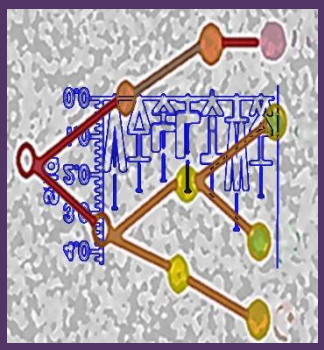
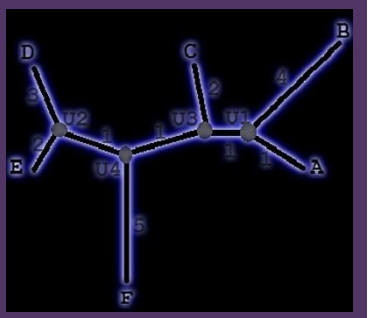
Cours de phylogénie moléculaire

Distances et constructions phylogénétiques

Support pédagogique de phylogénie moléculaire destiné aux étudiants du système LMD de Master ‘M1 et M2) et doctorants de Biotechnologie végétale, Biochimie et Microbiologie.





UNIVERSITE
CONSTANTINE 1

Faculté des Sciences de

la Nature et de la Vie

Pr DJEKOUN A.

Pr HAMIDECHI M. A.

Plan du cours

Partie 1 : LES DONNÉES DE LA PHYLOGÉNIE

1. **Les données phénotypiques**
2. **Les données moléculaires**
3. **La structure d'un arbre phylogénétique**
4. **Notion de distances**

Partie 2 : LES MÉTHODES DE CONSTRUCTIONS D'ARBRES PHYLOGÉNÉTIQUES

1- **Les méthodes phénétiques**
2- **Les méthodes cladistiques**
3- **Les méthodes du maximum de parcimonie**
4- **Le bootstrap**
5- **Les méthodes phylogénétiques par l'exemple**

Partie 3 : LES OUTILS DE LA PHYLOGÉNIE

Partie 4 : Exercices d'application

Préambule : La phylogénie moléculaire est une discipline qui connaît un essor grandissant étant donné l'avancement spectaculaire des techniques de la biologie moléculaire et du génie génétique que l'on peut appeler maintenant biotechnologies moléculaires. Ces techniques ont permis un nombre incalculables de données biomoléculaires telles que les séquences des différents gènes et protéines. Actuellement on peut recenser quelques 80 millions de séquences[1] sur le portail NCBI par exemple !

La phylogénie permet d'étudier les espèces végétales, animales et microbiennes, sur les deux plans phénotypique et génotypique, afin de les classer en fonction de leurs ressemblances et en fonction de leurs structures géniques (liens de parenté). La phylogénie étudie, en fait, les relations de parenté entre les individus et représente sous forme d'arbre le résultat de ces relations.

Donc face à ce tas de données, il y aura besoin d'outils adéquats pour pouvoir traiter toutes ces informations et tirer un meilleur profit. La manipulation correcte des données initiales va permettre d'aboutir à des interprétations et des concluions pertinentes : Grâce aux résultats de la phylogénie, le chercheur peut tirer des hypothèses sur les liens génétiques des espèces, les états ancestraux des caractères étudiés, la divergence ou la convergence des caractères.

Ce cours est destiné aux étudiants de Master et Doctorants de Biotechnologie végétale, Biochimie et Microbiologie. Ils trouveront non seulement des rappels des cours relatifs aux alignements multiples et aux notions de motifs, mais également des notions simplifiées sur la phylogénie et des méthodes qui permettent d'initier nos étudiants et doctorants (chacun à son niveau) aux différentes méthodes de constructions phylogénétiques et à leurs principes de base. C'est un support pédagogique qui va guider les étudiants dans leurs travaux grâce aux exercices corrigés et aux différents exemples cités dans tous les paragraphes.

Objectifs :

1- Connaître la nature des différentes données pour la phylogénie
2- Comprendre la structure des arbres phylogénétiques (ou les dendrogrammes)
3- Apprendre quelques méthodes de constructions phylogénétiques

Pré-requis :

1- Savoir réaliser un alignement multiple avec un programme informatique tel que MEGA
2- Savoir définir un motif moléculaire commun à un ensemble de séquences biologiques

[1] ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt

Partie 1 : LES DONNÉES DE LA PHYLOGENIE

Les progrès des biotechnologies moléculaires ont conduit à une vaste accumulation de nouvelles données biologiques principalement sous forme de séquences nucléiques (gènes ou ORFs, marqueurs moléculaires, …) et protéiques (enzymes du métabolisme énergétique, protéines de structures, etc…).

En parallèle, beaucoup de sites ont vu le jour sur le la grande toile du web pour permettre le stockage et la manipulation de toutes ces informations. L'acquisition de ces données et leur traitement nécessitent des méthodes et des outils adéquats.

La phylogénie se base sur le principe de la comparaison de caractères spécifiques pour un ensemble d'individus. Ces caractères sont en général homologues et appartiennent à des organismes contemporains. Sauf que leur comparaison, par le biais des méthodes phylogénétiques, va permettre de postuler des hypothèses quant à l'éventuelle histoire commune ou non entre ces individus du point de vue moléculaire et phénotypique.

On peut diviser les données qui vont nous servir pour la construction d'arbres phylogénétiques en deux groupes distincts :

- Les données liées aux caractères phénotypiques.
- Les données moléculaires telles que les séquences d'ADN ou de protéines.

En fait ces données concernent les caractères morphologiques, physiologiques, génétiques et génomiques.

Le traitement de l'une ou de l'autre catégorie de données va aboutir à un dendrogramme donné et nécessite des approches et des traitements différents.

Les données phénotypiques : comprennent les caractères observables (aux différents états : morphologiques, biochimiques et physiologiques) et les patterns binaires (de type présence d'un caractère donné / absence de ce même caractère). Dans le cas des bactéries, par exemple, les caractères peuvent être :

- Biochimiques et enzymatiques,
- Antigéniques
- Sensibilité vis-à-vis des antibiotiques
- Sensibilités aux phages,
- Profils électophorétiques de systèmes enzymatiques, …

Par exemple, nous pouvons avoir ce type de données qui servira à construire une phylogénie :

	Caractère1	Caractère2	Caractère3	Caractère4	Caractère5
Souche1	0	1	1	0	1
Souche2	1	1	0	0	1
Souche3	1	0	1	1	1
Souche4	1	0	1	0	0

On constate que chaque caractère existe sous deux états différents :

- L'état 1 qui signifie la présence de ce caractère.
- L'état 0 qui signifie l'absence de ce caractère.

Un état de caractère est donc un attribut observable et mesurable sur un individu et qui peut être d'une quelconque nature (moléculaire, physiologique, …) :

Caractères	Etats du caractère				
physiologique	Croissance à 5°C		Croissance à pH acide		
	1	0	1	0	
Couleur de la colonie	Blanche	Crème	Jaune	Orange	
Motif ATATA	1		0		
Ag O	1		0		
Nageoires	Oui		Non		
Séquence gène	A	C	G	T	Gap
Séquence protéine	20 Acides aminés ou Gap				

Les données moléculaires : Dans ce cas, ce sont des séquences biologiques de type acides nucléiques telles que les séquences de gènes particuliers, d'ARNm, RFLPs, Microsatellites, SNPs, IGS (ARNr et mitochondries), ITS(ARNr et mitochondries), séquences des cytochromes C, séquences des facteurs d'élongation alpha, ou encore des séquences de protéines enzymatiques ou de structure.
Les données les plus employées pour les constructions phylogénétiques sont les marqueurs suivants :

- ADNr 16S : Bactéries
- ADNr 18S, actine, EF1, RPB1 : Eucaryotes
- ADNr 18S, RBCL : Végétaux
- Animaux :
    - Niveau phylum, classe, ordre : ADNr 18S, génome mt
    - Niveau famille : RAG2, 12S, 16S mt
    - Niveau genre : ITS, protéines mt
    - Niveau intra spécifique : D-LOOP, Introns

Les séquences retenues pour la phylogénie doivent être alignées en utilisant un programme informatique adéquat. Pour cela, un large choix est disponible sur la grande toile :

1. http://bioweb.pasteur.fr/seqanal/interfaces/clustalw.html
2. http://www.ebi.ac.uk/Tools/clustalw2/
3. http://phylogenomics.berkeley.edu/cgi-bin/muscle/input_muscle.py

ATTENTION : *Il convient de bien analyser le résultat de l'alignement multiple avant de passer à la construction de l'arbre phylogénétique et de bien régler les paramètres du logiciel.*

A titre de rappel, nous allons procéder à l'alignement multiple du jeu de séquences en utilisant l'outil ClustalW[2] . Ces séquences appartiennent à la famille des facteurs de transcription du type "Basic Leucine Zipper". Ce sont des gènes qui codent pour des protéines qui régulent la transcription des ARNm (voir fin du document).

Le résultat de l'alignement multiple de cette série de séquences est le suivant :

```
Solanum.tuberosum1466pb          GGAATTTTGATTTTCAAGATTCCATTCAATTTTTCTTCTATGGGT----G 46
Triticum.monococcum1062pb        ----------------------------------------ATGG------- 4
Rattus.norvegicus1785pb          -------------------------------GTACATATGTGA------- 12
Zea.mays1236pb                   ------------------TGCGCACGCCACCGCGCTTCATTGG------- 25
Oryza.sativa1272pb               ----------ATGGCCATCATCCATGACACCTCTGATCAACAAG-----G 35
Xenopus.laevis1188pb             --------------------ATGAACGGCCTTGTTCCTTGCCCAAC----C 27
Arabidopsis.thaliana1489pb       ------------------GCGGCTCCACGTAAGACAAAAACAAA------ 26
Triticum.aestivum1585pb          -----------------CGTTCCGCGCTGTCCGTCACACACAGGATGGCG 33
Solanum.tuberosum1466pb          TTTA-GAAGGATTTAGGCTTTTAGAG-----TTTTGAAGTGGGGAAAAAA 90
Triticum.monococcum1062pb        --CA-GAGG---CCAGCCCT---------------------AGAACAGAA 27
Rattus.norvegicus1785pb          ---A-GGAAGAGTCAGCATG---AGG-----ACTAGACACCCAGAGAGAG 50
Zea.mays1236pb                   -CCACGCCGTTGCCATCACG--------------------CCGATTAAAC 54
Oryza.sativa1272pb               TACACCATAATTCTATTTCTTCCAAGCCAACTTTACATGCATGATCATAC 85
Xenopus.laevis1188pb             CCTGTGGATCCACCTGCTCCTCCTCA-----GGTACCTCCCACATTAGAG 72
Arabidopsis.thaliana1489pb       TTAAGGAAAGAGCGAGAGAGAGAGAGAGAGAGAAAAAAGAAGAAGAAGAA 76
Triticum.aestivum1585pb          TCCGCCATGGAGCTCTCCCTCCTCAGCC---CCGCAATGCACCACCACGG 80
                                                                                 *

Solanum.tuberosum1466pb          ----AGGTTTC---------TTGAAGACTT-TGCTGTT----GTTGGCTT 122
Triticum.monococcum1062pb        ----ACGTCAA---------CAGATGA----TACTGAT----GAAAATCT 56
Rattus.norvegicus1785pb          ----CTGTCTC---------TGAAGGG----TACAAATCAGAGCAGGTCA 83
Zea.mays1236pb                   ----TAGCCGA---------TCGATCG----CCCAGCTC---GCCTGCC- 83
Oryza.sativa1272pb               ----ATGCACC---------AAAAACA-TG-TACAATTC---ATTTAGTT 117
Xenopus.laevis1188pb             GCTTATGCTCCACCTCCACATGAAGAGATT-TCCCCATCA--ACATCTCT 119
Arabidopsis.thaliana1489pb       TTTGAGGCGAT---------TGGGTAAGCGATGCACATC---GTCTTCTC 114
Triticum.aestivum1585pb          --CATCGCGGC---------CAAGACGGCCTCCCACCTC----CCTGTTC 115
                                       *                        *   *
```

Vers la fin de l'alignement :

```
Solanum.tuberosum1466pb          ---AAACCC-----TTCTACAGATGTCC------TGCA----TAC----T 1143
Triticum.monococcum1062pb        ---AGACTC-----TTCAGCAGATGCAA------AGGA----TCC----T 902
Rattus.norvegicus1785pb          -CCACCGCC-----CCCACCAGATATACGACCTTTGGAGAGGTAC----T 1391
Zea.mays1236pb                   -CCAAGCAC-----ACACACACACACACG---TAAGCACGAATACG-AGT 1012
Oryza.sativa1272pb               ---GGCCTC-----TCCTGGTACGGCGC----CGAGGG----------T 1074
Xenopus.laevis1188pb             -TGAAACCC-----TTCTACAGATG--------GAGC----------T 1017
Arabidopsis.thaliana1489pb       ATGAAACCC-----TACCCCATTGGAATCA--CAAGCA----------T 1255
```

[2] http://www.ebi.ac.uk/Tools/msa/clustalw2/

```
Triticum.aestivum1585pb          TCTGGCTCCGTTGATTTTGCCGAGTTTGAGCCCAAGCTTGTCTACTGAAT 1274
                                       *                       *           *

Solanum.tuberosum1466pb          -GGCTGCAC-----ACCAAT-CAGCT--------CAGGGTC------TCC 1172
Triticum.monococcum1062pb        TGACCACAG-----GC-AGT-CTGCC--------CGTGCAC------TTC 931
Rattus.norvegicus1785pb          GGGCAGCCC-----ACCAG--CAGCTG-------CAGGAAGCTGATATCC 1427
Zea.mays1236pb                   TGGTAGCGG-----TC--AT-CAGCCC-------CGAGCGCACGGTGTAC 1047
Oryza.sativa1272pb               TGGTAG-AA-----GCTAG---AGCTT-------AGCTAGC--------- 1099
Xenopus.laevis1188pb             CGACAGCAACGACTGCTAA---AGTTGC------CGAAAGC--------- 1049
Arabidopsis.thaliana1489pb       TAACCAGAA--------AAA-GAGTCAT------TGGTTTT--------- 1281
Triticum.aestivum1585pb          TTGTAGAAGAAGGATCCATCTCTGCCTTTCTTCTCAGACATAGTCATGCA 1324
                                                       *

Solanum.tuberosum1466pb          TT------------GCCTTAGG------AGAGT----ACTTTAAACGTC- 1199
Triticum.monococcum1062pb        TT------------GTGATAAG------TGATT----ACTCATCCCGGC- 958
Rattus.norvegicus1785pb          TTAAACTGAGTCAGGCATCAAGA----CTAAGC----ACTCAGCAAGTG- 1468
Zea.mays1236pb                   ATA-----------GCTTTCAG-----TAGATCG--AATTCCAGGCATG- 1078
Oryza.sativa1272pb               ---------------TAGCGAG------AGAGTG--AGCTCAGCTAAGC- 1125
Xenopus.laevis1188pb             --------------GCAGCAGA-------GATCCCTAATACTATAAAAG- 1077
Arabidopsis.thaliana1489pb       -----------------GTGATT----TTGATTG---AGGTAACTATTG- 1306
Triticum.aestivum1585pb          TCATGCT--------CCTCGAGAGTCTCTGAATGAGCACATGATCCATGG 1366
                                                          *

Solanum.tuberosum1466pb          TTCG-----TGCTCTTA-----GCTCACTTTGGGC--------TGGTCGT 1231
Triticum.monococcum1062pb        TTCG-----TGCCCTAA-----GTTCTCTTTGG-C---------T--TTGC 987
Rattus.norvegicus1785pb          CTGGA---CTGGTTTGACTCTCGATTGCCCAAGCCAGCAGAAGTGGTAGT 1515
Zea.mays1236pb                   TCCA---------TCAACAAGCAGTTTCTTC------------TCGTCAT 1107
Oryza.sativa1272pb               TTAATTAGCTGGCTTGAT---TGCTTGCTTTG-----------TGGCTGG 1161
Xenopus.laevis1188pb             TAGG-----------GAT-----GTCCTTTTGATA---------CGTCAC 1102
Arabidopsis.thaliana1489pb       TCTG-----TATTTTTAT------TTACTGTATGACTCAGCGACGGTAAA 1345
Triticum.aestivum1585pb          TTAATTAACAGGATCTAC-----ATCCTCCTG-----------TGCTCAT 1400
                                             *       *
```

Cet alignement présente beaucoup de gap qui faussent l'interprétation. Ceci est dû au fait que nos séquences appartiennent à des individus dont la taxonomie est totalement différente. Nous avons aligné des séquences de grenouille, de blé, …

Nous allons reprendre cet alignement mais cette fois-ci avec les séquences du règne végétal uniquement :

L'ordre des individus qui apparaissent dans le résultat de l'alignement multiple est le suivant :

1. *Triticum aestivum*
2. *Oryza sativa*
3. *Zea mays*
4. *Arabidopsis thaliana*
5. *Solanum tuberosum*
6. *Triticum monococcum*

```
gi|62736387|gb|AY914051.1|       GCTTCTACACCACGGACTTCGACGAGATGGAGCAGCTGTTCAACGCCGAGATTAAC---A
gi|33943625|gb|AY346329.1|       GCGGCGGCGCGGCGGCGTACGAGGAGGAGGAGGAGGAGGTTGAGGACGACGACGGCGGCG
gi|308044466|ref|NM_001196644.1| GCGTGGCCATGGAGGGCGACGACGACGGCCCGGAGTGGATGATG---GAGGTGGGCGGCG
gi|334185982|ref|NM_001203162.1| ---------------CCACAGGCTTATCAA-TGAGTTGTCTGGTTCCGATTCGAGCCCTA
gi|575417|emb|X82544.1|          GATTCTTGAAGTCGAGAATTGCCTAGAGAA-CGAGATGCCTATTATGGAGAAAAGACTAG
gi|461682445|gb|JX424318.1|      -ATGGCAGAGGCCAGCCCTAGAACAGAAAC-GTCAACAGATGATACTGATGAAAATCTTA
                                                 *   :        ..  .   .     **    ..    .

gi|62736387|gb|AY914051.1|       AGCAGCTCAACCAGGACG-------------AGTTCGACGCGCTGCTGCAGGAGTTCAAG
gi|33943625|gb|AY346329.1|       GCGGCGGCGGCGGCGGCG-------------GCGGCGGTGGGGGGCTCGGGGAGAAGAAG
gi|308044466|ref|NM_001196644.1| CGGGCGCCACAGGGAAGG-------------GAAAAGGCGGCGCGCTGGACAAGAACAAG
gi|334185982|ref|NM_001203162.1| CGACTAACACAATCGAGAGATCACCTCCACCGGTTCAGTCTCTTTCGAGATTAGAAGAAA
gi|575417|emb|X82544.1|          AGAATGAGATAGAGGAAC---CATCACAAGTGACTGTTGGAATGTCTAACAGATATGAAC
gi|461682445|gb|JX424318.1|      TGCTTGAACCAGGGAATG---CTGCTCTTGCTGTTGTTTCTGACT---CTAGTGACAGAT
                                           .   ..                                    : :  .*

gi|62736387|gb|AY914051.1|       ACGGACTACAACCAGACCCACTTCATCCGCAACCCCGAGTTCAAGGAAGCTGCCGACAAG
gi|33943625|gb|AY346329.1|       CGGCGGCTGGCGGCGGAGC---AGGTGCGGGCGCTGGAGCGGAGCTTCGAGGCGGACAAC
gi|308044466|ref|NM_001196644.1| AAGCGCTTCAGCGAGGAGC---AGATCAAGTCTCTCGAGTCCATGTTCGCCACGCAGACC
```

```
gi|334185982|ref|NM_001203162.1|      CCGTTGACGAAACCGAAGATGTTGTTGAGATTCAGAAACCGC---------AGAATCATC
gi|575417|emb|X82544.1|               CTGAAACAA---CTAAACGTATTGATAAGGTGCGTAGACGCC---------TTGCACAAA
gi|461682445|gb|JX424318.1|           CCAGAGACA---AAAACGGAGATCAAAAGACAATGCGTCGGC---------TTGCTCAAA
                                      . .           ...     :  : ..       .:                : *

gi|62736387|gb|AY914051.1|            ATGCAGGGCCCGCTCCGCCAGATCTTCGTCGAGTTCCTCGAGCGCTCCTGCACCGCC---
gi|33943625|gb|AY346329.1|            AAGCTGGACCCGGAGCGGAAGGCCCGGATCGCCCGCGACCTTCGCCTCCACCCTCGC---
gi|308044466|ref|NM_001196644.1|      AAGCTGGAGCCGCGCCAGAAGCTGCAGCTGGCGCGGGAGCTCGGCCTGCAGCCGCGC---
gi|334185982|ref|NM_001203162.1|      GAC-------------------------------------------------------
gi|575417|emb|X82544.1|               ACCGCGAGGCTGCTCGTAAAAGTCGTTTACGGAAGAAGGCCTATGTCCAGCAGTTGGAAA
gi|461682445|gb|JX424318.1|           ATCGTGAGGCTGCTAGGAAAAGTCGTTTGAGGAAAAAGGCATATGTTCAACAATTGGAGA
                                      .

gi|62736387|gb|AY914051.1|            ---GAGTTCTCCGGGTTCCTCCTCTACAAGGAGCTCGGCCGCAGGCT------CAAGAAA
gi|33943625|gb|AY346329.1|            ---CAGG---------TCGCCGTCTGGTTCCAGAACCGCCGCGCGAG------GTGGAAG
gi|308044466|ref|NM_001196644.1|      ---CAGG---------TCGCCATCTGGTTCCAGAACAAGCGCGCGCG------CTGGAAG
gi|334185982|ref|NM_001203162.1|      ------------GGCTCCCTGTTGATGATCAAGGGAAGAATCGGAATCGTGCTCCGTCGT
gi|575417|emb|X82544.1|               ATAGTAAACTGAAGCTGCTTCAGTTGGAACAAGAACTAGAACGTAATAGACAACAGGGTC
gi|461682445|gb|JX424318.1|           ACAGCAGGCTAAAGCTTACCCAGCTAGAGCAGGAGTTGCAACGAGCTCGTCAACAAGGCA
                                                     .      :  :   .*    . . *. ..         .

gi|62736387|gb|AY914051.1|            ACCAACCCGGTGGTGGCTGAGATCTTCTCGCTCATGTCCAGGGACGAGGCCCGGCACGCT
gi|33943625|gb|AY346329.1|            ACCAAGCAGATCGAGCGCGACTTCGCCGCCCTCCGCTCCCGCCACGACGCCCTCCGCCTC
gi|308044466|ref|NM_001196644.1|      TCCAAGCAGCTGGAGCGCGACTACTCCGCGCTCCGCGACGACTACGACGCGCTCCTCTGC
gi|334185982|ref|NM_001203162.1|      CTGATCCGGTTGATTCTTCAGCTCCTGTTGTTGTTGATCCTAATCAGTATCATGCGATTC
gi|575417|emb|X82544.1|               TGTATGTAGGTGATGGTTT---------------AGATGCTAGTCAGATAGGTTGCTCTG
gi|461682445|gb|JX424318.1|           TTTTTATATCTAGTTCAGC---------------AGACCA--GTCCCATTCC--------
                                         ::     * .:                          :*

gi|62736387|gb|AY914051.1|            GGGTTCTTGAACAAGGGGCTGTCCGACTTCAACCTGGCTCTGGACCTCGGCTTCTTGACC
gi|33943625|gb|AY346329.1|            GAGTGCGAC-------------------------------------------------------
gi|308044466|ref|NM_001196644.1|      AGCTACGAG-------------------------------------------------------
gi|334185982|ref|NM_001203162.1|      TTAAGA-----GCAAGCTCGAGCTTGCTTGCGCTGCTGTTGCTCGTCGTGTGGGAACTGT
gi|575417|emb|X82544.1|               GAACCGCAAATTCAGGAATAGCTTCTTTTGAAATGGAGTACGGCCATTGGGTGGAAGAGC
gi|461682445|gb|JX424318.1|           --ATGAGTGGAAATGGGGCGTTGGCTTTTGACACAGAGTACGCACGGTGGTTGGAAGAAC

gi|62736387|gb|AY914051.1|            AAGGCTAGGAAGTACACCTTCTTCAAGCCAGAGTTCATCTTCTACGCCACATACCTGTCC
gi|33943625|gb|AY346329.1|            ------------------GCCCTCCGCC------------------------------GC
gi|308044466|ref|NM_001196644.1|      ------------------TCCCTCAAGA------------------------------AG
gi|334185982|ref|NM_001203162.1|      GAA------ACC-------GGAAGATTCGAGTGCTTCAGCTAGCAATCAAAAACAAGCT-
gi|575417|emb|X82544.1|               AAG------ATAGACAAACAGATGATTTAAGGAATGCTCTGAACTCCCAAATGGGTGAAA
gi|461682445|gb|JX424318.1|           ACA------ATCGACAAGTTAATGAGCTGAGAGCTGCAGTTAATGCTCATGCAGGCGATA
                                                        : .

gi|62736387|gb|AY914051.1|            GAGAAGATCGGCTACTGGAGGTACATCACCATCTTCAGGCACCTAAAGG---CCAACCCG
gi|33943625|gb|AY346329.1|            GA------------------CAAGGACGCCCTCGCCGCCGAGATCGCCG---ACCTCCGG
gi|308044466|ref|NM_001196644.1|      GA------------------GAAGCACACGCTCCTCAAGCAGCTGGAGA---AGCTAGCC
gi|334185982|ref|NM_001203162.1|      --CAAGGCTCCATTGTGGCACAAACCTCACCTGGTGCTTCATCTGTTAGATTTTCTCCCA
gi|575417|emb|X82544.1|               TAGAATTGCGCATTCTTGTCGAGAGTT--GCTTGAATCAC-TATTTTGATCTCTTTCGCT
gi|461682445|gb|JX424318.1|           CTGAGCTGCGTAGTGTTGTTGAGAAGA--TCATGTCACAC-TATGATGAGATTTTTAAGC
                                                     :.       .:          .*    .      :.

gi|62736387|gb|AY914051.1|            GAGTACCAGGTGTACCCCATCTTCAAGTACTTCGAGAACTGGTGTCAGGACGAGAACCGG
gi|33943625|gb|AY346329.1|            GACAGGGTGGACGGCCAGATGTCC---------GTCAAGCTGGAGGCCGTGGCCG---CG
gi|308044466|ref|NM_001196644.1|      GAGATGCTGCACGAGCCGCGGGGCAAGTACAGCGGCAATGCGGACGCCGCCGGCG---CC
gi|334185982|ref|NM_001203162.1|      CAACAAGCACGCAAAAGAAACCTGATGTTC---CAGCCAGACAAACTAGTATTTC---AT
gi|575417|emb|X82544.1|               TGAAAGCTACAGCCGCAAATGCTGATGTTC---TCTACCTTATGTCTGGCACATG-----
gi|461682445|gb|JX424318.1|           AAAAAGGAAATGCAGCCAAAGCAGATGTCT---TTCATGTGTTATCAGGCATGTG-----
                                       .     .     . .                 .     .    *

gi|62736387|gb|AY914051.1|            CATGGCGATTTCTTCTCCGCGCTGCTCAAGGCGCAGCCGCAGTTCCTCAATGACTGGAAG
gi|33943625|gb|AY346329.1|            GACGAACACCAGCCGCCTCCGCCGCCGCCGCCGCCGCCACTGGCGTATAACAGCAAGGTG
gi|308044466|ref|NM_001196644.1|      GGGGACGACGT---------------GCGCTCGGGCGTCGGCGGCATGAA-GGACGAGTTT
gi|334185982|ref|NM_001203162.1|      CACGAGATGATTCTGATGACGATGATCTTGATGGAGACGCAGATAAT-------------
gi|575417|emb|X82544.1|               ---GAAGACATCAGCTGAGCGTTTCTTCTTGTGGATTGGGGGATTT--------------
gi|461682445|gb|JX424318.1|           ---GAAGACACCAGCTGAGAGGTGTTTCCTATGGCTTGGAGGTTTC--------------
                                         *.  :                       * .   .  *

gi|62736387|gb|AY914051.1|            GCCAAGCTCTGGTCACGCTTCTTCTGCCTCTCGGTGTATATAAC---------CATGTAC
gi|33943625|gb|AY346329.1|            GTGGACGGCTCGACGGACAGCGACTCGAGCGCGGTGTTCAACGAGGAGGCGTCGCCGTAC
gi|308044466|ref|NM_001196644.1|      GCAGACGCCGGGGCCGCGCCCTACTCGTCCGAGGGCGGTGGCGGTGGCAAGTTCGCGCAC
gi|334185982|ref|NM_001203162.1|      -------------GGAGATCCTACTGATGTGAAGCGTGCTAGGA-----------GGATG
gi|575417|emb|X82544.1|               -------------CGCCCCTCCGAACTTCTAAAGGTTCTCACGC-----------CACAT
gi|461682445|gb|JX424318.1|           -------------CGACCTTCTGAGCTTTTAAAGCTTCTTTCGA-----------CCCAA
                                                     *  .       ..*        .              :
```

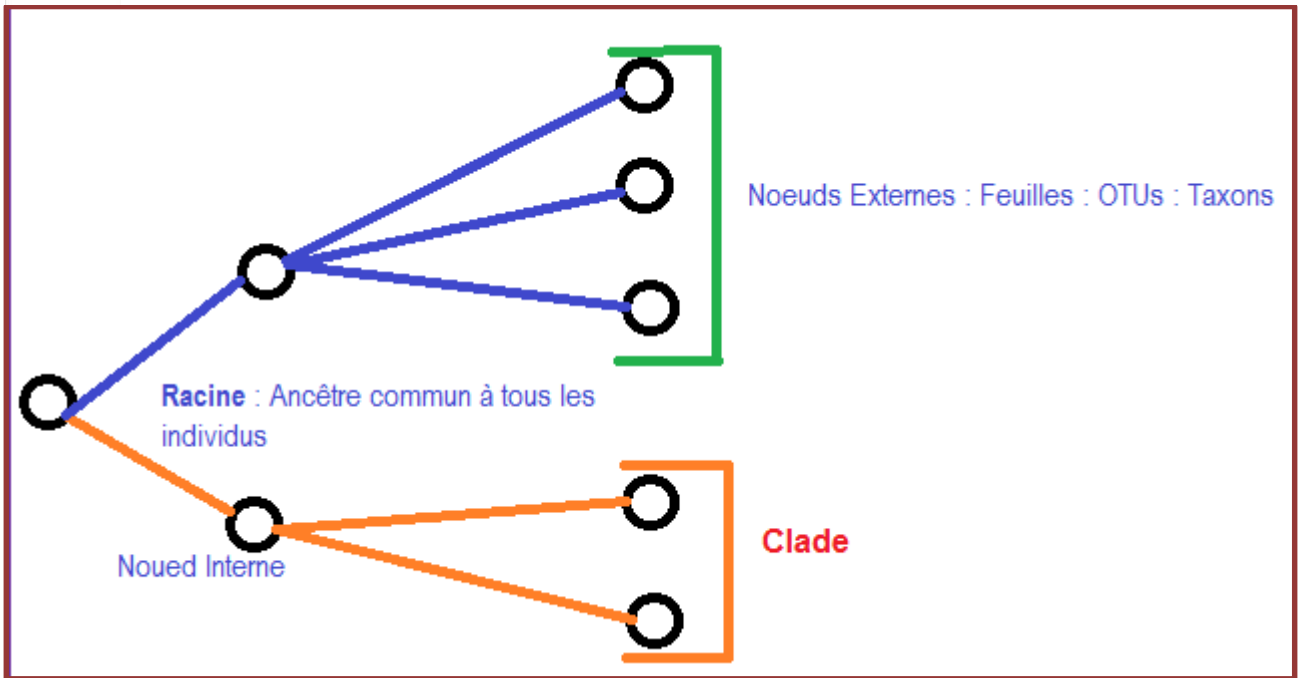
```
gi|62736387|gb|AY914051.1|          CTGAATGACTGCCAACGTAGTGCCTTCTACGAAGGAATTGGTCTCAACACCAAAGAATTC
gi|33943625|gb|AY346329.1|          T---CCGGCGCGGCCATCGACCACCACCACCACCAAACTCCGGCGAGCT----ACGACAC
gi|308044466|ref|NM_001196644.1|    TTCACGGACGACGACGTGGGAGCCCTCTTCCGGCCGTCGTCTCCGCAGC----CGAGCGC
gi|334185982|ref|NM_001203162.1|    CTCTCAAACCGAGAATCCGCTAGGCGCTCTAGGAG--AAGAAAGCAAGAACAAATGAATG
gi|575417|emb|X82544.1|             --GTGGAACCATTGTCAGATCAACAAATCCAGGAGGTTAGCAACCTCACCCAGTCTTGTC
gi|461682445|gb|JX424318.1|         --CTTGAACCCCTAACTGAGCAGCAGCTGTCAGGGATATGCAACCTTCAGCAATCATCAC
                                          ..*         .       .     .

gi|62736387|gb|AY914051.1|          GACATGCATGTCATCTATGAGACAAACCGCACGACGGCGAGGATCTTCCCT-----GCTG
gi|33943625|gb|AY346329.1|          GGCGGGGTTCACCTCCTTCTTCGCGCCATCCACCACGCTCACCTCGTCCCTCTCCTTCCC
gi|308044466|ref|NM_001196644.1|    CGCTGGCTTCACCTCGTCGGGGCCGCCGGAGCACCAGC---------CGTTCCAGTTCCA
gi|334185982|ref|NM_001203162.1|    AATTTGATACACAGGTAAAGTTTTT---ACCAATTGTAGCTGCTGTTGACAACAGAATTC
gi|575417|emb|X82544.1|             AG---CAGGCAGAAGATGCGTTGTC---CCAAGGAATGGTAAAACTCCATCAGATTCTTG
gi|461682445|gb|JX424318.1|         AA---CAAGCTGAGGATGCTCTTTC---ACAAGGAATGGAGGCTCTTCAGCAGTCTTTGG
                                     .       : .   :           . .
```

Constatons qu'il y a moins de gap et bien plus d'identités. Nous pouvons également utiliser des séquences protéiques pour réaliser un alignement multiple en vue d'une construction phylogénétique. Pour cela, il faut un jeu de séquences appartenant à la même famille protéique. Donc des séquences de forte homologie et de fonction identique : Prenons l'exemple d'une série de séquences de la nucléase. Le résultat de l'alignement multiple de cette série de séquences est à la fin de ce document.

L'alignement multiple peut être amélioré en éliminant les séquences de *Saccharomyces cerevisiae, Gallus gallus et Xenopus laevis.* Le résultat va donc concerner les espèces *Homo sapiens, Mus musculus, Rattus norvegicus* et *Bos taurus (voir fin du document).*

La structure d'un arbre phylogénétique

Définition : Un arbre est un graphe non cyclique constitué de plusieurs **nœuds** qui sont les unités taxonomiques évolutives (**OTUs : Operational Taxonomic Unit**) c'est-à-dire les individus à partir desquels nous avons récolté les données sous forme phénotypiques tels que les caractères biochimiques bactériens ou génotypiques comme les séquences d'un gène particulier. Ces nœuds sont reliés entre eux par des **branches** dont la signification est relative aux types de données utilisées pour tracer la phylogénie. Dans le cas de données moléculaires par exemple (séquences), la longueur est proportionnelle aux distances qui séparent ces individus. Les nœuds internes sont des unités inexistantes en fait, mais des ancêtres hypothétiques qui permettent l'interprétation de l'arbre.

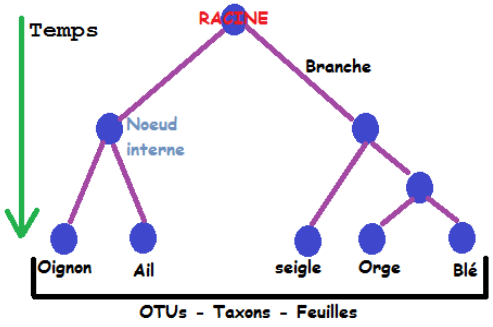


Dans cet arbre, il y a :

- Une racine : supposée l'ancêtre de tous les individus ou OTUs
- Deux nœuds internes qui représentent les ancêtres théoriques des OTUs
- Cinq OTUs qui sont les individus réels sur lesquels ont été collectées les données.

Les méthodes de construction d'arbres donnent des arbres sans racine. Mais il est toujours utile d'enraciner l'arbre.

Rôle de la racine dans un arbre ?
Il faut savoir que la racine fait partie des OTUs avec lesquelles nous construisons l'arbre. Ce n'est pas un ancêtre de quelques centaines de millions d'années. Sauf que cet individu est un peu "*spécial*" par rapport au reste des OTUs. En effet, l'individu qui représentera la racine de l'arbre est choisi de sorte qu'il soit *intermédiaire* : ni très différents ni très identique à l'ensemble des OTUs. Ainsi, toutes les OTUs seront comparées entre elles et avec la racine pour tracer l'arbre phylogénétique. Pour ces raisons, la racine est appelée le groupe externe ou out group. Dans un arbre non raciné, seule la comparaison entre les OTUs a lieu. En fait, il n'y a pas de dimension temporelle dans un arbre non raciné. La racine apporte à l'arbre une dimension temps au cours de laquelle les différents états des caractères auront apparus.



Nombre d'arbres : Le nombre théorique d'arbres phylogénétiques dépend du nombre d'OTUs qui entrent dans la construction phylogénétique.

$$N_{\text{arbres non-enracinés}} = \prod_{i=3}^{S} (2i - 5) \qquad N_{\text{arbres enracinés}} = \prod_{i=2}^{S} (2i - 3)$$

S : étant le nombre d'individus ou de taxons

Exemple pour S = 4 taxons

- N non raciné = (2x3 – 5) x (2x4 – 5) = (6-5) x (8-5) = 1x3 = 3 arbres possibles
- N raciné = (2x2 -3) x (2x3 – 3) x (2x4 – 3) = 1 x 3 x 5 = 15 arbres possibles

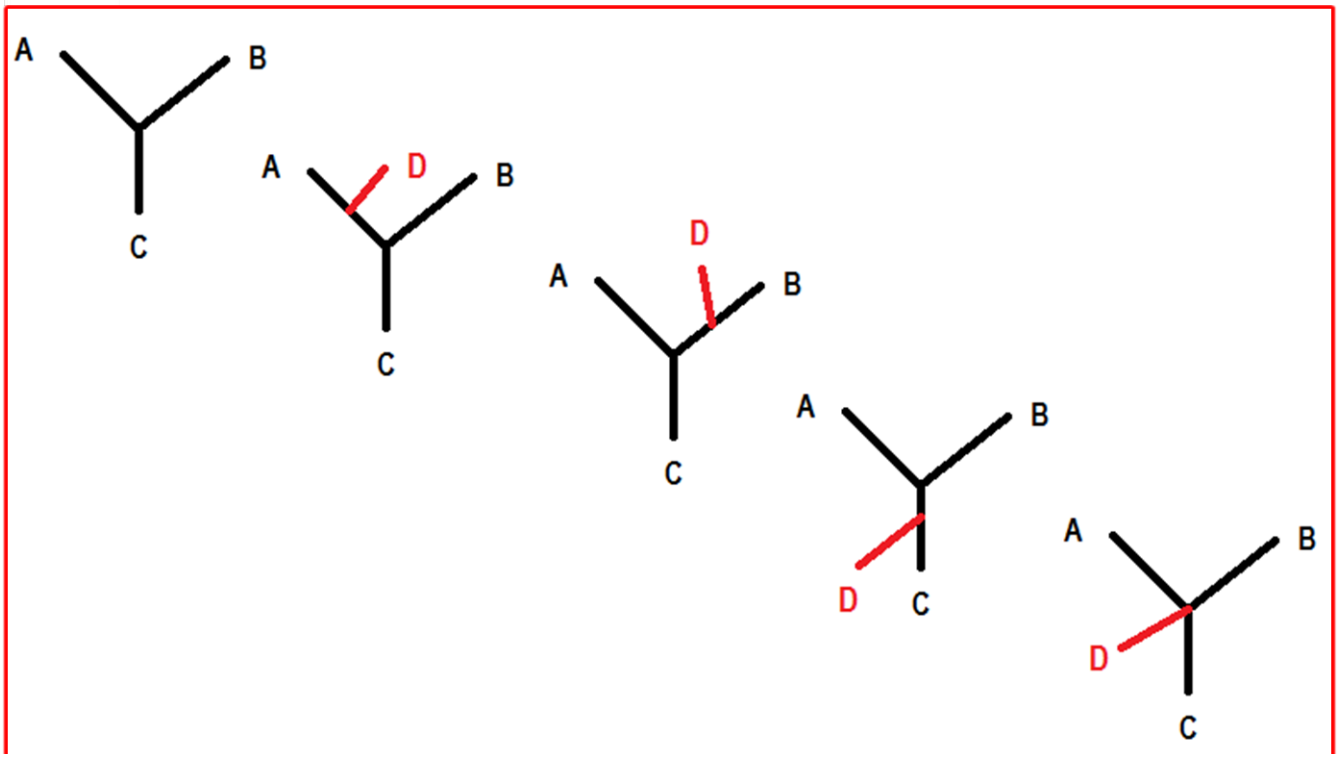
Exemple pour S = 5 taxons

- N non raciné = (2x3 – 5) x (2x4 – 5) x (2x5 – 5) = 1x3x5 = 15 arbres possibles
- N raciné = (2x2 -3) x (2x3 – 3) x (2x4 – 3) x (2x5 – 3) = 1 x 3 x 5 x 7 = 105 arbres possibles

Nombre d'OTUs	Nombre d'arbres non racinés	Nombre d'arbres racinés
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10 395
8	10 395	135 135
9	135 135	2 027 025
10	2 027 025	34 459 425

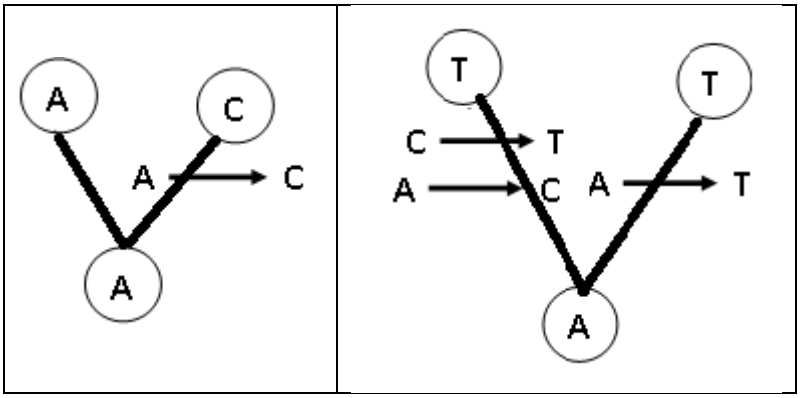
Le nombre d'arbres non racinés pour S = 4 est égal à trois. D'où viennent ces trois arbres ? En fait, c'est un nombre théorique qui exprime toutes les combinaisons ou toutes les topologies possibles entre les quatre individus ; sauf que le résultat sera un et un seul arbre après traitement des données dont nous disposons. L'arbre obtenu d'après nos données doit ressembler à l'un des trois combinaisons théoriques. Cela voudrait dire qu'avec un autre jeu de données et pour S=4 on peut obtenir une autre structure d'arbre qui fait partie toujours des combinaisons possibles : Il y a quatre possibilités pour insérer le quatrième taxon à partir d'un arbre de trois individus :

1. Sur la branche de A
2. Sur la branche de B
3. Sur la branche de C
4. Sur le nœud principal des trois branches A, B et C



Notions de distances : La distance évolutive (notée d) entre deux séquences est une fonction du temps et de la vitesse d'évolution λ de deux séquences[3]. L'unité de la distance évolutive est le nombre total de substitutions par site et rapportée à la longueur des deux séquenques alignées. La distance évolutive mesure la dissimilarité (la non identité) entre les séquences. Elle prend toujours des valeurs supérieures ou égales à zéro : $d \in [0 , \infty[$

Dans le cas des séquences nucléiques, l'information sur l'histoire des mutations de type substitutions, insertions ou déletions, n'est pas vraiment claire du fait que pour un site donné nous n'observons qu'une seule mutation alors que par rapport au temps d'évolution et sur ce même site, beaucoup de mutations auraient pu avoir lieu comme le montrent les schémas suivants:



[3] (http://www.frangun.org)

Dans le premier cas, les individus diffèrent avec moins de changements évolutifs comparés au deuxième cas dans lequel nous observons trois changements mais aucune différence chez les taxons ! Les deux individus possèdent un T dans le même site, mais leur histoire évolutive n'est pas la même. Il faut donc manipuler avec soin les données que nous manipulons pour tirer des interprétations adéquates.

La comparaison entre le nombre de substitutions réelles et celles observées est résumée sur cet exemple[4] :

	Séq 1	Séq 2	Substitutions observées	Substitutions réelles
Substitution unique	T	T → A	1	1
Substitutions multiples	A	A → G → T	1	2
Substitution par allèle	A → C	A → G	1	2
Substitution coïncidant au même site	T → A	A → T	0	2
Substitution convergentes	A → C → T	A → T	0	3
Substitutions inverses	T → A → T	T	0	2

Il existe plusieurs types de distances. Elles peuvent signifier un état de réarrangements génomiques telles que les substitutions ou les inversions. Les distances entre les séquences sont estimées de plusieyrs façons :

- Matrice de substitution Pam 250
- Matrice de substitution Blosum 62
- Distance de Hamming

La distance consiste à compter le nombre de nucléotides différents entre deux séquences, mais cela ne renseigne pas sur la vraie estimation des mutations en un site donné au cours du temps omme cela est indiqué dans l'exemple précédant. Il faut donc apporter des "*modifications*" aux valeurs de ces mesures : on parle alors de ***distance corrigée***.

Les méthodes de corrections basées sur l'hypothèse de régularité du processus évolutif[5] sont nombreuses mais présentent quelques traits de différences : Variation de la fréquence des nucléotides, Type de substitution, probabilité de substitution, …

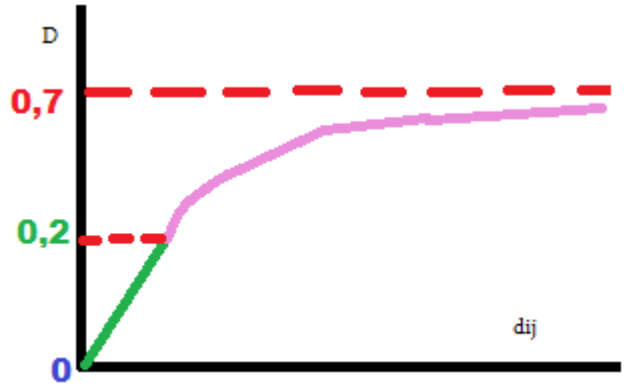
a. **Méthode de Jukes – Cantor (1969) :** suppose que les quatre nucléotides ont les mêmes fréquences (p(A) = p(C) = p(G) = p(T)) et leur substitutions équiprobables : La probabilité α de transition et la probabilité β de transversion sont égales : α = β.

$$dij = -\frac{3}{4} Ln(1 - \frac{4}{3} D)$$

[4] Abdoulaye Baniré Diallo. Université du Québec. Montréal.

[5] *www7.inra.fr/internet/Projets/agroBI/PHYLO/Gouy.pdf*

Où D : représente la distance observée et calculée à partir de l'alignement. Elle exprime le nombre de nucléotides différents entre les séquences i et j ou le pourcentage de différence entre ces séquences.
La correction de dij a une limite au-delà de laquelle la correction est quasi-impossible : lorsque D = 3/4 = 75%.



Exemple 1 : Soient les deux séquences suivantes

T	A	C	G	T	A	**A**	**G**	G	**T**	C	C	A	**G**	T
T	A	C	G	T	A	**C**	**A**	G	**C**	C	C	A	**T**	T

Il y a quatre différence sur une longueur de 15 sites. D = 4/15 = 0,267.

La correction de cette distance D est :

$$dij = -\frac{3}{4} Ln\left(1 - \frac{4}{3} D\right) = -\frac{3}{4} Ln(1 - \frac{3}{4}(0,267))$$

Dij = 0,167

La matrice des taux de substitution entre les quatre nucléotides est déduite de l'hypothèse que les sites de l'alignement sont équivalents et les mutations sont équiprobables :

	A	C	G	T
A	1-D			
C	D/3	1-D		
G	D/3	D/3	1-D	
T	D/3	D/3	D/3	1-D

Exemple 2 : Corriger la distance D retrouvée égale à 0,05 entre deux séquences alignées i et j.
Réponse : Dij = -3/4 Ln(0,9334) = 0,0517

Exemple 3 : Même question avec D observée égale à 0,5.
Réponse : Dij = -3/4 Ln(1/3) = 0,824

b. **Méthode de Tajima - Nei :** Les fréquences des nucléotides ne sont pas égales comme dans le modèle Jukes – Cantor (p(A) ≠ p(C) ≠ p(G) ≠ p(T)) ; mais les transitions et les tranvesrsions sont équivalentes (α= β) . La distance dij est corrigée selon la formule : $dij = -B\,Ln(1-\frac{D}{B})$**.** Le paramètre B dépond de la nature des bases qui composent les séquences $B = 1 - \Sigma q_i^2$ qi étant la somme des fréquences des nucléotides ($pA^2 + pC^2 + pG^2 + pT^2$)

c. **Modèle de Kimura à deux paramètres:** Dans ce modèle, les paramètres α et β sont différents (α≠ β) ; les transitions et les transversions ne sont pas équivalentes en termes de proportions. Les transitions étant plus nombreuses :

$$dij = -\frac{1}{2}Ln(1-2P-Q) + \frac{1}{4}Ln(1-2Q)$$ *Formule (a)*

La matrice de substitutions déduite du modèle de kimura avec : P = fréquence des transitions et Q = fréquence des transversions

	A	C	G	T
A	1-P-Q			
C	Q/2	1-P-Q		
G	P	Q/2	1-P-Q	
T	Q/2	P	Q/2	1-P-Q

Ou encore :

$$dij = -\frac{1}{2}Ln(1-2P-Q)\sqrt{(1-2Q)}$$ *Formule (b)*

Dans le cas des séquences protéiques, on suppose que les résidus des différents sites de la séquence protéique évoluent indépendamment les une des autres. La coorection des distances est donnée par l'approximation de Kimura :

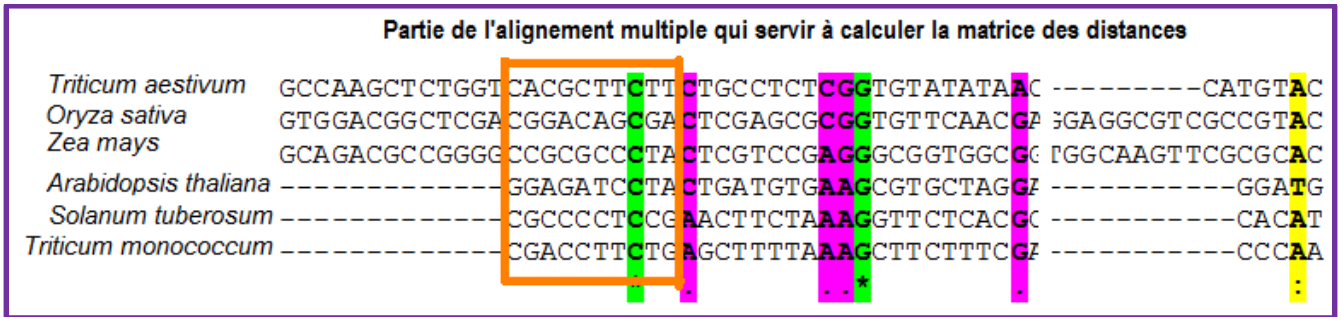
$$dij = -Ln(1-p-0,2\,p^2)$$ *Formule (c)*

Où p : Fraction des différences observées.
Exemple[6] : La correction des distances, au dessus de la diagonale, avec la méthode de Kimura (Formule a) donne :

	Ind 1	Ind 2	Ind 3	Ind 4	Ind 5
Ind 1		*0,088*	*0,103*	*0,160*	*0,181*
Ind 2	0,094		*0,106*	*0,170*	*0,189*
Ind 3	0,111	0,115		*0,166*	*0,189*
Ind 4	0,180	0,194	0,188		*0,188*
Ind 5	0,207	0,218	0,218	0,216	

[6] *www7.inra.fr/internet/Projets/agroBI/PHYLO/Gouy.pdf*

L'alignement multiple des séquences parmet d'obtenir une matrice de distances qui va servir à la construction de l'arbre phylogénétique : Nous allons choisir un segment, juste à titre d'exemple qui va servir pour le calcul de la matrice des distances :



Le segment d'alignement choisi (rectangle orange) comprend 10 positions et six espèces. Nous pouvons le reprendre comme suit :

	1	2	3	4	5	6	7	8	9	10
T.aestivum	C	A	C	G	C	T	T	C	T	T
O. sativa	C	G	G	A	C	A	G	C	G	A
Z. mays	C	C	G	C	G	C	C	C	T	A
A. thaliana	G	G	A	G	A	T	C	C	T	A
S. tuberosum	C	G	C	C	C	C	T	C	C	G
T. monococcum	C	G	A	C	C	T	T	C	T	G

Nous allons calculer la distance (en %) entre les séquence de *T.aestivum* et *O. sativa*. Pour cela déterminons le nombre de mutations entre les 10 positions de cet alignement, en sachant que la position numéro huit de cet alignement est un site conservé : tous les individus possèdent un C à cette position. Les mutations concernent sept sites: 2, 3, 4, 6, 7, 9 et 10. Il y a donc sept mutations sur un total de 10 sites ; cela fait 70% de différence en tre les deux individus *T.aestivum* et *O. sativa*. La distance d = 70%.

Avec le même raisonnement, calculons les distances entre l'ensemble des individus et portons le résultat sous forme d'une matice symétrique avec deux parties Nord et Sud par rapport à la diagonale. La distance d'un individu i avec lui-même étant nulle :

	T.aestivum	*O. sativa*	*Z. mays*	*A. thaliana*	*S. tuberosum*	*T. monococcum*
T.aestivum						
O. sativa	70					
Z. mays	70	60				
A. thaliana	60	70	60			
S. tuberosum	50	60	60	80		
T. monococcum	40	60	60	50	30	

Un autre type de distance peut être calculé mais en partant de données autres que les séquences alignées. C'est le cas des données qui concernet les caractères de nature phénotypique. Dans ce cas, le caractère ne possède que deux états codés 1 ou 0 pour

désigner la présence ou l'absence respectivement. Par exemple, ce tableau résume les états de quelques caractères pour un ensemble d'individus comme dans le cas d'une taxonomie numérique où on peut s'intéresser à la biotypie, la lysotypie, sérotypie, zymotypie, … :

	Mobilité	Aérobiose	Uréase	Lactose	Indole	Gaz	Arg	ODC	LDC	ONPG
Taxon 1	1	1	1	1	1	1	0	0	1	0
Taxon 2	0	0	1	0	1	1	1	1	0	1
Taxon 3	1	0	0	1	0	0	0	1	1	1
Taxon 4	1	1	0	0	1	0	0	0	1	1
Taxon 5	0	0	1	1	1	0	1	1	1	1

Ce système de codage donne des problèmes d'importance (utilisation de plus ou moins de chiffres). Or l'implication génétique n'est pas fortement liée à l'importance de ces caractères[7]. Il faut ajouter à cela la grande variabilité d'indices utilisés en fonction de la nature des données observées sur les individus. Ces données peuvent être de nature binaire, ordinale ou même quantitative et le traitement de ce type de données passe par le calcul d'indice de similarité (de "*ressemblance*" noté s) pour le convertir ensuite en matrice de distances selon la relation : d = 1 – s.

Le principe de calcul de la similarité pour les données binaires se base sur le dénombrement des différents états de caractères entre les différentes paires d'individus. En comparant ces états de caractères, on se rend compte des possibilités suivantes :

(1,1) : le caractère est présent chez les deux individus
(1,0) : le caractère est présent chez le premier individu et absent chez le deuxième
(0,1) : le caractère est absent chez le premier individu et présent chez le deuxième
(0,0) : le caractère est absent chez les deux individus

Dans l'exemple cité plus haut on peut mesurer entre le taxon 1 et le taxon 2:

- La paire (1,1) trois fois
- La paire (1,0) quatre fois
- La paire (0,1) trois fois
- La paire (0,0) zéro fois. Cependant, elle se répète trois fois entre les taxons 3 et 4.

Pour calculer l'indice s, deux méthodes se distinguent à cause du double zéro. Certains prennent en considération le nombre de fois où la paire (0,0) est observée entre deux individus : on parle alors d'indice symétrique ; alors que d'autres considèrent que cette double absence peut biaiser la similarité du fait que la présence (notée 1) est démontrable alors que l'absence est plus difficile à démonter donc moins pondérable et on parle dans ce cas d'indice asymétrique.

1. Indice de simple concordance (symétrique) : $S = \frac{(1,1)+(0,0)}{(1,1)+(1,0)+(0,1)+(0,0)}$
2. Indice de Jaccard (asymétrique) : $S = \frac{(1,1)}{(1,1)+(1,0)+(0,1)}$

[7] http://fdanieau.free.fr/cours/A1/microbiologie/chapitre9/Taxonomie.pdf

Si l'on considère l'équation de l'indice de Jaccard, la similarité entre les cinq taxons est :

	Taxon 1	Taxon 2	Taxon 3	Taxon 4	Taxon 5
Taxon 1					
Taxon 2	0,30				
Taxon 3	0,33	0,22			
Taxon 4	0,50	0,22	0,43		
Taxon 5	0,40	0,63	0,50	0,33	

Dans le cas des données quantitatives, l'indice asymétrique de similarité (indice de Steinhaus) est de la forme :

$$s = 2 * \frac{W}{A + B}$$

Avec :

- W : La somme des minima des valeurs mesurées entre deux individus
- A : La somme des valeurs observées pour l'individu 1
- B : La somme des valeurs observées pour l'individu 2

Soit l'exemple dans lequel on a mesuré l'abondance d'uen variété dans deux régions (individus) différentes sur plusieurs parcelles :

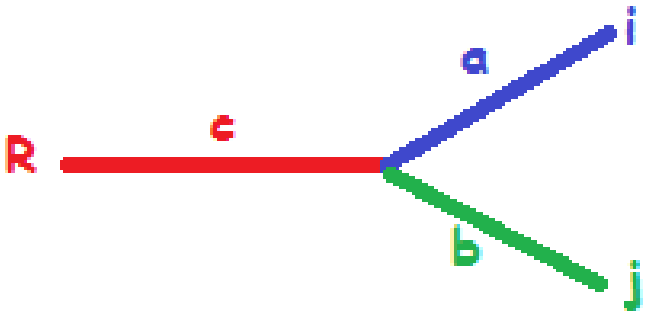
	Présence des variétés						A	B	W
Région 1	15	11	3	12	14	7	62		
Région 2	17	9	5	16	10	2		59	
Minima	15	9	3	12	10	2			51

$$s = 2 * \frac{51}{62 + 59}$$

s = 0,8429 soit 84,29 % de similarité dans la distribution variétale entre ces deux régions géographiques et d = 1 - 0,8429 = 0,1571 ou les deux régions sont distantes de 15,71 % par rapport à la distribution variétale.

Les types de distances :

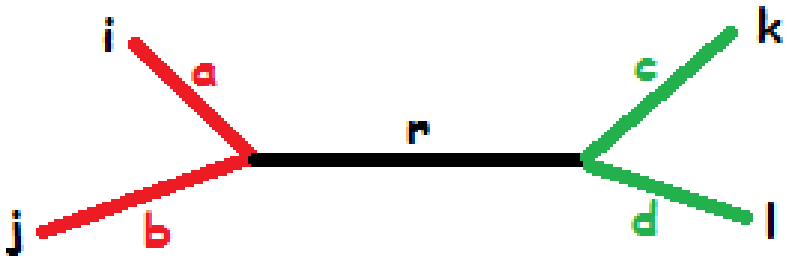
1. Distances ultramétriques : Ce modèle suppose que la distance entre un nœud (i,j) et la racine R est toujours la même : principe des trois points :



$d(i,j) \leq d(i,R) = d(j,R)$

$a+b \leq a+c = b+c$

2. Distances additives : Soient deux individus (i et j) appartenant à un même clade et deux autres individus (k et l) appartenant un deuxième clade. La distance entre cette paire de nœuds est égale à la somme de la longueur des branches séparant ces clades principe des quatre points :

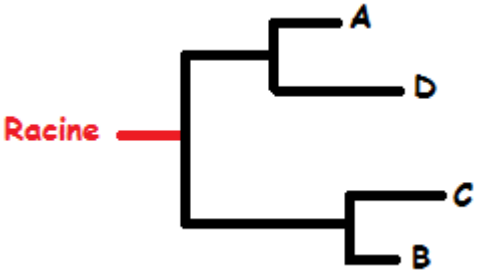
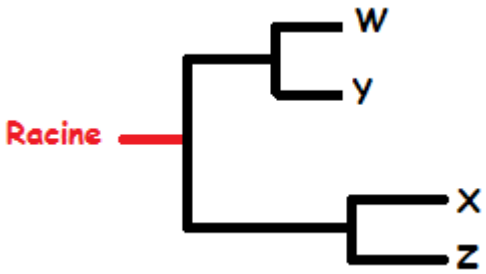


a, b, c, d et r sont les longueurs des différentes branches. Les individus i, j, k et l sont les OTUs.

$d(i,j) + d(k,l) \leq d(i,k) + d(j,l) = d(i,l) + d(j,k)$

$(a+b) + (c+d) \leq (a+r+c) + (b+r+d) = (a+r+d) + (b+r+c)$

En fonction des distances utilisées, on obtient les types d'arbres suivants :

Arbre Additif : la longueur des branches depuis la racine vers les OTUs n'est pas la même	**Arbre Ultramétrique :** la longueur des branches depuis la racine vers les OTUs est la même
	

Partie 2 : LES METHODES DE CONSTRUCTION D'ARBRES PHYLOGENETIQUES

La construction d'arbres se base sur deux méthodes différentes. Ces différences sont, en fait, dues à la nature des données dont on dispose et à la manière de les utiliser. Ainsi on peut distinguer les méthodes:

Les méthodes phénétiques qui se basent sur les ressemblances globales en comparant un maximum de caractères qui peuvent être de natures morphologiques codés en 1 et 0, des fréquences alléliques, des séquences de protéines ou de gènes, …

Le degré de parenté entre les taxons est apprécié par l'importance de la ressemblance globale entre ces individus pris deux à deux. A partir de cette similitude globale est apparu le concept fort important qui est la notion de distance. Si les données sont des séquences alignées, on s'intéresse au nombre de mutations (substitutions) comme on peut s'intéresser à la dissimilarité pour le cas des caractères phénotypiques binaires par exemple. Les concepts fondamentaux des méthodes phénétiques ont été établis par Sneath et Sokal dans Numerical taxonomy éditée en 1973 :

- Les relations entre les taxons sont phénétiques et non phylogénétiques,
- La qualité de la classification des taxons est fonction du nombre de caractères étudiés
- Les caractères étudiés ont tous la même pondération
- La construction de l'arbre se fait par un phénogramme
- Les inférences phylogénétiques se font sur la base d'un ensemble d'hypothèses sur les mécanismes évolutifs.

Les méthodes cladistiques (William Hennig dans les années 1960) : se basent sur l'analyse des caractères en identifiant leurs états plésiomorphe et apomorphe ; c'est-à-dire les caractères primitifs de ceux dérivés. Le principe de cette méthode vise à mesurer le degré de parenté entre deux individus sur la présence de caractères apomorphes qui soient communs à ces individus. Ces caractères communs issus d'un ancêtre commun sont donc homologues.

Dans la méthode cladistique, les caractères homologues sont déterminés en fonction de plusieurs critères :

- La similarité entre les paires de taxons
- Les caractères homologues ne peuvent coexister chez le même taxon
- L'homologie des caractères aboutit à des arbres superposables ou congruents.

En termes de séquences moléculaires (ADN ou protéines), l'homologie prend un autre sens. Faut-il parler de similitude ou d'homologie en cas de ressemblance entre les séquences ?
La réponse est donnée par l'analyse bioinformatique des séquences en passant par un alignement multiple, un Blast, recherche de motifs et de domaines communs, …

L'homologie est plutôt une propriété non mesurable contrairement à la similarité. Elle a une connotation génétique. Si le taux de similarité est important (significatif), les séquences peuvent alors être considérées comme homologues, donc supposées descendre d'un même ancêtre. Cependant, si ce même taux est faible, cela n'implique en rien une non-homologie ; car des séquences homologues peuvent bien être non similaires. Le minimum à retenir pour

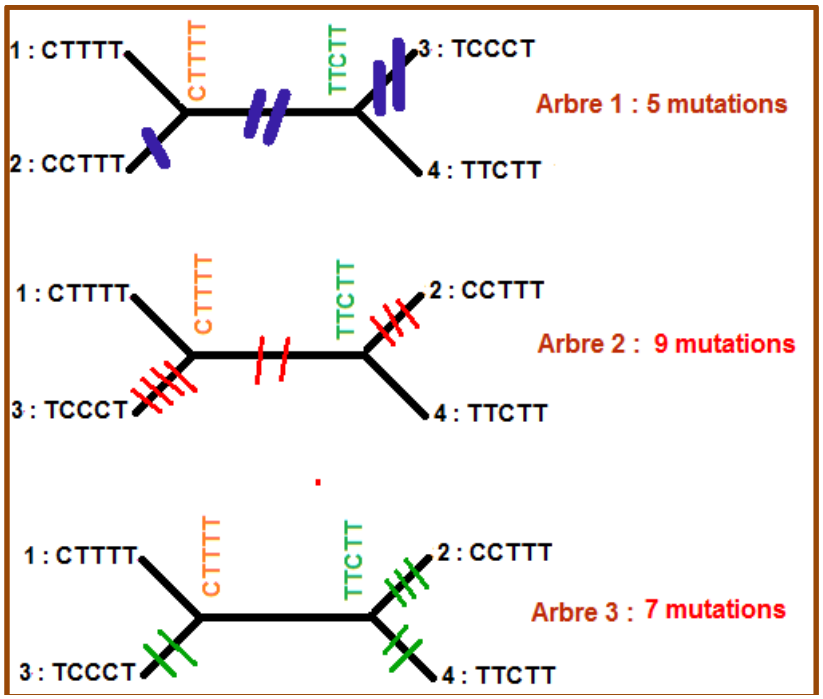
une homologie serait de 20 – 25%[8] de similarité calculée après alignement. La difficulté majeure avec les données moléculaires réside dans l'obtention d'alignements équivalents conduisant à des arbres différents. Pour cela, il faut savoir interpréter ces alignements en se basant sur les notions de délétion/Insertion et mutations.

Les méthodes du maximum de parcimonie

Le prinicipe fondamental des méthodes de parcimonie consiste à diminuer au maximum le nombre d'évènements génétiques (mutations, substitutions) pour comparer deux séquences ou deux individus. L'arbre qui en découle compte le moins de pas évolutifs, c'est-à-dire le moins de mutations possibles ; c'est l'arbre optimal.
Dans l'exemple suivant, nous allons rechercher l'arbre le plus parcimonieux parmi un ensemble d'arbres. Pour cela, on suppose que les quatre individus possèdent les séquences suivantes :

Ind1	C	C	T	C	C
Ind2	C	C	C	T	C
Ind3	C	C	C	C	T
Ind4	C	C	T	T	T

Le nombre d'arbres non racinés théorique entre ces quatre individus est égal à trois selon la formule donnée plus haut :



[8] http://rna.igmors.u-psud.fr/gautheret/cours/homologie.html

Pour déterminer l'arbre le plus parcimonieux, c'est-à-dire celui qui offre le moins de pas évolutifs (mutations) par rapport à la séquence ancestrale théorique. Dans notre cas, nous allons supposer les deux séquences CCTCC et CCTTT comme référence ancestrale avec lesquelles nous allons calculer le nombres de mutations par comparaison aux quatre individus :

Dans le cas de l'arbre 1, le nombre de mutations entre l'individu 2 et la séquence repère est égal à 1. Entre les deux séquences repères CTTTT et TTCTT, il y a deux mutations. Entre la séquence repère TTCTT et l'individu 3, il y a deux mutations. Au total, il y a :
1 + 2 + 2 = 5 mutations.

En refaisant les calculs pour les cas des arbres 2 et 3, le nombre de mutations est neuf pour l'arbre 2 et sept pour l'arbre 3. Donc l'arbre le plus parcimonieux, parmi les trois, est donc l'arbre 1. Il considère qu'il y a eu moins de mutations entre les individus ayant contribué à la topologie de cet arbre.

Application : trouvez l'arbre le plus parcimonieux pour les quatre séquences suivantes :

	1	2	3	4	5	6	7	8
1	G	A	A	T	A	G	C	C
2	G	C	T	T	C	G	C	A
3	G	G	T	T	C	G	C	A
4	G	T	T	T	G	G	C	A

Maximum de parcimonie et sites informatifs : L'alignement multiple effectué, on peut trouver des sites dans lesquels au moins deux individus sont dans un même état (même nucléotid) et deux autres sont dans un autre état différent des deux premiers. Dans les exemples suivants, les sites informatifs sont colorés.

Exemple 1 : Au niveau du troisième site qui est informatif, **Ind1 et Ind2** sont en état **E** et les **individus 4 et 5** sont dans l'état **K.** Trouvez les autres sites informatifs

Ind 1	L	A	E	V	E	D	V	Q
Ind 2	L	A	E	V	E	E	M	Q
Ind 3	L	G	E	I	D	E	V	Q
Ind 4	F	D	K	I	E	E	P	Q
Ind 5	W	G	K	I	G	D	P	Q

Exemple 2 : Les six sites de l'alignement sont informatifs

Ind 1	A	C	C	C	G	A
Ind 2	C	T	C	T	G	A
Ind 3	A	T	C	C	A	C
Ind 4	G	T	T	C	C	C
Ind 5	G	C	T	T	C	T
Ind 6	A	C	T	T	G	T

Exemple 3 :

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
I	L	I	I	I	V	L	L	L	I	C	I	C	I	V	L	L	I	L	V
C	M	M	V	L	I	I	V	C	M	C	V	M	C	L	I	C	C	M	M
V	C	M	C	M	L	C	C	C	L	I	V	M	I	I	L	C	I	I	I
C	M	C	M	V	M	L	L	I	M	I	C	L	C	L	M	I	M	M	C
M	V	C	C	I	C	V	M	V	I	C	I	M	M	C	C	V	L	I	V
		*								*	*		*					*	

Les sites informatifs sont utiles pour construire une phylogénie après un alignement multiple. Ils permettent de déduire les données de construction des différentes topologies pour déduire à la fin l'arbre le plus parcimonieux. Les sites informatifs 3, 11, 12 et 19 de cet exemple peuvent servir à construire une phylogénie au lieu des 20 positions de l'alignement.

Exemple 4 :A partir d'un alignement multiple, remplacez tous les nucléotides par la valeur 1 et les Indels par zéro. Vous obtenez ainsi une matrice binaire :

ACCTAAAGG---CC
AGCTGGAGA---AG
-TATTTTGATCTCT
-TATGATGAGATTT

1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	1	1	1	1	1	1	1	1	0	0	0	1	1
1	1	1	1	1	1	1	1	1	0	0	0	1	1
0	1	1	1	1	1	1	1	1	1	1	1	1	1
0	1	1	1	1	1	1	1	1	1	1	1	1	1

Certains sites de cette matrice codée sont informatifs. Lesquels?

Le Bootstrap : La construction des arbres phylogénétiques est basée sur hypothèses évolutionistes qui nécessite des corrections. Cela induit des erreurs au niveau topologiques et conduit à des erreurs d'interprétation. Pour y rémédier, il existe une méthode, le bootstrap, qui permet de vérifier la robustesse et la fiabilité de l'arbre obtenu avec telle ou telle méthode de construction.

Le prinicpe général du bootstrap est de réaliser des modifications sur les colonnes de l'alignement en les déplaçant aléatoirement de leurs sites initiaux pour retrouver le même arbre car en permutant les colonnes de leurs sites initiaux, nous supposons que ceux-ci évoluent de manière indépendante. A partir de l'alignement multiple principal, on construit de nouveaux alignements qui sont obtenus en permutant les colonnes (les sites). Pour chaque nouvel alignement, on construit une matrice des distances, puis l'arbre correspondant.

Les étapes principales du bootstrap sont les suivantes :

1. à partir des séquences réaliser l'alignement multiple global
2. construire l'arbre qui en découle et interpréter ses nœuds
3. à partir de l'alignement multiple global, remplacer, par permutations, plusieurs colonnes aléatoirement de leurs positions initiales.
4. Ces alignements nouveaux, déduits des permutations des colonnes vont permettre de construire autant d'arbres.
5. Calculer le nombre de fois où le nœud de l'arbre principal est retrouvé sur les arbres secondaires construits par les permutations de l'alignement principal. Le nombre de permutations réalisé au hazard à partir de l'alignement principal est 1000 ! d'où la nécessité d'un programme informatique.
6. Le % qui représente la concordance d'un nœud retrouvé en même temps sur l'arbre principal et sur les 1000 arbres construits à partir des permutations dans l'alignement principal représent la valeur de bootstrap. C'est le % de réplication d'un même nœud.
7. Les branches de l'arbre principal doivent être soutenues par les arbres secondaires par une valeur de bootstrap ≥ 90% pour quelles soient significatives.

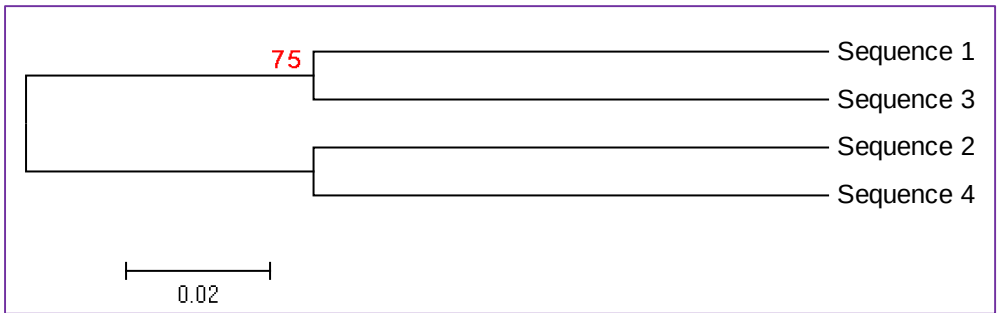
Le bootstrap par l'exemple :

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
G	N	L	V	R	T	E	P	V	C	D	G	Q	Y	L
G	S	L	I	R	T	E	H	V	C	D	G	Q	Y	L
G	S	L	V	R	T	E	P	V	C	D	G	H	Y	L
G	S	L	V	R	T	E	H	V	C	D	G	Q	Y	I

Arbre principal avec sa série de nœuds.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
G	N	L	V	R	D	G	Q	V	C	T	E	P	Y	L
G	S	L	V	R	D	G	Q	I	C	T	E	H	Y	L
G	S	L	V	R	D	G	H	V	C	T	E	P	Y	L
G	S	L	V	R	D	G	Q	V	C	T	E	H	Y	I

1000 arbres secondaires (artificiels) à partir de 1000 permutations aléatoires



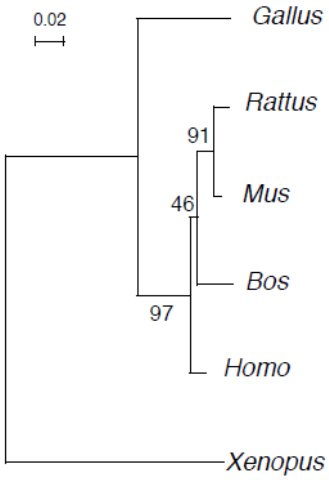
Dans cet arbre, la valeur de bootstrap du nœud formé par les individus 1 et 3 est égale à 75 %. Cela veut dire que sur les 1000 arbres artificiels construits par les permutations aléatoires, nous retrouvons le nœud ou clade (séquence 1,séquence 3) dans 750 fois. Cette valeur n'est synomyme d'une bonne significativité statistique.

Application : Construisez 10 matrices de distances après avoir réalisé des permutations aléatoires sur l'alignement multiple suivant :

1	V	I	N	L	R	H	F	K	D	V
2	V	L	A	L	P	N	L	K	D	A
3	V	I	N	L	P	N	F	K	D	V
4	V	I	T	L	P	N	L	K	D	V
	*	:		*		.	:	*	*	.

Remarque : vous pouvez utiliser le programme ***BIONJ*** du portail Mobyle de l'Institut Pasteur.

Application : Interprétez les valeurs du bootstrap de l'arbre[9] suivant :



[9] *www7.inra.fr/internet/Projets/agroBI/PHYLO/Gouy.pdf*

Les constructions phylogénétiques par l'exemple

Les jeux de données

Exemple 1 : Cas des données moléculaires. La méthode **UPGMA** (Unweight Pair Group Method with Arithmetic mean).

Cette méthode est utilisée pour construire des arbres phylogénétiques si les séquences ne sont pas trop divergentes. Cela veut dire que cette méthode repose sur la théotie de l'horloge moléculaire dans laquelle on suppose que tous les individus issus de parents communs subissent le même taux de mutation (substitutions, délétions et insertions) au cours de leur évolution. Par exemple, des protéines ayant la même fonction auraient subit les mêmes mutations et avec les mêmes taux.

Elle utilise un algorithme de clustérisation séquentielle car elle construit l'arbre pas à pas, au fur et à mesure que les clades sont définis. Un clade est une paire d'individus très similaires regroupés ensemble. Il y a identification du premier couple d'indiividus (1er clade) les plus proches qui sera considéré comme un seul individu pour la suite de la clustérisation.

Comprendre UPGMA par l'exemple : Le tableau suivant résume une portion d'un alignement multiple entre cinq séquences.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Taxon 1	A	T	G	G	C	T	A	T	T	C	T	T	A	T	A	G	T	A	C	G
Taxon 2	A	T	C	G	C	T	A	G	T	C	T	T	A	T	A	T	T	A	C	A
Taxon 3	T	T	C	A	C	T	A	G	A	C	C	T	G	T	G	G	T	C	C	A
Taxon 4	T	T	G	A	C	C	A	G	A	C	C	T	G	T	G	G	T	C	C	G
Taxon 5	T	T	G	A	C	C	A	G	T	T	C	T	C	T	A	G	T	T	C	G

La matrice (n x m) de distances (exprimée en %) entre les cinq taxons est donnée par :

$$Dij = \frac{nombre\ mutations\ entre\ i\ et\ j}{Taille\ de\ l'alignement} x100$$

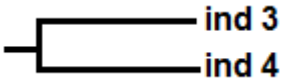
	Taxon 1	Taxon 2	Taxon 3	Taxon 4	Taxon 5
Taxon 1					
Taxon 2	20				
Taxon 3	50	40			
Taxon 4	45	55	15		
Taxon 5	40	50	40	25	

L'algorithme UPGMA repose sur quatre étapes principales :

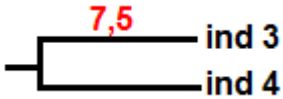
1- Identification des deux OTUs ayant la plus petite distance Dij
2- Mettre ces deux OTUs à égale distance du nœud formant le clade ij. La longueur ded la branche du clade est Dij / 2
3- Construire un nouvel ensemble en considérant le clade ij comme étant un individu à part qu'il faut comparer avec les reste des individus en calculant une nouvelle matrice de dimensions n-1 et m-1
4- Recommencer à partir de l'étape 1

Application 1

Etape 1 : Les deux individus ayant la plus faible distance sont les OTUs 3 et 4 car Dij=15. Ils constituent donc le premier clade de notre arbre :



Etape 2 : La longueur de chaque branche est la moyenne de la distance qui sépare ind 3 de ind 4 ; soit 15 / 2 = 7,5. Cette valeur est à mettre sur une seule branche car dans la méthode UPGMA les deux branches d'un clade donné sont égales.



Etape 3 : on suppose que ind3 et ind4 forment un seul individu du fait de la très forte ressemblance due à la faible distance qui les sépare. On lui donne le nom de U. Donc l'ensemble des individus sera E = {U, ind1, ind2, ind5}.
A cette étape, nous devons calculer les distances qui séparent la taxon U du reste des taxons qui sont : ind1, ind2, ind5

La distance entre U et ind1 est :

$$D(U, ind1) = (D(ind3, ind1) + D(ind4, ind1))/2$$

D (U,ind1) = ((50) + (45)) / 2 = 47,5
La distance entre U et ind2 est :

$$D(U, ind2) = (D(ind3, ind2) + D(ind4, ind2))/2$$

D (U,ind2) = ((40) + (55)) / 2 = 47,5
La distance entre U et ind5 est :

$$D(U, ind5) = (D(ind3, ind5) + D(ind4, ind5))/2$$

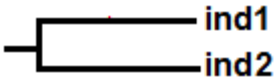
D (U,ind5) = ((40) + (25)) / 2 = 32,5 La nouvelle matrice de distances sera donc :

	Taxon U	Taxon 1	Taxon 2	Taxon 5
Taxon U				
Taxon 1	47,5			
Taxon 2	47,5	20		
Taxon 5	32,5	40	50	

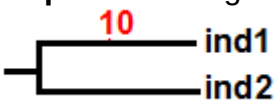
Etape 4 : A partir de cette nouvelle matrice, reprenons les étapes précédantes dans un deuxième cycle de calculs.

Cycle 2 :

Etape 1 : Les deux individus ayant la plus faible distance sont les OTUs 1 et 2 car Dij=20. Ils constituent donc le deuxième clade de l'arbre :



Etape 2 : La longueur de chaque branche est 20 / 2 = 10.



Etape 3 : Les taxons 1 et 2 forment un seul individu. On lui donne le nom de W. Donc l'ensemble des individus sera E = {U, W, ind5}.
La distance entre U et W est donnée par la relation :

$$D(U,W) = (D(U,ind1) + D(U,ind2))/2$$

D (U,W) = ((47,5) + (47,5)) / 2 = 47,5
La distance entre ind5 et W est donnée par la relation :

$$D(ind5,W) = (D(ind5,ind1) + D(ind5,ind2))/2$$

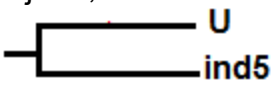
D (ind5,W) = ((40) + (50)) / 2 = 45
La nouvelle matrice de distances est alors :

	W	U	ind5
W			
U	47,5		
ind5	45	32,5	

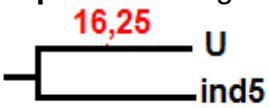
Etape 4 : A partir de cette nouvelle matrice, reprenons les étapes précédantes dans un troisième cycle de calculs.

Cycle 3 :

Etape 1 : Les deux individus ayant la plus faible distance sont les OTUs U et ind5 car Dij=32,5. Ils constituent donc le troisième clade de l'arbre :



Etape 2 : La longueur de chaque branche est 32,5 / 2 =16,25.

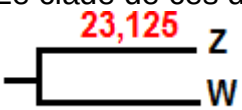


Etape 3 : Les taxons U et ind5 forment un seul individu. On lui donne le nom de Z. Donc l'ensemble des individus sera E = {Z, W}.
La distance entre ces deux OTUs est :

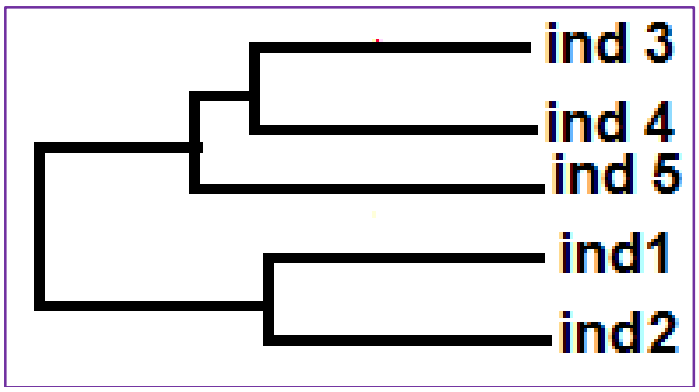
$$D(Z,W) = (D(U,W) + D(ind5,W))/2$$

D (Z,W) = ((47,5) + (45)) / 2 = 46,25

Le clade de ces deux OTUs est :



L'arbre UPGMA est le suivant :

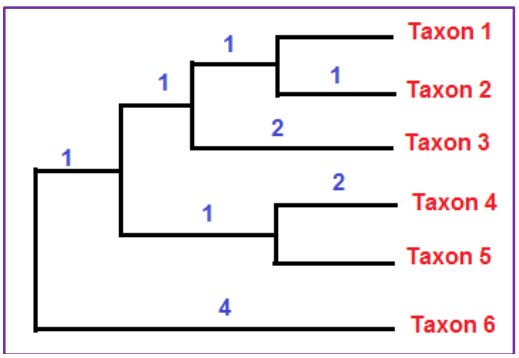


Application 2

On considère la matrice de distances de six taxons. Construisez l'arbre phylogénétique en utilisant la méthode UPGMA.

	Taxon 1	Taxon 2	Taxon 3	Taxon 4	Taxon 5	Taxon 6
Taxon 1						
Taxon 2	2					
Taxon 3	4	4				
Taxon 4	6	6	6			
Taxon 5	6	6	6	4		
Taxon 6	8	8	8	8	8	

Réponse :



Exemple 2 : L'analyse de cinq taxons a conduit aux résultats suivants

	1	2	3	4	5	6	7	8	9
A	0	1	0	0	0	0	1	0	1
B	1	0	1	1	1	0	0	0	1
C	1	0	0	1	1	1	1	1	0
D	1	1	1	1	1	1	1	0	1
E	0	0	0	1	0	1	1	1	1

1. Calculer la matrice de similarité en utilisant les deux indices symétrique et asymétrique.
2. Déterminer alors les deux matrices qui en découlent
3. En utilisant l'algorithme UPGMA, tracer les deux arbres et interpréter.

Solution :

Cas de l'indice symétrique : La formule de cet indice est données par la relation :

$$s = \frac{(1,1) + (0,0)}{(1,1) + (1,0) + (0,1) + (0,0)}$$

La similarité entre les taxon A et B est :

(1,1) : le nombre de fois où un des neuf caractères est retrouvé en même temps chez les deux taxons A et B. Dans ce cas, il est retrouvé **une seule fois** sur le neuvième caractère.

(1,0) : le nombre de cas dans lesquels un caractère est retrouvé chez A mais non chez B. on retrouve **deux fois** : les caractères 2 et 7 sont retrouvés chez A mais non chez B

(0,1) : le caractère est absent chez A mais présent chez B. Cette situation est retrouvée dans **quatre** caractères qui sont : le caractère 1, le caractère 3, le caractère 4 et le caractère 5.

(0,0) : Aucun des deux taxons A et B ne possède le caractère. Cette double absence est retrouvée **deux** fois et concerne les caractères 6 et 8.

Simple vérification : 1+2+4+2=9 c'est le nombre de caractères étudiés.
D'où le similarité entre A et B est :

$$s = \frac{1+2}{1+2+4+2} \qquad s = \frac{3}{9} \qquad s = 0{,}33$$

Remarque : L'indice S peut être exprimé en % : S=33,33%

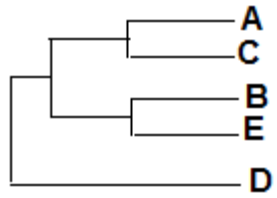
En applicant le même calcul pour le reste des taxons, on obtient la matrice de similarité suivante :

	A	B	C	D	E
A					
B	33,33				
C	22,22	44,44			
D	44,44	66,66	55,55		
E	55,55	33,33	66,66	44,44	

$\longrightarrow$ D=100-S

	A	B	C	D	E
A					
B	66,67				
C	77,78	55,56			
D	55,56	33,34	44,45		
E	44,45	66,67	33,34	55,56	

L'application de la méthode UPGMA conduit à la topologie suivante :



Cet arbre peut aussi être représenté selon le format de Newick : **(((A,C),(B,E)),(D))**

Cas de l'indice asymétrique : La formule de cet indice est données par la relation :

$$s = \frac{(1,1)}{(1,1) + (1,0) + (0,1)}$$

Dans ce cas, la double absence (0,0) n'est pas prise en compte. La matrice de similarité est :

	A	B	C	D	E
A					
B	14,29				
C	12,5	37,5			
D	37,5	62,5	55,55		
E	33,33	25	57,14	44,44	

$\longrightarrow$ D=100-S

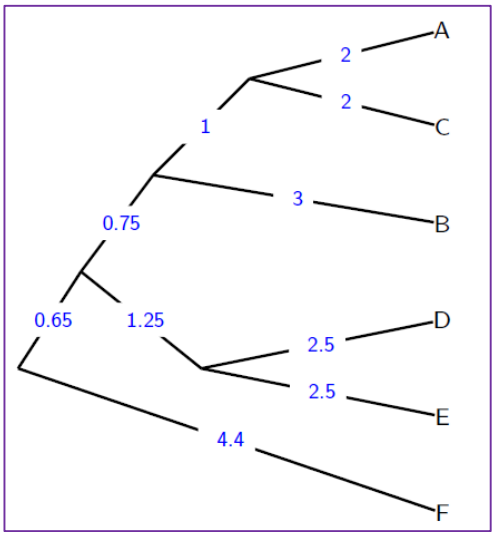
	A	B	C	D	E
A					
B	85,71				
C	87,5	62,5			
D	62,5	37,5	44,45		
E	66,67	75	42,86	55,56	

Montrez que, pour ce cas, la méthode UPGMA conduit à la même topologie

Application : Réaliser la phylogénie en se basant sur la matrice de distances[10] suivante:

	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

Solution :



[10] Aida Ouangraoua, Mathieu Giraud, Maude Pupin. Université de Lille 1. Mars 2011

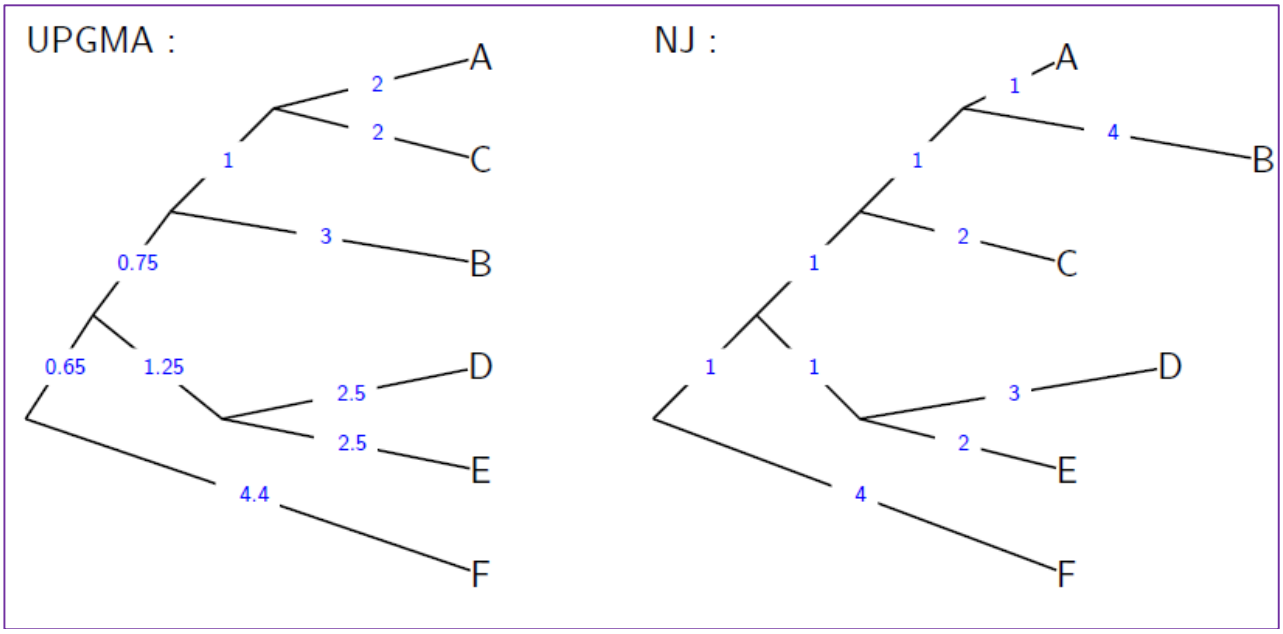
Application : TP Hhomologies et construction d'arbres

Dans ce TP, vous allez vous intéresser à un exemple particulier d'une famille d'enzymes : la phosphatase alcaline (PAL).

La PAL (E.C. 3.1.3.1) catalyse l'hydrolyse d'un substrat qui contient une liaison de type ester monophosphate ; c'est la déphosphorylation. Elle a une activité aux pH alcalins (pH > 7). C'est une enzyme présente chez divers individus : bactéries, vertébrés, …

Première partie : récupération de séquences et recherche d'homologie

1. Sur le site UniProt (http://www.uniprot.org/) introduisez votre requête pour la recherche des séquences de la PAL chez un ensemble de bactéries dont les codes d'accession sont les suivants : B8DBC3, A4SL54, F5ZJD7, F7YNR2, P00634, P19406, Q83SJ2 et P35483.
2. Regroupez toutes ces séquences dans un même fichier au format FAST
3. Identifiez les organismes sources de ces PAL
4. Réalisez un Blastp de la PAL P09923 contre SwissProt sur le site NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi)
5. Interprétez le résultat graphique, relatif aux domaines protéiques de cette PAL.



Deuxième partie : Construction phylogénétique
Allez au site http://mobyle.pasteur.fr/ et choisissez le menu ***phylogeny***. Plusieurs méthodes de construction vous sont alors proposées.

1. Commentez les méthodes proposées
2. En choisissant la méthode ***Distance***, puis le lien ***Distmat***, construisez votre arbre

La méthode du Neighbor – Joining (NJ)
Développée par Saitou et Nei (1987), elle est la plus souvent utilisée par les chercheurs. Contrairement à la méthode UPGMA, la méthode NJ autorise des taux de mutations différents donc des longueurs de branches inégales car ne se base pas sur la théorie de l'horloge moléculaire, mais tente plutôt de corriger UPGMA[11] puisqu'elle suppose que les caractères des taxons évoluent indépendamment les uns des autres.

Le principe de NJ consiste en le calcul des longueurs des branches de l'arbre de sorte qu'elles soient la plus petites possibles. A chaque étape d'agglomération, NJ préfère les taxons qi réduisent la longueur des branches de l'arbre.

A la différence de UPGMA, la méthode NJ ne choisit pas le couple de taxon ayant la plus petite distance pour faire de lui le premier clade de l'arbre, mais calcule dabord la divergence de chaque taxon par rapport aux autres.

Exemple : On considère la matrice de distances utilisée plus haut dans le cas de UPGMA. Appliquons les étapes de l'algorithme NJ sous forme de plusieurs cycles pour aboutir à la topologie de l'arbre :

	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

Cycle 1 :

Etape 1 : Calcul de la divergence " r " de taxon par rapport aux autres

r(A) = 5+4+7+6+8 = 30 r(B) = 5+7+10+9+11 = 42 r(C) = 4+7+7+6+8 = 32
r(D) = 7+10+7+6+8 = 38 r(E) = 6+9+6+5+8 = 34 r(F) = 8+11+8+9+8 = 44

[11] Aida Ouangraoua, Mathieu Giraud, Maude Pupin. Université de Lille 1. Mars 2011

Etape 2 : Calcul de la nouvelle matrice M(i,j) telle que :

$$M(i,j) = d(i,j) - \frac{(ri + rj)}{N - 2}$$

Avec :
d(i,j) : distance sur la matrice initiale entre les individus i et j
ri et rj : les divergences des individus i et j
N : Nombre d'individus, pour ce premier cycle, N = 6

Application numérique :

$$M(A,B) = d(A,B) - \frac{rA + rB}{6 - 2}$$

M(A,B) = -13

En appliquant la formule pour tous les couples i et, on obtient :

M(A,C)=-11,5	M(A,D)=-10	M(A,E)=-10	M(A,F)=-10,5	M(B,C)=-11,5	M(B,D)=-10	M(B,E)=-10
M(B,F)=-10,5	M(C,D)=-10,5	M(C,E)=-10,5	M(C,F)=-11	M(D,E)=-13	M(D,F)=-11,5	M(E,F)=-11,5

On peut écrire ce résultat sous forme d'une matrice carrée.

Etape 3 : Choix des deux individus avec M(i,j) la plus petite. On remarque que deux couple de taxons possèdent la plus faible valeur M(i,j) : (A,B) et (D,E) avec une valeur égale à -13. On choisit la paire (A,B) et on crée un nœud U1 qui regroupe cette paire d'individus.
Le calcul de la branche entre U1 et A et entre U1 et B est donné par :

$$S(A,U1) = \frac{d(A,b)}{2} + \frac{rA - rb}{2(N-2)} = \frac{5}{2} + \frac{30 - 42}{8} = 1$$

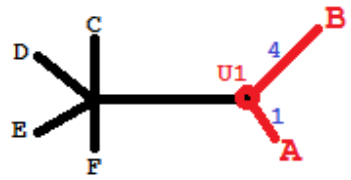
S(B, U1) = d(A,B) - S(A,U1) = 5 – 1 = 4

$$d(C,U1) = \frac{d(A,C) + d(B,C) - d(A,B)}{2} = \frac{4 + 7 - 5}{2} = 3$$

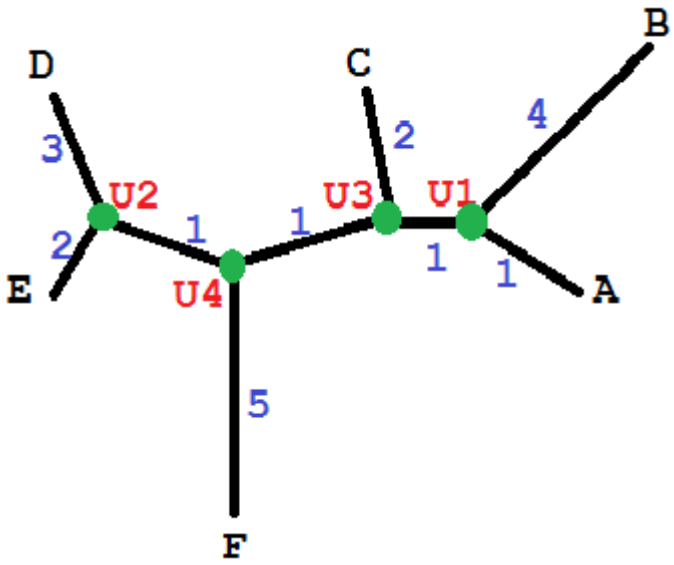
On obtient :

	U1	C	D	E
C	3			
D	6	7		
E	5	6	5	
F	7	8	9	8

A cette étape, on peut joindre A à B dans un même clade et porter le reste des taxons sous forme d'étoile :



Le deuxième cycle reprend avec N = 5 car A et B sont considéré comme un seul individ et à la fin de tous les cycles (5 en tout) on obtient la topologie suivante :



Partie 3 : LES OUTILS DE LA PHYLOGENIE

Pour réaliser une phylogénie, un bon nombre de programmes informatiques sont disponible sur la grande toile. On peut citer par exemple le portail Mobyle de l'Institut Pasteur, le programme MEGA, Phylowin, Philip, …

Les meilleurs sites que vous pouvez consulter pour réaliser vos phylogénies:

http://evolution.genetics.washington.edu/phylip/software.html
http://bioweb.pasteur.fr/seqanal/phylogeny/
http://pbil.univ-lyon1.fr/ird/IRD.html
http://bioweb2.pasteur.fr/
http://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi

Une liste des programmes les plus utilisés :

- PHYLIP
- PAUP*
- MEGA
- Phylo_win
- ARB
- DAMBE
- PAL
- Bionumerics
- Mesquite
- PaupUp
- BIRCH
- Bosque
- EMBOSS
- phangorn
- Bio++
- ETE
- DendroPy
- SeaView
- Crux

Concernant la méthode des distances, la liste des programmes est :

• PHYLIP • PAUP* • MEGA • MacT • ODEN • TREECON • DISPAN • RESTSITE • NTSYSpc • METREE • GDA • SeqPup • PHYLTEST • Lintre • Phylo_win • POPTREE2 • Gambit • gmaes • DENDRON • BIONJ • TFPGA • MVSP • ARB • Darwin	• T-REX • sendbs • nneighbor • DAMBE • weighbor • DNASIS • MINSPNET • PAL • Arlequin • PEBBLE • HY-PHY • Vanilla • GelCompar II • Bionumerics • qclust • TCS • Populations • Winboot • SYN-TAX • PTP • SplitsTree • FastME • APE • MacVector • QuickTree	• Simplot • ProfDist • START2 • STC • NimbleTree • CBCAnalyzer • PaupUp • Geneious • BIRCH • SEMPHY • FASTML • Rate4Site • SWORDS • IDEA • FAMD • Bosque • GAME • Bioinformatics_Toolbox • TreeFit • EMBOSS • phangorn • PC-ORD • Bio++ • UGENE • NINJA	• SeaView • Statio • TIMER • Crux • Ancestor • ANC-GENE • Bn-Bs

Les programmes de la méthode de maximum de parcimonie :

• PHYLIP • PAUP* • Hennig86 • MEGA • RA • NONA • CAFCA • PHYLIP • Phylo_win • sog • gmaes • LVB • GeneTree • ARB • DAMBE	• Mesquite • PAST • FootPrinter • BPAnalysis • Simplot • Parsimov • NimbleTree • PaupUp • Notung • BIRCH • IDEA • PSODA • PRAP • SeqState • Bosque • PhyloNet	• MALIGN • POY • Gambit • TNT • GelCompar II • Bionumerics • Network • TCS • GAPars • CRANN • EMBOSS • phangorn • Murka • Freqpars • SeaView • PAUPRat

Résultat total de l'alignement multiple des DNA hydrolases

```
SP|sp|P51530|DNA2_HUMAN|DNA2_HUMAN MEQLNELELLMEKSFWEEAELPAELFQKK-VVASFPRTVLSTGMDNRYLVLAVNTVQNKE 59
SP|sp|Q6ZQJ5|DNA2_MOUSE|DNA2_MOUSE MEPLDELDLLLLEEDGGAEAVPRVELLRKKADALFPETVLSRGVDNRYLVLAVETSQNER 60
SP|sp|E1BMP7|DNA2_BOVIN|DNA2_BOVIN MERVDELELLMEKSFWQEAEPSAELFQKKKVEASFSKIVLSRGMDNRYLVLAVDIVQSEE 60
SP|sp|D3ZG52|DNA2_RAT|DNA2_RAT     MEPLDDLDLLLLEEDSGAEAVPRMEILQKKADAFFAETVLSRGVDNRYLVLAVETKLNER 60
                                   ** :::*:**: :.           : :* . * * . *** *:*********:   .:.

SP|sp|P51530|DNA2_HUMAN|DNA2_HUMAN GNCEKRLVITASQSLENKELCILRNDWCSVPVEPGDIIHLEGDCTSDTWIIDKDFGYLIL 119
SP|sp|Q6ZQJ5|DNA2_MOUSE|DNA2_MOUSE GAEEKRLHVTASQDREHEVLCILRNGWSSVPVEPGDIVHLEGDCTSEPWIIDDDFGYFIL 120
SP|sp|E1BMP7|DNA2_BOVIN|DNA2_BOVIN GNHEKHLIITASQSLEYKELCILRNDWCSVPVEPGDIIHLEGDCISNTWIIDEDFGYLIL 120
SP|sp|D3ZG52|DNA2_RAT|DNA2_RAT     GAEEKHLLITVSQEGEQEVLCILRNGWSSVPVEPGDIIHIEGDCTSEPWIVDDDFGYFIL 120
                                   *  **:* :*.**. * : ****** *.*********:*:**** *: **:*.****:**

SP|sp|P51530|DNA2_HUMAN|DNA2_HUMAN YPDMLISGTSIASSIRCMRRAVLSETFRSSDPATRQMLIGTVLHEVFQKAINNSFAPEKL 179
SP|sp|Q6ZQJ5|DNA2_MOUSE|DNA2_MOUSE YPDMMISGTSVASSIRCLRRAVLSETFRGSDPATRQMLIGTILHEVFQKAISESFAPERL 180
SP|sp|E1BMP7|DNA2_BOVIN|DNA2_BOVIN YPDMLISGTSIASSIRCMRRAVLSETFRSSDPATRQMLIGTVLHEVFQKAVSDSFAPEKL 180
SP|sp|D3ZG52|DNA2_RAT|DNA2_RAT     SPDMLISGTSVASSIRCLRRAVLSETFRVSDTATRQMLIGTILHEVFQKAISESFAPEKL 180
                                    ***:*****:******:********** ** *********:********:.:*****:*

SP|sp|P51530|DNA2_HUMAN|DNA2_HUMAN QELAFQTIQEIRHLKEMYRLNLSQDEIKQEVEDYLPSFCKWAGDFMHKNTSTDFPQMQLS 239
SP|sp|Q6ZQJ5|DNA2_MOUSE|DNA2_MOUSE QELALQTLREVRHLKEMYRLNLSQDEILCEVEEYLPSFSKWAEDFMRKGPSSEFPQMQLS 240
SP|sp|E1BMP7|DNA2_BOVIN|DNA2_BOVIN QELASQTIQEIRHLKEMYRLKLNQDEIKQEVEEYLPSFSKWAGDFMHKHTSTDFPQMQLS 240
SP|sp|D3ZG52|DNA2_RAT|DNA2_RAT     QELALQTLREVRHLKEMYRLNLSQDEVRCEVEEYLPSFSKWADEFMHKGTKAEFPQMHLS 240
                                   **** **::*:*********:*.***:  ***:*****.*** :**:*  .::****:**

SP|sp|P51530|DNA2_HUMAN|DNA2_HUMAN LPSDNSKDNSTCNIEVVKPMDIEESIWSPRFGLKGKIDVTVGVKIHRGYKTKYKIMPLEL 299
SP|sp|Q6ZQJ5|DNA2_MOUSE|DNA2_MOUSE LPSDGSNRSSPCNIEVVKSLDIEESIWSPRFGLKGKIDVTVGVKIHRDCKMKYKVMPLEL 300
SP|sp|E1BMP7|DNA2_BOVIN|DNA2_BOVIN LPSDGSNSNSTCNIEVTNSLDIEESIWSPRFGLKGKIDVTVGVKIHRGCKTKYKIMPLEL 300
SP|sp|D3ZG52|DNA2_RAT|DNA2_RAT     LPSDSSDRSSPCNIEVVKSLDIEESIWSPRFGLKGKIDVTVGVKIHRDCKTKYKIMPLEL 300
                                   **** *. .* *****.: :************************** * ***:*****

SP|sp|P51530|DNA2_HUMAN|DNA2_HUMAN KTGKESNSIEHRSQVVLYTLLSQERRADPEAGLLLYLKTGQMYPVPANHLDKRELLKLRN 359
SP|sp|Q6ZQJ5|DNA2_MOUSE|DNA2_MOUSE KTGKESNSIEHRSQVVLYTLLSQERREDPEAGWLLYLKTGQMYPVPANHLDKRELLKLRN 360
SP|sp|E1BMP7|DNA2_BOVIN|DNA2_BOVIN KTGKESNSIEHRSQLVLYTLLSQERRADPEAGLLLYLKTGQMYPVPAKHLDKRELLRLRN 360
SP|sp|D3ZG52|DNA2_RAT|DNA2_RAT     KTGKESNSIEHRGQVILYTLLSQERREDPEAGWLLYLKTGQMYPVPANHLDKRELLKLRN 360
                                   ************.*::********** ***** **************:********:***

SP|sp|P51530|DNA2_HUMAN|DNA2_HUMAN QMAFSLFHRISKSATRQKTQLASLPQIIEEEKTCKYCSQIGNCALYSRAVEQQMDCSSVP 419
SP|sp|Q6ZQJ5|DNA2_MOUSE|DNA2_MOUSE WLAASLLHRVSRAAPGEEARLSALPQIIEEEKTCKYCSQIGNCALYSRAVEEQGDDASIP 420
SP|sp|E1BMP7|DNA2_BOVIN|DNA2_BOVIN QMAFSLFHRINKST-GEKTELAPLPQIIEEQQTCKYCSQMGNCALYSRAVEQQMEDSSVP 419
SP|sp|D3ZG52|DNA2_RAT|DNA2_RAT     QLAFSLLHRVSRAAAGEEARLLALPQIIEEEKTCKYCSQMGNCALYSRAVEQ-VHDTSIP 419
                                    :* **:**:.:::  :::.*  *******::*******:***********:  . :*:*

SP|sp|P51530|DNA2_HUMAN|DNA2_HUMAN IVMLPKIEEETQHLKQTHLEYFSLWCLMLTLESQSKDNKKNHQNIWLMPASEMEKSGSCI 479
SP|sp|Q6ZQJ5|DNA2_MOUSE|DNA2_MOUSE EAMLSKIQEETRHLQLAHLKYFSLWCLMLTLESQSKDNRKTHQSIWLTPASELEESGNCV 480
SP|sp|E1BMP7|DNA2_BOVIN|DNA2_BOVIN TSMWPKIKEETQHLKPIHLEYFSLWCLMLTLESQSKDNKRNYQHIWLMPASEMEESGSCI 479
SP|sp|D3ZG52|DNA2_RAT|DNA2_RAT     EGMRSKIQEGTQHLTRAHLKYFSLWCLMLTLESQSKDTKKSHQSIWLTPASKLEESGNCI 479
                                     *  **:* *:**   **:*****************.:::.:* *** ***::*:**.*:

SP|sp|P51530|DNA2_HUMAN|DNA2_HUMAN GNLIRMEHVKIVCDGQYLHNFQCKHGAIPVTNLMAGDRVIVSGEERSLFALSRGYVKEIN 539
SP|sp|Q6ZQJ5|DNA2_MOUSE|DNA2_MOUSE GNLVRTEPVSRVCDGQYLHNFQRKNGPMPATNLMAGDRIILSGEERKLFALSKGYVKKMN 540
SP|sp|E1BMP7|DNA2_BOVIN|DNA2_BOVIN GSLIRIEHVKTVCDGQYLHNFQRKNGAIPITNLMAGDRIILSGEERTLFALSRGYVKEIN 539
SP|sp|D3ZG52|DNA2_RAT|DNA2_RAT     GSLVRTEPVKRVCDGHYLHNFQRKNGPMPATNLMAGDRIILSGEERKLFALSKGYVKRID 539
                                   *.*:* * *. ****:****** *.* :* ********:*:*****.*****:****.::

SP|sp|P51530|DNA2_HUMAN|DNA2_HUMAN MTTVTCLLDRNLSVLPESTLFRLDQEEKNCDIDTPLGNLSKLMENTFVSKKLRDLIIDFR 599
SP|sp|Q6ZQJ5|DNA2_MOUSE|DNA2_MOUSE KAAVTCLLDRNLSTLPATTVFRLDREERHGDISTPLGNLSKLMESTDPSKRLRELIIDFR 600
SP|sp|E1BMP7|DNA2_BOVIN|DNA2_BOVIN STTVTCSLDRNLSGLPESTLFRLDQEEKNCDIDTPLGNLSKLMENTRASQKLRDLIIDFR 599
SP|sp|D3ZG52|DNA2_RAT|DNA2_RAT     TAAVTCLLDRNLSTLPETTLFRLDREEKHGDINTPLGNLSKLMENTDSSKRLRELIIDFK 599
                                    ::*** ****** ** :*:****:**:. **.***********.*  *::**:*****:

SP|sp|P51530|DNA2_HUMAN|DNA2_HUMAN EPQFISYLSSVLPHDAKDTVACILKGLNKPQRQAMKKVLLSKDYTLIVGMPGTGKTTTIC 659
SP|sp|Q6ZQJ5|DNA2_MOUSE|DNA2_MOUSE EPQFIAYLSSVLPHDAKDTVANILKGLNKPQRQAMKRVLLSKDYTLIVGMPGTGKTTTIC 660
SP|sp|E1BMP7|DNA2_BOVIN|DNA2_BOVIN EPQFISYLSSVLPHEAKDTVACILKGLNKPQRQAMKKVLLSKDYTLIVGMPGTGKTTTIC 659
SP|sp|D3ZG52|DNA2_RAT|DNA2_RAT     EPQFIAYLSSVLPHDAKDTVANILKGLNKPQRQAMKKVLLSKDYTLIVGMPGTGKTTTIC 659
                                   *****:********:****** **************:************************

SP|sp|P51530|DNA2_HUMAN|DNA2_HUMAN TLVRILYACGFSVLLTSYTHSAVDNILLKLAKFKIGFLRLGQIQKVHPAIQQFTEQEICR 719
SP|sp|Q6ZQJ5|DNA2_MOUSE|DNA2_MOUSE ALVRILSACGFSVLLTSYTHSAVDNILLKLAKFKVGFLRLGQSHKVHPDIQKFTEEEICR 720
SP|sp|E1BMP7|DNA2_BOVIN|DNA2_BOVIN TLVRILYACGFSVLLTSYTHSAVDNILLKLAKFKIGFLRLGQIQKVHPDIQKFTEEEICR 719
SP|sp|D3ZG52|DNA2_RAT|DNA2_RAT     ALVRILSACGFSVLLTSYTHSAVDNILLKLAKFKIGFLRLGQSHKVHPDIQKFTEEEMCR 719
                                   :***** ************************:******* :**** **:***:*:**

SP|sp|P51530|DNA2_HUMAN|DNA2_HUMAN SKSIKSLALLEELYNSQLIVATTCMGINHPIFSRKIFDFCIVDEASQISQPICLGPLFFS 779
SP|sp|Q6ZQJ5|DNA2_MOUSE|DNA2_MOUSE SRSIASLAHLEELYNSHPIVATTCMGINHPIFSRKTFDFCIVDEASQISQPVCLGPLFFS 780
```

```
SP|sp|E1BMP7|DNA2_BOVIN|DNA2_BOVIN SKSIKSLALLEELYNSQLIVATTCMGINHPIFSRKTFDFCIVDEASQISQPVCLGPLFFS 779
SP|sp|D3ZG52|DNA2_RAT|DNA2_RAT     LRSIASLAHLEELYNSHPVVATTCMGISHPMFSRKTFDFCIVDEASQISQPICLGPLFFS 779
                                    :** *** *******: :********.**:**** ***************:********

SP|sp|P51530|DNA2_HUMAN|DNA2_HUMAN RRFVLVGDHQQLPPLVLNREARALGMSESLFKRLEQNKSAVVQLTVQYRMNSKIMSLSNK 839
SP|sp|Q6ZQJ5|DNA2_MOUSE|DNA2_MOUSE RRFVLVGDHQQLPPLVVNREARALGMSESLFKRLERNESAVVQLTVQYRMNRKIMSLSNK 840
SP|sp|E1BMP7|DNA2_BOVIN|DNA2_BOVIN RRFVLVGDHQQLPPLVLNREARALGMSESLFKRLEQNKNAVVQLTVQYRMNSKIMSLSNK 839
SP|sp|D3ZG52|DNA2_RAT|DNA2_RAT     RRFVLVGDHKQLPPLVLNREARALGMSESLFKRLERNESAVVQLTIQYRMNRKIMSLSNK 839
                                    *********:******:******************:*:.******:***** ********

SP|sp|P51530|DNA2_HUMAN|DNA2_HUMAN LTYEGKLECGSDKVANAVINLRHFKDVKLELEFYADYSDNPWLMGVFEPNNPVCFLNTDK 899
SP|sp|Q6ZQJ5|DNA2_MOUSE|DNA2_MOUSE LTYAGKLECGSDRVANAVLALPNLKDARLSLQLYADYSDSPWLAGVLEPDNPVCFLNTDK 900
SP|sp|E1BMP7|DNA2_BOVIN|DNA2_BOVIN LTYEGKLECGSDKVANAVINLPNFKDVKLELEFYADYSENPWLIAAFEPNNPVCFLNTHK 899
SP|sp|D3ZG52|DNA2_RAT|DNA2_RAT     LTYEGKLECGSDRVANAVITLPNLKDVRL--EFYADYSDNPWLAGVFEPDNPVCFLNTDK 897
                                    *** ********:*****: * .:**.:*  ::*****:.*** ..:**:********.*

SP|sp|P51530|DNA2_HUMAN|DNA2_HUMAN VPAPEQVEKGGVSNVTEAKLIVFLTSIFVKAGCSPSDIGIIAPYRQQLKIINDLLAR-SI 958
SP|sp|Q6ZQJ5|DNA2_MOUSE|DNA2_MOUSE VPAPEQVENGGVSNVTEARLIVFLTSTFIKAGCSPSDIGVIAPYRQQLRIISDLLARSSV 960
SP|sp|E1BMP7|DNA2_BOVIN|DNA2_BOVIN VPAPEQVEKGGVSNIMEAKLVVFLTSVFIKAGCKPSDIGIIAPYRQQLKVISDLLAQSSV 959
SP|sp|D3ZG52|DNA2_RAT|DNA2_RAT     VPAPEQIENGGVSNVTEARLIVFLTSTFIKAGCSPSDIGIIAPYRQQLRTITDLLARSSV 957
                                    ******:*:*****: **:*:***** *:****.*****:********: *.****: *:

SP|sp|P51530|DNA2_HUMAN|DNA2_HUMAN GMVEVNTVDKYQGRDKSIVLVSFVRSNKDGTVGELLKDWRRLNVAITRAKHKLILLGCVP 1018
SP|sp|Q6ZQJ5|DNA2_MOUSE|DNA2_MOUSE GMVEVNTVDKYQGRDKSLILVSFVRSNEDGTLGELLKDWRRLNVALTRAKHKLILLGSVS 1020
SP|sp|E1BMP7|DNA2_BOVIN|DNA2_BOVIN GMVEVNTVDRYQGRDKSIVVVSFVRSNEDGTLGELLKDWRRLNVAITRAKHKLILLGCVP 1019
SP|sp|D3ZG52|DNA2_RAT|DNA2_RAT     GMVEVNTVDKYQGRDKSLILVSFVRSNEDGTLGELLKDWRRLNVAITRAKHKLILLGSVS 1017
                                    *********:*******:::*******:***:**************:***********.*

SP|sp|P51530|DNA2_HUMAN|DNA2_HUMAN SLNCYPPLEKLLNHLNSEKLIIDLPSREHESLCHILGDFQRE 1060
SP|sp|Q6ZQJ5|DNA2_MOUSE|DNA2_MOUSE SLKRFPPLGTLFDHLNAEQLILDLPSREHESLSHILGDCQRD 1062
SP|sp|E1BMP7|DNA2_BOVIN|DNA2_BOVIN SLSRYPPLRKLLNHLNSEKLIIDLPSGEHESLFHLLGDFQRK 1061
SP|sp|D3ZG52|DNA2_RAT|DNA2_RAT     SLKRFPPLEKLFDHLNAEQLISNLPSREHESLYHILGDCQRD 1059
                                    **. :*** .*::***:*:** :*** ***** *:*** **.
```

Liste des ADN Hydrolases

>sp|P51530|DNA2_HUMAN **DNA replication ATP-dependent helicase/nuclease** DNA2 OS=**Homo sapie**ns GN=DNA2 PE=1 SV=3 **1060AA**

MEQLNELELLMEKSFWEEAELPAELFQKKVVASFPRTVLSTGMDNRYLVLAVNTVQNKEGNCEKRLVITASQSLENKELCIL
RNDWCSVPVEPGDIIHLEGDCTSDTWIIDKDFGYLILYPDMLISGTSIASSIRCMRRAVLSETFRSSDPATRQMLIGTVLHE
VFQKAINNSFAPEKLQELAFQTIQEIRHLKEMYRLNLSQDEIKQEVEDYLPSFCKWAGDFMHKNTSTDFPQMQLSLPSDNSK
DNSTCNIEVVKPMDIEESIWSPRFGLKGKIDVTVGVKIHRGYKTKYKIMPLELKTGKESNSIEHRSQVVLYTLLSQERRADP
EAGLLLYLKTGQMYPVPANHLDKRELLKLRNQMAFSLFHRISKSATRQKTQLASLPQIIEEEKTCKYCSQIGNCALYSRAVE
QQMDCSSVPIVMLPKIEEETQHLKQTHLEYFSLWCLMLTLESQSKDNKKNHQNIWLMPASEMEKSGSCIGNLIRMEHVKIVC
DGQYLHNFQCKHGAIPVTNLMAGDRVIVSGEERSLFALSRGYVKEINMTTVTCLLDRNLSVLPESTLFRLDQEEKNCDIDTP
LGNLSKLMENTFVSKKLRDLIIDFREPQFISYLSSVLPHDAKDTVACILKGLNKPQRQAMKKVLLSKDYTLIVGMPGTGKTT
TICTLVRILYACGFSVLLTSYTHSAVDNILLKLAKFKIGFLRLGQIQKVHPAIQQFTEQEICRSKSIKSLALLEELYNSQLI
VATTCMGINHPIFSRKIFDFCIVDEASQISQPICLGPLFFSRRFVLVGDHQQLPPLVLNREARALGMSESLFKRLEQNKSAV
VQLTVQYRMNSKIMSLSNKLTYEGKLECGSDKVANAVINLRHFKDVKLELEFYADYSDNPWLMGVFEPNNPVCFLNTDKVPA
PEQVEKGGVSNVTEAKLIVFLTSIFVKAGCSPSDIGIIAPYRQQLKIINDLLARSIGMVEVNTVDKYQGRDKSIVLVSFVRS
NKDGTVGELLKDWRRLNVAITRAKHKLILLGCVPSLNCYPPLEKLLNHLNSEKLIIDLPSREHESLCHILGDFQRE

>sp|P38859|DNA2_YEAST DNA replication ATP-dependent helicase/nuclease DNA2 OS=**Saccharomyces cerevisiae 1522AA**

MPGTPQKNKRSASISVSPAKKTEEKEIIQNDSKAILSKQTKRKKKYAFAPINNLNGKNTKVSNASVLKSIAVSQVRNTSRTK
DINKAVSKSVKQLPNSQVKPKREMSNLSRHHDFTQDEDGPMEEVIWKYSPLQRDMSDKTTSAAEYSDDYEDVQNPSSTPIVP
NRLKTVLSFTNIQVPNADVNQLIQENGNEQVRPKPAEISTRESLRNIDDILDDIEGDLTIKPTITKFSDLPSSPIKAPNVEK
KAEVNAEEVDKMDSTGDSNDGDDSLIDILTQKYVEKRKSESQITIQGNTNQKSGAQESCGKNDNTKSRGEIEDHENVDNQAK
TGNAFYENEEDSNCQRIKKNEKIEYNSSDEFSDDSLIELLNETQTQVEPNTIEQDLDKVEKMVSDDLRIATDSTLSAYALRA
KSGAPRDGVVRLVIVSLRSVELPKIGTQKILECIDGKGEQSSVVVRHPWVYLEFEVGDVIHIIEGKNIENKRLLSDDKNPKT
QLANDNLLVLNPDVLFSATSVGSSVGCLRRSILQMQFQDPRGEPSLVMTLGNIVHELLQDSIKYKLSHNKISMEIIIQKLDS
LLETYSFSIIICNEEIQYVKELVMKEHAENILYFVNKFVSKSNYGCYTSISGTRRTQPISISNVIDIEENIWSPIYGLKGFL
DATVEANVENNKKHIVPLEVKTGKSRSVSYEVQGLIYTLLLNDRYEIPIEFFLLYFTRDKNMTKFPSVLHSIKHILMSRNRM
SMNFKHQLQEVFGQAQSRFELPPLLRDSSCDSCFIKESCMVLNKLLEDGTPEESGLVEGEFEILTNHLSQNLANYKEFFTKY
NDLITKEESSITCVNKELFLLDGSTRESRSGRCLSGLVVSEVVEHEKTEGAYIYCFSRRRNDNNSQSMLSSQIAANDFVIIS
DEEGHFCLCQGRVQFINPAKIGISVKRKLLNNRLLDKEKGVTTIQSVVESELEQSSLIATQNLVTYRIDKNDIQQSLSLARF
NLLSLFLPAVSPGVDIVDERSKLCRKTKRSDGGNEILRSLLVDNRAPKFRDANDDPVIPYKLSKDTTLNLNQKEAIDKVMRA

EDYALILGMPGTGKTTVIAEIIKILVSEGKRVLLTSYTHSAVDNILIKLRNTNISIMRLGMKHKVHPDTQKYVPNYASVKSY
NDYLSKINSTSVVATTCLGINDILFTLNEKDFDYVILDEASQISMPVALGPLRYGNRFIMVGDHYQLPPLVKNDAARLGGLE
ESLFKTFCEKHPESVAELTLQYRMCGDIVTLSNFLIYDNKLKCGNNEVFAQSLELPMPEALSRYRNESANSKQWLEDILEPT
RKVVFLNYDNCPDIIEQSEKDNITNHGEAELTLQCVEGMLLSGVPCEDIGVMTLYRAQLRLLKKIFNKNVYDGLEILTADQF
QGRDKKCIIISMVRRNSQLNGGALLKELRRVNVAMTRAKSKLIIIGSKSTIGSVPEIKSFVNLLEERNWVYTMCKDALYKYK
FPDRSNAIDEARKGCGKRTGAKPITSKSKFVSDKPIIKEILQEYES

>sp|Q6ZQJ5|DNA2_MOUSE DNA replication ATP-dependent helicase/nuclease DNA2 OS=**Mus musculus 1062AA**
MEPLDELDLLLLEEDGGAEAVPRVELLRKKADALFPETVLSRGVDNRYLVLAVETSQNERGAEEKRLHVTASQDREHEVLCI
LRNGWSSVPVEPGDIVHLEGDCTSEPWIIDDDFGYFILYPDMMISGTSVASSIRCLRRAVLSETFRGSDPATRQMLIGTILH
EVFQKAISESFAPERLQELALQTLREVRHLKEMYRLNLSQDEILCEVEEYLPSFSKWAEDFMRKGPSSEFPQMQLSLPSDGS
NRSSPCNIEVVKSLDIEESIWSPRFGLKGKIDVTVGVKIHRDCKMKYKVMPLELKTGKESNSIEHRSQVVLYTLLSQERRED
PEAGWLLYLKTGQMYPVPANHLDKRELLKLRNWLAASLLHRVSRAAPGEEARLSALPQIIEEEKTCKYCSQIGNCALYSRAV
EEQGDDASIPEAMLSKIQEETRHLQLAHLKYFSLWCLMLTLESQSKDNRKTHQSIWLTPASELEESGNCVGNLVRTEPVSRV
CDGQYLHNFQRKNGPMPATNLMAGDRIILSGEERKLFALSKGYVKKMNKAAVTCLLDRNLSTLPATTVFRLDREERHGDIST
PLGNLSKLMESTDPSKRLRELIIDFREPQFIAYLSSVLPHDAKDTVANILKGLNKPQRQAMKRVLLSKDYTLIVGMPGTGKT
TTICALVRILSACGFSVLLTSYTHSAVDNILLKLAKFKVGFLRLGQSHKVHPDIQKFTEEEICRSRSIASLAHLEELYNSHP
IVATTCMGINHPIFSRKTFDFCIVDEASQISQPVCLGPLFFSRRFVLVGDHQQLPPLVVNREARALGMSESLFKRLERNESA
VVQLTVQYRMNRKIMSLSNKLTYAGKLECGSDRVANAVLALPNLKDARLSLQLYADYSDSPWLAGVLEPDNPVCFLNTDKVP
APEQVENGGVSNVTEARLIVFLTSTFIKAGCSPSDIGVIAPYRQQLRIISDLLARSSVGMVEVNTVDKYQGRDKSLILVSFV
RSNEDGTLGELLKDWRRLNVALTRAKHKLILLGSVSSLKRFPPLGTLFDHLNAEQLILDLPSREHESLSHILGDCQRD

>sp|Q8QHA5|DNA2_XENLA DNA replication ATP-dependent helicase/nuclease DNA2 OS=**Xenopus laevis 1053AA**
MEPVSAECHLPPEDDLLEMMMEQSFTEPEEKSQDKPTRKIIPKTKLCKGVNNRYCVLHIKEVYAQREEKHLTITASQEGDDL
ELCILKDDWVALQIKPGDIIHLEGNCSVDNTWTISRDTGYLILYPDLLISGTSIANGIRCLRRSVLSEKFKVCDKGSRQMLN
GTMLHDIFQRATTCGFTDSVLQELAHHTVHGPKYLKEMYQLKLNQADVMGEIQEYLPSLSKWATDFMTHPLNQQQINRTKST
AGDPTETTKVSEFLDIEENIWSPRFGLKGKIDVTARVKIHQKSKAHLKIMPLELKTGKESNSIEHRSQVVLYTLLSQERRED
PEAGLLLYLKTGNMYTVPGNRLDRRELLKIRNELSYYLTNVLHKSDNGSKETTLASLPAMIADRQACKYCSQMRNCALYNRS
VEQQTENCYIPPEMIPVVQKETEHLTEDHLQYFRLWYLMCTLEANSKDSKMGRKNIWMMSSSEREEDGQCIGNLIRTGHVQT
ISDVQYLHSFQRRSGSVPATNLASGDRVVVSGEERFLALSTGYIKEVKDENITCILDRSLVKLPEDLLFRLDHEEGGGGLEF
HLGNLSRLMENSSVSEKLRKLIIDFSKPNFVQHLSSILPPDAKDIVASILRGLNKPQKQAMKRVLLSKDYTLIVGMPGTGKT
TTICTLVRILYACGFSVLLTSYTHSAVDNILLKLKKFQVGFLRLGRTQKLHPDVQEFSEEEICKAKSIKSLSALEELYNSQP
VVATTCMGVNHPIFTRRRFDFCIVDEASQISQPICLGPLFFADRFVLVGDHQQLPPLVKSAEARELGMSESLFKRLERNQEA
VVQLTVQYRMNSKIMALSNKLVYEGRLECASDRVSNAVVQLPHIKTLLLELEFRESQESMWIKDVLEPSNPVCFLNTEKIPA
LETEEKGGISNWIEAKLVFHLTKLYLKAGCRPSDIGIIAPYRQQLKMISNYFNSLSASAVEVNTVDKYQGRDKSVIIVSFVR
SNIDGKLGDLLKDWRRLNVALTRAKHKLIMLGCVPTLNRFDCLEQLICNLKTENQIYDLPEGAHEHFPV

>sp|E1BMP7|DNA2_BOVIN DNA replication ATP-dependent helicase/nuclease DNA2 OS=**Bos taurus 1061AA**
MERVDELELLMEKSFWQEAEPSAELFQKKKVEASFSKIVLSRGMDNRYLVLAVDIVQSEEGNHEKHLIITASQSLEYKELCI
LRNDWCSVPVEPGDIIHLEGDCISNTWIIDEDFGYLILYPDMLISGTSIASSIRCMRRAVLSETFRSSDPATRQMLIGTVLH
EVFQKAVSDSFAPEKLQELASQTIQEIRHLKEMYRLKLNQDEIKQEVEEYLPSFSKWAGDFMHKHTSTDFPQMQLSLPSDGS
NSNSTCNIEVTNSLDIEESIWSPRFGLKGKIDVTVGVKIHRGCKTKYKIMPLELKTGKESNSIEHRSQLVLYTLLSQERRAD
PEAGLLLYLKTGQMYPVPAKHLDKRELLRLRNQMAFSLFHRINKSTGEKTELAPLPQIIEEQQTCKYCSQMGNCALYSRAVE
QQMEDSSVPTSMWPKIKEETQHLKPIHLEYFSLWCLMLTLESQSKDNKRNYQHIWLMPASEMEESGSCIGSLIRIEHVKTVC
DGQYLHNFQRKNGAIPITNLMAGDRIILSGEERTLFALSRGYVKEINSTTVTCSLDRNLSGLPESTLFRLDQEEKNCDIDTP
LGNLSKLMENTRASQKLRDLIIDFREPQFISYLSSVLPHEAKDTVACILKGLNKPQRQAMKKVLLSKDYTLIVGMPGTGKTT
TICTLVRILYACGFSVLLTSYTHSAVDNILLKLAKFKIGFLRLGQIQKVHPDIQKFTEEEICRSKSIKSLALLEELYNSQLI
VATTCMGINHPIFSRKTFDFCIVDEASQISQPVCLGPLFFSRRFVLVGDHQQLPPLVLNREARALGMSESLFKRLEQNKNAV
VQLTVQYRMNSKIMSLSNKLTYEGKLECGSDKVANAVINLPNFKDVKLELEFYADYSENPWLIAAFEPNNPVCFLNTHKVPA
PEQVEKGGVSNIMEAKLVVFLTSVFIKAGCKPSDIGIIAPYRQQLKVISDLLAQSSVGMVEVNTVDRYQGRDKSIVVVSFVR
SNEDGTLGELLKDWRRLNVAITRAKHKLILLGCVPSLSRYPPLRKLLNHLNSEKLIIDLPSGEHESLFHLLGDFQRK

>sp|D3ZG52|DNA2_RAT DNA replication ATP-dependent helicase/nuclease DNA2 OS=**Rattus norvegicus 1059AA**

MEPLDDLDLLLLEEDSGAEAVPRMEILQKKADAFFAETVLSRGVDNRYLVLAVETKLNERGAEEKHLLITVSQEGEQEVLCI
LRNGWSSVPVEPGDIIHIEGDCTSEPWIVDDDFGYFILSPDMLISGTSVASSIRCLRRAVLSETFRVSDTATRQMLIGTILH
EVFQKAISESFAPEKLQELALQTLREVRHLKEMYRLNLSQDEVRCEVEEYLPSFSKWADEFMHKGTKAEFPQMHLSLPSDSS
DRSSPCNIEVVKSLDIEESIWSPRFGLKGKIDVTVGVKIHRDCKTKYKIMPLELKTGKESNSIEHRGQVILYTLLSQERRED
PEAGWLLYLKTGQMYPVPANHLDKRELLKLRNQLAFSLLHRVSRAAAGEEARLLALPQIIEEEKTCKYCSQMGNCALYSRAV
EQVHDTSIPEGMRSKIQEGTQHLTRAHLKYFSLWCLMLTLESQSKDTKKSHQSIWLTPASKLEESGNCIGSLVRTEPVKRVC
DGHYLHNFQRKNGPMPATNLMAGDRIILSGEERKLFALSKGYVKRIDTAAVTCLLDRNLSTLPETTLFRLDREEKHGDINTP
LGNLSKLMENTDSSKRLRELIIDFKEPQFIAYLSSVLPHDAKDTVANILKGLNKPQRQAMKKVLLSKDYTLIVGMPGTGKTT
TICALVRILSACGFSVLLTSYTHSAVDNILLKLAKFKIGFLRLGQSHKVHPDIQKFTEEEMCRLRSIASLAHLEELYNSHPV
VATTCMGISHPMFSRKTFDFCIVDEASQISQPICLGPLFFSRRFVLVGDHKQLPPLVLNREARALGMSESLFKRLERNESAV
VQLTIQYRMNRKIMSLSNKLTYEGKLECGSDRVANAVITLPNLKDVRLEFYADYSDNPWLAGVFEPDNPVCFLNTDKVPAPE
QIENGGVSNVTEARLIVFLTSTFIKAGCSPSDIGIIAPYRQQLRTITDLLARSSVGMVEVNTVDKYQGRDKSLILVSFVRSN
EDGTLGELLKDWRRLNVAITRAKHKLILLGSVSSLKRFPPLEKLFDHLNAEQLISNLPSREHESLYHILGDCQRD

>sp|Q5ZKG3|DNA2_CHICK DNA replication ATP-dependent helicase/nuclease DNA2 OS=**Gallus gallus 992AA**

MADPSNAALRSGLNNRYRVLEVRVVRGEGRDPEKHLAVSSDPSLGDTELCVLQNGWESVPVVPGDIVHLEGDCSSGTWVINE
QSGYLILYPDLLLSGTTISSSIRCMRKAVLSERFRGSECGSRQTLVGTILHEIFQQSVTNNLSPEKVEELAKKIVYGQKYLK
EMYHLKLKQTEIMQEIEEYLPSFFKWTEDFVRNPANQNKMQLKLPSENKTGDCSSTVEIVDILDIEENIWSPRFGLKGKIDV
TARVKIHRQGGIQSRIMPLELKSGKESNSVEHRSQVILYTLLNLERRVDPEAGFLLYLKTGTMYPVTGTRMDRRELIKLRNQ
VAFYLMHSTYKSAVGRQQSQLAALPPLIDDSQACKYCSQIHNCFLYSRAVEERMASVSFPPALIPIIEKETQHLKPSHLEYF
SLWYLMLTLEMQSGDSKKGYKNIWMIPSLEREKAGDCVGNMIRVDQVQEVSEGQYLHSFQRKNGAVPGANLLVGDRVVVSGE
ENGLLGLATGYVREISATKISCLLGRNLSKLPESTTFRLDHEEGDCSIGVPFENLSKLMKDSPVSEKLRNLIIDFHKPRFIQ
HLSSVLPPEAKEAVASILKGLNKPQKQAMKQVLLSRDYTLIVGMPGTGKTTTICALVRILSACGFSVLLTSFTHTAVDNILL
KLAKFKVGFLRLGRAQKVHPDIRKFTEEEICRSKSIKSVTDLEELYNSQPVVATACMGINHPIFVQKQFDFCIVDEASQISQ
PICLGPLFCSKRFVLVGDHQQLPPLVQNSEARDLGMSESLFKRLEQNQNAVVQLTVQYRMNSKIMSLSNKLVYEGKLECGSE
KVSNATANLPNLKMLKLEFADASKTWLKEVLEPDKPVCFLNTEKAGCRPSDIGIISPYRHQLKVITDLMARLKENRVEVNTI
DKYQGRDKSIIIVSFVRNSNDENLGALLKDWRRLNVAITRAKHKLIMVGCVPSLRRYPPLEKLLCHLQSEAMIFNLPPGAHE
SIHKCNIL

Liste Leucine Zipper

>gi|334185982|ref|NM_001203162.1| **Arabidopsis thaliana** basic leucine zipper 25 mRNA,**1489pb**

GCGGCTCCACGTAAGACAAAAACAAATTAAGGAAAGAGCGAGAGAGAGAGAGAGAGAGAAAAAAGAAGAAGAAGAATTTGAG
GCGATTGGGTAAGCGATGCACATCGTCTTCTCTGTCGATGATTTGACCGAATCCTTCTGGCCTGTTCCCGCTCCGGCGCCGT
CGCCGGGATCGTCATCGACTCCGTCACCGACGCAGAACGTGGCGGATGGGATGACTCGAAGTCAATCGGAGTGGGCGTTCCA
CAGGCTTATCAATGAGTTGTCTGGTTCCGATTCGAGCCCTACGACTAACACAATCGAGAGATCACCTCCACCGGTTCAGTCT
CTTTCGAGATTAGAAGAAACCGTTGACGAAACCGAAGATGTTGTTGAGATTCAGAAACCGCAGAATCATCGACGGCTCCCTG
TTGATGATCAAGGGAAGAATCGGAATCGTGCTCCGTCGTCTGATCCGGTTGATTCTTCAGCTCCTGTTGTTGTTGATCCTAA
TCAGTATCATGCGATTCTTAAGAGCAAGCTCGAGCTTGCTTGCGCTGCTGTTGCTCGTCGTGTGGGAACTGTGAAACCGGAA
GATTCGAGTGCTTCAGCTAGCAATCAAAAACAAGCTCAAGGCTCCATTGTGGCACAAACCTCACCTGGTGCTTCATCTGTTA
GATTTTCTCCCACAACAAGCACGCAAAAGAAACCTGATGTTCCAGCCAGACAAACTAGTATTTCATCACGAGATGATTCTGA
TGACGATGATCTTGATGGAGACGCAGATAATGGAGATCCTACTGATGTGAAGCGTGCTAGGAGGATGCTCTCAAACCGAGAA
TCCGCTAGGCGCTCTAGGAGAAGAAAGCAAGAACAAATGAATGAATTTGATACACAGGTAAAGTTTTTACCAATTGTAGCTG
CTGTTGACAACAGAATTCTAAGAGCTGATATCGAAACTTTGAGAACAAAGGTTAAGATGGCAGAGGAAACTGTGAAAAGAGT
GACAGGAGTGAACCCTTTGCATTGGTCAAGACCAAACATGGGCATACCATTCAGCAACACACCGAGTGCTTCCTCTAGTATT
CCGCCAAACTCTAACCACATTTTGAAACCAGCAAATTCAAGTACCAACACCAGTGCTGGTTTAGCACAAAATCAGAGAGTGG
AGACGGCAAATTTCTTGCCGGAACAGGTGAACCGTGAAGGCATGCAGAATCCATTTGCTCCTGATTCTAATCTGTATGAAAC
CCTACCCCATTGGAATCACAAGCATTAACCAGAAAAAGAGTCATTGGTTTTGTGATTTTGATTGAGGTAACTATTGTCTGTA
TTTTTATTTACTGTATGACTCAGCGACGGTAAACTCTTAATATTATATATATGATTGGACATGATGGAAGAATGTATAAGCT
ATCGTGCTTCTAAGGCAAATCGAAGTTGAAATGGTCGTAGACGTAATCACGTAAACATGTGCAGCAGCGTAAGCGTATAAAC
ACAAATGTATCAA

>gi|575417|emb|X82544.1| **Solanum tuberosum** mRNA leucine zipper transcription factor **1466pb**

GGAATTTTGATTTTCAAGATTCCATTCAATTTTTCTTCTATGGGTGTTTAGAAGGATTTAGGCTTTTAGAGTTTTGAAGTGG
GGAAAAAAAGGTTTCTTGAAGACTTTGCTGTTGTTGGCTTTAAGTCCAAAGGCAATGAATTCTTCAACATATACTCAATTTG
TTGCCTCTAAAAGGATGGGTATATGTGACCCAATCCATCAGATTGGCATGTGGGGAGATTTCAAAGGTAGCAGTTTCCCAGA
TACCTTGATTCTTGAAGTCGAGAATTGCCTAGAGAACGAGATGCCTATTATGGAGAAAAGACTAGAGAATGAGATAGAGGAA
CCATCACAAGTGACTGTTGGAATGTCTAACAGATATGAACCTGAAACAACTAAACGTATTGATAAGGTGCGTAGACGCCTTG
CACAAAACCGCGAGGCTGCTCGTAAAAGTCGTTTACGGAAGAAGGCCTATGTCCAGCAGTTGGAAAATAGTAAACTGAAGCT
GCTTCAGTTGGAACAAGAACTAGAACGTAATAGACAACAGGGTCTGTATGTAGGTGATGGTTTAGATGCTAGTCAGATAGGT
TGCTCTGGAACCGCAAATTCAGGAATAGCTTCTTTTGAAATGGAGTACGGCCATTGGGTGGAAGAGCAAGATAGACAAACAG
ATGATTTAAGGAATGCTCTGAACTCCCAAATGGGTGAAATAGAATTGCGCATTCTTGTCGAGAGTTGCTTGAATCACTATTT
TGATCTCTTTCGCTTGAAAGCTACAGCCGCAAATGCTGATGTTCTCTACCTTATGTCTGGCACATGGAAGACATCAGCTGAG
CGTTTCTTCTTGTGGATTGGGGGATTTCGCCCCTCCGAACTTCTAAAGGTTCTCACGCCACATGTGGAACCATTGTCAGATC
AACAAATCCAGGAGGTTAGCAACCTCACCCAGTCTTGTCAGCAGGCAGAAGATGCGTTGTCCCAAGGAATGGTAAAACTCCA
TCAGATTCTTGCTGAGGCTGTTGCAGCTGGTACACTAGGCGAAGGAATTATCCTTCCACAGATGACCGCTACCATTGAGAAG
TTGGAAGCTCTAGTTAGATTTGTAAATCAGGCAGATCATCTTCGCCAAGAAACCCTTCTACAGATGTCCTGCATACTGGCTG
CACACCAATCAGCTCAGGGTCTCCTTGCCTTAGGAGAGTACTTTAAACGTCTTCGTGCTCTTAGCTCACTTTGGGCTGGTCG
TCTTTCTGAACCGGCTTAAAGAAGTAAGAAACAACCATTATATATGCTACCTGGGGTTTTCAACTTTGCCAATCAAAGCTGT
GTGAGGTAAAATCTGCCTCGTGTTGCTCGTAAAAACTATGAGTTGGCGCTCATTCAGACACTACAAATCTGTATATAAGTTC
TGCTATGCTGTAACTGGAACTTGGAAATTGCGTTTATTTTCTTTTCACATTGGTTTTGATGATTGTAAATCA

>gi|4388769|gb|U37375.2| **Xenopus laevis** leucine zipper mRNA,**1188pb**

ATGAACGGCCTTGTTCCTTGCCCAACCCCTGTGGATCCACCTGCTCCTCCTCAGGTACCTCCCACATTAGAGGCTTATGCTC
CACCTCCACATGAAGAGATTTCCCCATCAACATCTCTTATTTCTGGCTCTGGCTTGCTCCTTGCTCGATCCCTGCTTGGCTG
TCGGAGCACTCCTTCCTCCCCTACCTCATCTACTTCCCCCGGCAATTCCCGGCGACGCAGGGAATTCACCCCAGATGAGAAG
AAAGATGATTGCTACTGGGACAAGAGGCGCAAGAACAATGAGGCTGCAAAACGCTCAAGAGAAAAGCGCCGTGCTGGAGATC
TTGCCCTGGAGGGCCGTGTCATTGCCCTTCTTGAAGAGAATGCCCGTCTCCGTGCTGAACTCCTTGCTCTACGTTTCCGCTT
TGGCCTGGTACGTGACCCATGTGAAGAAACTCGAGGAGGCTACTCAGCACAGTGTGGCCTCCATGAGCCTCCACCAGCCAAC
CCTCCTCCACCTCCTCCTCTCCCCCATTCAGAAGATTCTGGGTTCTCTACACCAAGTGTGGGCAGTCCTGTTTTCTTTGAAG
ACCGAGTGCCAGAACATGAACCTCAACCAATGCCTCATTCAGCTCTGTCTTATTATGGTCCCATAAATGGAGAAACTGTGGA
GAATCCTCGGGGAAGACTGGAGACTCTTGGAGATTGTTATAAGAGCCTTCCACATAAGCTGAGGTTTAAGGGAGGTGCTTCT
GGTGAAGAGGGGTCACATAGTATCATGCAAATGACAAGAGAAACAGGTGGTTTTGCCTCGGAGACTGCCACAGGGTTCTCAC
AGCAGCCTCCACCCTGTCCTCGCGGGCGATGGATGGCCCACAGTCAAGACGGAGTACCACCAGCTGCCTCTGAGAACTCAGA
GCTACGGTCCCAGCTGGCTTCTCTTTCTGCAGAGGTGGCACATTTAAAAAGAATTTTCTCCCAGCAGGTAGCGGCCCACGGG
GGCCATGAATGAAACCCTTCTACAGATGGAGCTCGACAGCAACGACTGCTAAAGTTGCCGAAAGCGCAGCAGAGATCCCTAA
TACTATAAAAGTAGGGATGTCCTTTTGATACGTCACATGTTCTGAACACAACACTTTCTAATGCCAGGCATTACAGCATCGA
CTACCAACTGTGTTGTTATATCCCATTTTTCTGTTGTGTC

>gi|78214358|ref|NM_152849.2| **Rattus norvegicus** homeobox and leucine zipper encoding (Homez), mRNA, **1785pb**

GTACATATGTGAAGGAAGAGTCAGCATGAGGACTAGACACCCAGAGAGAGCTGTCTCTGAAGGGTACAAATCAGAGCAGGTC
ATGAGTCCGAATAAAGACTCCAGCAGTCTCAACAGCTCGGGAGCAGGGCTTGTCTGCCTCCCGCCAGTCTCCGAGGAGCTAC
AGCTTGTGTGGACTCAAGCGGTCCAAACCAGTGAGCTGGACGGAAATGAGCACCTGCTACAGGCTTTCAGCTACTTCCCCTA
CCCAAGCCTGGCAGATATCGCTCTCCTCTGCCTGCGGCATGGGCTGCAGATGGAGAAGGTAAAGACCTGGTTCATGGCCCAA
CGTCTCCGCTGTGGCATTAGCTGGTCATCTGAAGAGATAGAAGAGACTCGAGCCAGAGTGGTGTACCACCGAGATCAGCTCC
TTTTCAAGTCTCTTCTGTCTTTCACACATCACGCAGTGAGGCCCCCACAGGAGATGCCTCCAGTGACCCCTCCAGAACAGGT
TGCTCTTGGACTGCGTCCTCTGGCGCTTAGCGAGCCCACTCAGATGAAAGGATTGAAGGTTGAGCCTGAGGAGCCCTTTCAG
GTATCACAGCTGCCACTGAATCATCAGAAAGTTAAGGAGCCTTTGATGATGGGCAGCAGAACATTGAACCACCAATCAGATT
GTCAGGATCTTCAGATCAGTGGCCTCTCTAAGGAGCAGGCAGGCGGGGGTCCTGACCAGTCATGTGGTGGAGGAACTGCTTC
CTGGAATCACTCCATAGCTGTCCATCAGCCAGATAAGTCCCCATTGATCTCATTACTTGATAATAGTTGTAAGGAGGAATCC
GAACCTAGTGGAACACCTCCATCTTCCTCTGCCTCTTCTCCTTTCCAGGTACTGGCTAATGGAACCACTGCCACCCCTAAAC
CCCTCCAGCCTTTGGGTTATATCCCACAGTCATTCTCACCTAGTGAGCAGGCACTGTCTCCACAAGTAGAACCACTCTGGTC
CCAAAGGCTACGGAATAACTCAGTACCGAGCAGGGTTGGCCCCACCGAATACCTTTCCCCAGATATGCAACACCAGCGAAAG
ACTAAGCGTAAAACCAAAGAACAGTTGGCCATCCTTAAATCCTTTTTCCTGCAGTGCCAATGGGCACGACGAGAAGATTACC
ATAAGTTAGAACAGATCACTGGTTTACCTCGCCCTGAGATCATTCAATGGTTTGGTGATACACGCTATGCCTTGAAGCATGG
ACAGCTGAAATGGTTTCGGGACAATGCAGTACTTGGTACTCCTAGTTTTCAAGATCCAGCCATTTCTACATCATCAACTCGT
TCCTTGAAAGAATGGGCCAAGACACCACCTCTACCAGCCCCACCGCCCCCACCAGATATACGACCTTTGGAGAGGTACTGGG
CAGCCCACCAGCAGCTGCAGGAAGCTGATATCCTTAAACTGAGTCAGGCATCAAGACTAAGCACTCAGCAAGTGCTGGACTG
GTTTGACTCTCGATTGCCCAAGCCAGCAGAAGTGGTAGTTTGTTTAGATGAAGAGGAAGAGGAGGAGGATGAAGAACTGCCA
GAAGATGGTGAGGAAGAAGAGGAGGAGGAGGAGGAAGAAGATGATGATGGTCTGTCTTGTGGACAGAAGGACCATGGTCTTC

TAACTAATGGTGCTGTAGAACGGGGATCACTGCTAACAGTTGGTTAGTCATCTGAAAGTAGAGTGTAAGCCCCAGTTGCTAA
GATAATAAAAAGGCCAAAAATAAAATTGTTTCTCTTCAAAAAAAAAAAAAAAAAAAAAAAAA

>gi|33943625|gb|AY346329.1| **Oryza sativa** homeodomain leucine-zipper protein Hox8 mRNA, **1272pb**
ATGGCCATCATCCATGACACCTCTGATCAACAAGGTACACCATAATTCTATTTCTTCCAAGCCAACTTTACATGCATGATCA
TACATGCACCAAAAACATGTACAATTCATTTAGTTTTTGTACGCATTTTGTAGCTAGTTGCTGAATTATGCGTGAATTTAGA
GACTTTTTTTTTGTTTTGCAATGGTAATTAACACGTAGAACATCATGTGTGTTGTGTTGCAGAGGACAACATGAGGTCGTAC
ATGGACGGCGGCGGCGCGGCGGCGTACGAGGAGGAGGAGGAGGAGGTTGAGGACGACGACGGCGGCGGCGGCGGCGGCGCG
GCGGCGGCGGTGGGGGGCTCGGGGAGAAGAAGCGGCGGCTGGCGGCGGAGCAGGTGCGGGCGCTGGAGCGGAGCTTCGAGGC
GGACAACAAGCTGGACCCGGAGCGGAAGGCCCGGATCGCCCGCGACCTTCGCCTCCACCCTCGCCAGGTCGCCGTCTGGTTC
CAGAACCGCCGCGCGAGGTGGAAGACCAAGCAGATCGAGCGCGACTTCGCCGCCCTCCGCTCCCGCCACGACGCCCTCCGCC
TCGAGTGCGACGCCCTCCGCCGCGACAAGGACGCCCTCGCCGCCGAGATCGCCGACCTCCGGGACAGGGTGGACGGCCAGAT
GTCCGTCAAGCTGGAGGCCGTGGCCGCGGACGAACACCAGCCGCCTCCGCCGCCGCCGCCGCCGCCACTGGCGTATAACAGC
AAGGTGGTGGACGGCTCGACGGACAGCGACTCGAGCGCGGTGTTCAACGAGGAGGCGTCGCCGTACTCCGGCGCGGCCATCG
ACCACCACCACCACCAAACTCCGGCGAGCTACGACACGGCGGGGTTCACCTCCTTCTTCGCGCCATCCACCACGCTCACCTC
GTCCCTCTCCTTCCCTTCCATGTTCCACGCGTCATCGCATTTCGATGGCCACCAAGAACTCCTCGTCGGCGGCGGCGGCGCC
GGCGCAGTGGCCGACGCCGACCTCGGAGGCGCCGGATTCTTCGCCGGCGACGAGCACGCCGGCGGCCTCTCCTGGTACGGCG
CCGAGGGTTGGTAGAAGCTAGAGCTTAGCTAGCTAGCGAGAGAGTGAGCTCAGCTAAGCTTAATTAGCTGGCTTGATTGCTT
GCTTTGTGGCTGGTTGGTGACGCCATTGTTGTAAATTACGGCATTGTTAAGTTGTACATGCATGCATGGGGGTAATTATAAC
TAAAGTCAATTTCTATAGTTTTTACTAAAAAAAAAAAAAAAA

>gi|62736387|gb|AY914051.1| **Triticum aestivum** putative leucine zipper protein (zip1) mRNA, **1585pb**
CGTTCCGCGCTGTCCGTCACACACAGGATGGCGTCCGCCATGGAGCTCTCCCTCCTCAGCCCCGCAATGCACCACCACGGCA
TCGCGGCCAAGACGGCCTCCCACCTCCCTGTTCTCCCCGCGCGCCGGGGCGCCGTCCGCTTCCGCGTGAGGGCCGCCGCCGC
GGCGCCGCCTGCACCCGCCGCGAAGCCCGGCTCGCCCAAGAAGCGGGGCAAGACGGAGGTCAACGAGTCGCTGCTCACGCCG
CGCTTCTACACCACGGACTTCGACGAGATGGAGCAGCTGTTCAACGCCGAGATTAACAAGCAGCTCAACCAGGACGAGTTCG
ACGCGCTGCTGCAGGAGTTCAAGACGGACTACAACCAGACCCACTTCATCCGCAACCCCGAGTTCAAGGAAGCTGCCGACAA
GATGCAGGGCCCGCTCCGCCAGATCTTCGTCGAGTTCCTCGAGCGCTCCTGCACCGCCGAGTTCTCCGGGTTCCTCCTCTAC
AAGGAGCTCGGCCGCAGGCTCAAGAAAACCAACCCGGTGGTGGCTGAGATCTTCTCGCTCATGTCCAGGGACGAGGCCCGGC
ACGCTGGGTTCTTGAACAAGGGGCTGTCCGACTTCAACCTGGCTCTGGACCTCGGCTTCTTGACCAAGGCTAGGAAGTACAC
CTTCTTCAAGCCAGAGTTCATCTTCTACGCCACATACCTGTCCGAGAAGATCGGCTACTGGAGGTACATCACCATCTTCAGG
CACCTAAAGGCCAACCCGGAGTACCAGGTGTACCCCATCTTCAAGTACTTCGAGAACTGGTGTCAGGACGAGAACCGGCATG
GCGATTTCTTCTCCGCGCTGCTCAAGGCGCAGCCGCAGTTCCTCAATGACTGGAAGGCCAAGCTCTGGTCACGCTTCTTCTG
CCTCTCGGTGTATATAACCATGTACCTGAATGACTGCCAACGTAGTGCCTTCTACGAAGGAATTGGTCTCAACACCAAAGAA
TTCGACATGCATGTCATCTATGAGACAAACCGCACGACGGCGAGGATCTTCCCTGCTGTACCGGATGTTGAGAACCCTGAAT
TCAAGAGGAAGCTAGACGGGATGGTAGATATCAACCTGAAGATCATTTCTATAGGAGAGTCCAACGACATGCCCCTGGTGAA
GAACCTGAAGAGGGTTCCTCTTATTGCCCAACTAGTGTCTGAGATCATCGCTGCGTACCTCATGCCCCCAATCGAGTCTGGC
TCCGTTGATTTTGCCGAGTTTGAGCCCAAGCTTGTCTACTGAATTTGTAGAAGAAGGATCCATCTCTGCCTTTCTTCTCAGA
CATAGTCATGCATCATGCTCCTCGAGAGTCTCTGAATGAGCACATGATCCATGGTTAATTAACAGGATCTACATCCTCCTGT
GCTCATCTGTAAAGTATTAAAACTCGGCAAGTTTTTGCTTAGTCACATCAACATGTAAGTGGCAGTGAACTGCATCAGAGCA
TGTGTTCCGGTTTTGCTTCAGGTGGAGAAGCATATCTGAAGTTGTTATGTAACTTGTGTCGATACTTGATTAAATAGCAATA
TAGCATCCGATTTTGTAAAAAAAAAAA

>gi|461682445|gb|JX424318.1| **Triticum monococcum** TGA-type basic leucine zipper protein mRNA, **1062pb**
ATGGCAGAGGCCAGCCCTAGAACAGAAACGTCAACAGATGATACTGATGAAAATCTTATGCTTGAACCAGGGAATGCTGCTC
TTGCTGTTGTTTCTGACTCTAGTGACAGATCCAGAGACAAAAACGGAGATCAAAAGACAATGCGTCGGCTTGCTCAAAATCG
TGAGGCTGCTAGGAAAAGTCGTTTGAGGAAAAAGGCATATGTTCAACAATTGGAGAACAGCAGGCTAAAGCTTACCCAGCTA
GAGCAGGAGTTGCAACGAGCTCGTCAACAAGGCATTTTTATATCTAGTTCAGCAGACCAGTCCCATTCCATGAGTGGAAATG
GGGCGTTGGCTTTTGACACAGAGTACGCACGGTGGTTGGAAGAACACAATCGACAAGTTAATGAGCTGAGAGCTGCAGTTAA
TGCTCATGCAGGCGATACTGAGCTGCGTAGTGTTGTTGAGAAGATCATGTCACACTATGATGAGATTTTTAAGCAAAAAGGA
AATGCAGCCAAAGCAGATGTCTTTCATGTGTTATCAGGCATGTGGAAGACACCAGCTGAGAGGTGTTTCCTATGGCTTGGAG
GTTTCCGACCTTCTGAGCTTTTAAAGCTTCTTTCGACCCAACTTGAACCCCTAACTGAGCAGCAGCTGTCAGGGATATGCAA
CCTTCAGCAATCATCACAACAAGCTGAGGATGCTCTTTCACAAGGAATGGAGGCTCTTCAGCAGTCTTTGGCAGAAACGTTG
GCTGGGTCTATCGGCTCTTCTGGATCTGGATCAACAGGAAATGTGGCAAACTACATGGGGCAAATGGGCCATGGCCATGGGA
AAGCTTGGAAACCCTTGAAAATTTCCTTAGTCAAGGCTGACACCCTGCGGCAGCAGACTCTTCAGCAGATGCAAAGGATCCT
TGACCACAGGCAGTCTGCCCGTGCACTTCTTGTGATAAGTGATTACTCATCCCGGCTTCGTGCCCTAAGTTCTCTTTGGCTT
TGCCTCGACCGAGATAAAAGGTTGGGCGCCGCCGACCAGCTTCTTGGTACAAAGTGGATCGATCAGGCCTGCCGATGA

>gi|308044466|ref|NM_001196644.1| **Zea mays** putative homeobox DNA-binding and leucine zipper domain family protein (LOC100502166), **1236pb**

TGCGCACGCCACCGCGCTTCATTGGCCACGCCGTTGCCATCACGCCGATTAAACTAGCCGATCGATCGCCCAGCTCGCCTGC
CTGTGATCGACCGGGTCGCTGCCTCCGATCCTCTTGCTGCCCAGCACCCTGGCTACTTCAGCCAGCTAGCCAGGTTGAGACC
GACTAGCTCGATCTAGCTGCTGAGGCGTGGCCATGGAGGGCGACGACGACGGCCCGGAGTGGATGATGGAGGTGGGCGGCGC
GGGCGCCACAGGGAAGGGAAAAGGCGGCGCGCTGGACAAGAACAAGAAGCGCTTCAGCGAGGAGCAGATCAAGTCTCTCGAG
TCCATGTTCGCCACGCAGACCAAGCTGGAGCCGCGCCAGAAGCTGCAGCTGGCGCGGGAGCTCGGCCTGCAGCCGCGCCAGG
TCGCCATCTGGTTCCAGAACAAGCGCGCGCGCTGGAAGTCCAAGCAGCTGGAGCGCGACTACTCCGCGCTCCGCGACGACTA
CGACGCGCTCCTCTGCAGCTACGAGTCCCTCAAGAAGGAGAAGCACACGCTCCTCAAGCAGCTGGAGAAGCTAGCCGAGATG
CTGCACGAGCCGCGGGGCAAGTACAGCGGCAATGCGGACGCCGCCGGCGCCGGGGACGACGTGCGCTCGGGCGTCGGCGGCA
TGAAGGACGAGTTTGCAGACGCCGGGGCCGCGCCCTACTCGTCCGAGGGCGGTGGCGGTGGCAAGTTCGCGCACTTCACGGA
CGACGACGTGGGAGCCCTCTTCCGGCCGTCGTCTCCGCAGCCGAGCGCCGCTGGCTTCACCTCGTCGGGGCCGCCGGAGCAC
CAGCCGTTCCAGTTCCACTCCGGCTGCTGGCCATCGTCGACGGAGCAGACCTGCAGCAGCTCGCAGTGGTGGGAGTTCGAGT
CCCTCAGTGAGTGAGTGTCTGAGTGATCGATCGCCAGACCATGCGACGGCGGGTCACTCGGTTCCAACTTCCAAGCACACAC
ACACACACACGTAAGCACGAATACGAGTTGGTAGCGGTCATCAGCCCCGAGCGCACGGTGTACATAGCTTTCAGTAGATCGA
ATTCCAGGCATGTCCATCAACAAGCAGTTTCTTCTCGTCATCGATCATCATGCATGCAAAAGAAAATTTTCTCTCCCCCATT
GTCGTCGCCGCTACCAGATCATGTAATCCAGGAACATGTAGAGAAAGATCAAACGAGCTTATAGAGAAGGGAGGTCACATGT
TCGATC

PARTIE 4 : EXERCICES D'APPLICATION

Exercice 1 : Interprétation d'un alignement
Deux séquences nucléiques ont été alignées. Quel est l'alignement optimal parmi les deux résultats obtenus.

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

1. Indiquez les homologies des deux séquences
2. Déterminez l'alignement optimal en utilisant les critères suivants : Coût de Substitution=2, Coût InDel=1 et Coût ouverture de gap = 4

Exercice 2 : Méthode UPGMA
Allez sur le site de GenBank et téléchargez, au format Fasta, les séquences ARNr 5S des bactéries suivantes : *Bacillus subtilis, Bacillus stearothermophilus, Lactobacillus viridescence, Acholeplasma modicum* et *Micrococcus luteus*.

1. Réalisez un alignement multiple des toutes ces séquences.
2. Construire l'arbre selon UPGMA.
3. Que signifie la théorie de l'horloge moléculaire sur laquelle se base UPGMA pour construire un arbre ?

Exercice 3 : Modèle Kimura

1. En s'appuyant sur la matrice de distances obtenue dans l'exercice 2, réalisez les transformations selon Kimura.
2. A quoi vont vous servir de telles transformations ?

Exercice 4 : Maximum de parcimonie
Une portion d'un alignement multiple sur 10 positions concerne cinq taxons.

	1	2	3	4	5	6	7	8	9	10
Taxon 1	C	G	T	T	A	T	C	A	T	A
Taxon 2	C	G	T	C	A	G	C	A	T	G
Taxon 3	C	C	T	T	A	G	C	T	T	A
Taxon 4	C	G	T	A	A	G	C	C	T	G
Taxon 5	C	C	T	T	A	T	C	T	T	C

1. Quels les sites informatifs de cet alignement multiple ?
2. A l'aide de la méthode UPGMA, tracez l'arbre en utilisant la longueur totale de cet alignement
3. En utilisant que les sites informatifs, reconstruisez l'arbre par la méthode du maximum de parcimonie.

Exercice 5 : Méthode Neighbor Joining

Soit l'alignement multiple réalisé sur quatre individus

Ind 1	TCA	CAA	GAA	CTG
Ind 2	TTA	TAA	GAA	CTG
Ind 3	TCG	TAA	GAG	CAG
Ind 4	CCA	CAA	GAA	CTG

1. A partir de cet alignement, déduisez la matrice de distances entre les quatre individus
2. Quelle est la différence principale entre NJ et UPGMA
3. Que signifie la valeur calculée dans la première étape de NJ ?
4. Construisez l'arbre selon NJ en précisant toutes les étapes de tous les cycles.
5. Calculez la longueur totale de cet arbre et concluez.

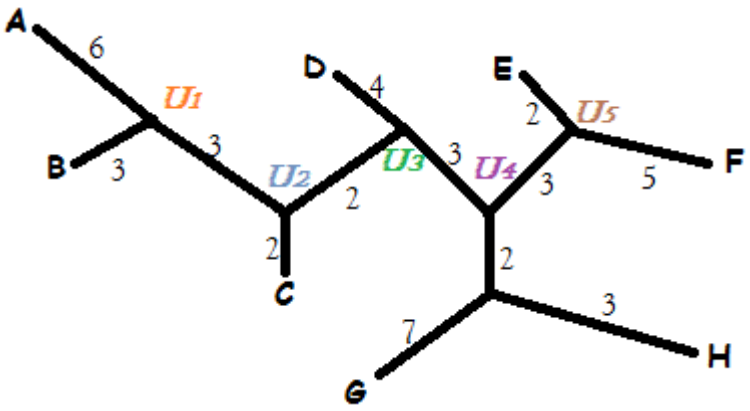
Exercice 6 : Indices de similarité et de distance

Les caractères d'identification pour quelques genres d'Entérobactéries sont portés sur le tableau suivant :

	Lac	ONPG	Ind	VP	Cit	Mob	Uréase	PDA	H_2S
Escherichia	+	+	+	-	-	+	-	-	-
Citrobacter	+	+	-	-	+	+	-	-	+
Enterobacter	+	+	-	+	+	+	-	-	-
Klebsiella	+	+	+	+	+	-	+	-	-
Serratia	-	+	-	+	+	+	-	-	-
Salmonella	-	-	-	-	+	+	-	-	+
Shigella	-	+	+	-	-	-	-	-	-
Proteus	-	-	+	-	+	+	+	+	+
Providencia	-	-	+	-	+	+	-	+	-
Yersinia	-	+	+	+	-	+	+	-	-

1. Calculez les indices de similarité symétrique et asymétrique.
2. Déduisez les deux matrices de distances.
3. Construisez les deux arbres qui en découlent selon UPGMA. Conclusion ?

Exercice 7 : Format de Newick. Soit la topologie de l'arbre suivant :



1. Calculez la distance entre les nœuds (A,B) et (E,F)
2. Ecrire l'arbre au format de Newick
3. A partir du format de Newick suivant, dessinez l'arbre : (((B,C),A),D).

Exercice 8 : Comparaison UPGMA et Maximum de parcimonie
On donne les séquences de quatre individus. Déterminez les sites informatifs et déduisez l'arbre avec le minimum de mutations. Reprenez les séquences et déduisez la matrice de distances pour tracer l'arbre UPGMA.
Les différences entre les deux méthodes sont – elles justifiées ?

Ind 1 : C T A T A A

Ind 2 : C T G T C T

Ind 3 : G T A A G T

Ind 4 : G A G A T C

Exercice 9 : On donne la matrice de distances entre cinq taxons. Corrigez ces distances selon Kimura à deux paramètres.

	Taxon 1	Taxon 2	Taxon 3	Taxon 4	Taxon 5
Taxon 1					
Taxon 2	0,20				
Taxon 3	0,50	0,40			
Taxon 4	0,45	0,55	0,15		
Taxon 5	0,40	0,50	0,40	0,25	

Références

Z. YANG. 2006. Computational Molecular Evolution. Oxford series in Ecology and Evolution. Oxford University Press. Oxford. U.K. 357p.

J. FELSENSTEIN. 2004. Inferring Phylogenies. Sinauer Associates, Inc. Sunderland. MA, U.S.A. 664p.

M. NEI and S. KUMAR. 2000. Molecular Evolution and Phylogenetics. Oxford University Press. Oxford, U.K. 333p.

J. GU and P. E. BOURNE. 2009. Structural Bioinformatics. Wiley & Sons Inc. Hoboken. New Jersey. U.S.A. 1035p.

G. NUEL et B. PRUM. 2007. Analyse statistique des séquences biologiques. Modélisation markovienne, alignements et motifs. Lavoisier. Paris. 361p.