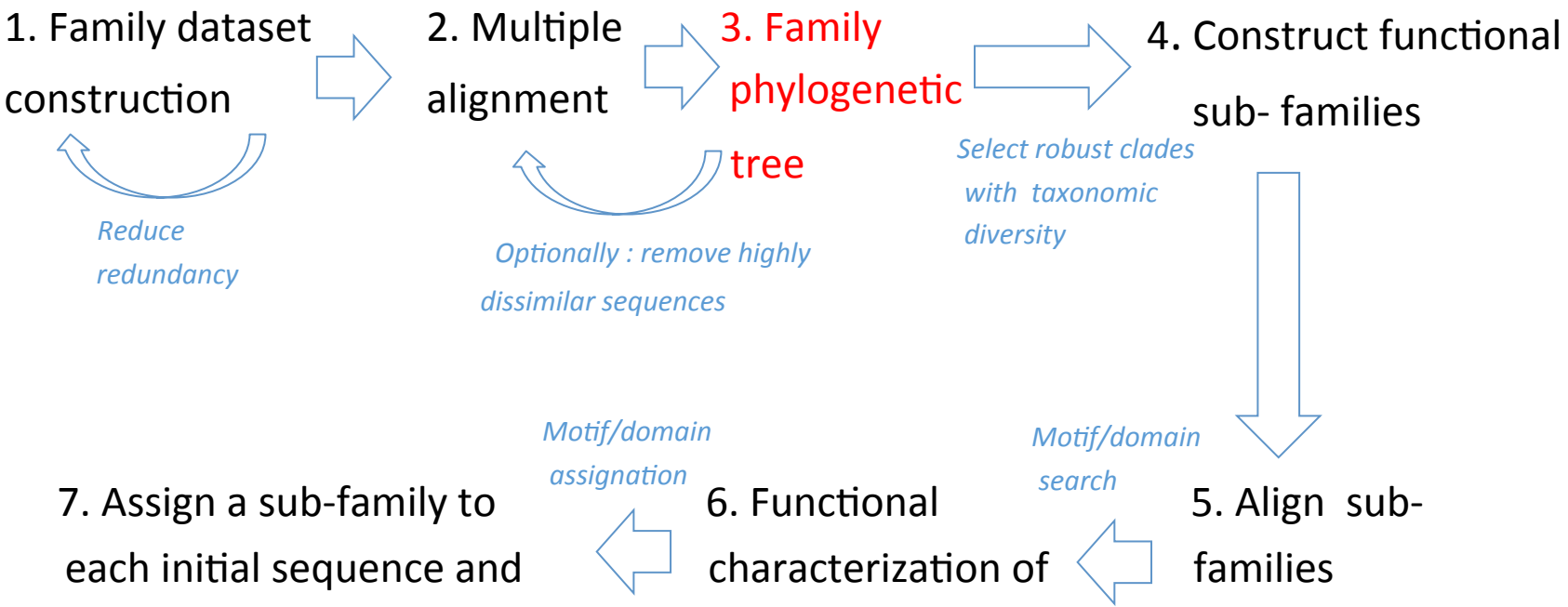
Quelle matrice doit-on utiliser?

- Les matrices BLOSUM sont les plus souvent proposées comme matrices par défaut car les fréquences de substitution sont directement calculées à partir de l'alignement.
- La BLOSUM62 (ou PAM120) est utilisée comme matrice par défaut car elle offre un bon compromis quand les distances évolutives entre les séquences ne sont pas connues.
- La BLOSUM80 (ou PAM40) donnera de meilleurs résultats pour des séquences proches dans l'évolution. Elle tend à trouver des alignements courts fortement similaires.
- La BLOSUM30 (ou PAM350) donnera de meilleurs résultats pour des séquences éloignées dans l'évolution. Elle trouvera de plus longs alignements locaux de faible conservation.

Partie 2 : Construction d'arbres phylogénétiques

Concepts de base

Une démarche bioinfo générique



Adapted from Mewis et al. App. And Env. Microb. 2016

Partie 2 - la stratégie utilisée:

Material and Methods - Mewis et al. App. And Env. Microb. 2016
"Manual separation of subfamilies was decided based on phylogenetic distances in this reduced tree. Subfamilies were required to contain at least 5 sequences found in this reduced tree in order to generate a proper multiple-sequence alignment"

Adaptation : define subfamilies using the refined phylogenetic tree and the three following criteria

- Sequences from robust and fully-resolved clades MAFFT
- Enough sequence number
- Large enough taxonomic diversity (at the class level) NCBI

La reconstruction d'arbre phylogénétique : plan

- Introduction et concepts de base (arbres, distances)

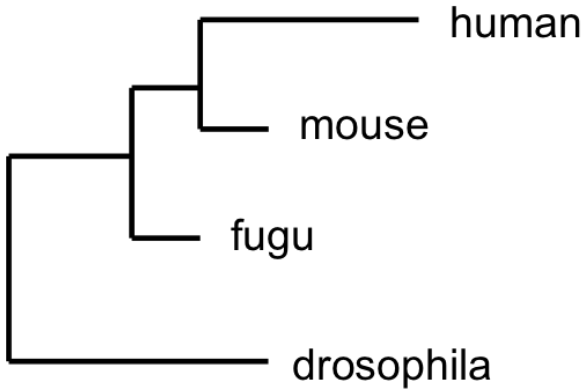
- Focus sur les méthodes de distance
  - UPGMA
  - NJ

- Interprétation
  - Test de la robustesse d'une topologie
  - Problèmes fréquents

Qu'est-ce qu'un arbre phylogénétique ?

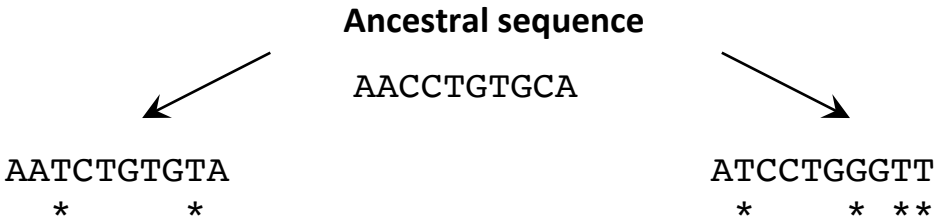
En **phylogénie moléculaire, lorsqu'on analyse une famille de gènes**

- un arbre phylogénétique est **une représentation de l'histoire évolutive de ces gènes** au cours du temps
- Cette représentation est **simplifiée**
- L'histoire des gènes peut être différente de l'histoire évolutive des unités taxonomiques qui portent ces gènes



Le point de départ: les sites d'un alignement multiple

Hypothèse de départ: les gènes dont les séquences sont alignées sont **homologues**, cad hérités verticalement d'un ancêtre commun



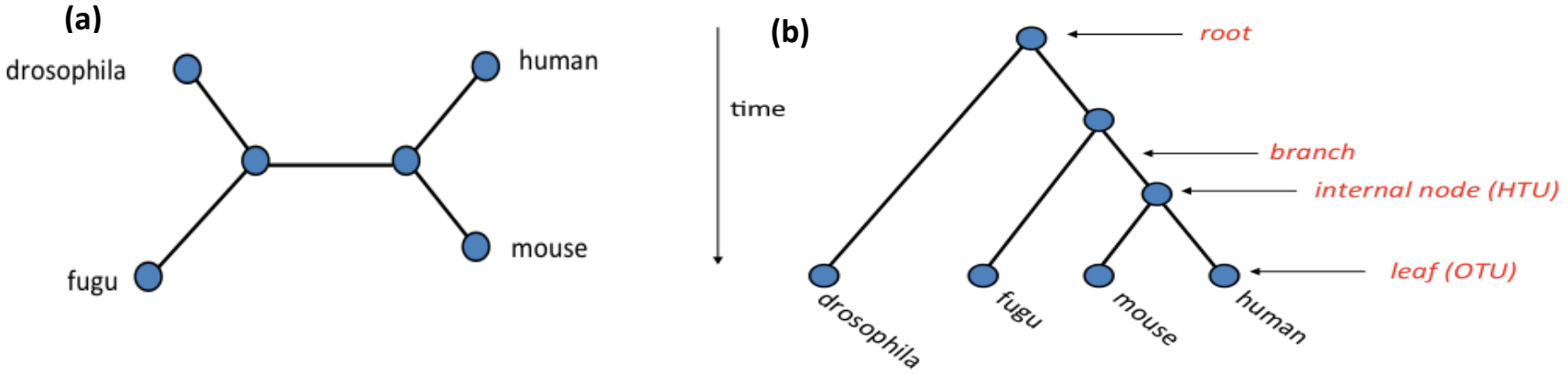

```
Seq1  AATCTGTGTA
Seq2  ATCCTGGGTT
       **   *  *
```

⇦ **1 site = 1 position de l'alignement**

Les méthodes phylogénétiques utilisent **les sites des alignements pour représenter l'histoire évolutive des séquences analysées**

Les arbres phylogénétiques: notions de base

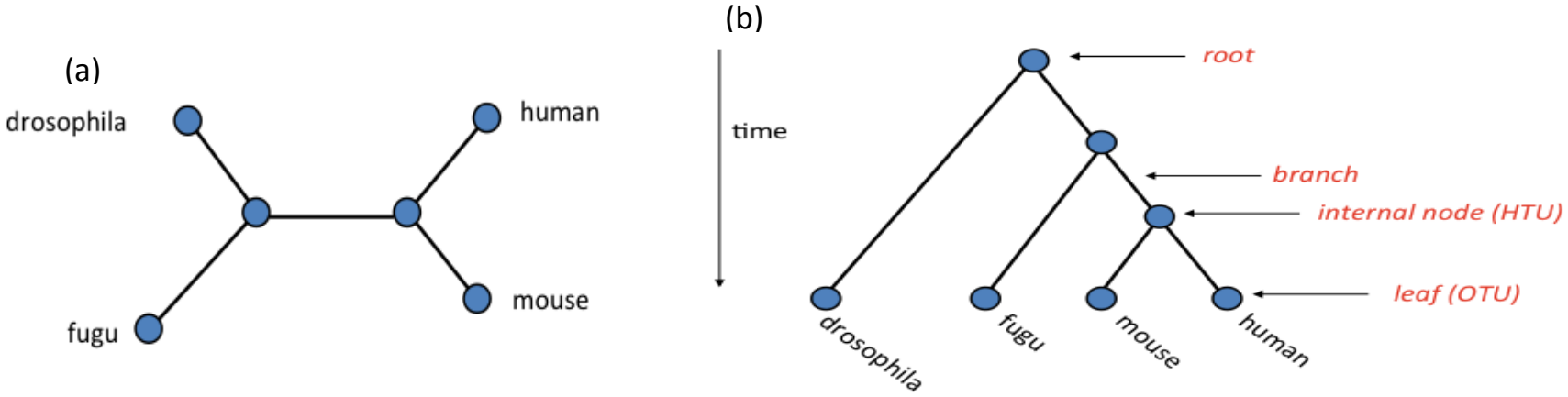
- Un arbre phylogénétique est défini par sa ***topologie*** et ses **l*ongueurs de branche***
- Un arbre phylogénétique peut-être ***non raciné (a)*** ou ***raciné (b)***



- Un arbre contient des ***branches***, des ***noeuds internes*** et des ***feuilles***
- Les termes usuels pour les taxa sont : ***OTU-*** Operational Taxonomic Units et ***HTU***- Hypothetical Taxonomic Units

Les arbres phylogénétiques: notions de base

•Un arbre phylogénétique peut-être ***non raciné (a)*** ou ***raciné (b)***



- La très grande majorité des méthodes phylogénétiques s'intéressent aux arbres **bifurquants** : chaque noeud interne a un degré 3

- La plupart des méthodes phylogénétiques produisent des **arbres non-racinés**

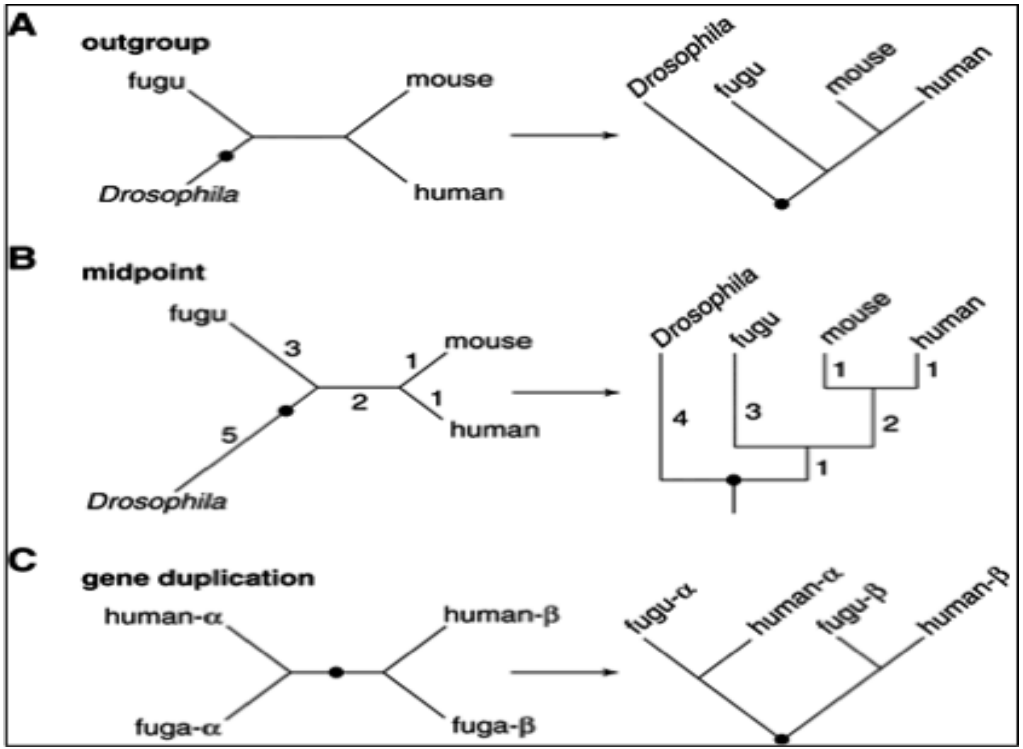
Comment enraciner un arbre ?

- Trois méthodes existent

A. **Outgroup** rooting

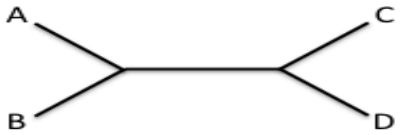
B. **Midpoint** rooting

C. Usage of external knowledge (ex. ancestral **gene duplication**)



Les formats d'arbres phylogénétique

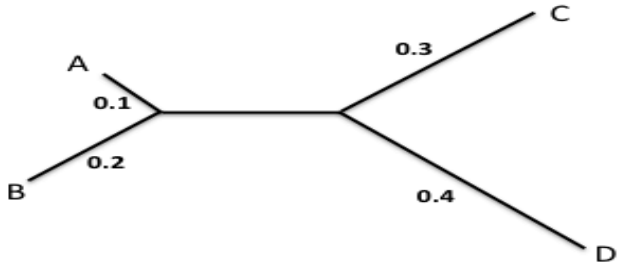
- **Deux formats très utilisés: NEWICK and NEXUS**



NEWICK: ((A,B), (C,D))

```
#NEXUS
BEGIN TAXA;
  TAXLABELS A B C D;
END;

BEGIN TREES;
  TREE tree1 = ((A,B),(C,D));
END;
```



NEWICK: ((A:0.1,B:0.2):0.2, (C:0.3,D:0.4))

```
#NEXUS:
Begin trees;
Translate
1 A,
2 B,
3 C,
4 D,
;
Tree tree2= [&U] ((1:0.1,2:0.2):0.2, (3:0.3,3:0.4));
End;
```

Les distances génétiques (ou évolutives)

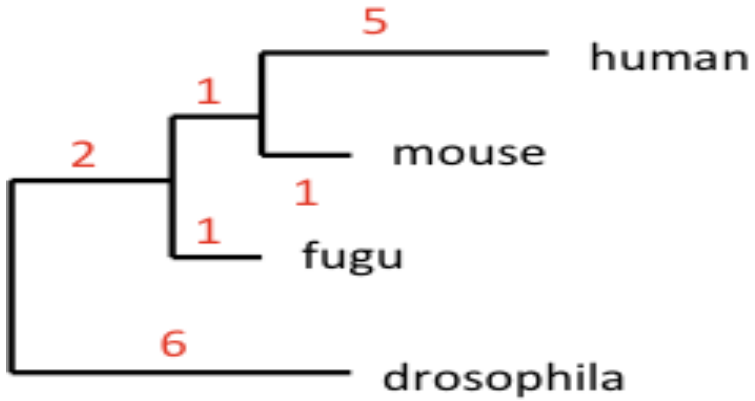
- Une distance génétique (évolutive) est u**ne mesure de la divergence entre deux séquences génétiques**

- Les **modalités de calcul de la distance génétique** entre deux séquences constituent une étape clé d'une analyse phylogénétique
  - C'est la première étape des méthodes basées sur des matrices de distance (UPGMA, NJ)
  - Dans les autres méthodes (maximum de vraisemblance, méthodes bayésiennes) les distances sont utilisées à travers les modèles de substitutions (nucléotides, AA) qui servent à calculer la vraisemblance d'un arbre

Distances et arbres

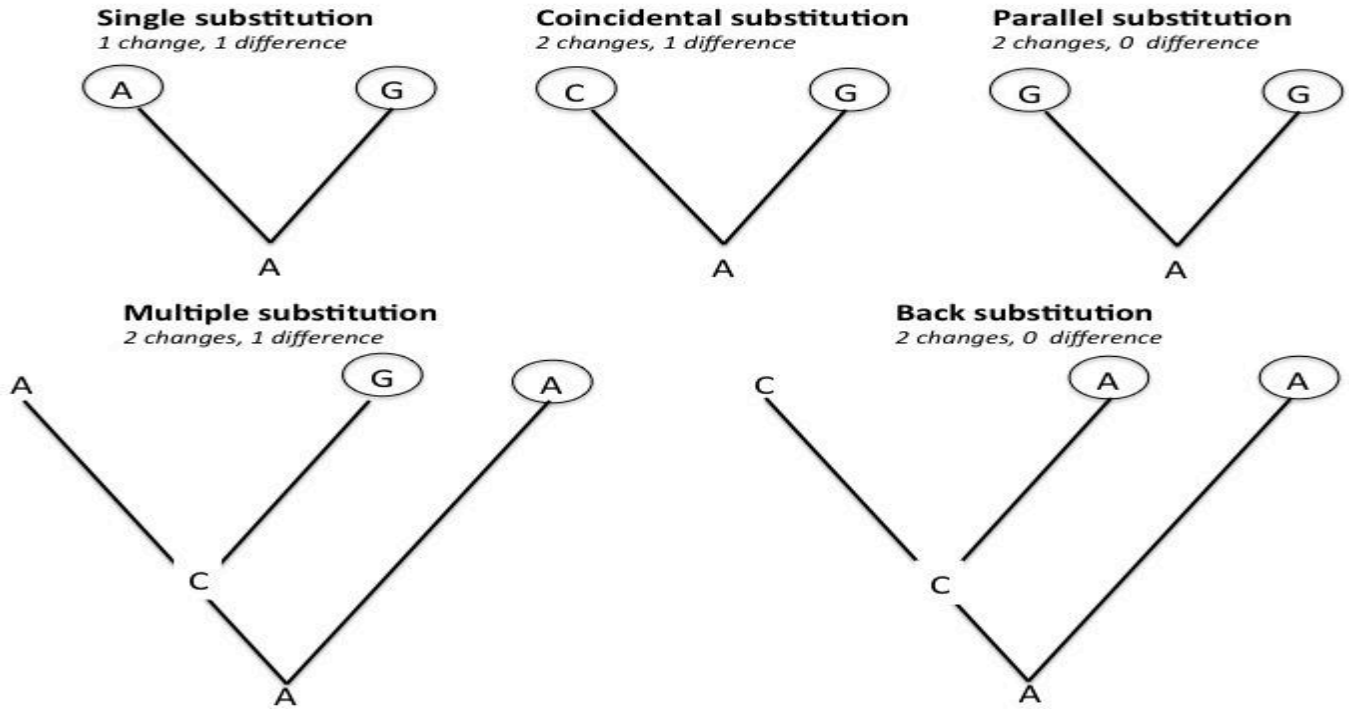
- Pour les séquences liées par un arbre, les **longueurs de branches représentent la distance entre les noeuds (séquences) dans l'arbre**
- Sous l'hypothèse d'horloge moléculaire, la d**istance génétique est linéairement proportionnelle au temps écoulé**

human	x			
mouse	6	x		
fugu	7	3	x	
drosophila	14	10	9	x
	human	*mouse*	*fugu*	*drosophila*



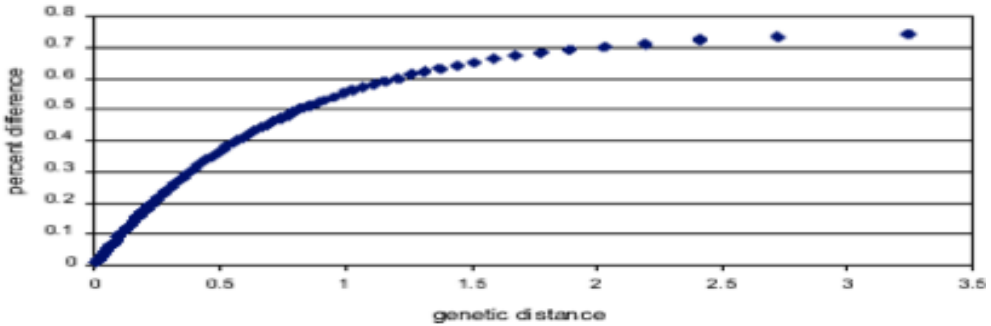
Distances génétiques vs Distances observées

La distance basée sur le nombre de substitutions observées n'est pas toujours informative !



Distances génétiques vs Distances observées

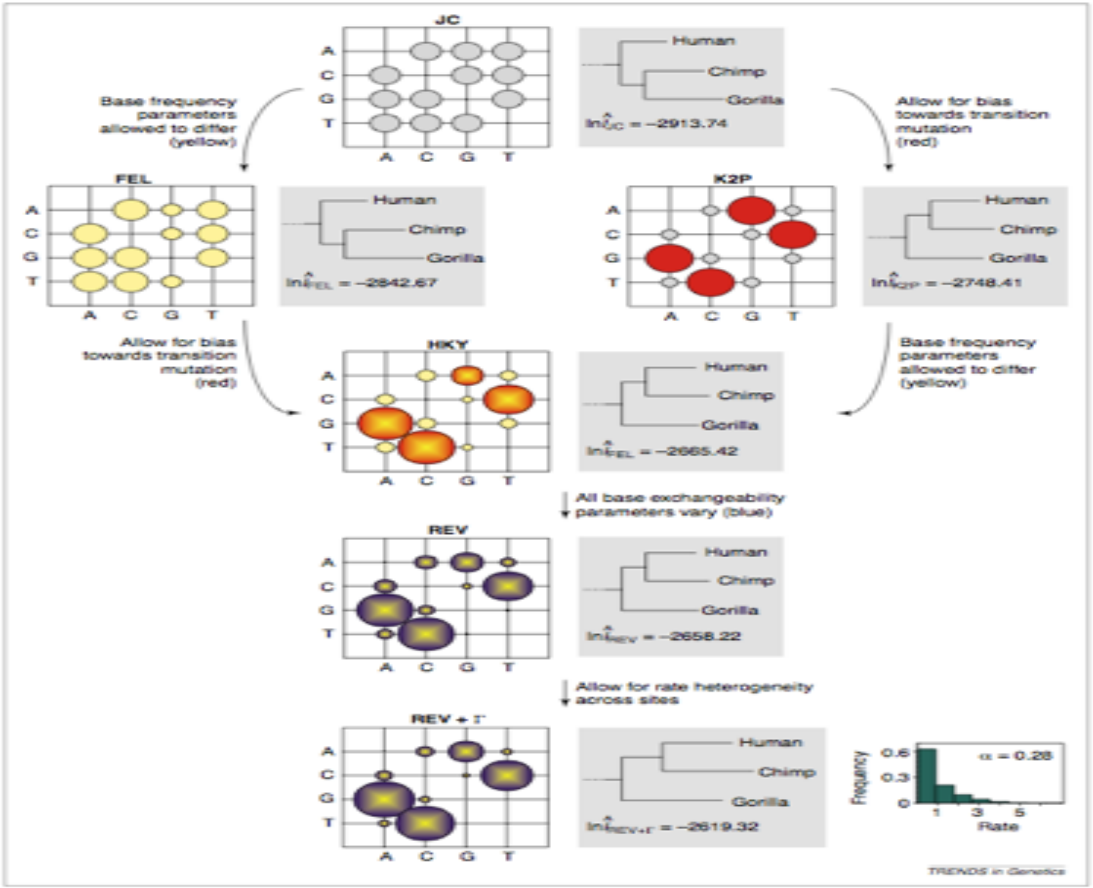
- La distance observée peut être calculée en comptant le nombre de sites où deux séquences diffèrent : elle est exprimée par le **nombre de différences nucléotidiques par site (p distance)**
- La distance observée est une sous-estimation de la distance génétique due aux substitutions multiples par site et à la saturation : des **modèles de substitution** sont utilisés.



Les modèles de substitution nucléotidique

- Une substitution de nucléotide peut être modélisée comme un **événement aléatoire** ;
- Les **processus de Markov en temps continu** sont particulièrement adaptés pour décrire les taux de substitution à chaque site, avec les hypothèses suivantes :
  - Chaque site évolue indépendamment ;
  - La substitution d'un nucléotide i en un nucléotide j est indépendante du nucléotide présent avant i (propriété de Markov) ;
  - Les fréquences des 4 nucléotides sont à l'équilibre (stationnarité).
- Les substitutions à un site donné sont décrites par une chaine de Markov dont les états sont les 4 nucléotides (A,T,C,G) et les changements de nucléotides sont décrits par la matrice des probabilités de transition P(t).

Les modèles de substitution nucléotidique : un aperçu



Choix d'un modèle nucléotidique

- L'utilisation d'un modèle donnée détermine les **longueurs des branches** de l'arbre et peut parfois changer sa topologie
- **Conseils**
  - Aller du modèle le plus simple vers le plus complexe
  - le modèle **GTR (General Time Reversible)** est le plus complexe (9 paramètres)
  - On peut effectuer un test statistique pour choisir le modèle le plus pertinent (LRT, AIC, BIC)

Il est recommandé d'expérimenter !

Les modèles de substitution protéiques

- **Concept similaire : les substitutions multiples d'acides aminés conduisent à une sous-estimation des distances d'évolution entre deux protéines homologues.**

- La fréquence de substitution des acides aminés dépend de l'AA: elle est plus élevée entre les acides aminés proches en terme de propriétés physiques (polarité, hydrophobicité, ...)

- Trop de paramètres (190) pour estimer les paramètres du modèle probabiliste => des **modèles empiriques** sont utilisés

- Le taux de transition entre les acides aminés sont estimés à partir de grands alignements de référence obtenus par concaténation de plusieurs protéines homologues

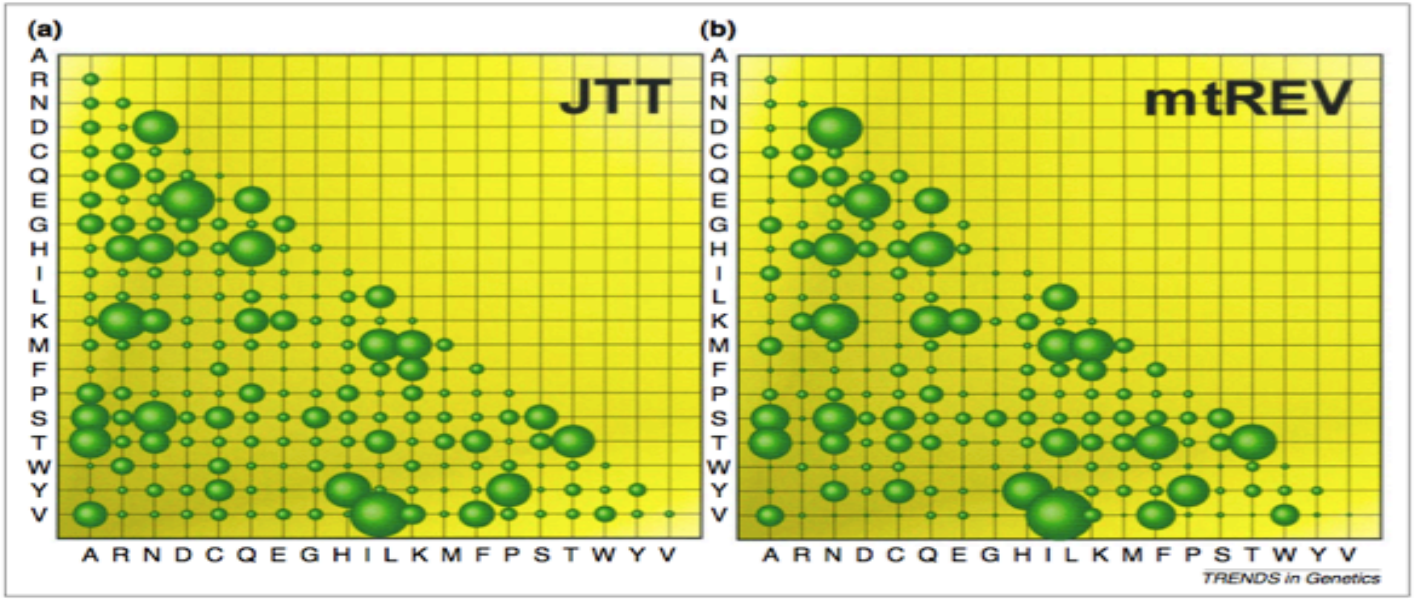
Les principaux modèles d'évolution protéique

Model	Dataset	Ref
Poisson	Poisson process	Zuckerkandl, 1965
PAM	1300 protein sequences from 71 homolog families	Dayhoff 1978
Blosum	Extension of PAM dataset	Henikoff 1992
JTT	16 300 sequences	Jones 1992
mtREV	Mitochondrial DNA	Adachi 1996
WAG & LG	**Likehood methods**	**Whelan 2001**

- Les modèles WAG et LG sont les modèles les plus utilisés
- On peut effectuer un test statistique pour choisir le modèle le plus pertinent (LRT, AIC, BIC)

Les modèles protéiques : exemple

JTT (1992, 16 300 sequences) vs mtREV (for mitochondrial proteins)

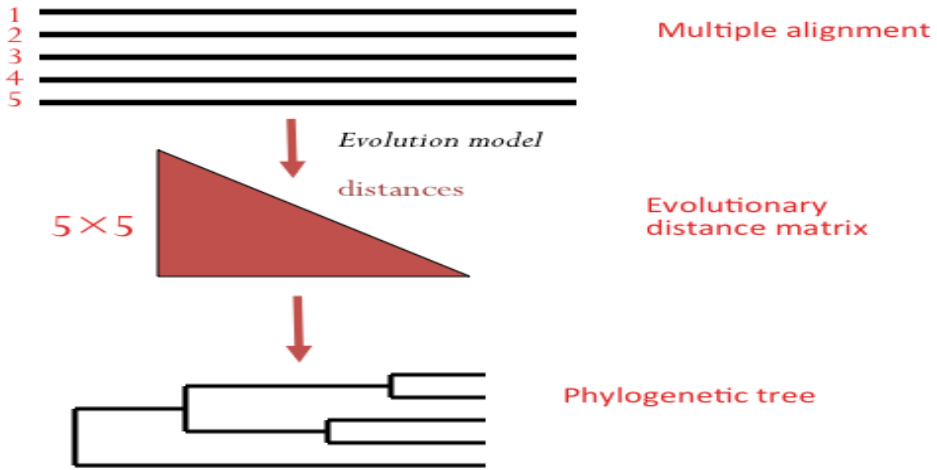


Les méthodes d'inférence phylogénétique

Input data	Method	Principle of the algorithms
Distance matrix	Unweighted Pair Group Method (UPGMA)	clustering
	Neighbor-Joining (NJ)	clustering
Character state	*Maximum Parsimony (MP)*	*Search for the tree(s) of minimum character changes*
	Maximum Likehood (ML)	*Search for the tree(s) that maximizes the probability of observing the character states giving a tree topology and a model of evolution*
	Bayesian Inference	*Target a probability distribution of trees (set of possible trees for the data)*

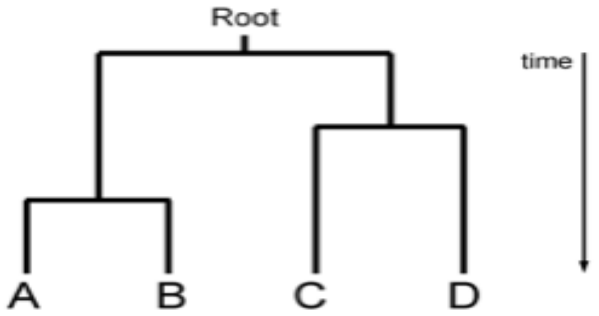
Les méthodes de distances pour inférer un arbre phylogénétique

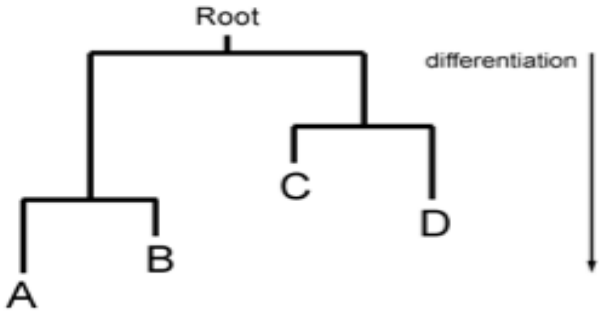
- Introduites en 1960, les plus anciennes !
- Objectif : représenter une matrice de distance génétique par un arbre
- Nécessitent un modèle d'évolution



Les méthodes de distance

- Objectif des méthodes de distance : faire que les distances patristiques générées dans l'arbre représentent au mieux les distances de départ
- Deux méthodes principales
  - **UPGMA**: a clustering method that produced ultrametric trees
  - **Neighbor-Joining**: use a greedy algorithm to compute the Minimal Evolution tree *i.e.* the optimal topology is the one which minimizes the tree length





La méthode UPGMA

Unweighted Pair Group Method with Arithmetic Mean (Sneath & Sokal 1973)

Cette méthode est la plus simple, elle est aussi appelée Classification ascendante hiérarchique

Elle suppose que les taux de substitution entre séquences sont à peu près homogènes dans toutes les lignées (hypothése d'horloge moléculaire)

Elle permet d'estimer des arbres ultramétriques à partir d'une matrice de distance

L'algorithme est itératif : il regroupe les séquences séparées par les distances les plus courtes, par ordre de similarité

UPGMA : principe de l'algorithme

• **Données de départ :** matrice des distances $D_{ij}=\{d_{ij}\}$ entre toutes les paires de séquences i et j.

	A	B	C	D	E
B	2				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8

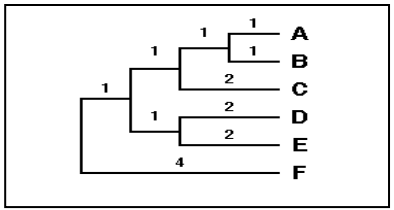
Principe de l'algorithme

Algorithme itératif:

1. Déterminer les deux clusters i et j pour lesquels d_{ij} est minimal.
2. Créer un nouveau cluster $C_k = C_i \cup C_j$ et calculer d_{kl} pour tous les l.
3. Stocker C_i et C_j comme fils droit et gauche dans C_k. Stocker la hauteur de C_k : $d_{ij} / 2$.
4. Ajouter k à la liste des clusters à traiter et supprimer i et j de cette liste.

Terminaison: quand il ne reste plus d'un seul cluster

• Résultat : arbre ultramétrique

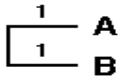


UPGMA : exemple

- Si on a 6 UTOs séparées par une distance donnée (matrice) :

	A	B	C	D	E
B	2				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8

- On commence par unir les 2 UTOs les plus similaires, A et B




```
dist(A,B),C = (distAC + distBC) / 2 = 4
dist(A,B),D = (distAD + distBD) / 2 = 6
dist(A,B),E = (distAE + distBE) / 2 = 6
dist(A,B),F = (distAF + distBF) / 2 = 8
```

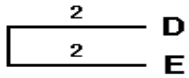
- Par la suite, A et B vont être considérées comme une unité (A,B)

UPGMA : exemple

Les cycles suivants :

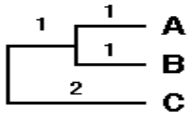
2ème cycle :

	A,B	C	D	E
C	4			
D	6	6		
E	6	6	4	
F	8	8	8	8



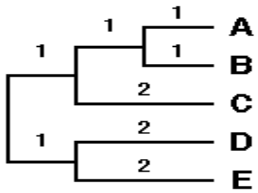
3ème cycle :

	A,B	C	D,E
C	4		
D,E	6	6	
F	8	8	8



4ème cycle :

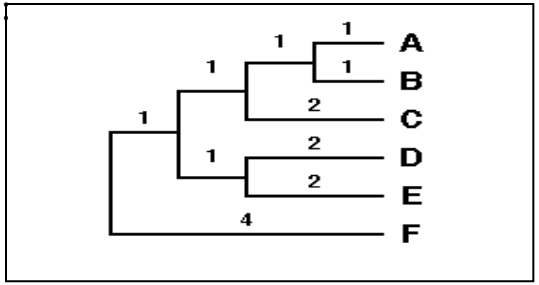
	AB,C	D,E
D,E	6	
F	8	8



5ème cycle :

	ABC,DE
F	8

Fin

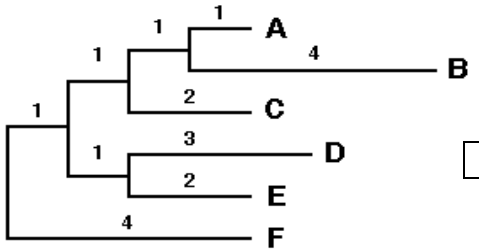


On obtient la topologie correcte

UPGMA : les limites

- La méthode UPGMA produit un arbre raciné
- Elle ne marche que pour les **distances ultramétriques** (principe d'horloge moléculaire stricte) : UPGMA échoue si le taux d'évolution varie entre les UTO

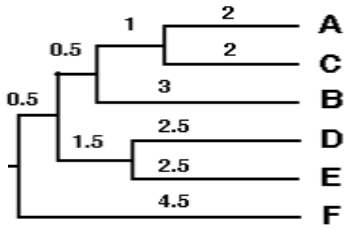
Arbre réel



Matrice de distances

	A	B	C	D	E
B	5				
C	4	7			
D	7	10	7		
E	6	9	6	5	
F	8	11	8	9	8

Arbre UPGMA



Topologie incorrecte

Conclusion : cette méthode est très rapide mais n'est quasiment plus employée en phylogénie moléculaire !

L'algorithme du Neighbor-Joining

- La méthode a été proposée par **Saitou and Nei en 1987**
- Cette méthode utilise un **algorithme itératif**, elle est **très rapide**
- Cette approche produit des **arbres sans racine**
- La méthode est basée sur le **principe d'évolution minimale** :

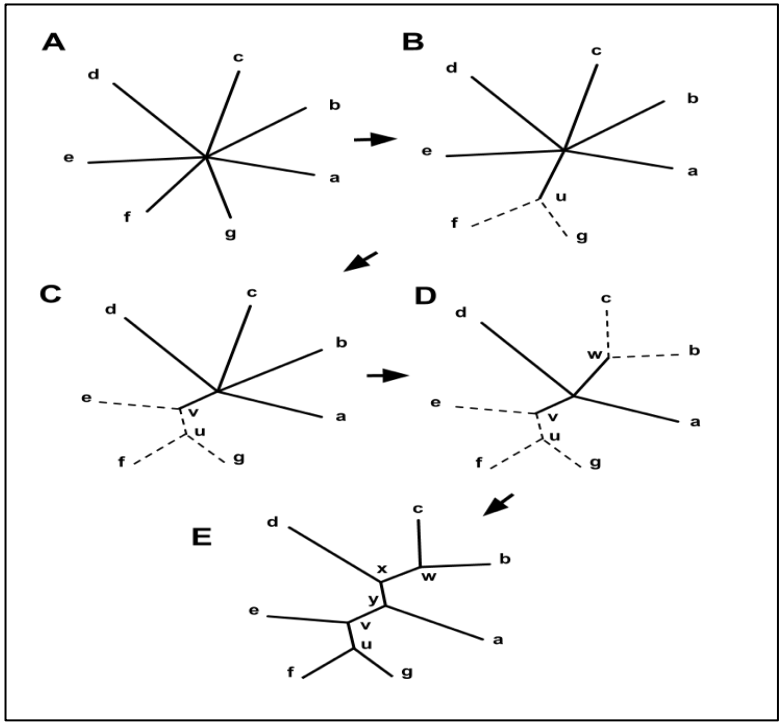
Si S= somme de toutes les branches de l'arbre avec T= nombre de branches et b_i = longueur de branche i, la meilleure topologie est celle qui correspond à $S=\sum b_i$ minimum

- Contrairement à UPGMA, cette méthode **construit des arbres non ultramétriques**, traduisant une possible hétérogénéité dans le nombre de mutations accumulées

Le Neighbor-Joining (NJ)

Principe de l'algorithme :

- Start with **a star tree** (A)
- Compute the **matrix Qij*** and find the pair of taxa with lowest value (here f and g)
- Join f and g and **create a new internal node, u**, as shown in (B)
- Compute the distances from node u to the nodes a-e
- **Repeat the process** : u and e are joined to the newly created v, as shown in (C).
- Two more iterations lead first to (D), and then to (E).



*Qij : matrice des distances corrigées

Le Neighbor-Joining (NJ) : bilan

Avantages :

- L'algorithme est constructif : si les distances de départ sont patristiques, l'arbre obtenu est correct
- Méthode très rapide et performante
- Mesure globale – l'arbre de longueur totale minimale

Inconvénients :

- Donne des mauvais résultats en cas de fort taux de substitution ou de variations de taux entre les branches
- Pas recommandé pour des séquences très divergentes
- Problème d'attraction des branches longues

Le Neighbor-Joining en pratique

Conseils :

- **Tester plusieurs modèles d'évolution** pour construire la matrice de distances (en particulier pour les séquences assez divergentes)
- Lorsque les taux de substitutions sont élevés ou variables entre les UTO, utiliser **l'algorithme BioNJ***, une variante de l'algorithme du NJ qui donne une meilleure précision de topologie

Logiciels :

- **MAFFT** (UPGMA, NJ) **:** http://mafft.cbrc.jp/alignment/server/
- **NJ and BioNJ** http://www.atgc-montpellier.fr/fastme/ or http://www.phylogeny.fr/one_task.cgi?task_type=bionj
- **Seaview** (NJ and BioNJ, standalone program) http://doua.prabi.fr/software/seaview

**Gascuel Molecular Biology and Evolution 1997*

Comment évaluer un arbre phylogénétique ?

Problème de confiance

- Quelle confiance avoir en l'arbre inféré?
- Quelles parties de l'arbre sont fiables / non fiables?
- Comment pouvons-nous valider l'arbre ?

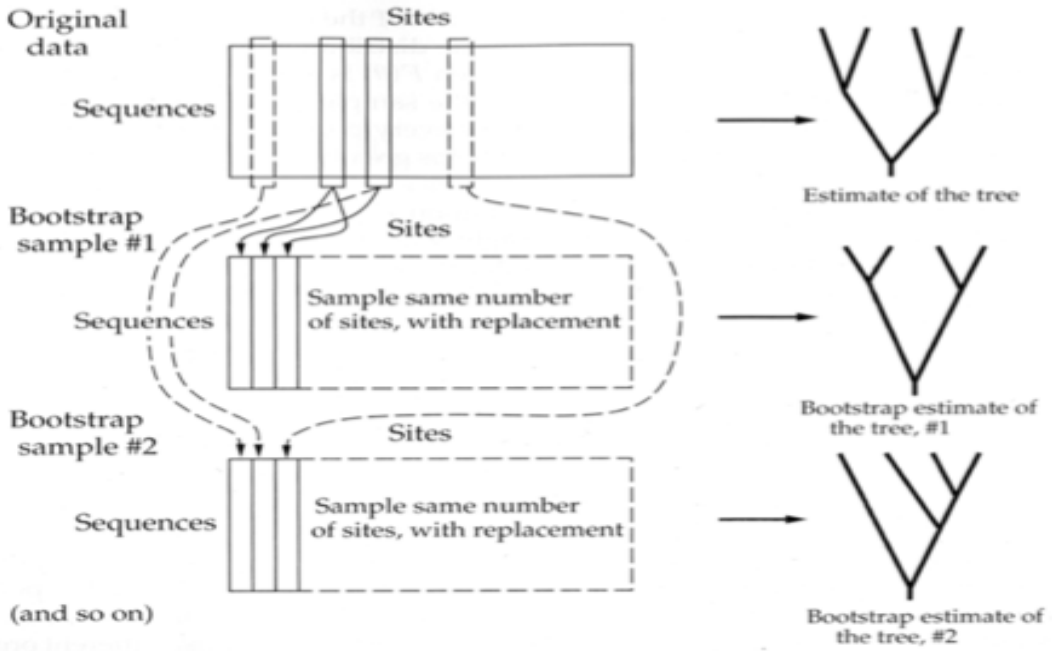
Problème: le vrai arbre est inconnu !

Solution :

- Utilisez le bootstrap pour évaluer la fiabilité de l'arbre inféré et des clades
- Combiner les sous-échantillons et les arbres de consensus pour obtenir des valeurs de soutien sur les branches

Évaluer la topologie d'un arbre

Principe du Bootstrap: on ré-échantillonne les positions d'un alignement



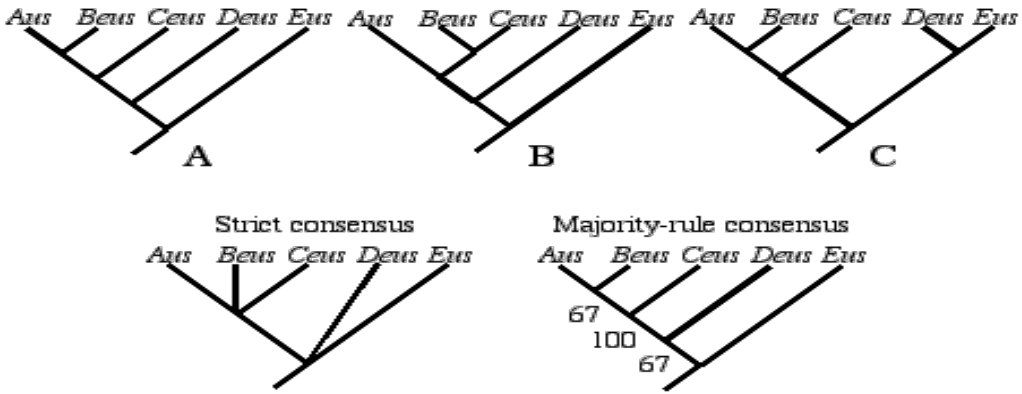
Principe du bootstrap et de l'arbre consensus

- Inférer plusieurs (~100) arbres en utilisant des techniques de **rééchantillonnage**;
- Identifier et conserver uniquement les clades présents dans de nombreux arbres ;
- Combinez les différents arbres pour produire un **arbre de consensus** qui est compatible avec la plupart les arbres .
- En général, l'arbre de consensus n'a pas de longueur de branche et une résolution inférieure à celle de l'arbre d'origine.
- Superposer les valeurs de bootstrap sur l'arbre d'origine

Comment construire un arbre consensus ?

Règles de consensus :

- **Strict Consensus**: clades présents dans tous les arbres;
- **Règle de majorité:** les clades sont présents dans au moins la moitié des arbres;
- **Règle de majorité étendue:** clades sont présents dans au moins la moitié des arbres et un peu plus jusqu'à ce que l'arbre soit résolu.



Le bootstrap en pratique

Attention à l'interprétation des valeurs de bootstrap :

- Les valeurs de bootstrap n'ont pas d'interprétation statistique claire;
- Une valeur bootstrap de 95% ne signifie pas que le clade correspondant a 95% de chance d'être "vrai";
- Les valeurs de bootstrap sont difficiles à interpréter quantitativement.

Cependant, les valeurs de Bootstrap sont (assez) faciles à interpréter **qualitativement**:

- **Plus la valeur bootstrap est élevée, plus vous pouvez être confiant dans votre clad**e;
- 95%, 90% et 66% constituent le seuil traditionnel pour être confiants dans un clade.

Quelques derniers conseils

- **Attraction des longues branches:** Les longues branches ont tendance à se regrouper dans l'arbre
- Solution : "décomposer" de longues branches en ajoutant quelques UTO à l'analyse;

- **Saturation:** Les caractères ont évolué depuis si longtemps qu'ils sont presque aléatoires
- Solution : Supprimer les sites saturés et / ou taxons; Lorsqu'elles sont disponibles, utiliser des séquences protéiques au lieu des séquences nucléiques;

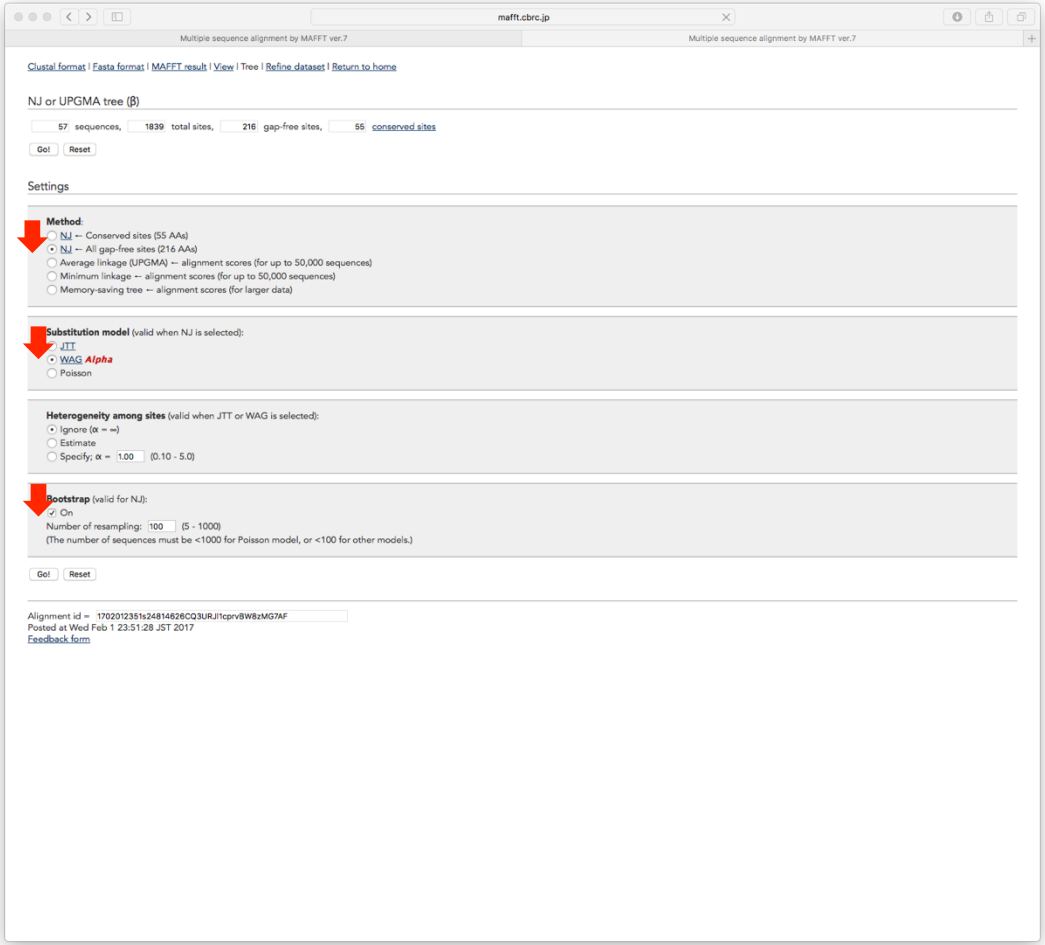
- **Filtrage des alignements** : pour les phylogénies basées sur un seul gène / protéine, le filtrage des régions avec “gaps” n'améliore pas la reconstruction des arbres (ref)
- Solution : Ne pas filtrer les alignements single gene/protein !

Travaux pratiques

http://mafft.cbrc.jp/alignment/server/

- **Exercice 1 : Construire un arbre UPGMA et un arbre de Neighbor-Joining à partir de l'alignement des 57 séquence de la famille GH131**

Commenter les différences

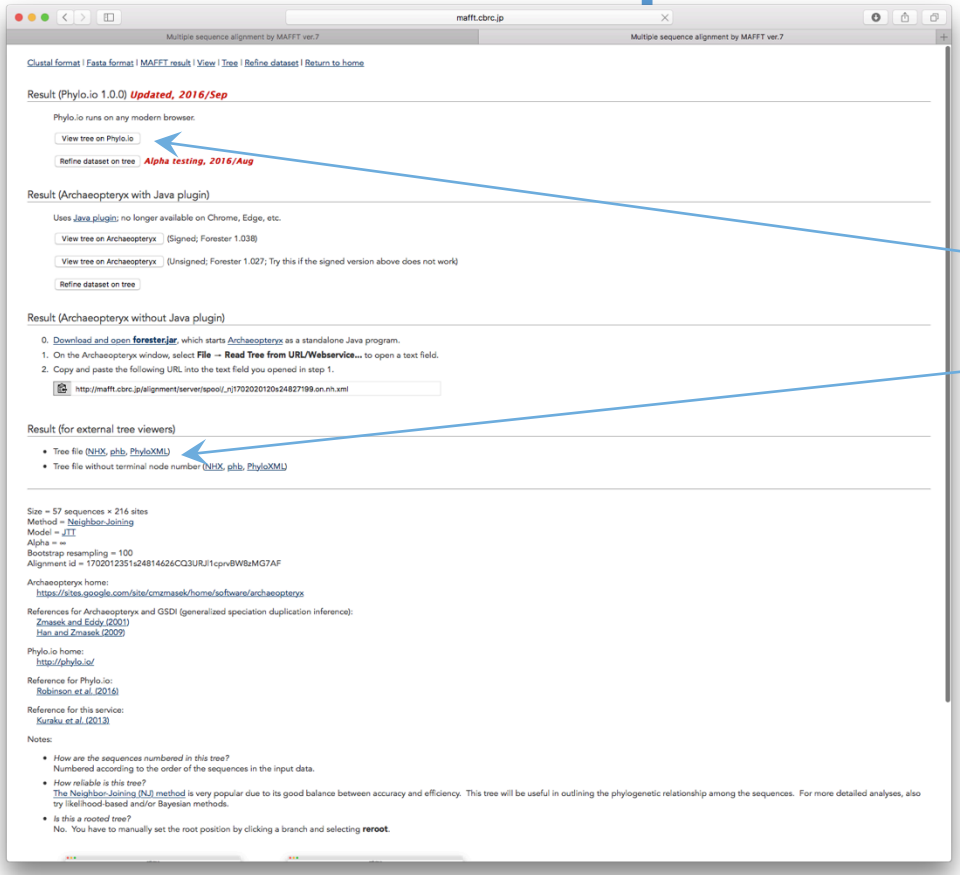


Pour le NJ utiliser les paramètres suivant :

- All gap-free sites
- **Substitution model : WAG**
- Heterogenity among sites : Ignore (default)
- Bootstrap : yes

Visualiser l'arbre et interpréter

Arbre des 57 séquences

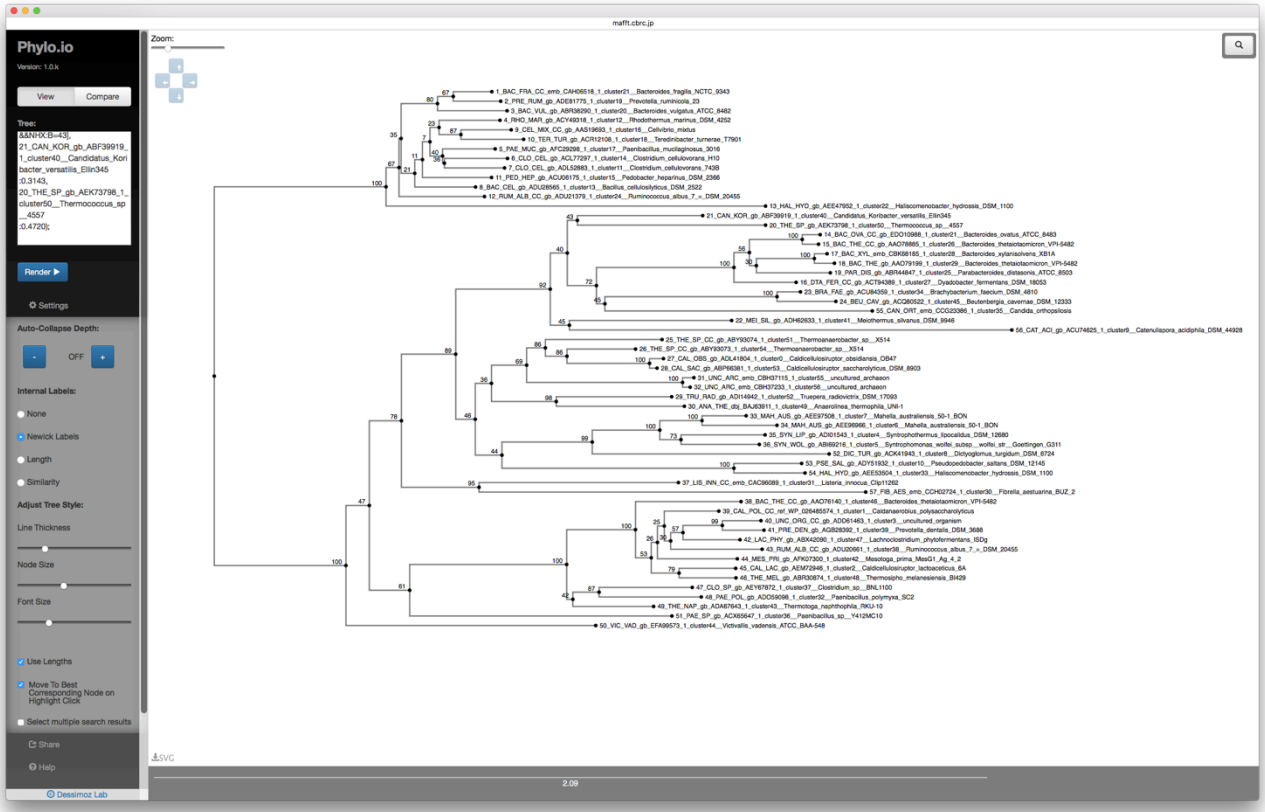


Explore the tree using Phylo.io

Select “view tree on Phylo.io”

Export/save tree in a standart format

Arbre des 57 séquences



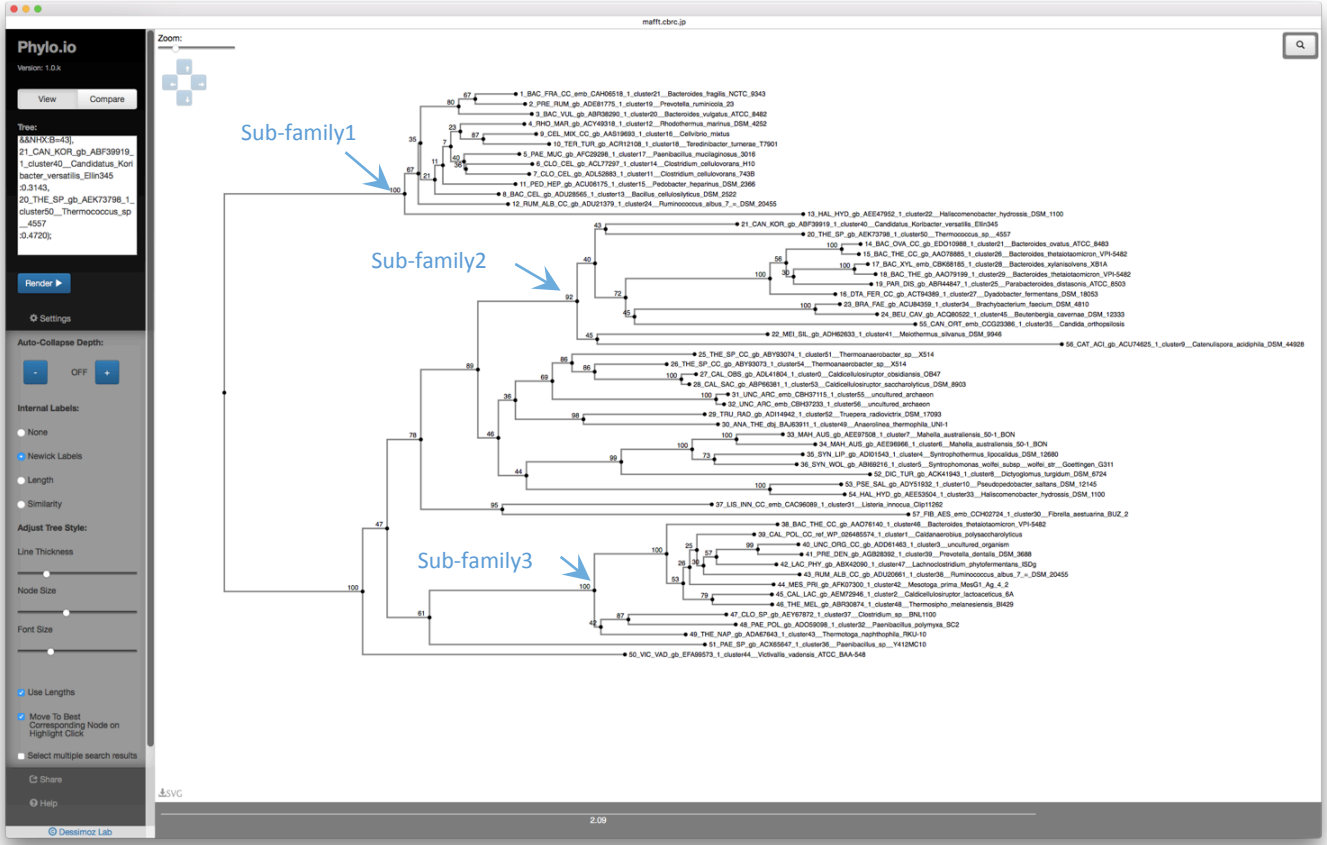
Exercise 2: select 3 sub-families meeting the 3 following criteria:

1. Bootstrap clade values ≥ 90
2. At least 12 sequences
3. Large enough taxonomic diversity (at the **class level)**

DOMAIN
Kingdom
Phylum
Class
Order
Family
Genus
species
subspecies

Mnemonic (memory aid)
D _____
K _____
P _____
C _____
O _____
F _____
G _____
S _____

Arbre des 57 séquences



Subfamilies definition

- Sub_family 1 : sequence 1 to 13
- Sub_family 2 : sequence 14 to 24 + 55 + 56
- Sub_family 3 : sequence 38 to 49

For taxonomic diversity, use https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi with species name

Définition des sous-familles

Sub-Family 1 : 13 seqs

>BAC_FRA_CC_emb_CAH06518.1 cluster21 [Bacteroides fragilis NCTC 9343] Phylum : Bacteroidetes Class : Bacteroidia;

>PRE_RUM_gb_ADE81775.1 cluster19 [Prevotella ruminicola 23] Phylum : Bacteroidetes Class : Bacteroidia;

>BAC_VUL_gb_ABR38290.1 cluster20 [Bacteroides vulgatus ATCC 8482] Phylum : Bacteroidetes Class : Bacteroidia;

>RHO_MAR_gb_ACY49318.1 cluster12 [Rhodothermus marinus DSM 4252] Phylum : Bacteroidetes Class : ?

>PAE_MUC_gb_AFC29298.1 cluster17 [Paenibacillus mucilaginosus 3016] Phylum : Bacilli; Class : Bacillales

>CLO_CEL_gb_ACL77297.1 cluster14 [Clostridium cellulovorans H10] Phylum : Firmicutes; Class : Clostridia

>CLO_CEL_gb_ADL52883.1 cluster11 [Clostridium cellulovorans 743B] Phylum : Firmicutes; Class : Bacilli

>BAC_CEL_gb_ADU28565.1 cluster13 [Bacillus cellulosilyticus DSM 2522] Phylum : Firmicutes; Class : Bacilli

>CEL_MIX_CC_gb_AAS19693.1 cluster16 [Cellvibrio mixtus] Phylum : Proteobacteria; Class : Gammaproteobacteria;

>TER_TUR_gb_ACR12108.1 cluster18 [Teredinibacter turnerae T7901] Phylum : Proteobacteria; Class : Gammaproteobacteria;

>PED_HEP_gb_ACU06175.1 cluster15 [Pedobacter heparinus DSM 2366] Phylum : Bacteroidetes; Class

Définition des sous-familles

Sub-Family 2 : 13 seqs

>BAC_OVA_CC_gb_EDO10988.1 cluster21 [Bacteroides ovatus ATCC 8483]
>BAC_THE_CC_gb_AAO78885.1 cluster26 [Bacteroides thetaiotaomicron VPI-5482]
>DTA_FER_CC_gb_ACT94389.1 cluster27 [Dyadobacter fermentans DSM 18053]
>BAC_XYL_emb_CBK68185.1 cluster28 [Bacteroides xylanisolvens XB1A]
>BAC_THE_gb_AAO79199.1 cluster29 [Bacteroides thetaiotaomicron VPI-5482]
>PAR_DIS_gb_ABR44847.1 cluster25 [Parabacteroides distasonis ATCC 8503]
>THE_SP_gb_AEK73798.1 cluster50 [Thermococcus sp. 4557]
>CAN_KOR_gb_ABF39919.1 cluster40 [Candidatus Koribacter versatilis Ellin345]
>MEI_SIL_gb_ADH62633.1 cluster41 [Meiothermus silvanus DSM 9946]
>BRA_FAE_gb_ACU84359.1 cluster34 [Brachybacterium faecium DSM 4810]
>BEU_CAV_gb_ACQ80522.1 cluster45 [Beutenbergia cavernae DSM 12333]
>CAN_ORT_emb_CCG23386.1 cluster35 [Candida orthopsilosis]
>CAT_ACI_gb_ACU74625.1 cluster9 [Catenulispora acidiphila DSM 44928]

Combien de phyla ? Combien de classes ?

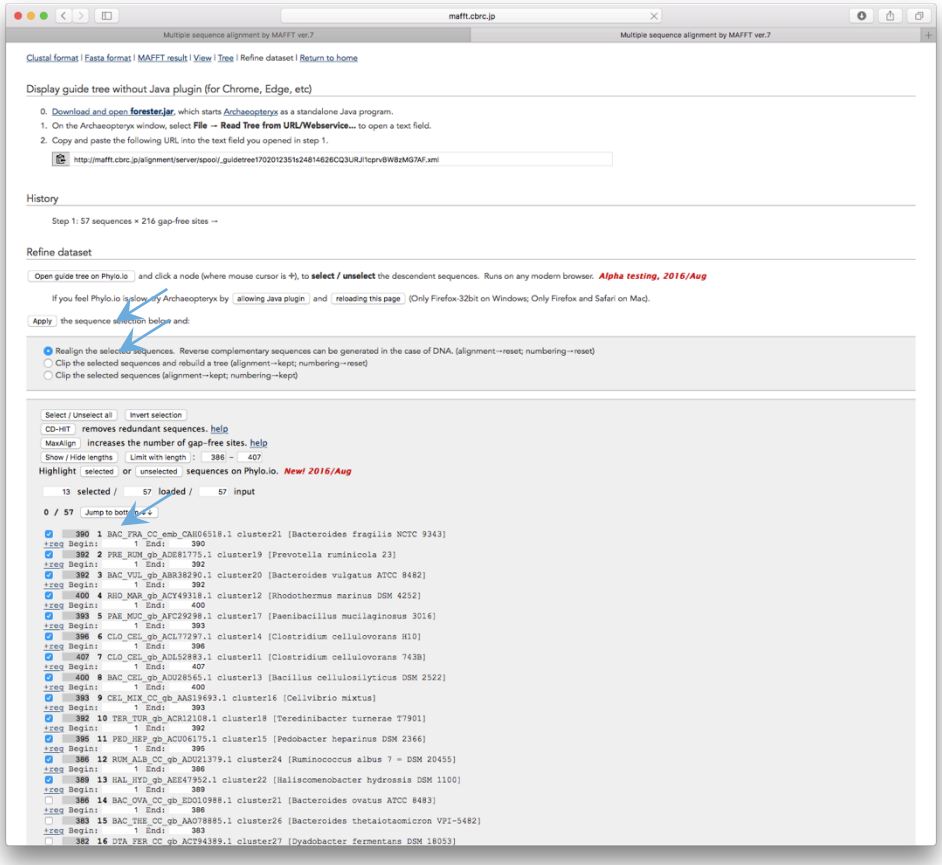
Définition des sous-familles

Sub-Family 3 : 12 seqs

>BAC_THE_CC_gb_AAO76140.1 cluster46 [Bacteroides thetaiotaomicron VPI-5482]
>CAL_POL_CC_ref_WP_026485574.1 cluster1 [Caldanaerobius polysaccharolyticus]
>RUM_ALB_CC_gb_ADU20661.1 cluster38 [Ruminococcus albus 7 = DSM 20455]
>UNC_ORG_CC_gb_ADD61463.1 cluster3 [uncultured organism]
>PRE_DEN_gb_AGB28392.1 cluster39 [Prevotella dentalis DSM 3688]
>LAC_PHY_gb_ABX42090.1 cluster47 [Lachnoclostridium phytofermentans ISDg]
>MES_PRI_gb_AFK07300.1 cluster42 [Mesotoga prima MesG1.Ag.4.2]
>CAL_LAC_gb_AEM72946.1 cluster2 [Caldicellulosiruptor lactoaceticus 6A]
>THE_MEL_gb_ABR30874.1 cluster48 [Thermosipho melanesiensis BI429]
>CLO_SP_gb_AEY67872.1 cluster37 [Clostridium sp. BNL1100]
>PAE_POL_gb_ADO59098.1 cluster32 [Paenibacillus polymyxa SC2]
>THE_NAP_gb_ADA67643.1 cluster43 [Thermotoga naphthophila RKU-10]

Combien de phyla ? Combien de classes ?

Alignement des sous-familles



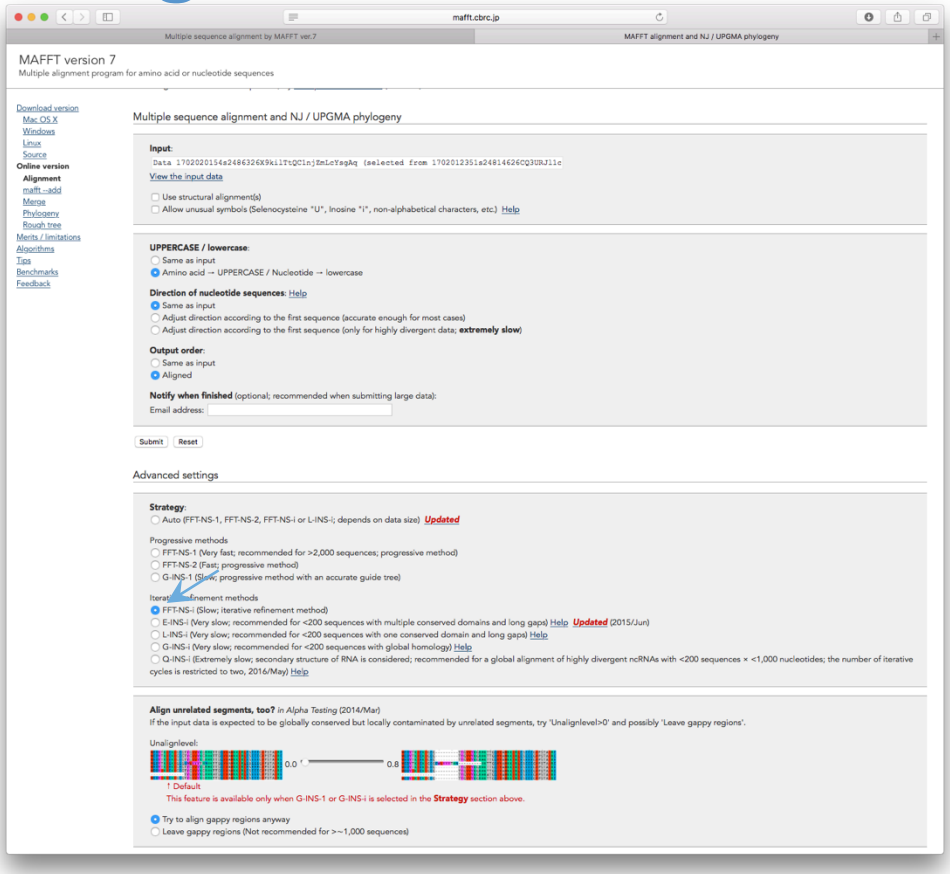
- **Exercise 3:** Refine the dataset to build three sub-alignments corresponding to the three subfamilies

  Subfamilies definition

  - Example : Sub_family 1

  - Select sequences
  - Select ‘Realign the selected sequences’

Alignement des sous-familles

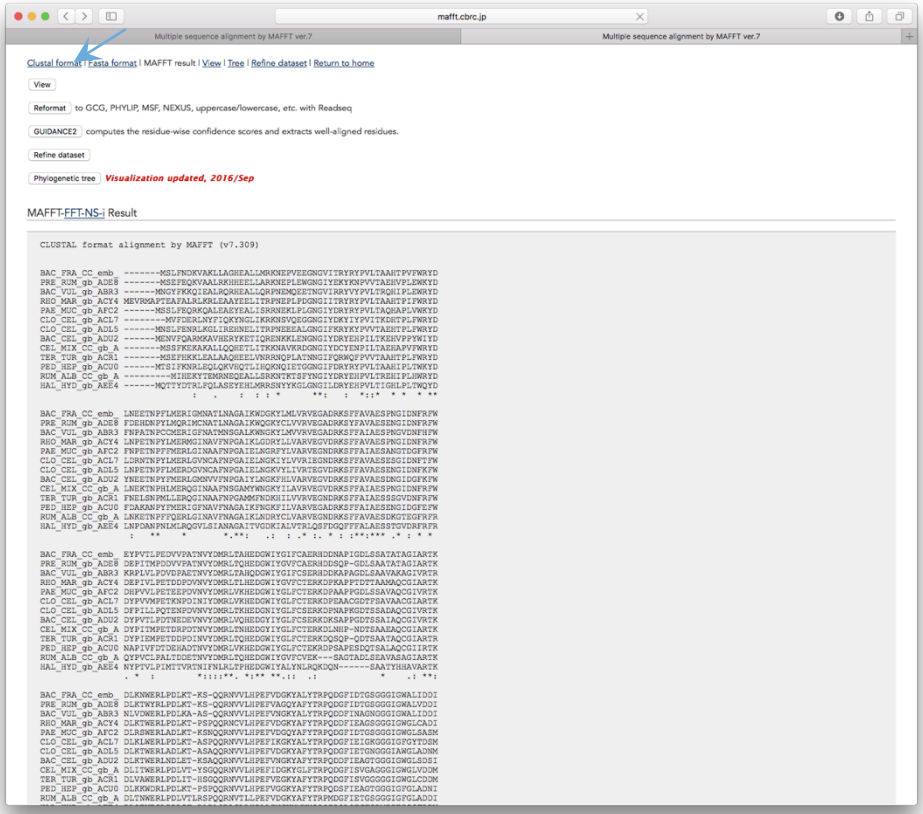


Subfamily 1 alignment

- Example : Sub_family 1

Strategy :
Select “FFT-NS-I iterative refinement method “

Alignement des sous-familles



Subfamily 1 alignment

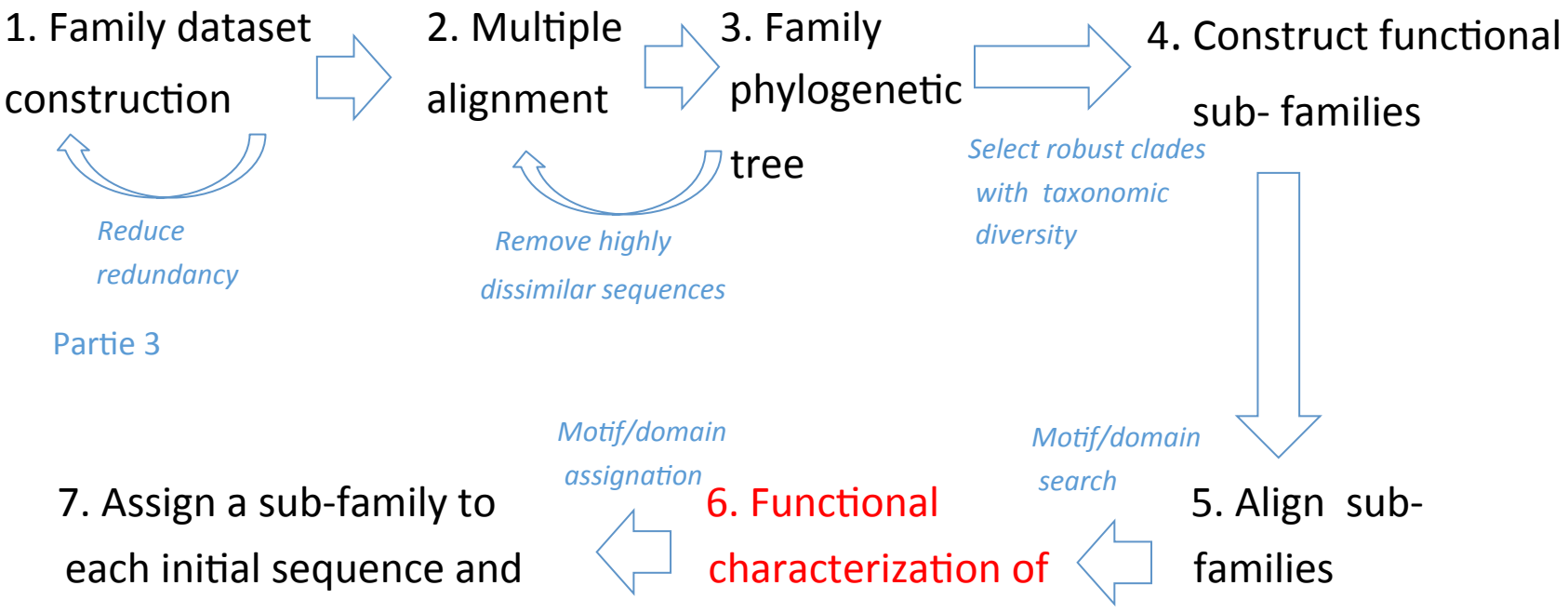
- Example : Sub_family 1 : sequence 1 to 13

Export alignment of subfamily 1 in clustal format

Same for subfamilies 2 and 3!

Motifs dans les séquences

Et une démarche bioinfo générique



Adapted from Mewis et al. App. And Env. Microb. 2016

Partie 3 - la stratégie utilisée:

Material and Methods - Mewis et al. App. And Env. Microb. 2016

- *Hidden Markov models (HMMs) were created for each subfamily, as well as for the complete GH43 subfamily, by using HMMer3* Hmmer2 sans ligne de commande
- *All GH43 sequences were compared to these HMMs by use of HMMer3 to assign a subfamily to each sequence. Each sequence was compared to all FIG 1 Phylogenetic tree of the GH43 family, showing the 37 subfamilies as colored branches.* Hmmer2 sans ligne de commande

Ajout de contenu :

- détecter *de novo* des motifs et les rechercher dans d'autres séquences MEME, hmmer ...
- Annoter une séquence avec des bases de données de motifs et de domaines connus Interpro

Recherche de profils de novo : MEME

Travaux pratiques

http://meme-suite.org/tools/meme

http://alternate.meme-suite.org/

•**Exercice 1 : Rechercher des motifs conservés dans le jeu de données** http://genoweb.toulouse.inra.fr/~formation/AlignClusteringLISBP/data/GH130InputRenamed.fasta.

*Paramètres : laisser les paramètres par défauts sauf monter à **13** le nombre de motifs recherché.*

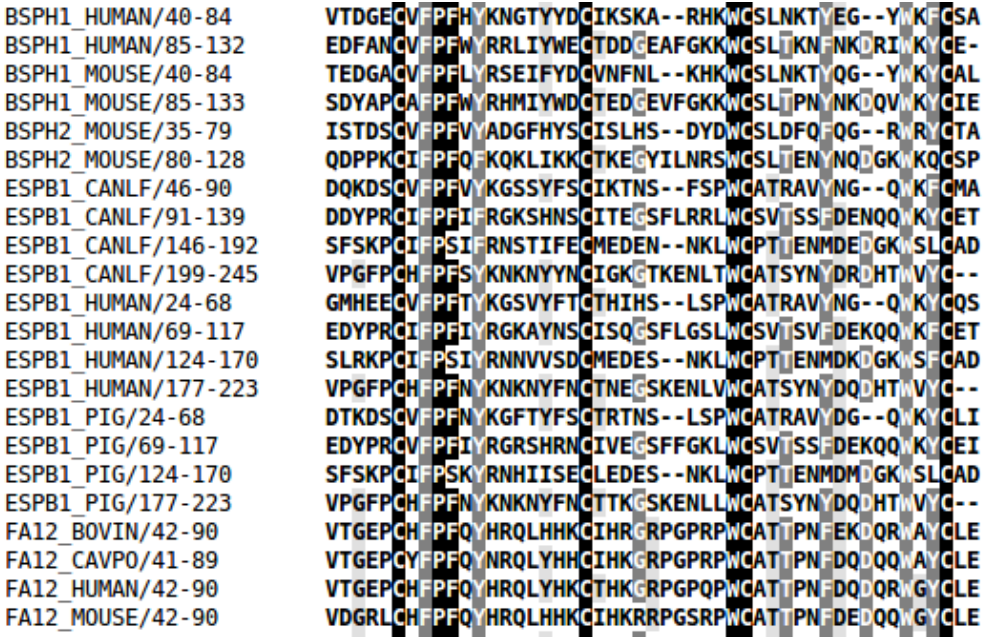
Cela va être un peu long...

A quoi ça sert ?

- **rechercher des éléments fonctionnels**
- rechercher des sites de restriction
- **déterminer la signature d'une famille ou d'une sous-famille protéique pour caractériser ses membres**

A quoi ça sert ?

- Exemples : la TATA box dans les séquences promotrices des gènes.
- Chez les protéines les motifs sont constitués de résidus conservés pas nécessairement consécutifs. Ex dans prosite : FN2_2, PS51092; Fibronectin type-II collagen-binding domain profile



Comment les décrit-on ?

Exemple d'un motif dans des séquences de maltose binding proteins :

```
Yvfk_Bs        PTPNIPEMNEIW
Yvfk_Bs        PTPNIPEMAEVW
MalX_Sp        PLPNISQMSAVW
MalE_Sc        PRPALPEYSSLW
MalE_Tm        PMPNVPEMAPVW
CymE_Ko        AMPSIPEMGYLW
MalE_Ea        IMPNIPQMSAFW
MalE_Sy        IMPNIPQMSAFW
MalE_Ec        IMPNIPQMSAFW
consens : PMPNIPEMSAXW
PROSITE:[PAI]-[TLRM]-P-[NAS]-[ILV]-[PS]-[EQ]-[MY]-[NASG]...
```

Expression régulière ou patterns (format Prosite)
Syntaxe:
- : séparation des éléments
x : n'importe quel acide aminé
(i,j) : nombre d'occurrences entre i et j avec i<j
[NHG] : alternative entre N H et G pour un même site

Comment les décrit-on ?

Exemple d'un motif dans des séquences de maltose binding proteins :

```
Yvfk_Bs        PTPNIPEMNEIW
Yvfk_Bs        PTPNIPEMAEVW
MalX_Sp        PLPNISQMSAVW
MalE_Sc        PRPALPEYSSLW
MalE_Tm        PMPNVPEMAPVW
CymE_Ko        AMPSIPEMGYLW
MalE_Ea        IMPNIPQMSAFW
MalE_Sy        IMPNIPQMSAFW
MalE_Ec        IMPNIPQMSAFW
consens : PMPNIPEMSAXW
PROSITE:[PAI]-[TLRM]-P-[NAS]-[ILV]-[PS]-[EQ]-[MY]-[NASG]...
```

Convient bien pour des motifs très conservés

Pour tenir compte de la répartition des différents acides aminés à chaque position : profils ou matrices de fréquences

Comment les décrit-on ?

Les matrices de fréquences ou de probabilité ou profils :

AATAGTCGC

GGTAGTCTA

ATTAGTCGA

GCTAGTCGG

position frequency matrix (PFM)
position probability matrix (PPM)

le profil correspondant :

	1	2	3	4	5	6	7	8	9
A:	0.5	0.25	0	1	0	0	0	0	0.5
T:	0	0.25	1	0	0	1	0	0.25	0
G:	0.5	0.25	0	0	1	0	0	0.75	0.25
C:	0	0.25	0	0	0	0	1	0	0.25

Comment les décrit-on ?

Les matrices de fréquences ou de probabilités ou profils :

	1	2	3	4	5	6	7	8	9
A:	0.5	0.25	0	1	0	0	0	0	0.5
T:	0	0.25	1	0	0	1	0	0.25	0
G:	0.5	0.25	0	0	1	0	0	0.75	0.25
C:	0	0.25	0	0	0	0	1	0	0.25

Ces matrices de fréquences peuvent être modifiées pour tenir compte de la fréquence de chaque résidus : elles deviennent ainsi des matrices de poids.

De plus une transformation (log2((f(x)+ε)/f(b))) est très souvent utilisée pour enlever les 0 dans la matrice et donc autoriser de nouvelles séquences qui peuvent correspondre au profil.

Comment les décrit-on ?

On obtient alors des pseudo-comptages.

Ces matrices sont appelées PWM : position weight matrix, PSWM : position-specific weight matrix ou PSSM : position-specific scoring matrix.

Le score d'un mot dans le modèle est la somme des poids puisqu'on est en log.

Comment les décrit-on ?

PWM : position weight matrix
PSWM : position-specific weight matrix
PSSM : position-specific scoring matrix

Les matrices de poids :

le profil correspondant :

	1	2	3	4	5	6	7	8	9
A:	0.5	0.25	0	1	0	0	0	0	0.5
T:	0	0.25	1	0	0	1	0	0.25	0
G:	0.5	0.25	0	0	1	0	0	0.75	0.25
C:	0	0.25	0	0	0	0	1	0	0.25

log2((f(x) + 0.05)/0.25))

	1	2	3	4	5	6	7	8	9
A:	1.14	0.26	-2.32	2.07	-2.32	-2.32	-2.32	-2.32	1.14
T:	-2.32	0.26	2.07	-2.32	-2.32	2.07	-2.32	0.26	-2.32
G:	1.14	0.26	-2.32	-2.32	2.07	-2.32	-2.32	1.68	0.26
C:	-2.32	0.26	-2.32	-2.32	-2.32	-2.32	2.07	-2.32	0.26

Score d'un mot dans ma séquence ?

Calcul d'un score d'une séquence :

	1	2	3	4	5	6	7	8	9
A:	1.14	0.26	-2.32	2.07	-2.32	-2.32	-2.32	-2.32	1.14
T:	-2.32	0.26	2.07	-2.32	-2.32	2.07	-2.32	0.26	-2.32
G:	1.14	0.26	-2.32	-2.32	2.07	-2.32	-2.32	1.68	0.26
C:	-2.32	0.26	-2.32	-2.32	-2.32	-2.32	2.07	-2.32	0.26

Score : CAGTGACCC ????

Score d'un mot dans ma séquence ?

Calcul d'un score d'une séquence :

	1	2	3	4	5	6	7	8	9
A:	1.14	0.26	-2.32	2.07	-2.32	-2.32	-2.32	-2.32	1.14
T:	-2.32	0.26	2.07	-2.32	-2.32	2.07	-2.32	0.26	-2.32
G:	1.14	0.26	-2.32	-2.32	2.07	-2.32	-2.32	1.68	0.26
C:	-2.32	0.26	-2.32	-2.32	-2.32	-2.32	2.07	-2.32	0.26

Score : CAGTGACCC ????

-2.32+0.26-2.32-2.32+2.07-2.32+2.07+0.26 = -4.62

score > 0 : la séquence a plus de chance de correspondre au motif qu'une séquence aléatoire,

score < 0 : séquence plus probablement aléatoire

Ce score est-il significatif ?

Il est nécessaire ensuite de comparer le score calculé aux scores max et min possibles avec le profil et de choisir le seuil de significativité.

⇒ pas de p-value.

Autres raffinements possibles du profil

Il est possible de tenir compte des taux de substitution / conservation observées en générale dans les séquences protéiques (matrice PAM / BLOSUM) et de rajouter des indels (en rajoutant une colonne pour les gaps).

Ce type de formalisme convient donc pour des motifs moins bien conservés et permet d'utiliser des connaissances biologiques pour améliorer le profil.

Recherche de profils de novo : MEME

Le programme MEME de la suite du même nom découvre des motifs non gappés dans un ensemble de séquence.

- Représente les motifs par une PPM (Position Probability Matrix) mais rajoute aussi l'information de la fréquence des nucléotides observées.
- Randomise les données en entrée (ou prend un set de données contrôle) pour trouver les motifs significativement enrichis dans les séquences.
- Plutôt pour de l'ADN (ex : promoteurs) mais certains logiciels fonctionnent aussi sur des séquences protéiques

Recherche de profils de novo : MEME

Utilise :

- des méthodes statistiques (algorithme EM) pour choisir automatiquement la longueur, le nombre d'occurrences et la description de chaque motif.
- Un modèle de background (fréquence des résidus, des mots de taille 2, 3 selon le paramétrage…) pour calculer la vraisemblance du motif, sa significativité (e-value) et la PSPM.

Recherche de profils de novo : MEME

Les scores :

- Le **log likelihood ratio** (LLR) du motif = log (Pr(occurrence | motif) / Pr(occurrence | back)). C'est une mesure de la différence du locus avec le modèle de background.
- **E-value** du motif : estimation du nombre de motif de même taille et ayant le même nombre d'occurrences qui aurait obtenu un LLR égal ou supérieur si les séquences avaient été obtenues aléatoirement selon le background modèle d'ordre 0.

Recherche d'un ensemble de motifs : MAST

MAST recherche un ensemble de motifs non gappés dans des séquences non alignées.

- calcul de la p-value pour chaque occurrence de motif
- calcul de la p-value globale (produit des p-values des occurrences de chaque motifs dans la séquence)
- Les résultats sont triés sur la e-value globale (p-value corrigée pour les tests multiples).
- Les p-values sont calculées en utilisant un modèle de background obtenu à partir d'une vieille version de la base de données nr du NCBI.

http://meme-suite.org/doc/mast.html?man_type=web

Travaux pratiques

En attendant ….

http://emboss-genotoul.toulouse.inra.fr/

http://meme-suite.org/doc/overview.html?man_type=web

- **jetons un coup d'oeil aux différents outils de la suite MEME et au manuel**
- **Jetons aussi un coup d'oeil à la suite emboss.**

En particulier les sections EDIT et PROTEINS PROFILE

Travaux pratiques

http://meme-suite.org/tools/meme

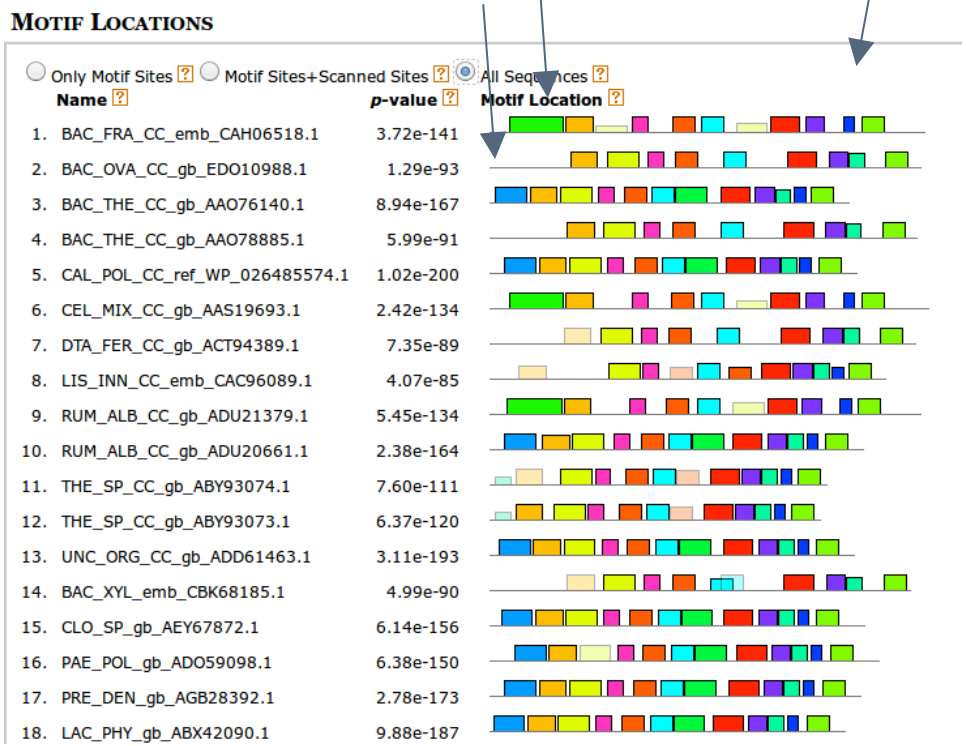
- **Exercice 1 : Rechercher des motifs conservés dans le jeu de données GH130InputRenamed.fasta**

Utilisez les pages de résultats HTML de MEME et de MAST

Comment ces motifs sont ils répartis ?

- *dans l'arbre ?*
- *selon les fonctions connues des protéines connues ?*

Regardons les motifs 7, 11 et 13 de plus près



Travaux pratiques

http://meme-suite.org/tools/meme; http://emboss-genotoul.toulouse.inra.fr/

- **Exercice 2 : télécharger et explorer ces motifs**

Faire 3 groupes : un sur le motif 7, un autre sur le 11 et le dernier sur le 13

1) *Télécharger le logo et le fasta des occurrences du motif*
2) *Passer dans emboss : Utiliser Prophecy pour générer une matrice de fréquence à partir de ce fichier. La sauvegarder en utilisant un nom explicite. Réutiliser Prophecy pour générer une matrice de poids de Grisbkov. La sauvegarder sous un autre nom explicite.*
3) *Utiliser Profit pour rechercher le motif sous la forme d'une matrice de fréquence simple dans les séquences GH130InputRenamed.fasta et Prophet pour chercher le motif sous forme de matrice de poids, augmenter la pénalité d'ouverture de gap à 3.*
4) *Comparer. Que remarquez-vous ? Choisir un seuil de significativité pour Prophet.*

Travaux pratiques

En option : si vous avez le temps

http://www.cazy.org/GH130_characterized.html

- **Exercice 2 : télécharger et explorer ces motifs**

Diviser la salle en 3 groupes : un sur le motif 7, un autre sur le 11 et le dernier sur le 13

5) Une des séquences présentant ces motifs a-t-elle une structure pdb ? Si oui, ouvrir une des fiches pdb correspondante et regarder où se situe le motif dans la structure secondaire et dans la structure tertiaire.

Comment les décrit-on ?

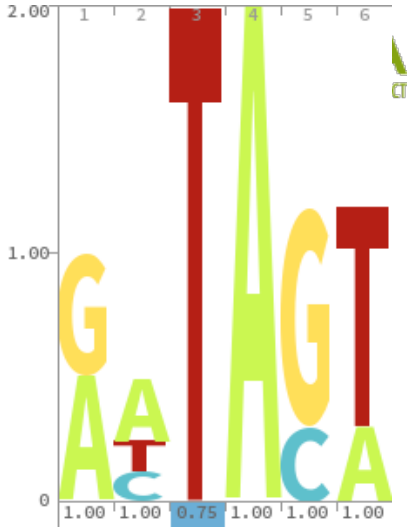
Il y a trois grands types de formalismes pour représenter un motifs :

Comment les décrit-on ?

Il y a trois grands types de formalismes pour représenter un motifs :

- les patterns (expression régulières ou format Prosite)
- les profils ou matrices de fréquence ou de poids
- les HMM

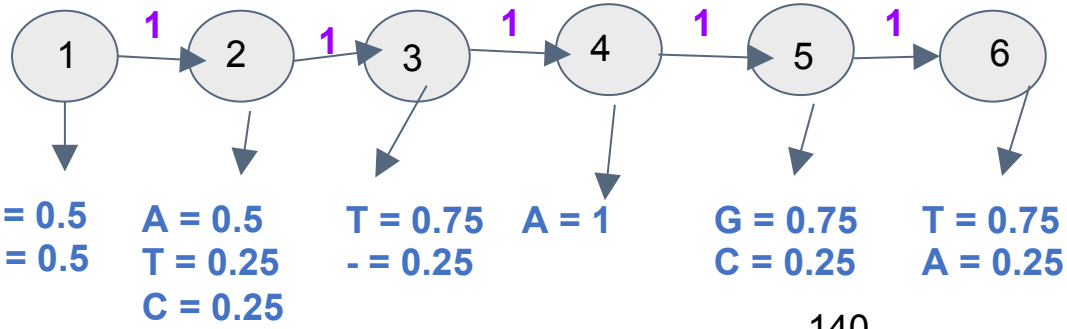
Comment les décrit-on ? Les HMM



Les HMM : Chaînes de markov cachées
Basiquement, il s'agit d'un modèle statistique représentant le motif en utilisant la probabilité d'avoir un résidu après un autre.
Ex :
AATACT
GA-AGT
ATTAGA
GCTAGT

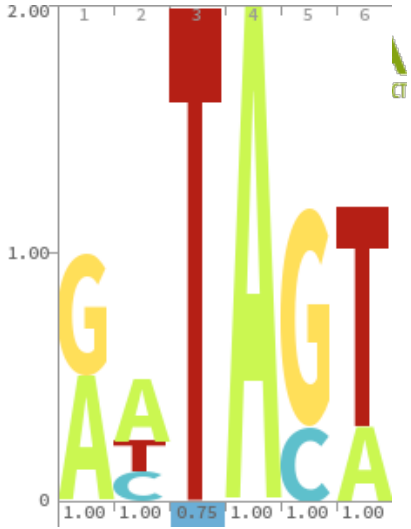
Si pas de gap

Probabilités de transition sans gap = 1



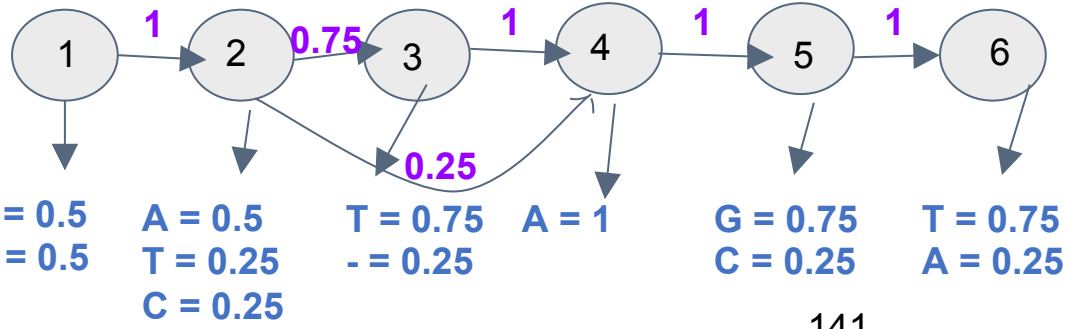
Probabilités d'émission

Comment les décrit-on ? Les HMM



Les HMM : Chaînes de markov cachées
Basiquement, il s'agit d'un modèle statistique représentant le motif en utilisant la probabilité d'avoir un résidu après un autre.
Ex :
AATACT
GA-AGT
ATTAGA
GCTAGT

Probabilités de transition sans gap = 1



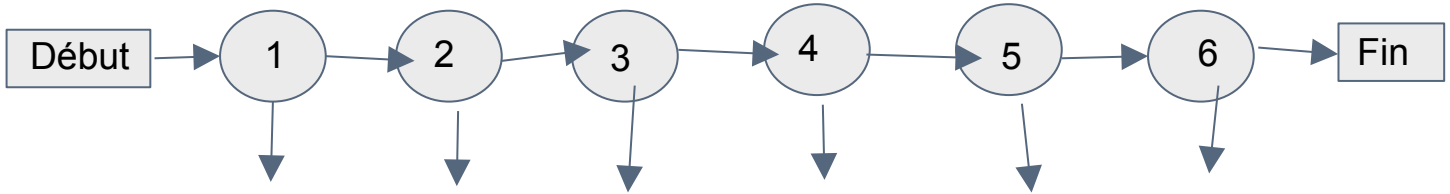
Probabilités d'émission

Comment les décrit-on ? Les HMM

Les HMM : Chaînes de markov cachées
Si on généralise l'exemple précédent :

Types de probabilités de transitions :
match -> match
begin -> match
match -> end

Match / mismatch

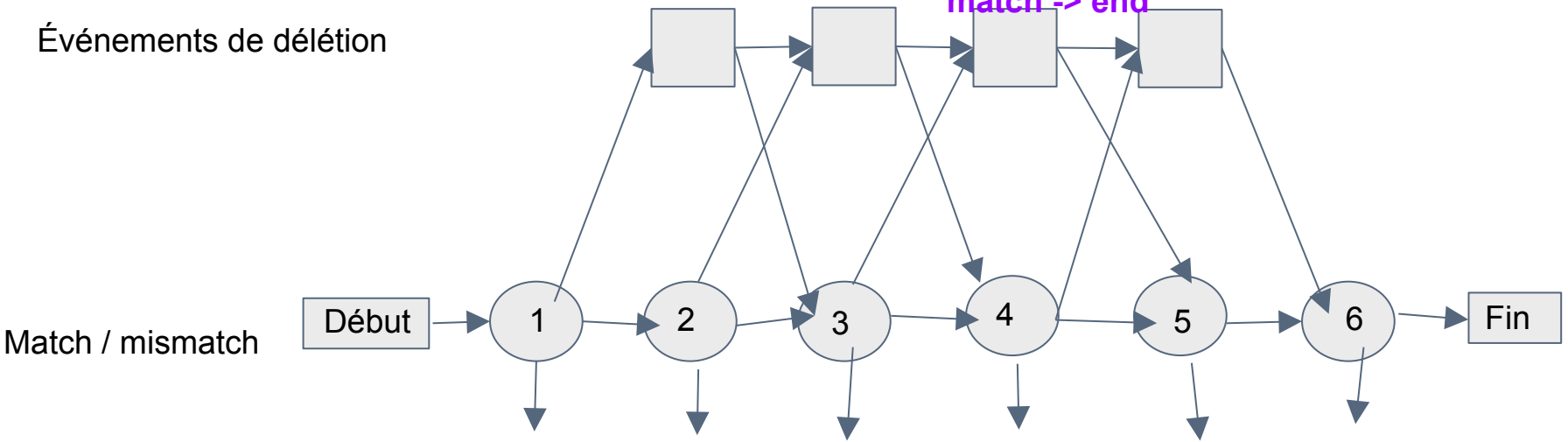


Probabilités d'émission

Comment les décrit-on ? Les HMM

Les HMM : Chaînes de markov cachées
Si on généralise l'exemple précédent :

Types de probabilités de transitions :
match -> match **delete -> delete**
match -> delete **begin -> match**
match -> end

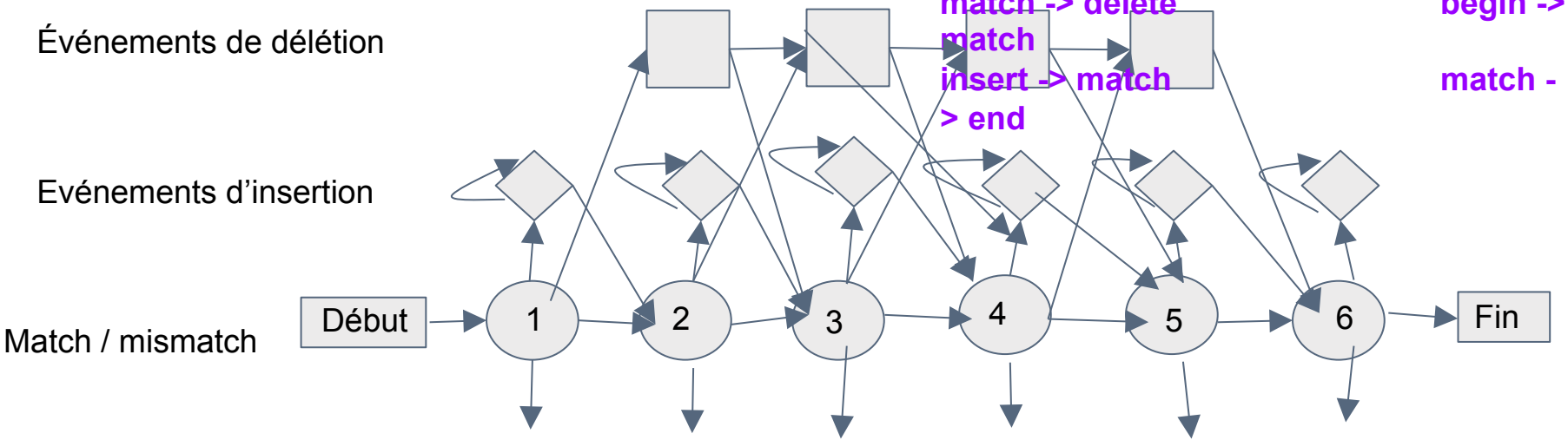


Comment les décrit-on ? Les HMM

Les HMM : Chaînes de markov cachées
Si on généralise l'exemple précédent :

Types de probabilités de transitions :

match -> match
match -> insert
match -> delete
insert -> match
insert -> insert
delete -> delete
begin -> match
match -> end



Probabilités d'émission

Comment les décrit-on ? Les HMM

Les HMM sont donc utiles pour :

- représenter des motifs complexes.
- permettre l'utilisation d'algorithmes utilisés dans des problèmes similaires.
- comme pour les profils, il est possible de rajouter de la "connaissance biologique" dans le motif en modifiant les probabilités par exemple en étant plus tolérant aux substitutions entre acides aminés fréquemment rencontrées dans les protéines en générale.
- retrouver des occurrences de motifs dans des séquences très divergentes.

Comment les décrit-on ? Les HMM

- ATTENTION :

Un profil HMM peut être aligné sur une séquence globalement ou localement (seulement une partie du profil) mais cela est choisi lors de la construction du profil et non lors de son utilisation.

Comment utilise-t-on les motifs HMM ?

Ma séquence contient-elle ce motif ?

Pour chercher si une séquence contient un profil HMM connu, le programme va chercher le chemin le plus probable de la séquence à travers le modèle en utilisant les probabilités de transition et d'émission pour calculer le score (S = Pobs / Pattendu).

Comment utilise-t-on les motifs HMM ?

A partir du score S, et grâce au modèle probabiliste, le logiciel calcule la p-value.

p-value : probabilité d'obtenir par chance un score au moins égal au score S.

Comment utilise-t-on les motifs HMM ?

Quelles sont les séquences contenant mon profil HMM dans cet ensemble de séquences (database) ?

- la p-value doit être corrigée pour les tests multiples ⇒ e-value
- la e-value est le nombre de hits qu'on aurait obtenus avec un aussi bon score contre une database de même taille contenant des séquences aléatoires.
- le score est un score log transformé appelé aussi log odds score ou bit score. Il ne dépend pas de la taille de la database mais seulement du modèle et de la séquence cible.

Pourquoi utilise-t-on les motifs HMM ?

- **Créer un profil HMM représentatif** d'une famille de protéines d'intérêt ou d'un domaine protéique d'intérêt. Puis utiliser ce profil pour rechercher d'autres séquences appartenant à cette famille ou contenant ce domaine. ⇒ hmmerBuild, HmmerSearch (Hmmer2), hmmbuild, hmmsearch (hmmer3)

- **Annoter une séquence** : chercher des motifs connus ou des domaines fonctionnels qu'on ne trouverait pas avec un blast standard car trop peu similaires ⇒ Interpro, hmmerPfam (hmmer2), hmmScan (hmmer3)

- **Construire un alignement multiple** très rapidement en utilisant le profil HMM comme graine. ⇒ hmmerAlign (hmmer2), hmmalign (hmmer3)

Travaux pratiques

http://sequenceconversion.bugaco.com/converter/biology/sequences/fasta_to_stockholm.php

- **Exercice 3 : caractérisons nos sous-familles par un profil HMM**

Diviser la salle en 3 : un sur chaque sous-famille

1) *Convertir l'alignement multiple clustal en alignement au format stockolm*

Si besoin : http://genoweb.toulouse.inra.fr/~formation/AlignClusteringLISBP/data/SubFamilies

Travaux pratiques

http://emboss-genotoul.toulouse.inra.fr/

- **Exercice 3 : caractérisons nos sous-familles par un profil HMM**

Diviser la salle en 3 : un sur chaque sous-famille

2) Construire un profil HMM global représentant la sous-famille
Utiliser en entrée l'alignement multiple en format stockolm
Attention : mettre obligatoirement un nom pour le HMM
Laisser le reste par défaut
Sauvegarder le fichier hmmfile avec un nom explicite et jetter un coup d'oeil au fichier.
Que sont les lignes et les colonnes de la matrice ?

Travaux pratiques

http://skylign.org

- **Exercice 3 : caractérisons nos sous-familles par un profil HMM**

Diviser la salle en 3 : un sur chaque sous-famille

3) Générer un logo pour ce profil HMM, le sauvegarder avec un nom explicite.

Travaux pratiques

http://emboss-genotoul.toulouse.inra.fr/

- **Exercice 3 : caractérisons nos sous-familles par un profil HMM**

Diviser la salle en 3 : un sur chaque sous-famille

4) Calibrer le profil HMM avec ehmmcalibrate.

Ceci permet d'augmenter la sensibilité de la recherche de ce motif contre un ensemble de séquence. Pour cela il génère un grand nombre de séquences aléatoires et compare le profil HMM original contre elles. Il calcule ainsi des paramètres de valeurs extrêmes qu'il écrit dans le profil HMM modifié qu'il génère.

Sauvegarder le fichier outhmmfile avec un nom explicite

Travaux pratiques

http://emboss-genotoul.toulouse.inra.fr/

- **Exercice 4 : recherchons les membres de la sous-famille dans le dataset d'origine (GH130InputRenamed.fasta) en utilisant le motif HMM.**

Diviser la salle en 3 : un sur chaque sous-famille

5) utiliser le bon programme pour rechercher le motif dans l'ensemble des séquences, en utilisant les paramètres par défaut.

Pourquoi y a-t-il deux sections de résultats ?

Qu'en concluez-vous ?

Travaux pratiques

En option : si vous avez le temps

https://www.ebi.ac.uk/Tools/hmmer/search/hmmsearch

- **Exercice 5 : utilisons hmmer3 pour rechercher des protéines similaires dans des banques**

Diviser la salle en 3 : un sur chaque sous-famille

1) utiliser l'alignement de la sous-famille en entrée
faire la recherche contre "swissprot",
dans les paramètres avancés, customised results cliquer sur tout.
Explorer les résultats. Comparer les scores avec ceux que nous avions obtenus.
N'oubliez pas la possibilité de cliquer sur ">".

Travaux pratiques

En option : si vous avez le temps

https://www.ebi.ac.uk/Tools/hmmer/search/hmmsearch

- **Exercice 5 : utilisons hmmer3 pour rechercher des protéines similaires dans des banques**

Diviser la salle en 3 : un sur chaque sous-famille

2) utiliser l'alignement de la sous-famille en entrée

faire la même chose que précédemment contre la banque "ensembl genome bacteria"

Que remarquez-vous ?

Petit bilan intermédiaire

- **Créer un profil HMM représentatif** d'une famille de protéines d'intérêt ou d'un domaine protéique d'intérêt. Puis utiliser ce profil pour rechercher d'autres séquences appartenant à cette famille ou contenant ce domaine. ⇒ hmmerBuild, HmmerSearch (Hmmer2), hmmbuild, hmmsearch (hmmer3)

OK

Conseil : Penser à enrichir le modèle avec les nouvelles protéines trouvées (processus itératif)

- **Annoter une séquence** : chercher des motifs connus ou des domaines fonctionnels qu'on ne trouverait pas avec un blast standard car trop peu similaires ⇒ Interpro, hmmerPfam (hmmer2), hmmScan (hmmer3)

Analyse fonctionnelle et structurale de séquences protéiques

Interpro : http://www.ebi.ac.uk/interpro/

- classifie en famille (si elles sont connues)
- prédit la présence de domaine (au sens de interpro : unité structurelle et / ou fonctionnelle)
- de sites annotés comme importants (site actif, site de liaison…)
- et de domaines structuraux (structure super secondaire comme l'hélice - boucle - hélice des facteurs de transcription
- pour cela il intègre de nombreux outils et base de données.

Travaux pratiques

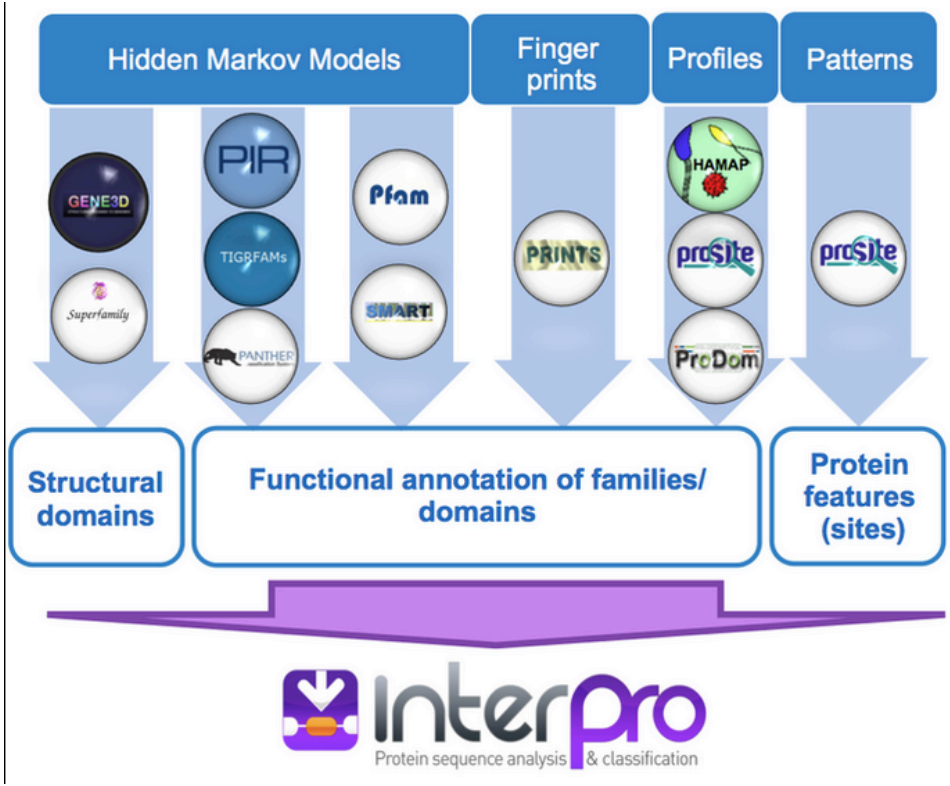
http://www.ebi.ac.uk/interpro/

- **Exercice 7 : utiliser Interpro**

1) Choisir la séquence de votre choix dans la sous-famille étudiée pour interroger interpro

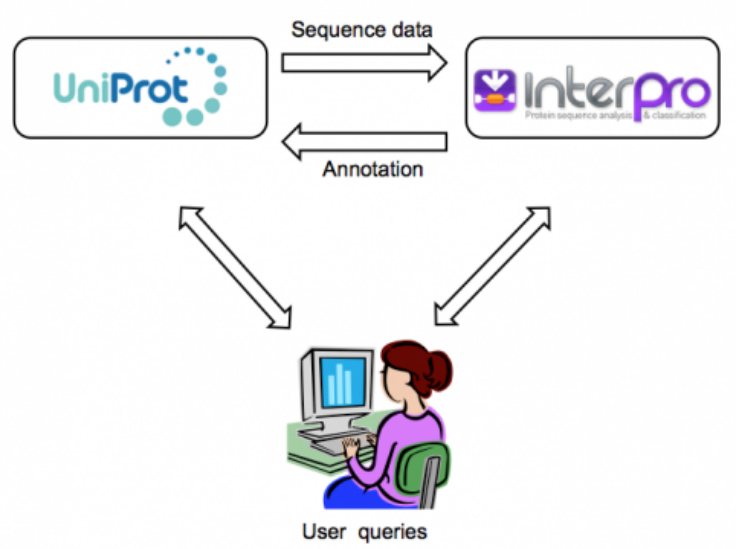
Cela va être un peu long...

Analyse fonctionnelle et structurale de séquences protéiques



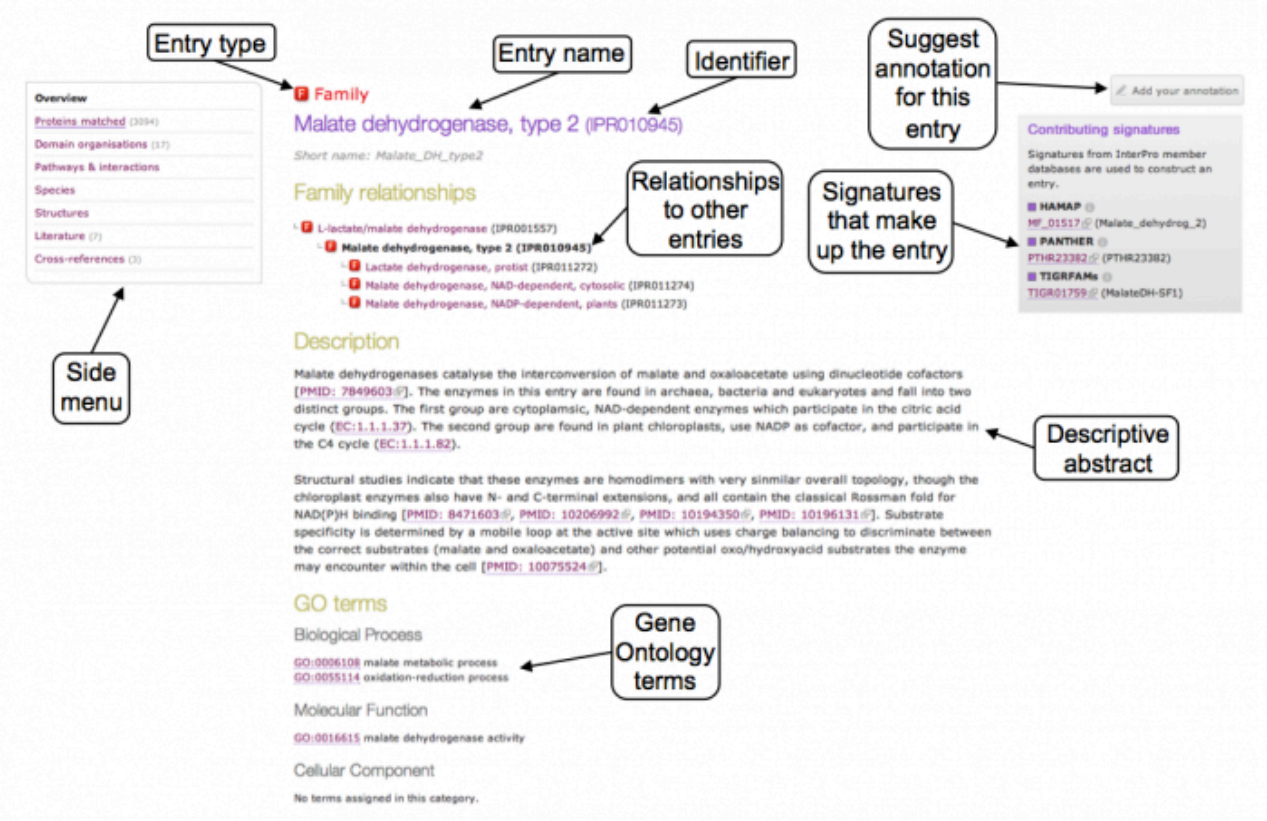
Analyse fonctionnelle et structurale de séquences protéiques

La base de donnée UniproKB est régulièrement analysée avec Interpro pour en améliorer l'annotation.



Analyse fonctionnelle et structurale de séquences protéiques

Intégration de différentes informations caractéristiques de la même famille protéique, du même domaine ou de la même caractéristique de séquence.
Ex de la fiche Interpro de la famille des malate déshydrogénase de type 2

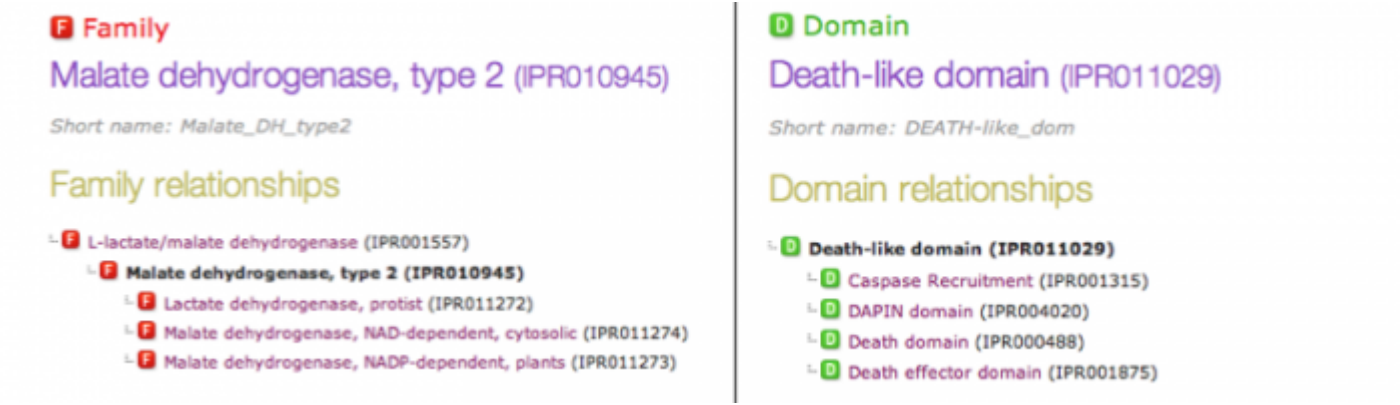


Analyse fonctionnelle et structurale de séquences protéiques

Les 4 catégories figurant dans Interpro

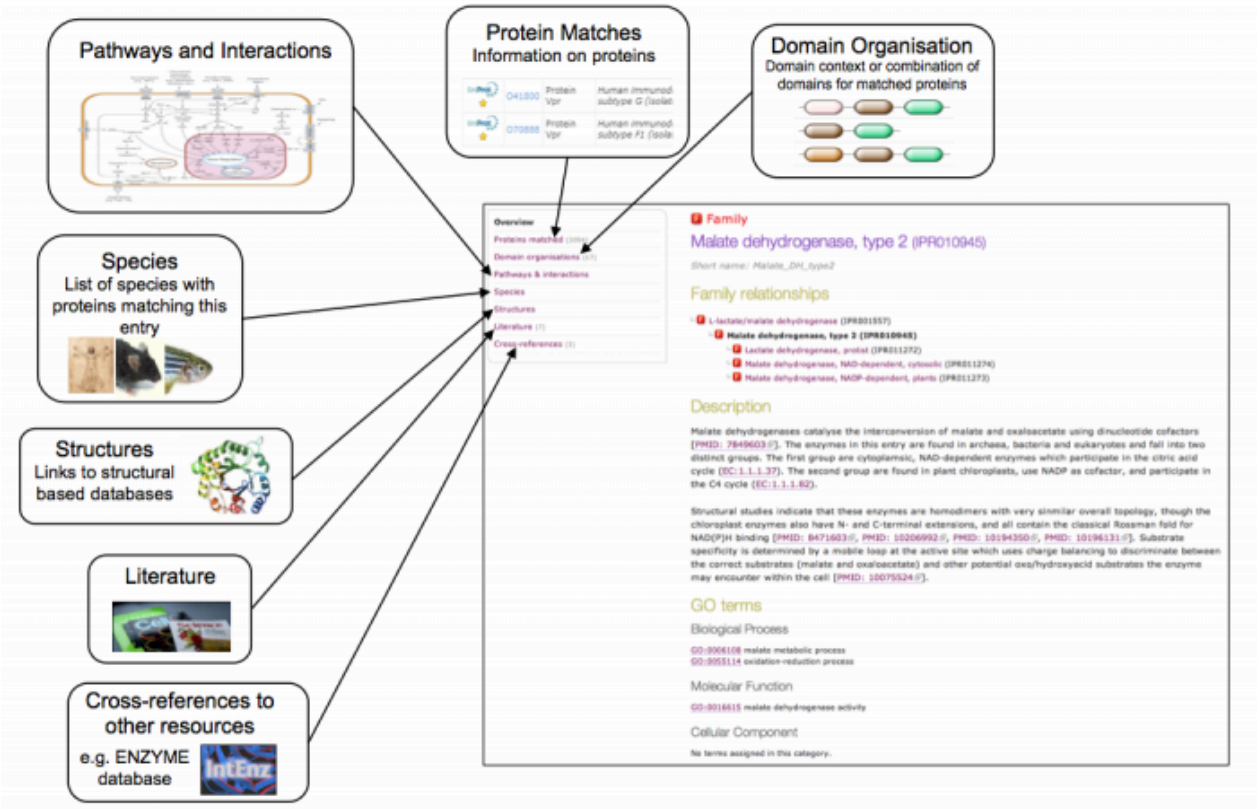
Analyse fonctionnelle et structurale de séquences protéiques

Les hiérarchies entre les entrées interpro des familles et des domaines.
Pas d'overlap entre les deux.

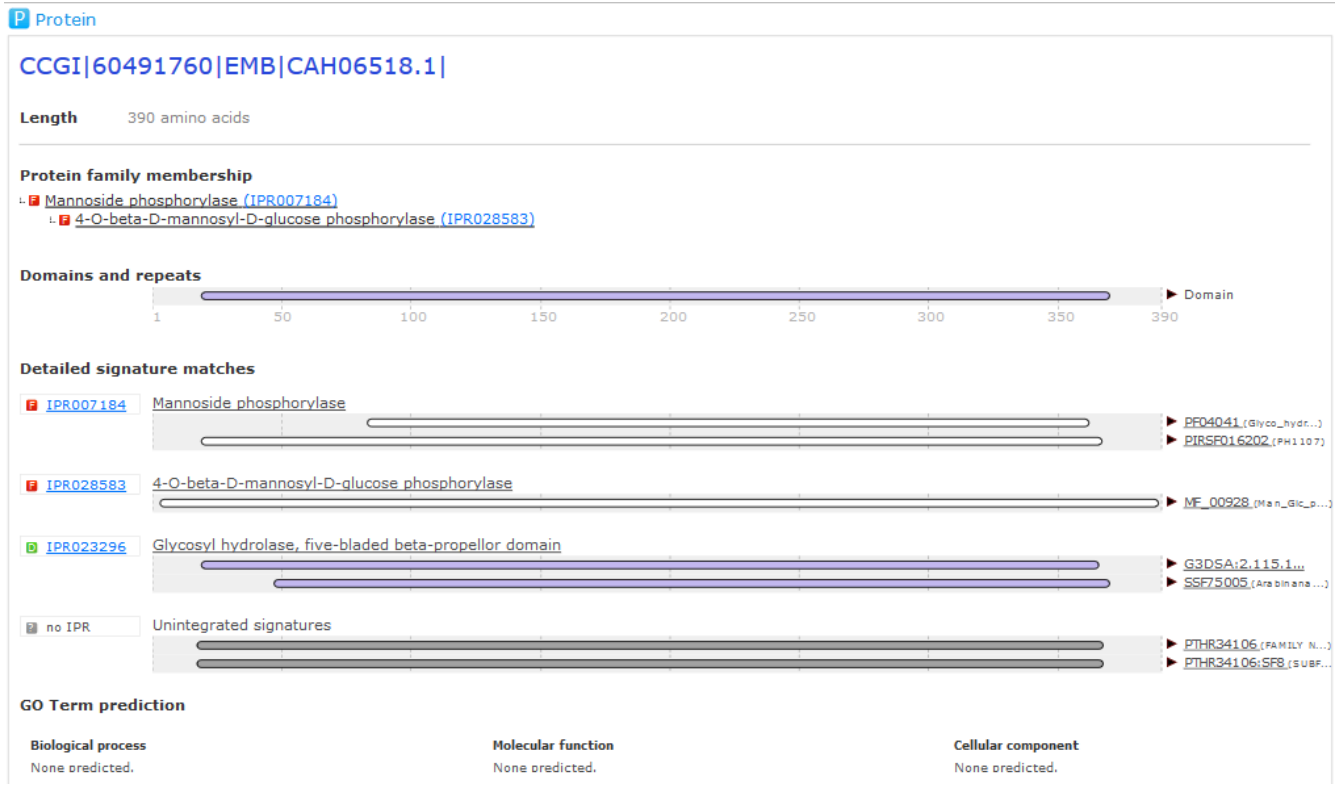


Analyse fonctionnelle et structurale de séquences protéiques

Menu en haut à gauche d'une entrée Interpro



Analyse fonctionnelle et structurale de séquences protéiques



Famille prédite par Interpro

Domaines et répétitions prédits comme contenus dans la séquence

Information détaillée concernant les motifs trouvés sur la séquence

Les GO (Gene Ontology) termes prédits à partir des informations précédentes

Travaux pratiques

http://www.ebi.ac.uk/interpro/

- **Exercice 6 : utiliser Interpro**

2) Explorer les résultats. Passer la souris sur les différents motifs trouvés. Ouvrir les fiches Interpro et des autres bases de données liées.

Conclusion

Les motifs peuvent :

- représenter des zones conservées dans des familles de protéines et servir ainsi de “portrait robot” de la famille.
- décrire des domaines ou sites fonctionnels ou structuraux.

Nous avons manipulé plusieurs outils permettant de détecter *de novo* des motifs ou de rechercher des motifs connus dans des séquences.

Identifier des domaines dans une séquence est une étape essentielle de l’annotation fonctionnelle surtout si notre séquence est loin des séquences présentent dans la banque de données curées manuellement.

Pour cela la meilleure stratégie est d’interroger et d’intégrer plusieurs bases de données complémentaires.

Quelques liens utiles….

http://skylign.org/
http://eddylab.org/software/hmmer3/3.1b2/Userguide.pdf

Conclusion Générale

Aujourd'hui nous avons appris à faire des alignements, des arbres et à rechercher des motifs conservés sans ligne de commande.

Mais vous aurez sans doute besoin de l'aide d'un bioinfo pour modifier les header en input du fichier fasta.

Il faudra enlever la redondance et peut-être sous-échantillonner pour certaines interfaces.

Tous les outils présentés existent en ligne de commande et permettent ainsi une utilisation à grande échelle.

MERCI !

Feedback :
https://enquetes.inra.fr/index.php?sid=84236&newtest=Y