UNIVERSITÉ DU QUÉBEC À MONTRÉAL

DÉVELOPPEMENT DE L'INTERFACE WEB DU LOGICIEL T-REX (TREE AND RETICULOGRAM RECONSTRUCTION)

MÉMOIRE
PRÉSENTÉ
COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN INFORMATIQUE

PAR

ADEL YOUNES

DÉCEMBRE 2006

REMERCIEMENTS

Ce mémoire se n'aurait vu le jour sans la confiance, la patience et la générosité de mon directeur de recherche, Monsieur Vladimir Makarenkov, que je veux vivement remercier. La pleine confiance qu'il m'a accordée, m'a permis d'élaborer un plan de travail personnel et propre à mes aspirations. Je voudrais aussi le remercier pour le temps et la patience qu'il m'a accordé, d'avoir cru en mes capacités et de m'avoir fourni d'excellentes conditions logistiques et financières.

Je remercie également mes collègues Alix Boc, Abdoulaye Baniré Diallo, Pablo Zentilli, Alpha Boubacar Diallo et Rabah Djema pour leurs points de vue et soutient qui ont été d'une grande utilité.

Je termine par remercier ma femme Soumaya El Mekki, ma mère, mon frère, ma sœur et mes amis pour leur soutien moral et leurs encouragements.

À tout ceux qui ont contribue de près ou de loin à la réalisation de ce projet, qu'ils trouvent ici toute l'expression de ma gratitude et ma profonde reconnaissance.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABRÉVIATIONS

ADN	Acide DésoxyriboNucléique
ARN	Acide RiboNucléique
DNA	Deoxyribonucleic acid
HGT	Horizontal Gene Transfer
LGT	Lateral Gene Transfer
ML	Maximum Likelihood
NJ	Neighbor Joining
OTU	Operational taxonomic unit
RBCL	RuBisCo Large subunit
RNA	RiboNucleic Acid
rRNA	Ribosomal RiboNucleic Acid
THG	Transfert Horizontal de Gène
TLG	Transfert Lateral de Gène
UPGMA	Unweighted Pair Group Method with Arithmetic mean

RÉSUMÉ

Depuis son apparition, le logiciel T-Rex (tree and reticulogram recontruction) est un outil puissant pour la reconstruction et la visualisation d'arbres phylogénétiques et de réticulogrammes. Un réticulogramme est un réseau phylogénétique permettant de représenter les phénomènes d'évolution réticulée tels que l'hybridation, la recombinaison génétique et le transfert latéral de gènes. La reconstruction peut être faite à partir des matrices de distances complètes, des matrices de distances incomplètes et des séquences moléculaires. Dans le but de permettre aux biologistes et aux bioinformaticiens de bénéficier des différentes options, modes de calcul, ainsi que des diverses nouveautés de T-Rex, nous avons développé la version Web de ce logiciel. Dans ce mémoire, nous décrivons les différentes fonctions, méthodes et outils inclus dans T-Rex Web. Nous décrivons aussi les nouveaux programmes ajoutés à T-Rex Web, tels que ClustalW, Calcul de la distance topologique de Robinson et Foulds, et Species Taxonomy. La dernière option permet de générer une matrice de distances d'arbre et de reconstruire des arbres phylogénétiques à partir des listes des lignées des espèces données. La version Web du logiciel T-Rex est disponible a l'adresse URL suivante : www.trex.uqam.ca.

Mots clés : T-Rex, arbre phylogénétique, matrice de distance, réticulogramme, interface Web, taxonomie d'espèces.

CHAPITRE I

1.1 Introduction

Dans les années 1980, grâce à EMBL et GenBank, premières banques de données moléculaires nous avons assisté à la naissance de la bioinformatique qui offre des méthodes et des logiciels permettant la gestion, l'organisation, l'analyse et l'exploitation des informations génétiques et génomiques pour l'élaboration de nouveaux concepts, ainsi que pour la prédiction et la production de nouvelles connaissances.

En biologie, les banques de données contribuent d'une façon directe à la construction de nouvelles connaissances en fournissant les données primaires à la bioinformatique.

On distingue essentiellement deux types des bases de données généralistes: bases de données généralistes offrant des informations hétérogènes suite à une collecte des données exhaustive et bases de données spécialisées offrant des informations homogènes établies autour d'une thématique.

Parmi les banques biologiques généralistes, il existe trois qui sont connues au niveau mondial, GenBank du NCBI, la banque de données d'ADN du Japon (DDBJ) et la banque du laboratoire de biologie moléculaire européen (EMBL). Ces banques de données ont des structures équivalentes, et font toutes partie de l'International Nucléotide Séquence Databases. Ils s'échangent des données quotidiennement pour offrir un ensemble de données cohérent et complet. Cette collaboration à pour conséquence que ces banques ont le contenu presque identique, qui pousse à réfléchir sur l'utilité de maintenir ces trois banques en même temps, et d'envisager peut-être un projet de fusion d'EMBL, de GenBank et de DDBJ .

1.2 Genbank (Genetic Sequence Bank)

GenBank est une banque de données publique de séquences d'acides nucléiques. Elle a été créée en 1988 par la société IntelliGenetics et diffusée par le NCBI (*National Centre for Biotechnology Information*, Los Alamos). NCBI a été fondée grâce à une législation spéciale de 1988 pour développer des systèmes d'information dans le domaine de la biologie moléculaire et pour soutenir la communauté des chercheurs dans le domaine biomédicale.

Depuis sa création, GenBank ne s'arrête de croître, d'une façon très exceptionnelle, elle double presque tous les 14 mois depuis sa création et offre une mise à jour tous les deux mois. En Août 2005, elle a atteint 100 milliards de paires de bases provenant d'environ de 165 000 organismes.

Il y a approximativement 51.674.486.881 bases dans 46.947.388 séquences dans les divisions traditionnelles de GenBank et 53.346.605.784 bases dans 10.276.161 séquences dans sa division de GT (Août 2005). Pour faciliter la consultation et l'archivage des données, la base a été répartie sur plusieurs divisions.

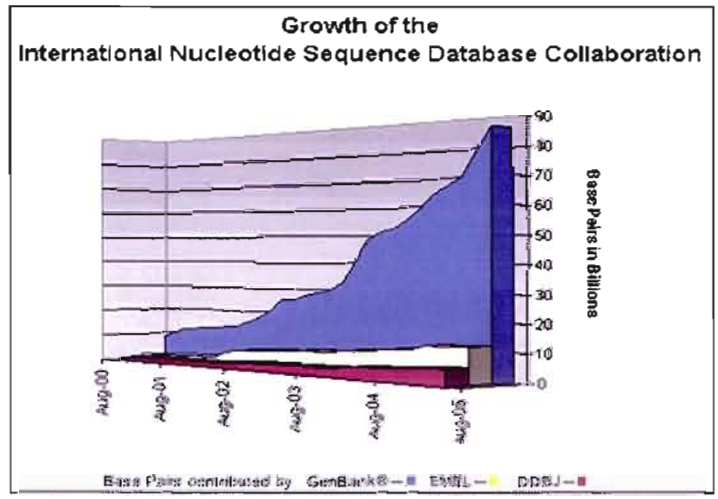


Figure 1.1 L'accroissement récent de la base de données Genbank

(Repris de la page Web http://www.ncbi.nlm.nih.gov/Genbank/index.html)

1.2.1 Format de stockage

Genbank est distribuée sous la forme d'un ensemble de fichiers texte, dans lesquels les séquences sont regroupées selon le critère taxonomique (e.g. virus, procaryotes, etc.) ou selon le critère d'origine (e.g. brevets, EST et STS).

Chaque séquence correspond a une entrée à laquelle on associe un nom (provenant du produit codé par la séquence et du nom de l'organisme d'origine), et d'un numéro d'accession, ainsi que d'autres informations liées à la séquence telles que sa structure, son rôle biologique, la nature de la molécule qui a été séquencée, etc. Ces informations sont structurées suivant un format précis.

1.2.1.1 Exemple d'entrée de Genbank : (gène de saccharomyces cerevisiae)

```
LOCUS       SCU49845     5028 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required
for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  PUBMED    7871890
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requires Axl2p, a novel
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
  PUBMED    8846915
REFERENCE   3  (bases 1 to 5028)
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University,
New
            Haven, CT, USA
FEATURES             Location/Qualifiers
     source          1..5028
                     /organism="Saccharomyces cerevisiae"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
                     /map="9"
```

```
     CDS             <1..206
                     /codon_start=3
                     /product="TCP1-beta"
                     /protein_id="AAA98665.1"
                     /db_xref="GI:1293614"

/translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEA
                     AEVLLRVDNIIRARPRTANRQHM"
     gene            687..3158
                     /gene="AXL2"
     CDS             687..3158
                     /gene="AXL2"
                     /note="plasma membrane glycoprotein"
                     /codon_start=1
                     /function="required for axial budding pattern of S.
                     cerevisiae"
                     /product="Axl2p"
                     /protein_id="AAA98666.1"
                     /db_xref="GI:1293615"

/translation="MTQLQISLLLTATISLLHLVVATPYEAYPIGKQYPPVARVNESF

TFQISNDTYKSSVDKTAQITYNCFDLPSWLSFDSSSRTFSGEPSSDLLSDANTTLYFN

VILEGTDSADSTSLNNTYQFVVTNRPSISLSSDFNLLALLKNYGYTNGKNALKLDPNE

VFNVTFDRSMFTNEESIVSYYGRSQLYNAPLPNWLFFDSGELKFTGTAPVINSAIAPE

TSYSFVIIATDIEGFSAVEVEFELVIGAHQLTTSIQNSLIINVTDTGNVSYDLPLNYV

YLDDDPISSDKLGSINLLDAPDWVALDNATISGSVPDELLGKNSNPANFSVSIYDTYG

DVIYFNFEVVSTTDLFAISSLPNINATRGEWFSYYFLPSQFTDYVNTNVSLEFTNSSQ

DHDWVKFQSSNLTLAGEVPKNFDKLSLGLKANQGSQSQELYFNIIGMDSKITHSNHSA

NATSTRSSHHSTSTSSYTSSTYTAKISSTSAAATSSAPAALPAANKTSSHNKKAVAIA

CGVAIPLGVILVALICFLIFWRRRRENPDDENLPHAISGPDLNNPANKPNQENATPLN

NPFDDDASSYDDTSIARRLAALNTLKLDNHSATESDISSVDEKRDSLSGMNTYNDQFQ

SQSKEELLAKPPVQPPESPFFDPQNRSSSVYMDSEPAVNKSWRYTGNLSPVSDIVRDS

YGSQKTVDTEKLFDLEAPEKEKRTSRDVTMSSLDPWNSNISPSPVRKSVTPSPYNVTK

HRNRHLQNIQDSQSGKNGITPTTMSTSSSDDFVPVKDGENFCWVHSMEPDRRPSKKRL
                     VDFSNKSNVNVGQVKDIHGRIPEML"
     gene            complement(3300..4037)
                     /gene="REV7"
     CDS             complement(3300..4037)
                     /gene="REV7"
                     /codon_start=1
                     /product="Rev7p"
```

```
                     /protein_id="AAA98667.1"
                     /db_xref="GI:1293616"

/translation="MNRWVEKWLRVYLKCYINLILFYRNVYPPQSFDYTTYQSFNLPQ

FVPINRHPALIDYIEELILDVLSKLTHVYRFSICIINKKNDLCIEKYVLDFSELQHVD

KDDQIITETEVFDEFRSSLNSLIMHLEKLPKVNDDTITFEAVINAIELELGHKLDRNR

RVDSLEEKAEIERDSNWVKCQEDENLPDNNGFQPPKIKLTSLVGSDVGPLIIHQFSEK
                     LISGDDKILNGVYSQYEEGESIFGSLF"
ORIGIN
        1 gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac ggaaccattg
       61 ccgacatgag acagttaggt atcgtcgaga gttacaagct aaaacgagca gtagtcagct
      121 ctgcatctga agccgctgaa gttctactaa gggtggataa catcatccgt gcaagaccaa
      181 gaaccgccaa tagacaacat atgtaacata tttaggatat acctcgaaaa taataaaccg
      241 ccacactgtc attattataa ttagaaacag aacgcaaaaa ttatccacta tataattcaa
      301 agacgcgaaa aaaaaagaac aacgcgtcat agaacttttg gcaattcgcg tcacaaataa
      361 attttggcaa cttatgtttc ctcttcgagc agtactcgag ccctgtctca agaatgtaat
      421 aatacccatc gtaggtatgg ttaaagatag catctccaca acctcaaagc tccttgccga
      481 gagtcgccct cctttgtcga gtaattttca cttttcatat gagaacttat tttcttattc
      541 tttactctca catcctgtag tgattgacac tgcaacagcc accatcacta gaagaacaga
      601 acaattactt aatagaaaaa ttatatcttc ctcgaaacga tttcctgctt ccaacatcta
      661 cgtatatcaa gaagcattca cttaccatga cacagcttca gatttcatta ttgctgacag
      721 ctactatatc actactccat ctagtagtgg ccacgcccta tgaggcatat cctatcggaa
      781 aacaataccc cccagtggca agagtcaatg aatcgtttac atttcaaatt tccaatgata
      841 cctataaatc gtctgtagac aagacagctc aaataacata caattgcttc gacttaccga
      901 gctggctttc gtttgactct agttctagaa cgttctcagg tgaaccttct tctgacttac
      961 tatctgatgc gaacaccacg ttgtatttca atgtaatact cgagggtacg gactctgccg
     1021 acagcacgtc tttgaacaat acataccaat ttgttgttac aaaccgtcca tccatctcgc
     1081 tatcgtcaga tttcaatcta ttggcgttgt taaaaaacta tggttatact aacggcaaaa
     1141 acgctctgaa actagatcct aatgaagtct tcaacgtgac ttttgaccgt tcaatgttca
     1201 ctaacgaaga atccattgtg tcgtattacg gacgttctca gttgtataat gcgccgttac
     1261 ccaattggct gttcttcgat tctggcgagt tgaagtttac tgggacggca ccggtgataa
     1321 actcggcgat tgctccagaa acaagctaca gttttgtcat catcgctaca gacattgaag
     1381 gattttctgc cgttgaggta gaattcgaat tagtcatcgg ggctcaccag ttaactacct
     1441 ctattcaaaa tagtttgata atcaacgtta ctgacacagg taacgtttca tatgacttac
     1501 ctctaaacta tgtttatctc gatgacgatc ctatttcttc tgataaattg ggttctataa
     1561 acttattgga tgctccagac tgggtggcat tagataatgc taccatttcc gggtctgtcc
     1621 cagatgaatt actcggtaag aactccaatc ctgccaattt ttctgtgtcc atttatgata
     1681 cttatggtga tgtgatttat ttcaacttcg aagttgtctc cacaacggat ttgtttgcca
     1741 ttagttctct tcccaatatt aacgctacaa ggggtgaatg gttctcctac tatttttgc
     1801 cttctcagtt tacagactac gtgaatacaa acgtttcatt agagtttact aattcaagcc
     1861 aagaccatga ctgggtgaaa ttccaatcat ctaatttaac attagctgga gaagtgccca
     1921 agaatttcga caagctttca ttaggtttga aagcgaacca aggttcacaa tctcaagagc
     1981 tatattttaa catcattggc atggattcaa agataactca ctcaaaccac agtgcgaatg
     2041 caacgtccac aagaagttct caccactcca cctcaacaag ttcttacaca tcttctactt
     2101 acactgcaaa aatttcttct acctccgctg ctgctacttc ttctgctcca gcagcgctgc
     2161 cagcagccaa taaaacttca tctcacaata aaaaagcagt agcaattgcg tgcggtgttg
     2221 ctatcccatt aggcgttatc ctagtagctc tcatttgctt cctaatattc tggagacgca
     2281 gaagggaaaa tccagacgat gaaaacttac cgcatgctat tagtggacct gatttgaata
     2341 atcctgcaaa taaaccaaat caagaaaacg ctacaccttt gaacaacccc tttgatgatg
     2401 atgcttcctc gtacgatgat acttcaatag caagaagatt ggctgctttg aacactttga
     2461 aattggataa ccactctgcc actgaatctg atatttccag cgtggatgaa aagagagatt
     2521 ctctatcagg tatgaataca tacaatgatc agttccaatc ccaaagtaaa gaagaattat
     2581 tagcaaaacc cccagtacag cctccagaga gcccgttctt tgacccacag aataggtctt
     2641 cttctgtgta tatggatagt gaaccagcag taaataaatc ctggcgatat actggcaacc
```

```
     2701 tgtcaccagt ctctgatatt gtcagagaca gttacggatc acaaaaaact gttgatacag
     2761 aaaaactttt cgatttagaa gcaccagaga aggaaaaacg tacgtcaagg gatgtcacta
     2821 tgtcttcact ggacccttgg aacagcaata ttagcccttc tcccgtaaga aaatcagtaa
     2881 caccatcacc atataacgta acgaagcatc gtaaccgcca cttacaaaat attcaagact
     2941 ctcaaagcgg taaaaacgga atcactccca caacaatgtc aacttcatct tctgacgatt
     3001 ttgttccggt taaagatggt gaaaattttt gctgggtcca tagcatggaa ccagacagaa
     3061 gaccaagtaa gaaaaggtta gtagattttt caaataagag taatgtcaat gttggtcaag
     3121 ttaaggacat tcacggacgc atcccagaaa tgctgtgatt atacgcaacg atattttgct
     3181 taattttatt ttcctgtttt atttttatt agtggtttac agatacccta tattttattt
     3241 agtttttata cttagagaca tttaatttta attccattct tcaaatttca tttttgcact
     3301 taaaacaaag atccaaaaat gctctcgccc tcttcatatt gagaatacac tccattcaaa
     3361 attttgtcgt caccgctgat taatttttca ctaaactgat gaataatcaa aggccccacg
     3421 tcagaaccga ctaaagaagt gagttttatt ttaggaggtt gaaaaccatt attgtctggt
     3481 aaattttcat cttcttgaca tttaacccag tttgaatccc tttcaatttc tgctttttcc
     3541 tccaaactat cgaccctcct gtttctgtcc aacttatgtc ctagttccaa ttcgatcgca
     3601 ttaataactg cttcaaatgt tattgtgtca tcgttgactt taggtaattt ctccaaatgc
     3661 ataatcaaac tatttaagga agatcggaat tcgtcgaaca cttcagtttc cgtaatgatc
     3721 tgatcgtctt tatccacatg ttgtaattca ctaaaatcta aaacgtattt ttcaatgcat
     3781 aaatcgttct ttttattaat aatgcagatg gaaaatctgt aaacgtgcgt taatttagaa
     3841 agaacatcca gtataagttc ttctatatag tcaattaaag caggatgcct attaatggga
     3901 acgaactgcg gcaagttgaa tgactggtaa gtagtgtagt cgaatgactg aggtgggtat
     3961 acatttctat aaaataaaat caaattaatg tagcatttta agtataccct cagccacttc
     4021 tctacccatc tattcataaa gctgacgcaa cgattactat ttttttttc ttcttggatc
     4081 tcagtcgtcg caaaaacgta taccttcttt ttccgacctt ttttttagct ttctggaaaa
     4141 gtttatatta gttaaacagg gtctagtctt agtgtgaaag ctagtggttt cgattgactg
     4201 atattaagaa agtggaaatt aaattagtag tgtagacgta tatgcatatg tatttctcgc
     4261 ctgtttatgt ttctacgtac ttttgattta tagcaagggg aaaagaaata catactattt
     4321 tttggtaaag gtgaaagcat aatgtaaaag ctagaataaa atggacgaaa taaagagagg
     4381 cttagttcat ctttttttcca aaaagcaccc aatgataata actaaaatga aaaggatttg
     4441 ccatctgtca gcaacatcag ttgtgtgagc aataataaaa tcatcacctc cgttgccttt
     4501 agcgcgtttg tcgtttgtat cttccgtaat tttagtctta tcaatgggaa tcataaattt
     4561 tccaatgaat tagcaatttc gtccaattct ttttgagctt cttcatattt gctttggaat
     4621 tcttcgcact tcttttccca ttcatctctt tcttcttcca aagcaacgat ccttctaccc
     4681 atttgctcag agttcaaatc ggcctctttc agtttatcca ttgcttcctt cagtttggct
     4741 tcactgtctt ctagctgttg ttctagatcc tggtttttct tggtgtagtt ctcattatta
     4801 gatctcaagt tattggagtc ttcagccaat tgctttgtat cagacaattg actctctaac
     4861 ttctccactt cactgtcgag ttgctcgttt ttagcggaca aagatttaat ctcgttttct
     4921 ttttcagtgt tagattgctc taattctttg agctgttctc tcagctcctc atatttttct
     4981 tgccatgact cagattctaa ttttaagcta ttcaatttct ctttgatc
//
```

Repris de la page Web http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html

1.2.1.2 Description des champs

a) LOCUS : Un locus définit l'emplacement d'un allèle ou d'un gène sur un chromosome. Ce champ contient un certain nombre d'informations. Chaque élément important est décrit ci-dessous.

. Locus Nene : (Nom du lieu). Dans cet exemple, il s'agit de SCU49845. Ce champ a été conçu pour faciliter des entrées de groupe avec des séquences semblables :

Habituellement, les trois premiers caractères désignent l'organisation; le quatrième et le cinquième caractère sont utilisés pour désigner d'autres groupes de désignations, (tel que le produit de gène). Pour les entrées segmentées, le dernier caractère vient d'une série de nombres entiers séquentiels.

Cependant, les dix caractères ne sont plus suffisants pour représenter la quantité d'information qui doit être contenue dans le nom de lieu. La seule règle appliquée est l'assignation d'un nom de lieu unique. Par exemple, pour les enregistrements de Genbank qui ont un numéro d'accession à six caractères (par exemple, U12345), le nom de lieu est habituellement la première lettre du genre et du nom d'espèce, suivie du numéro d'accession. Pour les numéros d'accessions à 8 caractères (par exemple, AF123456), le nom de lieu est le numéro d'accession.

La base de données, RefSeq, des références de séquences assignées à chaque enregistrement, un nom de lieu formel, basé sur le symbole de gène. RefSeq est séparé de la base de données GenBank, mais contient des références croisées à des enregistrements correspondants dans GenBank.

Champ De Recherche : Numéro D'Accession [ACCN].

Astuces de recherche : Il est préférable de rechercher par le numéro d'accession réel plutôt que par le nom de lieu, les premiers sont stables alors que les noms de lieu peuvent changer.

. Sequence Length : (Longueur de séquence) : Est le nombre de paires de base (bp). Dans l'exemple (Figure numéro), la longueur de la séquence est de 5028 bp. La taille minimale d'une séquence soumise à GenBank est de 50 bp. Il n'y a aucune limite sur la taille maximale. On peut soumettre à GenBank un génome en entier, si on a des morceaux contigus d'une séquence d'une molécule simple. Cependant, il y' a une limite maximale de 350 KB pour un enregistrement de GenBank.

Champ De Recherche : Longueur de séquence [SLEN].

Astuces de recherche : (1) Pour rechercher un enregistrement dans une marge de longueurs on utilise « : » comme opérateur (par exemple 2500:2600[SLEN]).

(2) Pour rechercher un enregistrement plus court qu'un certain nombre, on choisit 2 comme limite inférieure, par exemple, 2:100[SLEN].

(3) Pour rechercher tous les enregistrements plus longs qu'un certain nombre, on utilise une série de 9 comme limite supérieure, (par exemple, 25000:99999999[SLEN]).

. **Molécule Type** (Type de molécule) : dans cet exemple il s'agit d'ADN comme type de molécule. Chaque enregistrement de GenBank doit contenir des données de séquence contiguës d'un type de molécule simple. Les divers types de molécules sont décrits dans la documentation de Sequin et peuvent inclure l'ADN génomique, l'ARN génomique, l'ARN de précurseur, le mRNA (ADN), l'ARN ribosomal, l'ARN de transfert, le petit ARN nucléaire, et le petit ARN cytoplasmique.

. **GenBank Division** (division de GenBank) : C'est la division de GenBank à laquelle l'enregistrement appartient, elle est indiquée avec une abréviation de trois lettres. Dans cet exemple, la division de GenBank est PLN.

La base de données GenBank est Divisée en 18 divisions :

1. PRI - primate sequences : (séquences des primates).
2. ROD - rodent sequences : (séquences des rongeurs).
3. MAM - other mammalian sequences : (séquences des autres collagènes de mammifère).
4. VRT - other vertebrate sequences : (séquences des autres vertèbres).
5. INV - invertebrate sequences : (séquences des invertébrés).
6. PLN - plant, fungal, and algal sequences : (séquences des plantes, champignons, et des algues).

7. BCT - bacterial sequences : (séquences des bactéries).
8. VRL - viral sequences : (séquences des virus).
9. PHG - bacteriophage sequences (séquences de bactériophage).
10. SYN - synthetic sequences (séquences des synthétiques).
11. UNA - unannotated sequences (séquences non annotées).
12. EST - EST sequences (expressed sequence tags) : (séquences des EST).
13. PAT - patent sequences : (séquences brevetés).
14. STS - STS sequences (sequence tagged sites) : (séquences des STS).
15. GSS - GSS sequences (genome survey sequences) : (séquences d'enquête de génome).
16. HTG - HTG sequences (high-throughput genomic sequences) : (séquences génomique de haut sortie).
17. HTC - unfinished high-throughput cDNA sequencing:
18. ENV - environmental sampling sequences.

Certaines divisions contiennent des séquences provenant de certains groupes d'organisations, tandis que d'autres (EST, GSS, HTG, etc.) contiennent des données générées par des technologies de séquençage spécifiques provenant de beaucoup d'autre organisations différentes. Pour la recherche des séquences provenant des organisations particulières on devrait utiliser le Navigateur de taxonomie de NCBI.

Une division, appelée CON, a été ajoutée dans la version 115.0 (décembre 1999) mais n'est pas énumérée ci-dessus parce qu'elle se trouve encore au stade expérimental. Dans cette division, les enregistrements ne contiennent aucune donnée de séquences, mais contiennent des instructions sur la façon de construire des contigs à partir d'enregistrements multiples de GenBank.

Champ de recherche : Propriétés [PROP].

Astuces de recherche : les formats sont les suivants : gbdiv_pri, gbdiv_est servent à éliminer toutes les séquences d'une division donnée, telle que tout les EST. On peut utiliser une question booléenne formatée comme suit : human [ORGN] NOT gbdiv_est [PROP].

Pour les raisons citées ci-dessus, il ne faut pas utiliser les divisions de GenBank pour rechercher toutes les séquences d'un organisme spécifique. Il est recommandé d'utiliser le navigateur de taxonomie de NCBI à la place.

. **Modification date** (date de modification) : La date dans le champs locus est la date de dernière modification. Dans cet exemple elle est le 21 juin 1999. Dans certains cas, la date de modification correspond à celle de publication, mais il n'y a aucune manière de le savoir rien qu'en lisant l'enregistrement. Si on veut connaître la date de publication, il faut envoyer un message à info@ncbi.nlm.nih.gov.

Champs de recherche : Date de modification [MDAT].

Astuces de recherche : (1) Le format de date de recherche doit être sous la forme de yyyy/mm/dd, par exemple, 1999/07/25.

(2) Pour rechercher des enregistrements modifiés entre deux dates, on utilise « : » pour séparer la date de début et la date de fin (par exemple 1999/07/25 :1999 /0 7/ 31 [MDAT].)

- **DEFINITION :** Courte description de la séquence ; inclut l'information telle que l'organisme source, le nom du gène ou de la protéine, ou une certaine description de la fonction de la séquence.

Champs de recherche : Mot du titre [TITL].

Astuces de recherche : Bien que les lignes de définition de nucléotide suivent un format structuré, GenBank n'emploie pas un vocabulaire contrôle. Les auteurs déterminent le contenu de leurs enregistrements eux-mêmes. Par conséquent, si une recherche n'aboutit, pas les auteurs peuvent la relancer avec des termes qu'ils ont utilisés comme des synonymes, des abréviations, etc.

- **Accession:** C'est l'identifiant unique d'un enregistrement, il est habituellement une combinaison de chiffres et de lettres, comme une lettre suivie de cinq chiffres (par exemple, U12345) ou de deux lettres suivies de six chiffres (par exemple, AF123456). Certains numéros d'accessions pourraient être plus longs que les autres, en fonction du type d'enregistrement de la séquence.

Les numéros d'accession ne changent pas, même si les informations contenues dans l'enregistrement changent suite à la demande de l'auteur. Par contre, un numéro d'accession primaire pourrait devenir secondaire pour un nouveau numéro d'accession si les auteurs font une nouvelle soumission qui combine des séquences précédentes.

Les enregistrements de la base de données de RefSeq des références de séquences ont un format différent de numéro d'accession qui commence par deux lettres suivies d'une barre de soulignement et de six chiffres ou plus. Par exemple

NT_123456 constructed genomic contigs
NM_123456 mRNAs
NP_123456 proteins
NC_123456 chromosomes

Repris de la Page : http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html

Champs de recherche : Numéro d'accession [ACCN]

Astuces de recherche : Les lettres dans le numéro d'accession peuvent être écrites en majuscule ou en minuscule. Les numéros d'accessions de RefSeq doivent contenir une barre de soulignement entre les lettres et les nombres (par exemple, NM_002111).

- **Version :** C'est un numéro d'identification qui utilise un format spécifique mis en application en février 1999 par GenBank/EMBL/DDBJ. S'il y' a n'importe quel changement aux données de la séquence (même une base simple), le numéro de version s'incrémente de un (par exemple, U12345.1, U12345.2), mais le numéro d'accession reste fixe.

. GI : ("GenInfo Identifier") : Numéro d'identifiant de la séquence, dans ce cas-ci, pour une séquence de nucléotide. Si la séquence change, un nouveau nombre de GI lui sera assigné.

Il existe un champs pour la révision de l'historique des séquences, et pour dépister les divers nombres de GI. Le nombre des versions, et des dates de mise à jour pour les séquences est apparu dans un enregistrement spécifique de GenBank.

- KEYWORDS (Mots-clés) : Mot ou expression décrivant la séquence. Si l'entrée ne contient aucun mot-clé, alors le champ contient seulement une période. Ce champ existe principalement pour des raisons historiques.

- SOURCE : Cette information a un format libre, elle contient une forme abrégée du nom d'organisme, parfois suivie d'un type de molécule.

. Organisme : Le nom scientifique formel de l'organisme de source (genre et espèces) et sa lignée, basée sur la classification phylogénétique utilisée dans la base de données de taxonomie de NCBI.

Si la lignée complète d'un organisme est très longue, une lignée abrégée la remplacera dans l'enregistrement de Genbank, et la lignée complète sera disponible dans la base de données de taxonomie.

Champs de recherche : Le nom de l'organisme [ORGN].

Astuces de recherche : On peut chercher le champ nom d'organisme par n'importe quel noeud dans la hiérarchie taxonomique. Par exemple, on peut chercher le terme "saccharomyces cerevisiae", "Saccharomycetales", "Ascomycota", etc. pour trouver toutes les séquences des organismes dans un taxon particulier.

- **RÉFÉRENCE :** Les publications réalisées par les auteurs des séquences. Les références sont automatiquement triées selon la date de publication en débutant par les plus anciennes.

Quelques séquences n'ont pas étés publiés dans des revues officielles et montrent un statut de "non publié" ou "en cours d'impression". Quand le numéro d'accession et/ou les données des séquences apparaissent dans la copie, les auteurs de séquences devraient envoyer la citation complète de l'article à update@ncbi.nlm.nih.gov et le personnel de Genbank mettra l'enregistrement à jour.

. Authors (Auteurs) : Liste d'auteurs selon l'ordre d'apparition dans l'article cité.

Champs de recherche : Auteur [AUTH].

Astuces de recherche : Permet d'omettre les initiales, et d'utiliser des mots tronqués pour trouver tout les noms qui commencent par les caractères introduits (par exemple, Richards * ou Boguski M *).

. Title (Titre) : Titre du travail édité ou titre expérimental d'un travail non publié.

Parfois on retrouve l'expression "Direct Submission" qui veut dire "soumission directe" à la place d'un titre d'article.

. Journal : Abréviation du nom de la revue.

.PubMed : Identifiant de (PMID). Ce sont des références de correspondance pour les enregistrements de PubMed.

.Direct Submission : Informations de celui qui a fait la soumission, telle que l'institut, le département et adresse postale. Certains anciens enregistrements ne contiennent pas cette information. Le champ d'auteurs contient le nom de la personne qui a fait la soumission, le titre contient les mots "soumission directe". La date dans le sous champ de journal est la date de préparation de la soumission. Dans plusieurs cas, c'est également la date de réception de la séquence par le personnel de GenBank, qui diffère de la date de première publication.

. **Features** (Dispositif) : Informations sur les gènes et les produits de gènes.

. **Source :** Dispositif obligatoire dans chaque enregistrement, il récapitule la longueur de la séquence, le nom scientifique de l'organisme source, et le nombre d'identification du Taxon. Il peut aussi inclure d'autres informations telle que l'endroit de carte, clone, type de tissu, etc.,

. **Taxon :** Numéro d'identification unique stable pour le taxon de l'organisme source. Un nombre d'identification taxonomique est assigné à chaque taxon (espèces, genre, famille, etc...) dans la base de données de taxonomie de NCBI.

.CDS : Séquence programmée ; la région des nucléotides qui correspond à la séquence des acides aminés dans une protéine (endroit inclut des codons de début et d'arrêt). Le dispositif de CDS inclut une traduction d'acide aminé. Les auteurs peuvent indiquer la nature des CDS en employant le qualificateur "/evidence=experimental" ou /evidence = not_experimental".

. **<1..206** : Envergure basse du dispositif biologique indiqué vers la gauche, dans ce cas-ci, un dispositif de CDS. (Le dispositif de CDS est décrit ci-dessus, et son envergure basse inclut les codons de début et d'arrêt.) Les dispositifs peuvent être complets, partiels sur l'extrémité 5', partiels sur l'extrémité 3', et/ou sur la rive complémentaire. Exemples :

1. Le dispositif complet est écrit comme n..m exemple : 687...3158 Le dispositif étendue de la base 687 à la base 3158 dans l'ordre montré.

2. Dispositif partiel sur la cinquième extrémité. Exemple : 1..206, Le dispositif se prolonge de la base 1 à la base 206 et il est partiel sur la cinquième extrémité.

3. Dispositif partiel sur la troisième extrémité. Exemple : 4821...5028. Le dispositif se prolonge de la base 4821 à la base 5028 et est partiel sur la troisième extrémité.

4. Le complément indique que le dispositif se retrouve sur la rive complémentaire.

Exemple : complément (3300..4037).

Le dispositif se prolonge de la base 3300 à la base 4037, mais il est réellement sur la rive complémentaire.

- **Protein_id :** Numéro d'identification de la séquence de protéines, dans notre cas pour la traduction de la protéine. Les identifications de protéine se composent de trois lettres suivies de cinq chiffres, d'un point, et d'un numéro de version. Pour tout changement des données des séquences (même un acide aminé simple), le numéro de version s' incrémente, mais la partie du numéro d'accession demeure stable (par exemple, AAA98665.1 change en AAA98665.2).

La version du format du numéro d'accession du numéro d'identification de la séquence de protéine a été mis en application par GenBank/EMBL/DDBJ en février 1999 et fonctionne parallèlement au système de numération de GI.

- **GI** "GenInfo Identifier" : Numéro d'identification de la séquence, dans ce cas il s'agit de la synthèse de protéine. Le système GI d'identification de séquence fonctionne parallèlement avec le système de version des numéros de séquences qui a été mis en application par GenBank, EMBL et DDBJ en février 1999. Par conséquent, si la séquence de protéine change, elle recevra un nouveau numéro de GI, et le suffixe du « protein_id » sera incrémenté de un.

- **Translation :** La synthèse d'acide aminé correspondant à la séquence programmée de nucléotide (CDS). Dans plusieurs cas, les synthèses sont conceptuelles. Notons bien que les auteurs peuvent indiquer si les CDS sont basés sur l'évidence expérimentale ou non expérimentale.

- **Gene :** Une région d'intérêt biologique identifiée comme gène pour laquelle un nom a été assigné. L'envergure basse pour le dispositif de gène dépend des autres dispositifs 5' et 3'. D'autres exemples d'enregistrements qui montrent le rapport entre les dispositifs de gène et d'autres dispositifs tels que le mRNA et les CDS telles que AF165912 et AF090832.

- **Complement :** Indique que le dispositif est situé sur la rive complémentaire.

. **Other Features :** Exemples d'autres enregistrements qui montrent une variété de dispositifs biologiques. Un format graphique est également disponible pour chaque enregistrement et représente visuellement les dispositifs annotés.

-Origin (Origine) : L'origine peut être laissée en blanc, peut apparaître comme "Unreported," qui veut dire "non rapportée," ou peut donner un pointeur local au début de la séquence, impliquant habituellement un emplacement expérimentalement déterminé de fendage de restriction ou le lieu génétique (si disponible). Cette information est présentée seulement dans les anciens enregistrements.

1.2.2 Soumission à GenBank

De nombreuses revues exigent la soumission des informations sur les séquences à une base de données avant la publication de sorte qu'un numéro d'accession puisse apparaître dans le papier. L'outil Internet des soumissions, appelé BankIt (http://www.ncbi.nlm.nih.gov/BankIt/), pour la soumission commode et rapide des données des séquences. D'un autre coté Sequin (http://www.ncbi.nlm.nih.gov/Sequin/index.html), logiciel autonome de soumission de NCBI pour MAC (Macintosh), PC, et les plateformes d'UNIX, est disponible par FTP. En utilisant Sequin, les dossiers résultats de la soumission directe devraient être envoyés à GenBank par courrier électronique.

La source la plus importante de nouvelles données pour GenBank est la soumission directe par les chercheurs scientifiques. NCBI fournit le traitement opportun et précis et l'examen biologique de nouvelles entrées et mises à jour aux entrées existantes.

1.2.2.1 Réception de numéro d'accession pour le manuscrit

La plupart des revues s'attendent à ce que les séquences d'ADN et d'acide aminé qui apparaissent en articles soient soumises à une base de données de séquences avant la publication. Peu après la soumission, l'auteur reçoit un numéro d'accession de la base de données qu'il peut mentionner dans son article comme référence à sa séquence. L'échange de

données entre GenBank, EMBL et DDBJ se fait quotidiennement. Des données de séquence soumises avant la publication peuvent être maintenues confidentielles si on le demande.

Il y a diverses manières de soumettre des séquences d'ADN à GenBank. Essentiellement, il y en a deux principales, BankIt et Sequin. BankIt est un outil de soumission par Internet. Il est recommandé pour des soumissions simples. Dans BankIt, l'usager peut indiquer des régions de codage sur un mRNA avec un nom de produit et de gène. Pour un bon contrôle des annotations des entrées, des enregistrements segmentés, ou des entrées très longues, Sequin, est l'outil le plus approprié.

GenBank fournira à l'auteur un numéro d'accession pour identifier sa séquence, habituellement dans deux jours ouvrables, si la soumission est reçue par l'intermédiaire d'un courrier électronique. Ce numéro d'accession sert comme confirmation de réceptions des données soumises. Le numéro d'accession devrait être inclus dans le manuscrit, de préférence dans une apostille à la première page de l'article, ou selon les exigences des différentes procédures des revues.

1.2.2.2 BankIt : Soumission via le net

BankIt permet d'écrire les informations sur les séquences selon un format précis, de les éditer si nécessaire, et d'ajouter des annotations biologiques (par exemple, codifications des régions, des dispositifs de mRNA). BankIt transforme les données en format de GenBank pour les revues et quand l'enregistrement est prêt, il peut être soumis directement à GenBank. Il y' a l'option d'ajouter des informations en utilisant des boîtes de textes pour décrire d'une façon personnelle la source de la séquence et ses dispositifs biologiques. Le personnel d'annotation de GenBank passe en revue l'information textuelle soumise, l'incorpore dans des champs structurés appropriés, et renvoie l'enregistrement par e-mail à la revue.

1.2.2.3 Sequin

Si on n'a pas accès au web, NCBI présente un programme autonome de soumission appelé Sequin qui est rendu actuellement à sa version 6.00. C'est un logiciel pour Mac, PC/Windows et UNIX, il est interactif et orienté vers le graphique. Sequin est conçu pour simplifier le processus de soumission des séquences, fournir le visionnement graphique et des

options d'édition. Il incorpore une vérification robuste des erreurs et est adapté à des séquences très longues et à des annotations complexes.

1.2.2.4 SequinMacroSend

Cet outil a été conçu pour télécharger de grands fichiers .sqn directement de GenBank plutôt que d'envoyer de grands attachements par courrier éléctronique. On introduit simplement l'information sur l'interface de SequinMacroSend, téléchargent le dossier .sqn, et la soumission sera envoyée directement au personnel de soumission de GenBank.

1.2.2.5 TBL2ASN

Tbl2asn est un programme de ligne de commande qui automatise la création des enregistrements de séquence à soumettre à GenBank. Il utilise plusieurs fonctions similaires à celles de Sequin mais est généralement guidé par des fichiers de données. Tbl2asn (http://www.ncbi.nlm.nih.gov/Genbank/tbl2asn2.html) génère des dossiers de sqn à soumettre à GenBank. L'édition additionnelle de manuel n'est pas exigée avant la soumission.

Exemples :

- Soumission simple : une séquence par dossier de fsa

```
tbl2asn -t template.sbt -p path_to_files –v
```

1.2.2.6 Soumissions Spéciales : Génomes, séquences en lots, alignements

Sequin peut être utilisé pour la soumission d'une ou d'un nombre restreint de séquences. Cependant, il a été également conçu pour faciliter le traitement des types spéciaux de soumissions, et devrait être employé au lieu de BankIt pour les types suivants de soumissions : génomes et d'autres séquences très longues ; séquences multiples tels que des soumissions en lots et des ensembles segmentés, et études de population/phylogénétique/ mutation. En préparant la soumission d'un génome, on peut importer la séquence complète d'un génome dans Sequin aussi bien qu'un dossier contenant les traductions d'acide aminé dans le format FASTA, si disponibles. Puisque le dossier final de soumission (*. sqn) sera

trop grand, on peut l'envoyer au personnel de GenBank par l'intermédiaire du ftp plutôt que par courriel.

1.2.2.7 Envoi des données à GenBank

En utilisant BankIt, les entrées préparées des séquences sont soumises directement à GenBank via Internet. En utilisant Sequin, les fichiers résultats devraient être envoyés à GenBank par courrier électronique à : gb-sub@ncbi.nlm.nih.gov ou en téléchargeant des dossiers à SequinMacroSend.

1.2.2.8 Soumission de SNPs et d'autres données de polymorphisme

Des données sur la variation génétique des humains et d'autres organismes peuvent être soumises à la base de données de NCBI des polymorphismes simples de nucléotide (dbSNP).

DbSNP est une ressource séparée de la base de données de GenBank, et les soumissions ne reçoivent pas des numéros d'accessions de GenBank. Cependant, les entrées de dbSNP reçoivent des identifiants de dbSNP et contiennent des liens aux enregistrements associés de GenBank.

1.3 Données génétiques

La génétique est la science de l'hérédité qui étudie les caractères héréditaires des individus, leur transmission au fil des générations et leurs variations (mutations). En effet on se demande souvent pourquoi certains traits physiques tels la couleur des cheveux ou la taille semblent se retrouver chez les différents membres d'une même famille. Cette question touche au domaine de la génétique et plus précisément, de l'hérédité qui se réfère à la façon dont tout système vivant transmet ses propres caractères, ses empreintes, à la génération suivante. Lorsque les parents transmettent une copie de leur propre ADN à leurs enfants, cette information sert à générer des cellules qui à leur tour produisent les tissus, les organes et les structures physiologiques de l'enfant à naître.

1.3.1 Génome

La combinaison de mots « gène » et « chromosome » forme ce qu'on appelle « génome » qui désigne l'ensemble complet du matériel génétique d'un organisme vivant.

C'est une copie de tout l'ADN comprenant à la fois les chromosomes à l'intérieur du noyau et l'ADN dans les mitochondries. La plupart des cellules contiennent une copie du génome.

Par exemple le génome humain est composé d'environ 3 000 000 000 de paires de bases, assemblées en 23 paires de chromosomes. Contrairement aux bactéries qui peuvent n'avoir que 1 500 000 paires de bases dans un unique chromosome

1.3.2 Les gènes

Les gènes sont des segments d'ADN ou d'ARN (certains virus) qui portent une information génétique il est situé à un endroit précis sur un chromosome. Chaque personne contient dans ses cellules le même ensemble unique et complet de gènes.

1.3.3 Les acides nucléiques

Le nucléotide résulte d'un groupement phosphate, d'un glucide et de l'une des cinq molécules suivantes : (A) adénine, (G) guanine, cytosine (C), (T) thymine, (U) uracile. Les acides nucléiques sont des chaînes résultant d'une liaison le groupement phosphate et le glucide de deux nucléotides voisins. Ces chaînes sont de deux types :

- Acide désoxyribonucléique (ADN) : découvert en 1953 par James Watson et Francis Crick. C'est le produit chimique formant la base du matériel génétique et contenant les instructions génétiques pour la fabrication de tous les organismes vivants. C'est une longue molécule assemblée en chromosomes, dans laquelle l'information génétique est encodée sous forme d'une double hélice et la liaison entre les paires de bases assure sa cohésion.

- Acide Ribonucléique (ARN) : Est représentée sous forme de molécules courtes à un seul brin (contrairement aux ADN, longues à deux brins) et qui sont des copies de petites séquences de l'ADN. Dans l'ARN l'uracile représenté par la lettre U est substitué par la thymine représentée par la lettre T.

1.4 Données Taxonomiques

Le terme taxonomie vient du terme grec *taxis* défini comme étant un groupe d'organismes ayant certains caractères communs et descendant d'un même ancêtre. Pour classer les êtres vivants, on adopte une classification arborescente qui se présente sous la forme d'un arbre, allant de la racine incluant tous les êtres vivants existants ou ayant existé, jusqu'aux individus. Chaque nœud de cet arbre représente un taxon, qui regroupe tous les sous taxons qu'engendre le nœud. Cette classification comme tout autre classification, est liée à un état donné d'avancement des connaissances, doit évoluer et donc varier avec l'évolution des espèces.

1.4.1 Les classifications populaires

De nos jours elle conserve encore son importance bien qu'elle soit la première en date, c'est elle qui a permis de distinguer les genres et les espèces. Elle ne se soucie pas des données scientifiques, et se base sur des critères simples tel que l'apparence, les cris, etc. Elle procède par assimilation et/ou extension devant l'inconnu : par exemple, pour les chinois le kiwi du fait qu'il est couvert de poils, est assimilable à une souris végétale....

1.4.2 La classification scientifique

Est issue de celle de Carl Linnaeus (1707-1778), connu comme Carl von Linné, auteur d'une citation célèbre en 1755 qui est la suivante « *Si tu ignores le nom des choses, même leur connaissance disparaît* ». Cette classification est encore importante et très souvent utilisée. Elle divise le monde vivant en cinq règnes.

- les procaryotes (bactéries et archéobactéries),
- les protistes (eucaryotes unicellulaires),
- les champignons (eucaryotes multicellulaires),
- les végétaux (eucaryotes multicellulaires),
- les animaux (eucaryotes multicellulaires).

1.4.3 La classification phylogénétique

Elle est basée sur les caractères génétiques, et non sur les caractères morphologiques visibles ou les préférences nutritionnelles. Cette classification illustre les principes d'évolutions et de parenté des espèces et bouleverse les classifications fixistes qui considèrent que toutes les espèces sont apparues en même temps et que celles-ci étaient fixes.

Toutes ces classifications ont leurs limites, mais sont complémentaires. Par exemple deux « espèces » peuvent être morphologiquement identiques mais incapables de se reproduire entre elles. Ceci conduit a conclure qu'il y'a une différence génétique entre elles. L'évolution de la classification en deux groupes (végétal / animal) donne naissance a cinq règnes du vivant dans laquelle on distingue essentiellement deux groupes les eucaryotes et les procaryotes. Les procaryotes constituent le premier règne, ils sont unicellulaires, leur multiplication est scissiparité, possédant des enzymes localisés dans la paroi cellulaire. Ils n'ont pas de noyau qui enferme leur matériel génétique. Les eucaryotes sont formés de tout le reste des organismes qui peuvent être unicellulaires ou pluricellulaires. Le deuxième règne est constitué des protistes qui sont des eucaryotes unicellulaires. Les champignons, les végétaux chlorophylliens (métaphytes) et les animaux pluricellulaires (métazoaires) forment les trois règnes restants, ils sont tous des eucaryotes pluricellulaires

1.4.4 La classification biologique

Elle pourrait être la classification de l'avenir sachant qu'elle a surpris les scientifiques en rapprochant des espèces que l'on croyait éloignées, mais qui sont en fait très proches biologiquement. Cette discipline est basée essentiellement sur des critères de laboratoire.

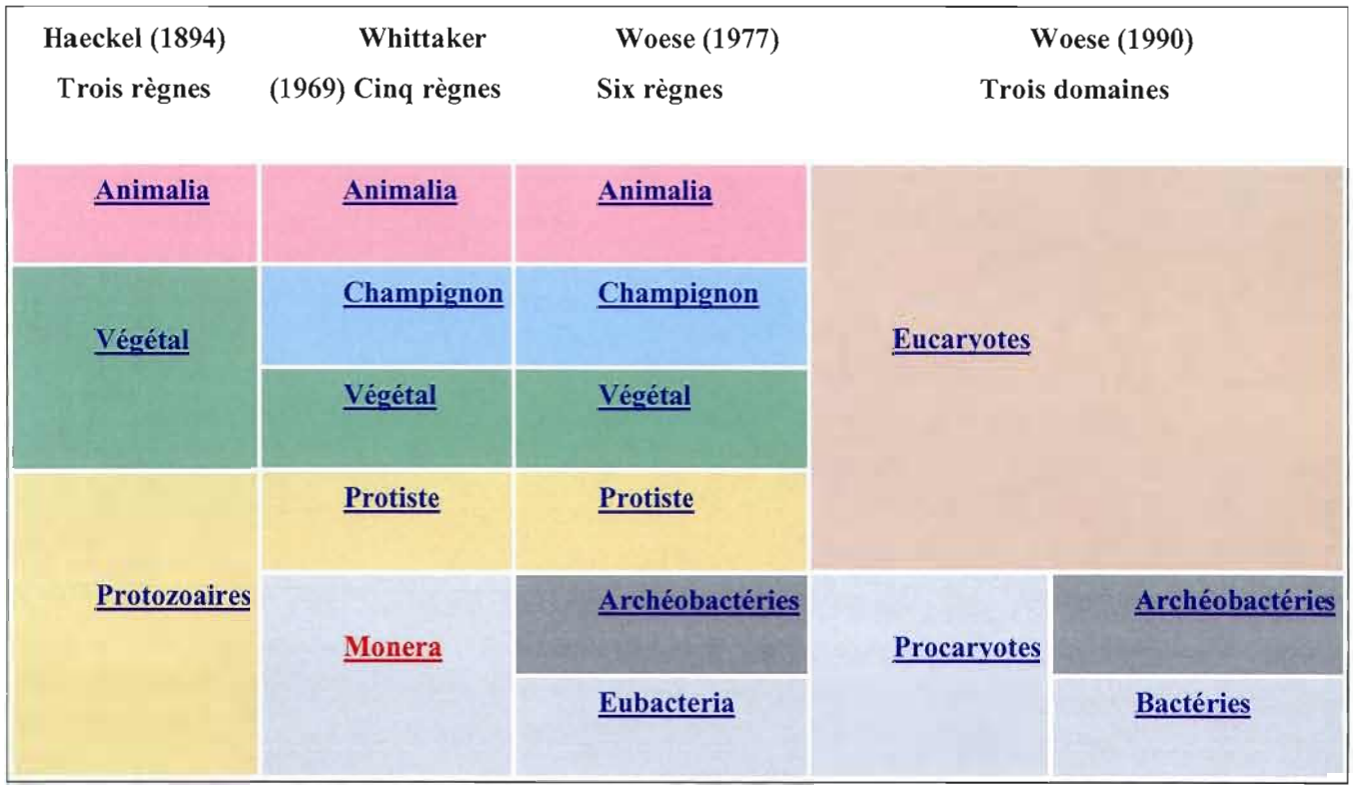


Tableau 1.2 Tableau illustrant l'évolution des systèmes de classification et des règnes

Référence prise sur le site http://fr.wikipedia.org/wiki/R%C3%A8gne

CHAPITRE II

2.1 Arbres phylogénétiques

2.1.1 Définition

Un arbre phylogénétique est un arbre qui montre les relations de parenté entre les espèces ou d'autres entités supposées avoir un ancêtre commun. Chacun des nœuds de l'arbre représente l'ancêtre commun de ses descendants. L'arbre peut être enraciné ou pas, selon qu'on est parvenu à identifier l'ancêtre commun à toutes les feuilles.

En informatique cet arbre n'est autre qu'une structure de données dont on distingue deux catégories d'éléments, les feuilles et les nœuds :

- les feuilles : elles représentent les unités taxonomiques dont les informations ont servi à la construction de l'arbre.
- les nœuds : ils représentent les ancêtres ; la racine est un nœud particulier qui représente l'ancêtre commun de toutes les espèces représentées pour cet arbre.

La relation entre les nœuds est une relation de descendance formée par les branches de l'arbre.

2.1.2 Exemple de représentation d'un arbre phylogénétique

- Nœuds : A, B, C…..I (nombre de nœuds = 2m-1)
- Branches : FA,FB,IG…. (nombre de branches = 2m-2)

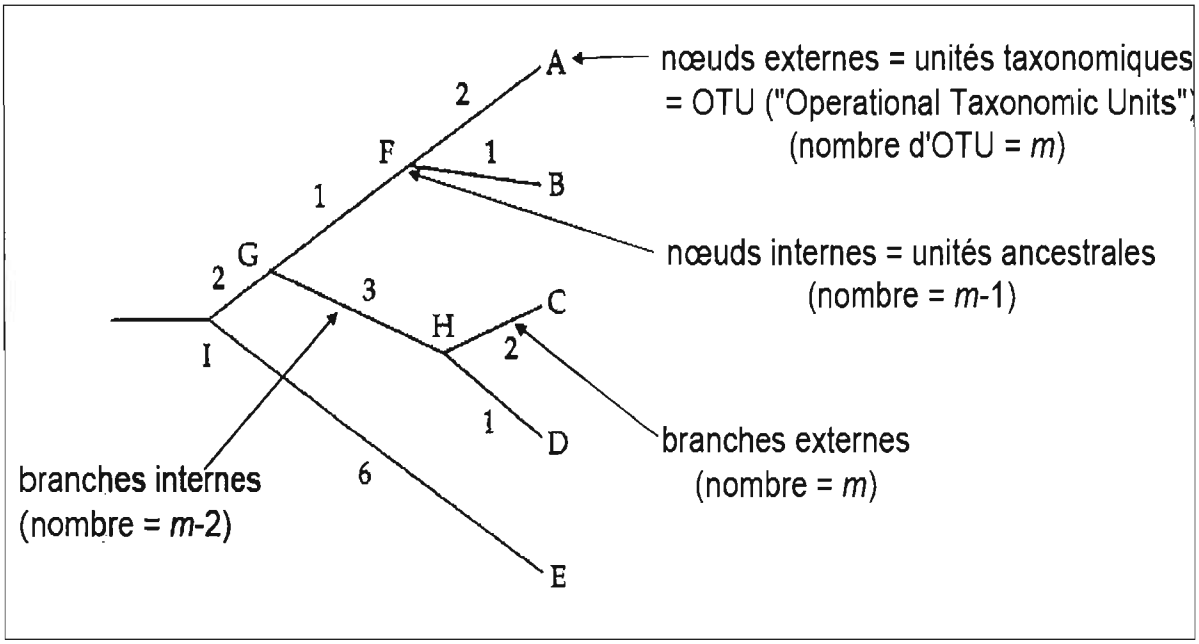


Figure 2.1 Un arbre phylogénétique à 5 feuilles (A,B,C,D et E)

Reprise du site http://www.univ-lille1.fr/gepv/downloads/enseignements/M1-S8-Vekemans-Chap3-Phylo_molec.pdf

Il existe plusieurs types de tracé d'arbre phylogénétiques (hiérarchique, radicale et axiale. Pour plus de détails voir Barthélemy et Guénoche (1991). Le logiciel T-REX de Vladimir Markarenkov (2001) implémente quatre de ces types de tracé.

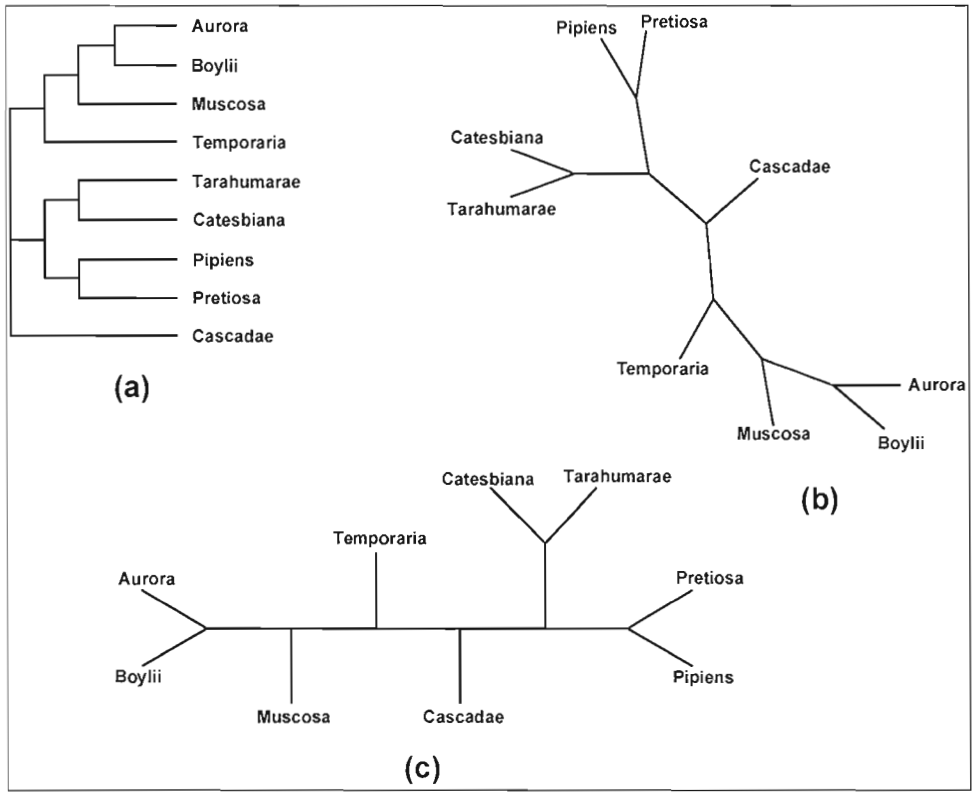


Figure 2.2 Trois façons de représenter un arbre phylogénétique

(a) Hiérarchique

(b) Radicale

(c) Axiale

2.1.3 Arbre enraciné et arbre non enraciné

Un arbre enraciné est considéré comme un graphe orienté. On peut distinguer un sommet particulier appelé la racine et signifiant l'identification d'un ancêtre commun formant ainsi un arbre enraciné. L'orientation est celle du temps d'évolution des espèces, la relation entre les nœuds est une relation de descendance. Par contre, les arbres non enracinés

sont des représentations intemporelles des relations phylogénétiques dont on ignore la notion du temps et d'ancêtre.

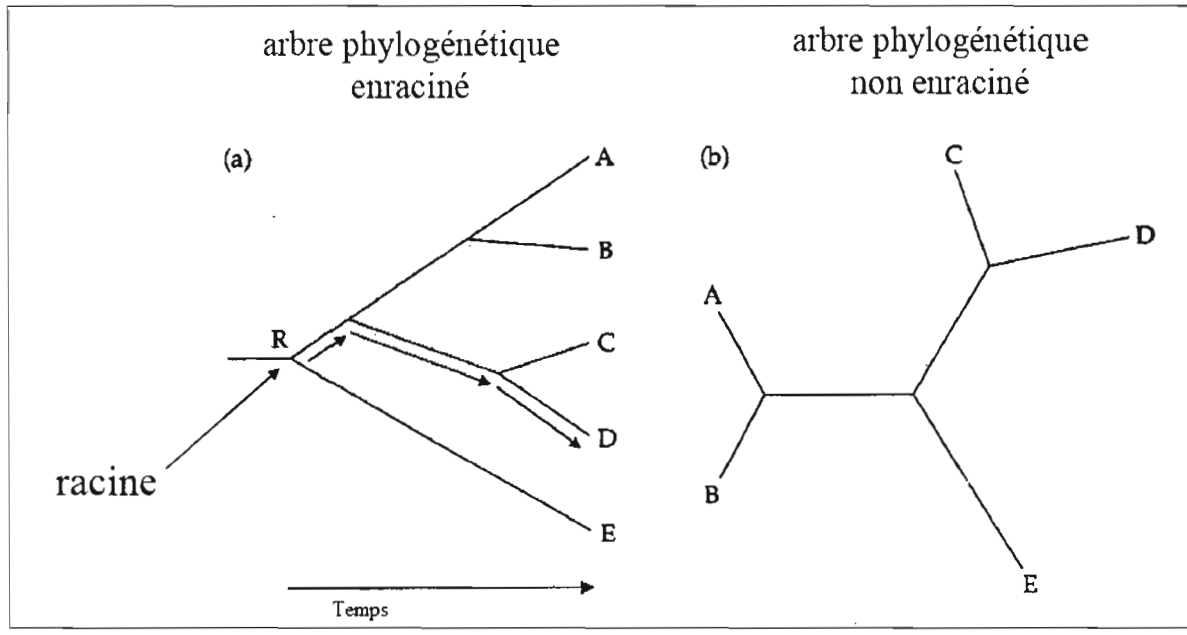


Figure 2.3 Représentation d'arbre enraciné et arbre non enraciné

Dans un arbre enraciné la direction du chemin évolutif est déterminée. On peut coder un arbre sous la forme des chaînes des parenthèses : (((A,B), (C,D)),G) par exemple, chaîne Newick. Le nombre d'arbres augmente exponentiellement en fonction du nombre de taxa, et pour un espèce donné le nombre d'arbres enracinés est différent du celui des arbres non enracinés.

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

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

où s est le nombre d'espèces.

Nombre de taxons	Nombre d'arbres non enracinés	Nombre d'arbres enracinés
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

Tableau 2.1 Correspondance entre le nombre de taxons et le nombre d'arbres enracinés et non enracinés.

2.1.4 Méthodes de reconstitution des arbres phylogénétiques

Grâce aux découvertes récentes, l'étude de la diversité biologique en vue de sa classification, adopte une classification phylogénétique regroupant les êtres vivants en fonction de leurs liens de parenté. Il existe actuellement deux techniques de reconstitution des arbres phylogénétiques : Les méthodes basées sur les distances et les méthodes basées sur les caractères.

2.1.4.1 Alignement des séquences

C'est une opération qui consiste à disposer les unes en dessous des autres des portions de séquences similaires en minimisant leurs différences. Les séquences peuvent être celles d'ADN, d'ARN, ou des séquences protéiques.

taxon1	ACCAG-TCGTACTGCCAGTAC-CTGACATGCCAGTCAGA
taxon2	ACCAG-TCGTGCTGCC-CAT--CTGACATGACA-TCAGA
taxon3	ACCTG-TCGTGCAGCCGCGT--CTGTCCTGCCAGTCGGA
taxon4	ACCTGGTCGTACTGCC-CATA-CTGGCCTGTCAGTCAGA
taxon5	ACTTG-TCGTACTGCCGTCGAACTGGCCTGTCAGTCAGA

Tableau 2.2 Exemple de 5 séquences d'ADN alignées

L'alignement de séquences d'ADN ou d'acides aminés a pour but d'identifier des zones conservées entre séquences. L'alignement sert notamment à :

- identifier des sites fonctionnels
- prédire la ou les fonctions d'une protéine
- prédire la structure secondaire (voire tertiaire) d'une protéine
- établir une phylogénie

On distingue deux types d'alignements :

- l'alignement par paires qui consiste à aligner deux séquences peut être réalisé grâce à un algorithme de complexité polynomiale. Il est possible de réaliser un alignement :
    - global, c'est-à-dire entre les deux séquences sur toute leur longueur
    - local, entre une séquence et une partie de l'autre séquence

- l'alignement multiple, qui est un alignement global, consiste à aligner plus de deux séquences et nécessite un temps de calcul et un espace de stockage exponentiels en fonction de la taille des données.

2.1.4.2 Approche basée sur les distances

Les méthodes de distances reconstruisent des arbres en partant des ressemblances observées entre les taxons. Elles calculent une mesure de distance, dite dissimilarité puis cherchent le meilleur arbre de prédiction pour ces distances.

La procédure de reconstruction peut donc se résumer somme suit :

Alignement → matrice de distance → arbre phylogénétique.

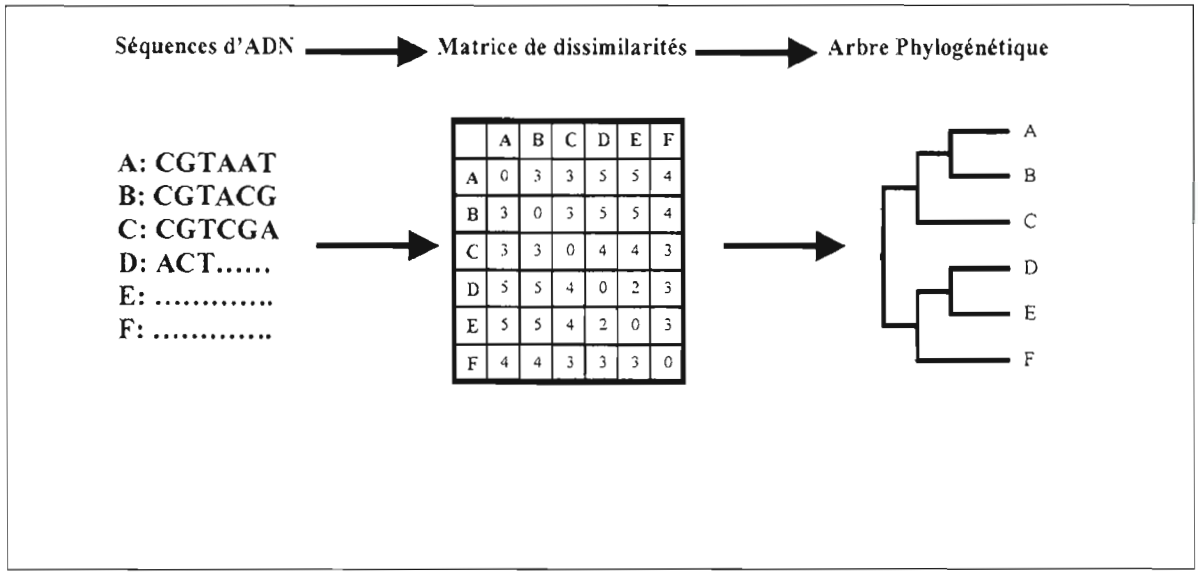


Figure 2.4 Procédure de reconstruction des arbres phylogénétiques

2.1.4.3.1 Calcul des distances

a) Distance observée

Pour comparer deux séquences, la manière la plus simple est d'évaluer leur similitude ou leur différence. Avoir un pourcentage de similitude entre les séquences peut aider à établir une relation entre ces dernières. Similarité entre deux séquences peut être calculée comme le nombre de sites synonymes divisé par la longueur de la séquence.

$$S = M/L$$

(S) Similarité, (M) nombre de sites synonymes, (L) Longueur de la séquence.

La distance observée entre deux séquences est donnée par $\mathbf{D = 1 - S.}$

b) Distance évolutive

La distance observée est le plus souvent une sous-estimation de la distance évolutive. Plusieurs événements, qui ont pu éventuellement se produire ne sont pas pris en considération dans son calcul. En effet les événements suivant : La convergence, la réversion et les changements multiples sont le plus souvent ignorées. La distance évolutive entre deux séquences est égal au nombre de substitutions qui se sont produites sur les deux lignées évolutives depuis l'ancêtre commun / nombre de sites. Elle pourrait être est égale à la distance observée uniquement si les séquences sont très proches et le nombre de substitutions observées correspond au nombre de substitutions qui se sont réellement produites.

2.1.4.4.2 Calcul des distances

a) Modèle de Jukes-Cantor (JC)

Le modèle à un paramètre de Jukes et Cantor (1969) assume que les 4 bases ont les mêmes fréquences et suppose que tous les types de substitutions (A -> C, A -> G, G -> A, etc.) sont équiprobables. Ayant α comme est la probabilité de transition et β comme la probabilité de transversion et les paramètres suivants :

$$\pi A = \pi T = \pi G = \pi C$$

$$\alpha = \beta$$

La formule est : $\mathbf{d_{xy} = -3/4 \ln (1-4/3 D)}$

D : distance observée entre deux séquences X et Y.

Les termes 3/4 et 4/3 correspondent aux quatre types de nucléotides et aux trois possibilités que le deuxième nucléotide peut être différent du premier.

b) Modèle de Tajima-Nei (TN) et de Felsenstein (F81)

Ce modèle prend en compte la variation des fréquences des bases. Il assume que la probabilité du changement vers une base donnée ne dépend pas de la base qui change. Ainsi, le changement de A vers T a la même probabilité que le changement de G ou C vers T.

Les paramètres sont les suivants :

$$\pi A \neq \pi T \neq \pi G \neq \pi C$$

$$\alpha = \beta$$

La formule est la suivante :

$$\mathbf{d_{xy} = - B \ln (1 - D/B)}$$

La valeur du paramètre **B** varie avec la composition en base.

$B = 1 - \Sigma q_i^2 \times q_i$ est la somme des fréquences des 4 nucléotides ($\pi A^2 + \pi T^2 + \pi G^2 + \pi C^2$).

Pour une valeur particulière q_i égal à 1/4, B égal à 3/4, et la formule devient le modèle de Jukes et Cantor.

c) Modèle de Kimura (K2P)

Ce modèle tient compte de la proportion entre le nombre de transitions (α) et transversions (β). Les transitions sont généralement plus fréquentes que les transversions.

Paramètres :

$$\pi A = \pi T = \pi G = \pi C$$

$$\alpha \neq \beta$$

Formule :

$$\mathbf{d}_{xy} = -\ 1/2 \ln (1-2P-Q) + 1/4 \ln (1-2Q)$$

Où P = la proportion de transitions, et Q = la proportion de transversions.

d) LogDet

C'est une méthode pour corriger le biais de composition en bases entre les séquences comparées. Quand on analyse les séquences qui différent significativement par leur taux GC, les séquences de même composition en bases ont tendance à être groupées ensemble.

$$\mathbf{d}_{xy} = -\ \ln [\det \mathbf{F}_{xy}],$$

Où **det** est un déterminant de la matrice $\mathbf{F}_{xy}$.

e) Gamma

La distribution gamma est utilisée pour corriger les variations de substitutions entre les différents sites. Le paramètre α est évalué à partir des données et correspond à l'intervalle de variations. Plus α est petit, plus l'intervalle des variations est grand. On peut inclure la distribution gamma dans différents modèles de changements évolutifs, Par exemple :

- JC + Γ $\mathbf{d}_{xy} = 3/4\ \alpha\ [\ (1 - 4/3\ p)^{-1/\alpha} - 1\ ]$,
- K2P + Γ $\mathbf{d}_{xy} = 1/4\ \alpha\ [\ 2\ (1 - 2P - Q)^{-1/\alpha} + (1 - 2Q)^{-1/\alpha} - 3\ ]$.

2.1.4.5.3 Construction d'arbre phylogénétique

Plusieurs méthodes ont été développées pour construire un arbre phylogénétique à partir d'une matrice de distance.

a) **UPGMA** (Unweight Pair Group Method with Arithmetic Mean) **:**

Cette méthode est utilisée pour reconstruire des arbres phylogénétiques à partir des matrices de distances si les séquences ne sont pas trop divergentes. Elle impose que les distances soient ultramétriques. Les séquences évoluent donc à une vitesse constante (hypothèse d'horloge moléculaire). UPGMA utilise un algorithme de regroupement séquentiel dans lequel les relations sont identifiées dans l'ordre de leur similarité et la reconstruction de l'arbre se fait pas à pas grâce à cet ordre.

Étapes :

- choisir une paire des séquences les plus proches (A, B) et calculer leur distance ultra métrique,
- calculer les distances moyennes entre le taxon composé (A, B) et les autres séquences,
- former une nouvelle matrice et répéter l'opération.

Cette méthode est rapide, mais présente certaines faiblesses du fait qu'elle suppose que tout le temps d'évolution de toutes les séquences a été le même. C'est-à-dire s'il existe une relation linéaire entre les distances évolutives et le temps de divergence.

b) **Neighbor Joining :**

C'est la méthode de distances la plus souvent utilisée. Elle a été introduite par Saitou et Nei (1987) afin de corriger la méthode UPGMA et d'autoriser un taux de mutation variable sur les branches. La méthode Neighbor Joining consiste à calculer les longueurs des branches, telles que les distances déduites de l'arbre soient les plus proches des distances

mesurées entre les séquences, et ensuite à calculer la longueur totale de l'arbre, égale à la somme des longueurs de ses branches. Le principe de cette méthode est la suivante :

- Regrouper toutes les espèces en un noeud :

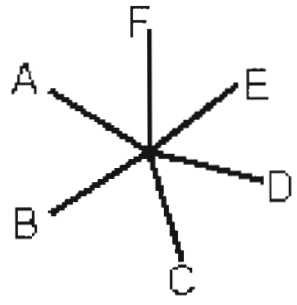


Figure 2.5 Toutes les espèces regroupées en un seul noeud

- Choisir au hasard deux taxons pour former un nouveau nœud, (la longueur totale des branches est recalculée), et reprendre l'opération avec tous les couples de séquences.

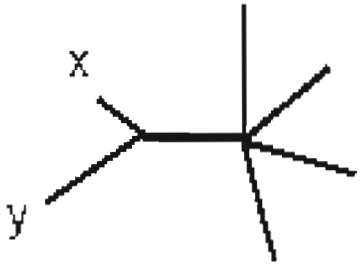


Figure 2.6 Choix au hasard de deux taxons pour former un nouveau nœud

- Les séquences jointes sont celles qui minimisent la longueur totale des branches.

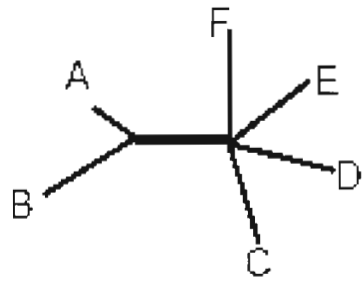


Figure 2.7 Les séquences jointes sont celles qui minimisent la longueur totale des branches

- Les séquences jointes sont fusionnées en une nouvelle feuille et les distances sont recalculées et on retourne à l'étape 1.

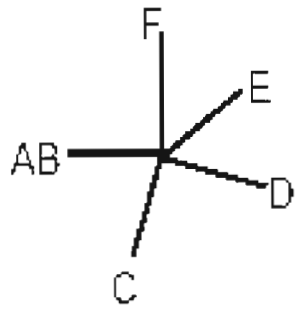


Figure 2.8 Les séquences jointes sont fusionnées en une nouvelle feuille

Étapes :

- calculer la divergence nette r(i) de chaque séquence par rapport aux autres,
- calculer la matrice des distances modifiées selon la formule $M_{ij} = d_{ij} - (r_i + r_j)/N-2$ (les distances sont négatives),
- choisir les deux séquences pour lesquelles la valeur de Mij est la plus petite,
- calculer la distance de chacune de deux séquences au nœud **u** :

$$S_{iu} = d_{ij}/2 + r_i\text{-}r_j/2(N\text{-}2) \text{ et } S_{ju} = d_{ij} - S_{iu}$$

$$S_{iu} = Sij / 2$$

- calculer les distances entre **u** et tous les noeuds terminaux,
- créer une nouvelle matrice et répéter l'opération.

2.1.4.6.4 Avantages et inconvénients des méthodes de distances

Pour analyser certains types de données, tel que la distance immunologique, les distances d'hybridation d'acides nucléiques, il s'avère que les méthodes de distance sont les seules disponibles. Elles sont rapides et permettent l'analyse de grandes bases de données et le test d'un grand nombre d'hypothèses alternatives, ainsi que l'intégration des modèles de changements évolutifs que d'autres méthodes sont incapables d'intégrer. Mais ces méthodes présentent certains désavantages comme la perte d'informations à cause de la réduction de la matrice des caractères à une matrice de distance, et sont incapables de combiner dans une même matrice des caractères de nature différente (par exemple, des caractères morphologiques et des séquences d'ADN).

2.1.4.3 Approche basée sur les caractères

2.1.4.3.1 Principe de la cladistique

Mise au point par un entomologiste allemand, Willi Hennig, dès l'année 1950. La méthode cladistique est basée sur la notion d'homologie. On n'établit des relations de parenté que sur la base du partage des états évolués des caractères (ou homologies). Cette méthode repose sur quelques principes : seuls les caractères spécialisés partagés, hérités de l'ancêtre commun qui les a acquis (synapomorphies), doivent être pris en compte pour l'établissement des relations de parenté. Les résultats de l'analyse des caractères dérivés communs sont exprimés graphiquement sous la forme d'arbres dichotomiques (cladogrammes). Pour chaque groupe monophylétique (c'est-à-dire issu d'un même ancêtre), on recherche le groupe frère, c'est-à-dire le groupe le plus proche parent.

Une règle importante (le principe de parcimonie) : en partant du postulat que la nature est économique, on retient parmi les cladogrammes celui qui comporte le moins de pas, c'est-à-dire le moins de transformations de caractères.

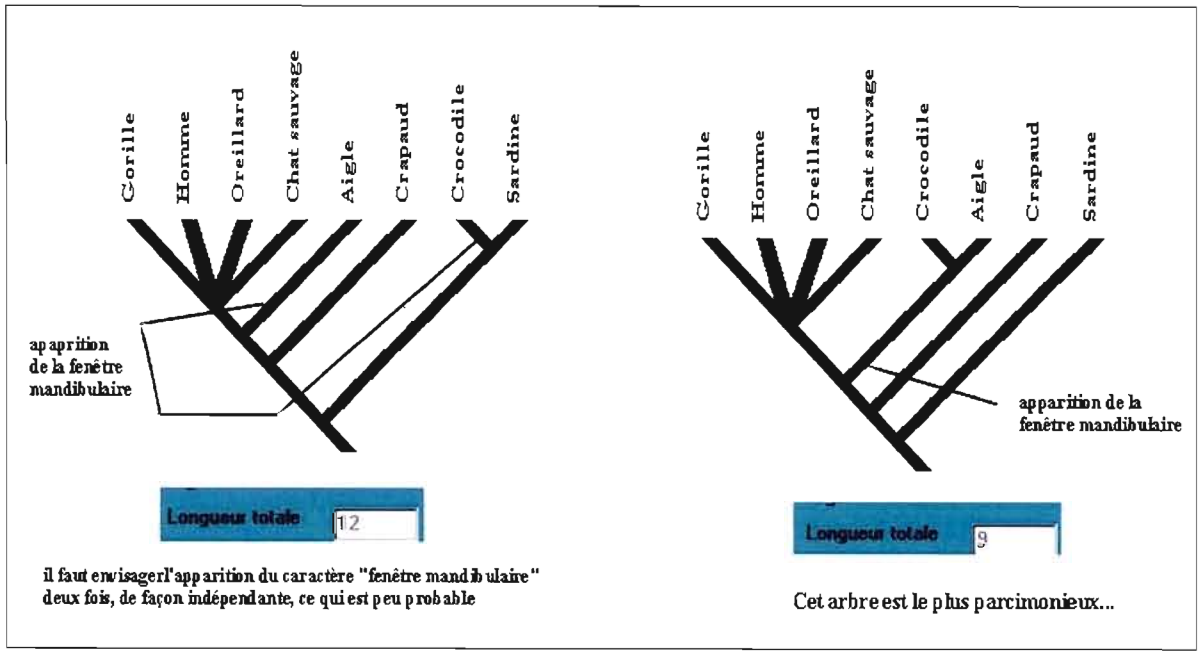


Figure 2.9 Choix de l'arbre le plus parcimonieux

Repris du site

http://www.inrp.fr/Acces/biotic/evolut/phylogene/documentation/html/parcim.htm

2.1.4.3.2 Méthode de parcimonie

Le principe de la parcimonie consiste à n'utiliser que le minimum de changements évolutifs pour expliquer des relations phylogéniques. Elle minimise le nombre de "pas" (mutations / substitutions) nécessaires pour passer d'une séquence à une autre dans une topologie de l'arbre. La longueur de l'arbre **L** est égale à la somme du nombre de changements **l** pour chacun des **k** sites.

$$\mathbf{L} = \sum_{i=1}^{k} L_i$$

Cette méthode s'appuie sur deux hypothèses principales :

- les sites évoluent indépendamment les uns des autres (la séquence peut être considérée comme une suite de caractères non ordonnés).

- la vitesse d'évolution est lente et constante au cours du temps.

La méthode de maximum de parcimonie recherche toutes les topologies possibles afin de trouver l'arbre optimal (minimum) et le temps nécessaire pour cette exploration croît rapidement avec le nombre de séquences. Pour chercher l'arbre le plus parcimonieux, il existe plusieurs méthodes :

a) Recherche exhaustive

L'analyse exhaustive consiste à calculer la longueur de tous les arbres de topologie différente et choisir l'arbre dont la longueur est minimale. C'est une méthode qui garantit de trouver un ou plusieurs arbres optimaux. Cependant, comme il y a plus de deux millions d'arbres pour dix taxons, une telle recherche n'est généralement possible qu'au-dessous de la dizaine de taxons. Lorsque le nombre de taxons à étudier est faible, la recherche exhaustive de l'arbre le plus parcimonieux consiste à examiner tous les arbres possibles en vue de connecter trois premiers taxons pour former un arbre. Puis on connecte le quatrième taxon sur chacune des trois branches ce qui donne trois arbres possibles. Puis on connecte le cinquième taxon sur les différentes branches des trois arbres, ce qui donne cinq arbres possibles pour chacun des trois arbres établis avec quatre taxons, soit quinze reconstructions possibles non enracinées représentées à la figure 2.10 .

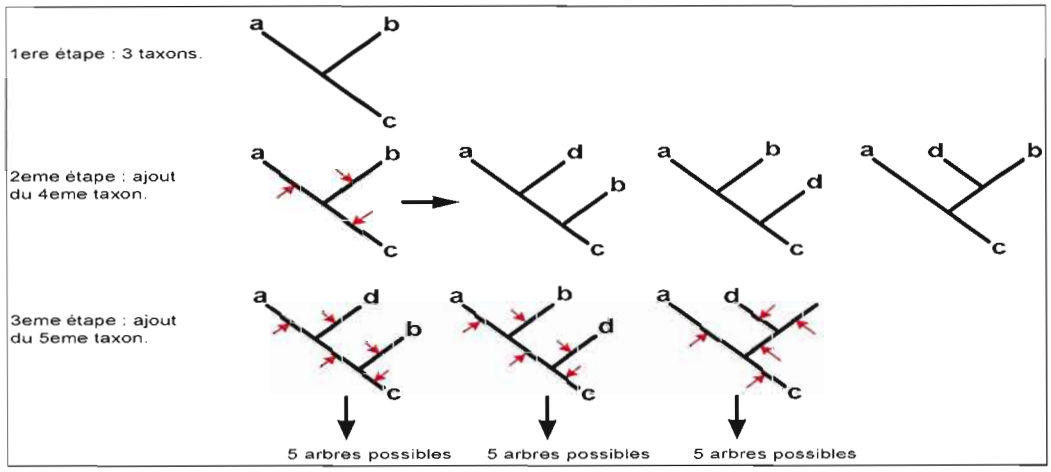


Figure 2.10 Longueur d'arbre pour les trois topologies possibles.

b) Branch and Bound (Algorithme de séparation et d'évaluation)

Cette méthode est dérivée du maximum de parcimonie, elle garantit de trouver le meilleur arbre mais sans évaluer tous les arbres possibles. Au départ on choisit un arbre de référence obtenu par une méthode heuristique, puis la recherche se fait par rapport à cet arbre. On compare la longueur L de cet arbre à celle des autres arbres au fur et mesure de leur construction. Si l'on rencontre un arbre plus long que L, on quitte le chemin qui a conduit à cet arbre et on en explore un autre. Si l'on rencontre un arbre aussi long que L, l'arbre est un arbre optimal possible. Si l'on rencontre un arbre plus court que L, cet arbre est le meilleur obtenu et devient la nouvelle référence. Plus l'arbre initial est court, plus rapidement on terminera l'opération. Quand l'ensemble des chemins a été exploré, tous les arbres de longueur minimale ont été identifiés.

C) Méthodes heuristiques

Ces algorithmes trouvent soit le meilleur arbre, soit une topologie qui est voisine de la topologie optimale. Elles sont utilisées lorsque la matrice des données est trop importante

pour l'usage d'algorithmes exacts (nombre élevé de taxa et de caractères). Généralement cette méthode consiste à la création d'un arbre initial qui est ensuite réarrangé de manière à diminuer sa longueur. L'arbre initial est obtenu par une addition des taxa pas à pas. La procédure d'addition pas à pas produit généralement des arbres qui correspondent à un optimum local. Des réarrangements des branches permettent de rechercher des arbres plus courts. Il se peut qu'il existe un arbre de longueur inférieure à celui identifié par la méthode heuristique. En effectuant la recherche heuristique plusieurs fois et en changeant l'ordre des taxons, on augmente la probabilité de trouver l'arbre optimal. Il existe plusieurs stratégies heuristiques dites de "réarrangement des branches" :

- Le réarrangement local (nearest-neighbor interchanges),
- Le réarrangement global (subtree pruning and regrafting),
- Le réarrangement par bissection et reconnexion (tree bisection and reconnection - TBR).

2.1.4.3.3 Maximum de vraisemblance

C'est une approche probabiliste introduite en phylogénie en 1971 par Jerzy Neyman. Elle consiste à la reconstitution de l'arbre en terme de probabilités, l'ordre des branchements et de la longueur des branches d'un arbre sous un modèle évolutif donné. L'arbre trouvé est renvoyé avec une probabilité que sa topologie explique les données observées. La vraisemblance est la probabilité d'observer les données **D** sous l'hypothèse **H**

$$\mathbf{L = Pr\,(D \mid H)}$$

La démarche consiste donc à rechercher la vraisemblance des données **D** sous différentes hypothèses évolutives **H** d'un modèle **M** et à retenir les hypothèses qui rendent cette vraisemblance maximale. En phylogénie **D** représente des séquences comparées, et **H** est l'arbre phylogénétique. Nous cherchons à trouver l'arbre dont la vraisemblance est maximale, étant donné les séquences observées et le modèle d'évolution choisi. Pour chaque site les bases ou acides aminées de toutes les séquences sont considérées séparément en utilisant un

modèle de probabilité. On calcule le log de la vraisemblance pour une topologie donnée. On cumule ce log de la vraisemblance, tous les sites et la somme est maximisée pour estimer la longueur de branches de l'arbre. On répète cette procédure pour toutes les topologies possibles et on choisit la topologie ayant la plus haute vraisemblance. La vraisemblance de l'arbre pour tous les sites est égale au produit des probabilités **Pr (D□T, M, r)** pour chaque site

$$\mathbf{L = L_1 \times L_2 \times L_3 \ldots \times L_N}$$

La valeur de vraisemblance étant très petite on l'exprime sous forme logarithmique. La somme de ln L pour chaque site donne la vraisemblance de l'arbre.

$$\mathbf{\ln L = \ln L_1 + \ln L_2 + \ln L_3 \ldots + L_N} = \sum_{j=1}^{N} \ln L_j$$

[La valeur de **ln L** est négative car la probabilité calculée est inférieure à 1]

Parmi les méthodes phylogénétiques, la méthode de maximum de vraisemblance est considérée comme la plus fiable conduisant aux résultats les plus proches de l'arbre évolutif réel. Elle est moins sensible aux effets de l'attraction de longues branches que la méthode de parcimonie. Elle permet d'appliquer les différents modèles d'évolution comme le modèle de Kimura qui tient compte de différences entre transitions et transversion et d'estimer la longueur des branches en fonction de changement évolutif. Par contre, elle n'est utilisée que pour les petits nombres de séquences car elle demande la plus grosse puissance de calcul (pas plus que 100 espèces peuvent être analysées a l a fois).

2.2 Présentation du logiciel T-REX

Le Logiciel T-Rex (*tree and reticulogram reconstruction*) est très simple d'utilisation, et assez complet. Il comporte de nombreuses options et modes de calculs pour reconstruire et visualiser des arbres et réseaux phylogénétiques. Il a été développé par Vladimir Makarenkov (2001), avec la collaboration d'un groupe de professeurs et d'étudiants. C'est un logiciel supporté par plusieurs plateformes et systèmes d'exploitation. Il existe plusieurs versions de ce logiciel, pour MSDOS, Windows, Macintosh de même qu'une version WEB. La version Windows a été développée par Vladimir Makarenkov en langage C++. L'implémentation de la version Macintosh a été réalisée par Vladimir Makarenkov et Philippe Casgrain en C++ et Pascal.

J'ai été impliqué dans le développement de la version web de T-Rex qui est accessible a l'adresse URL suivante http://www.trex.uqam.ca. À partir de ce site l'utilisateur peut télécharger les versions une version DOS 32-bits, ainsi que le code source en C/C++ pour les systèmes UNIX.

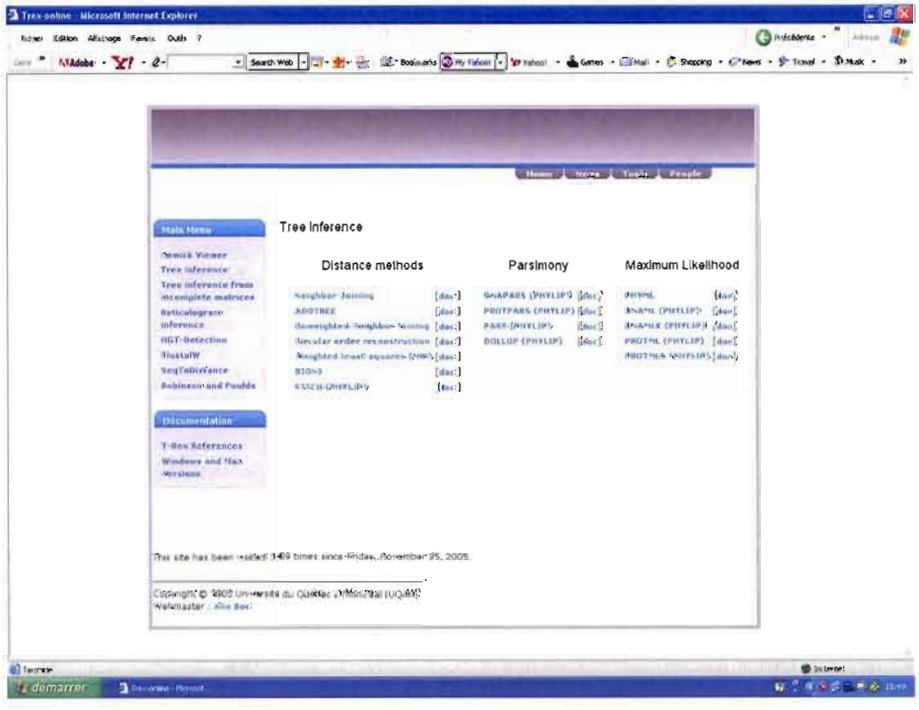


Figure 2.11 Interface de la version Web du logiciel T-Rex.

T-Rex permet la reconstruction des arbres phylogénétiques et des réseaux réticulés à partir d'une matrice de distance ou à partir des séquences. Il inclut un certain nombre de méthodes populaires pour l'inférence d'arbres phylogénétiques ainsi que de nouvelles méthodes d'analyse phylogénétique par exemple : la reconstruction d'arbres par les moindres carrés pondérés, l'inférence d'arbres des matrices de distance incomplètes ou la modélisation d'un réseau réticulé pour un groupe d'objets ou d'espèces. T-Rex permet de modifier le mode de calcul des écarts des moindres carrés et de fixer le nombre de réticulations à ajouter. Les arbres ou les structures en réseaux peuvent être visualisés d'une manière interactive en utilisant une représentation hiérarchique (verticale ou horizontale), radiale ou axiale. L'utilisateur peut afficher, en même temps que l'arbre, des statistiques et des résultats des calculs (telles que la matrice additive par exemple).

2.2.1 Format de soumission des données à T-Rex

L'utilisateur peut soumettre les données à T-Rex en format de matrice de distance, de séquences ou en format Newick.

a) Matrice de distance :

Cette façon de présenter la matrice est la plus économique. Elle contient pour chaque paire de séquences une distance entre elles, obtenue à partir d'un alignement multiple de ces séquences. La matrice présentée peut être une matrice incomplète incluant des valeurs manquantes. La première ligne est formée d'un nombre indiquant le nombre d'espèces formant la matrice (i.e format PHYLIP standard).

9 (taille de la matrice)

Aurora	0	10	13	12	57	22	86	89	97
Boylii	10	0	7	7	50	9	65	67	72
Cascadae	13	7	0	7	40	11	54	66	79
Muscosa	12	7	7	0	45	15	48	49	67
Temporaria	57	50	40	45	0	48	85	83	107
Pretiosa	22	9	11	15	48	0	54	55	60
Catesbiana	86	65	54	48	85	54	0	54	59
Pipiens	89	67	66	49	83	55	54	0	48
Tarahumarae	97	72	79	67	107	60	59	48	0

Tableau 2.3 Matrice de distances immunologiques entre les paires distinctes de neuf espèces des grenouilles du genre Rana

9

Aurora	0	-99	13	12	57	22	86	89	97
Boylii	-99	0	7	-99	50	-99	65	67	-99
Cascadae	13	7	0	7	40	11	-99	66	79
Muscosa	12	-99	7	0	45	15	48	49	67
Temporaria	57	50	40	45	0	48	85	83	107
Pretiosa	22	-99	11	15	48	0	54	-99	60
Catesbiana	86	65	-99	48	85	54	0	54	-99
Pipiens	89	67	66	49	83	-99	54	0	48
Tarahumarae	97	-99	79	67	107	60	-99	48	0

Tableau 2.4 Matrice de distances avec valeurs manquantes

entre les paires distinctes de neuf espèces des grenouilles du genre Rana. Les valeurs -99

b) Séquences :

T-Rex permet l'entrée des données sous format de séquences. Plusieurs formats de fichier sont acceptées dans certains options de T-Rex (exemple FASTA, PHYLIP, NBRF/PIR, EMBL/Swiss-Prot, GDE, Clustal, GCG/MSF et RSF). Les formats de fichiers les plus utilisés dans les options de T-Rex sont FASTA et PHYLIP.

- FASTA : Le format FASTA est un standard en bioinformatique pour représenter les séquences. Une séquence en format FASTA commence par une ligne de titre (nom, définition ...), suivi par les lignes de données de la séquence. La ligne de titre se distingue de la séquence par un symbole plus grand que (">") en début de la ligne. Il est recommandé de mettre la séquence sous forme de lignes de 80 caractères maximum. La séquence est finie si on rencontre le symbole (">") indiquant le début d'une autre séquence.

```
10 705
Cow       ATGGCATATCCCATACAACTAGGATTCCAAGATGCAACATCACCAATCATAGAAGAACTA
Carp      ATGGCACACCCAACGCAACTAGGTTTCAAGGACGCGGCCATACCCGTTATAGAGGAACTT
Chicken   ATGGCCAACCACTCCCAACTAGGCTTTCAAGACGCCTCATCCCCCATCATAGAAGAGCTC
Human     ATGGCACATGCAGCGCAAGTAGGTCTACAAGACGCTACTTCCCCTATCATAGAAGAGCTT
Loach     ATGGCACATCCCACACAATTAGGATTCCAAGACGCGGCCTCACCCGTAATAGAAGAACTT
Mouse     ATGGCCTACCCATTCCAACTTGGTCTACAAGACGCCACATCCCCTATTATAGAAGAGCTA
Rat       ATGGCTTACCCATTTCAACTTGGCTTACAAGACGCTACATCACCTATCATAGAAGAACTT
Seal      ATGGCATACCCCCTACAAATAGGCCTACAAGATGCAACCTCTCCCATTATAGAGGAGTTA
Whale     ATGGCATATCCATTCCAACTAGGTTTCCAAGATGCAGCATCACCCATCATAGAAGAGCTC
Frog      ATGGCACACCCATCACAATTAGGTTTTCAAGACGCAGCCTCTCCAATTATAGAAGAATTA
```

Figure 2.12 Exemple de format de fichier FASTA pris du site Wokshop on Molecular Evolution

(Repris de http://workshop.molecularevolution.org/resources/fileformats/fasta_dna_al.php)

- PHYLIP : Ce format est le plus souvent utilise dans T-Rex. Dans ce format la première ligne contient le nombre de séquences suivi du nombre de sites (égal au nombre de positions) par séquence. Les séquences sont supposées être préalablement alignées. Le nom de chaque séquence doit prendre dix caractères exactement (en incluant les espaces). Deux formats de bases sont proposés :

1- Format intercalé (= "interleaved")

Exemple :

```
   5    42
Turkey     AAGCTNGGGC ATTTCAGGGT
Salmo gairAAGCCTTGGC AGTGCAGGGT
H. SapiensACCGGTTGGC CGTTCAGGGT
Chimp      AAACCCTTGC CGTTACGCTT
Gorilla    AAACCCTTGC CGGTACGCTT

GAGCCCGGGC AATACAGGGT AT
GAGCCGTGGC CGGGCACGGT AT
ACAGGTTGGC CGTTCAGGGT AA
AAACCGAGGC CGGGACACTC AT
AAACCATTGC CGGTACGCTT AA
```

2- Format séquentiel

Les séquences se suivent (dans leur totalité) les unes après les autres.

Exemple :

```
   5    42
Turkey     AAGCTNGGGC ATTTCAGGGT
GAGCCCGGGC AATACAGGGT AT
Salmo gairAAGCCTTGGC AGTGCAGGGT
GAGCCGTGGC CGGGCACGGT AT
H. SapiensACCGGTTGGC CGTTCAGGGT
```

```
ACAGGTTGGC CGTTCAGGGT AA
Chimp      AAACCCTTGC CGTTACGCTT
AAACCGAGGC CGGGACACTC AT
Gorilla    AAACCCTTGC CGGTACGCTT
AAACCATTGC CGGTACGCTT AA
```

b) Newick :

La norme de Newick pour représenter les arbres en forme lisible par les ordinateurs sert de correspondance entre les arbres et les chaînes de parenthèses. Le codage des arbres en formes des chaînes de parenthèses a été introduit en 1857 par le célèbre mathématicien anglais Arthur Cayley. Par exemple l'arbre enracine suivant

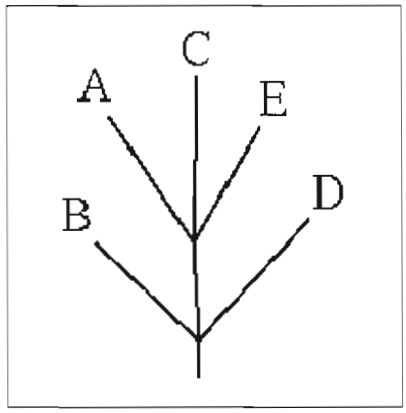


Figure 2.13 Arbre en format Newick

peut être représenté sous le format de chaînes de parenthèses (i.e Newick) suivant : (B,(A,C,E),D);

Sur la figure les extrémités de l'arbre sont représentées par des lettres. Les noeuds intérieurs sont représentés par une paire des parenthèses assorties. Entre eux figurent les représentations des noeuds qui sont des descendants immédiat de ce noeud, séparées par des

virgules. Dans l'arbre ci-dessus figure 2.13, les descendants immédiats sont B, un autre noeud intérieur, et D. L'autre noeud intérieur est représenté par une paire de parenthèses, enfermant des représentations de ses descendants immédiats, A, C et E.

Des longueurs de branches peuvent être ajoutées à un arbre en mettant un nombre réel après un noeud précédé de deux points. Ceci représente la longueur de la branche qui se retrouve immédiatement au-dessous de ce noeud. Ainsi l'arbre ci-dessus pourrait avoir des longueurs représentées comme suit : (B:6.0, (A:5.0, C:3.0, E:4.0):5.0, D:11.0) ;

Un autre exemple d'une chaîne Newick a 8 espèces :

((raccoon:19.19959,bear:6.80041):0.84600,((sea_lion:11.99700, seal:12.00300):7.52973,((monkey:100.85930,cat:47.14069):20.59201, weasel:18.87953) :2.09460):3.87382, dog:25. 46154);

2.2.2 Méthodes de reconstruction d'arbres à partir des matrices de distance dans T-Rex

Les six méthodes suivantes pour la reconstruction d'un arbre phylogénétique à partir d'une matrice de dissimilarité sont supportées par T-Rex

1. ADDTREE de Sattath et Tversky (1977);
2. Neighbor-joining (NJ) méthode de Saitou et Nei (1987);
3. BioNeighbor-joining (BioNJ) de Gascuel (1997);
4. Unweighted neighbor-joining method (UNJ) de Gascuel (1997);
5. Reconstruction par un ordre circulaire de Makarenkov, Leclerc (1997), et Yushmanov (1984);
6. La méthode des moindres carrés pondérées de Makarenkov (1999), et Makarenkov et Leclerc (1999).

Les deux premières méthodes pour l'inférence d'arbres phylogénétiques à partir des distances évolutives, ADDTREE et NJ, sont les méthodes les plus connues. Elles débutent la reconstruction de la structure d'un arbre phylogénétique par un arbre en étoile a n feuilles associées aux objets et n -1 branches. On ajoute au fur et a mesure de nouveaux nœuds internes jusqu'à l'obtention d'un arbre binaire a $2n$ - 2 noeuds et $2n$- 3 branches. La troisième et la quatrième méthode appelée respectivement BioNJ et UNJ utilisent à chaque étape les mêmes critères de choix, mais une différente méthode d'évaluation et des formules de réduction que NJ pour inférer l'arbre. La cinquième méthode a été décrite par Makarenkov et Leclerc (1997) inspiré d'un article de Yushmanov (1984), qui a introduit la notion d'ordre circulaire des objets correspondant à une lecture circulaire des feuilles d'un arbre. La sixième (Weighted least-squares method) comporte elle-même des options telles qu'utiliser une matrice de poids particulière et de l'appliquer localement ou globalement. Cette méthode permet la recherche du meilleur arbre phylogénétique en fonction des matrices de dissimilarité et de poids fournies par l'utilisateur. Elle tient compte des poids choisis arbitrairement, ou selon un des modèles classiques proposés dans la littérature.

Le transfert latéral des gènes (TLG) ou encore transfert horizontal de gène (THG), peut influencer l'évolution de différents groupes d'organismes. Par exemple entre des végétaux et des bactéries. Par ce mécanisme des gènes d'une espèce donnée sont transférés au génome d'une autre espèce sans qu'elles aient de liens de parenté proches. On retrouve le plus souvent ce phénomène chez les bactéries. On peut aussi le l'observer chez certaines espèces d'eucaryotes. Ainsi, la construction de réseaux réticulés mettrait en évidence des événements de spéciation par transfert de gène.

Dans T-Rex, lors de la reconstruction de réticulogrammes de THG (transfert horizontal de gène) l'arbre de gène est inscrit dans l'arbre d'espèces en utilisant le critère des moindres carrés. Les transferts horizontaux du gène considéré sont alors ajoutés dans l'arbre d'espèces, (pour plus de détail voir Boc et Makarenkov (2003)). T-Rex permet également la reconstruction d'un arbre additif à partir d'une matrice de dissimilarité incomplète (contenant des valeurs manquantes).

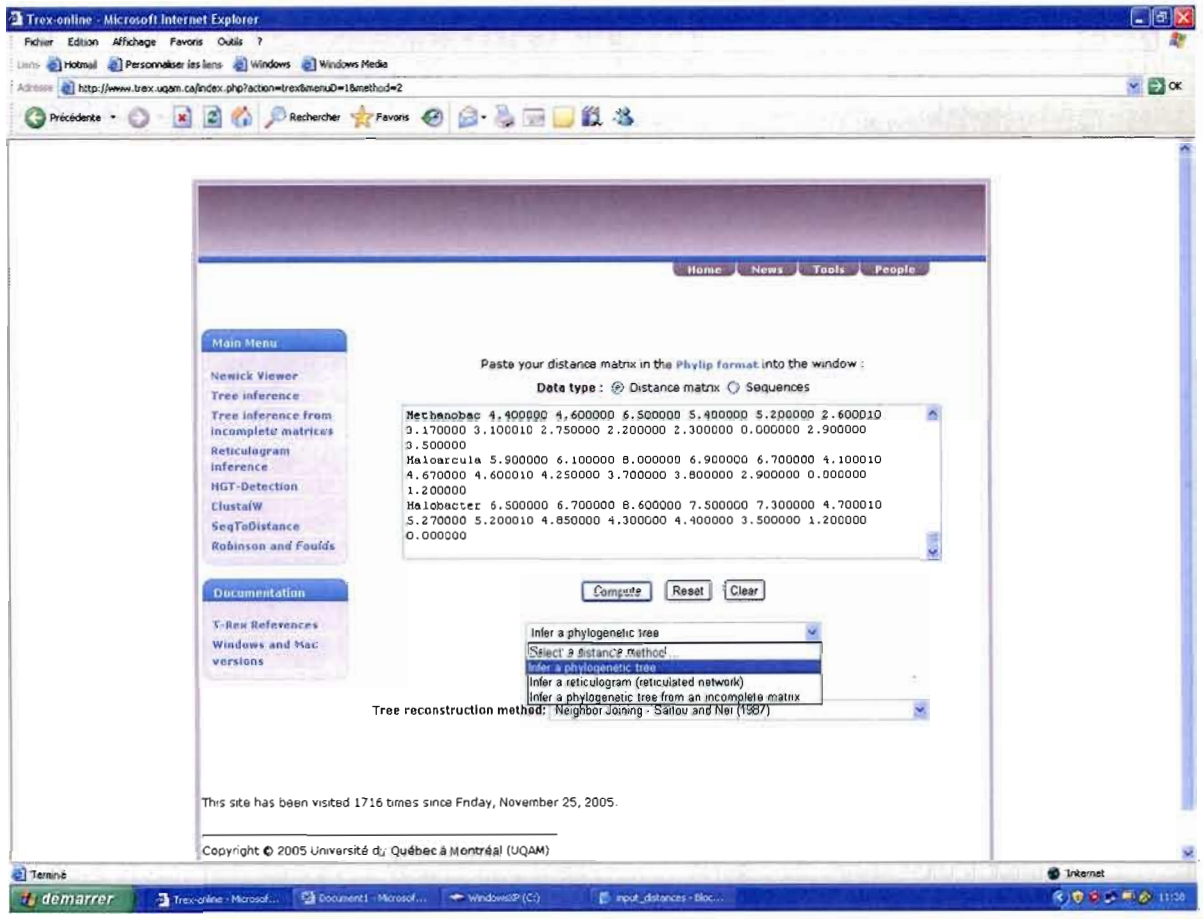


Figure 2.14 Écran de T-Rex montrant la boite de dialogue permettant de choisir, une méthode de reconstruction d'arbres phylogénétiques.

2.2.3 Pour analyser les matrices incomplètes T-Rex inclut les quatre méthodes suivantes :

(1) La Méthode des Triangles de Guénoche et Leclerc (2001).

(2) La Procédure Ultramétrique pour l'estimation des valeurs manquantes de Soete (1984) et Lapointe et Landry (1997) suivie par la méthode NJ.

(3) La Procédure Additive pour l'estimation des valeurs manquantes de Lapointe et Landry (1997) suivie par NJ.

(4) La Méthode de poids (MW^*) de Makarenkov et Lapointe (2004).

La première méthode, appelée Méthode des Triangles, permet de reconstruire, un arbre non enraciné à partir d'au moins $2n$-3 valeurs de distances entre les n éléments de X. Cette construction est basée sur une relation entre X-arbres et 2-arbres valués sur un ensemble de sommets X. Le lecteur est référé a l'article de Guénoche et Leclerc (2001). La deuxième méthode, appelée estimation Ultramétrique + NJ, se base sur l'exécution de l'algorithme mettant en application l'inégalité ultramétrique pour évaluer toutes les valeurs manquantes dans la matrice de distances (une description plus détaillée se trouve dans De Soete (1984) et Lapointe et Landry (1997)), suivie de la méthode NJ. La troisième méthode disponible est appelée estimation Additive + NJ. Cette méthode ne diffère de la seconde que par sa première étape qui se base sur une estimation additive, dans laquelle la condition des quatre points est utilisée pour évaluer les valeurs manquantes dans la matrice de dissimilarité (une description plus détaillée se trouve dans De Soete (1984) and Lapointe and Landry (1997)). La quatrième méthode est une version modifiée de la méthode de poids (MW) dans laquelle on affecte le poids 0 à des entrées manquantes de la matrice de dissimilarité, le poids 1 à des entrées existantes et des poids 0.5 à des valeurs reestimées.

2.2.4 Format de sortie dans T-Rex :

Le logiciel T-Rex permet d'afficher les statistiques d'inférence, ainsi que l'arbre obtenu en format Newick.

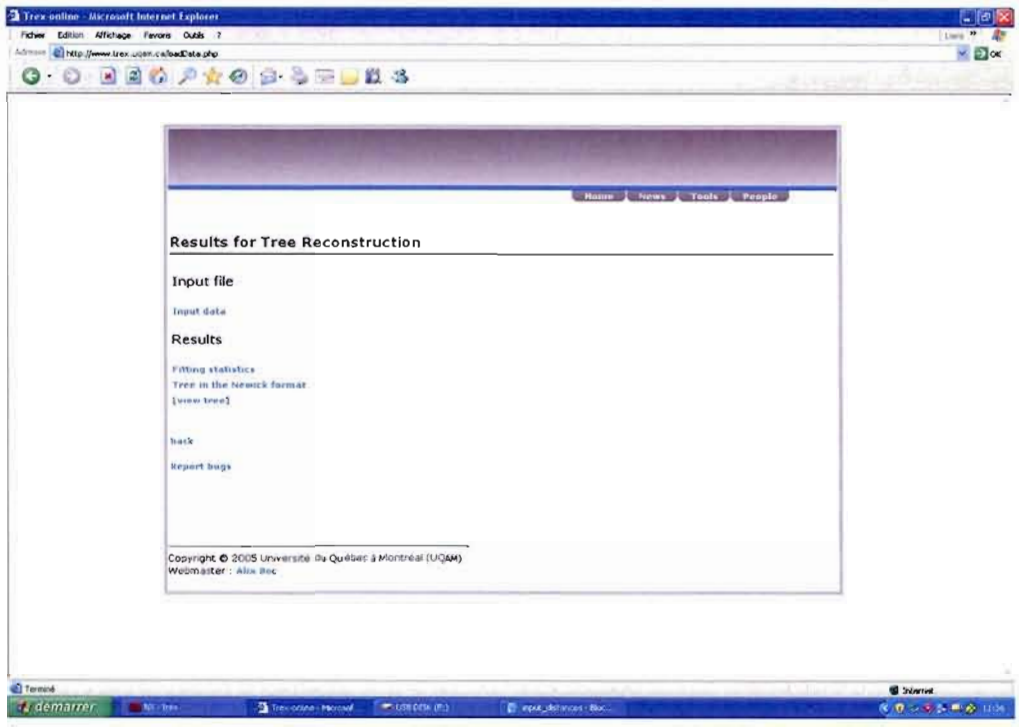


Figure 2.15 Options de sortie dans T-Rex

a) Statistiques :

T-Rex génère un fichier texte nomme Output ayant le format de l'exemple ci-dessous :

```
Tree reconstruction method - Neighbor Joining

TREE METRIC (ADDITIVE DISTANCE) MATRIX (AD)


Ferroplasm   0.00 2.00 7.00 6.00 7.00 7.00 7.00 6.00 5.00 4.00 5.00 5.00 6.00 6.00

Thermoplas   2.00 0.00 7.00 6.00 7.00 7.00 7.00 6.00 5.00 4.00 5.00 5.00 6.00 6.00

Aeropyrum   7.00 7.00 0.00 3.00 2.00 6.00 6.00 5.00 6.00 7.00 8.00 6.00 9.00 9.00
```

Pyrobaculu 6.00 6.00 3.00 0.00 3.00 5.00 5.00 4.00 5.00 6.00 7.00 5.00 8.00 8.00

Sulfolobus 7.00 7.00 2.00 3.00 0.00 6.00 6.00 5.00 6.00 7.00 8.00 6.00 9.00 9.00

Pyrococcus 7.00 7.00 6.00 5.00 6.00 0.00 2.00 3.00 6.00 7.00 8.00 6.00 9.00 9.00

Pyrococcu0 7.00 7.00 6.00 5.00 6.00 2.00 0.00 3.00 6.00 7.00 8.00 6.00 9.00 9.00

Pyrococcu1 6.00 6.00 5.00 4.00 5.00 3.00 3.00 0.00 5.00 6.00 7.00 5.00 8.00 8.00

Methanococ 5.00 5.00 6.00 5.00 6.00 6.00 6.00 5.00 0.00 5.00 6.00 2.00 7.00 7.00

Archaeoglo 4.00 4.00 7.00 6.00 7.00 7.00 7.00 6.00 5.00 0.00 3.00 5.00 4.00 4.00

Methanosar 5.00 5.00 8.00 7.00 8.00 8.00 8.00 7.00 6.00 3.00 0.00 6.00 3.00 3.00

Methanobac 5.00 5.00 6.00 5.00 6.00 6.00 6.00 5.00 2.00 5.00 6.00 0.00 7.00 7.00

Haloarcula 6.00 6.00 9.00 8.00 9.00 9.00 9.00 8.00 7.00 4.00 3.00 7.00 0.00 2.00

Halobacter 6.00 6.00 9.00 8.00 9.00 9.00 9.00 8.00 7.00 4.00 3.00 7.00 2.00 0.00

THE FOLLOWING STATISTICS ARE AVAILABLE FOR

A GIVEN DISSIMILARITY (D) AND AN OBTAINED TREE METRIC (AD)

Least-squares coefficient $\sum_{i<j} (Dij-ADij)^2 = 0.0000000000$

Average absolute difference $\sum_{i<j} |Dij-ADij|/(n(n-1)/2) = 0.0000000000$

Maximum absolute difference $\max_{i,j} |Dij-ADij| = 0.0000000000$

Total length of the tree L = 25.000000

TREE EDGES WITH THEIR LENGTHS

15--2 1.000000

16--4 1.000000

17--3	1.000000
18--16	1.000000
16--17	1.000000
17--5	1.000000
19--6	1.000000
20--19	1.000000
19--7	1.000000
21--18	1.000000
18--20	1.000000
20--8	1.000000
22--9	1.000000
23--21	1.000000
21--22	1.000000
22--12	1.000000
24--10	1.000000
25--11	1.000000
26--13	1.000000
1--15	1.000000
15--23	1.000000
23--24	1.000000
24--25	1.000000
25--26	1.000000
26--14	1.000000

Les premières lignes décrivent le traitement choisi, par exemple reconstruction d'arbre a partir d'une matrice incomplète, la méthode utilisée (dans l'exemple ci-dessus la méthode utilisée est Neighbor Joining) et matrice additive obtenue.

Le fichier intègre des informations statistiques comme par exemple, le cœfficient des moindres carrés, la différence moyenne absolue, la différence maximum absolue, longueur totale de l'arbre et la liste des branches avec longueur.

b) Arbre en format Newick :

T-Rex génère un fichier texte nomme Newickfile. Il contient l'arbre en format Newick. Le nombre réel, après un nom de l'espèce précédé par des deux points représente la longueur de la branche correspondante.

(((Pyrobaculu:1.0000,(Aeropyrum:1.0000,Sulfolobus:1.0000):1.0000):1.0000,(Pyrococcu1:1.0000,(Pyrococcus:1.0000,Pyrococcu0:1.0000):1.0000):1.0000):1.0000,(Methanococ:1.0000,Methanobac:1.0000):1.0000,((Ferroplasm:1.0000,Thermoplas:1.0000):1.0000,(Archaeoglo:1.0000,(Methanosar:1.0000,(Haloarcula:1.0000,Halobacter:1.0000):1.0000):1.0000):1.0000):1.0000);

Figure 2.16 Exemple d'arbre en format Newick généré par le logiciel T-Rex

c) Affichage de l'arbre :

On peut choisir d'afficher l'arbre en format radiale, axial ou hiérarchique : l'arbre peut être enraciné ou non, on peut garder les noms des espèces ou non et numéroter les nœuds internes.

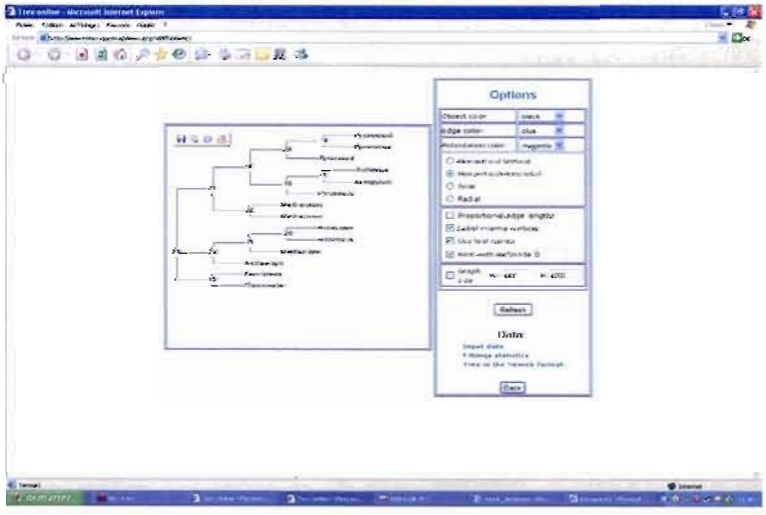


Figure 2.17 Représentation hiérarchique horizontale de l'arbre (version Web)

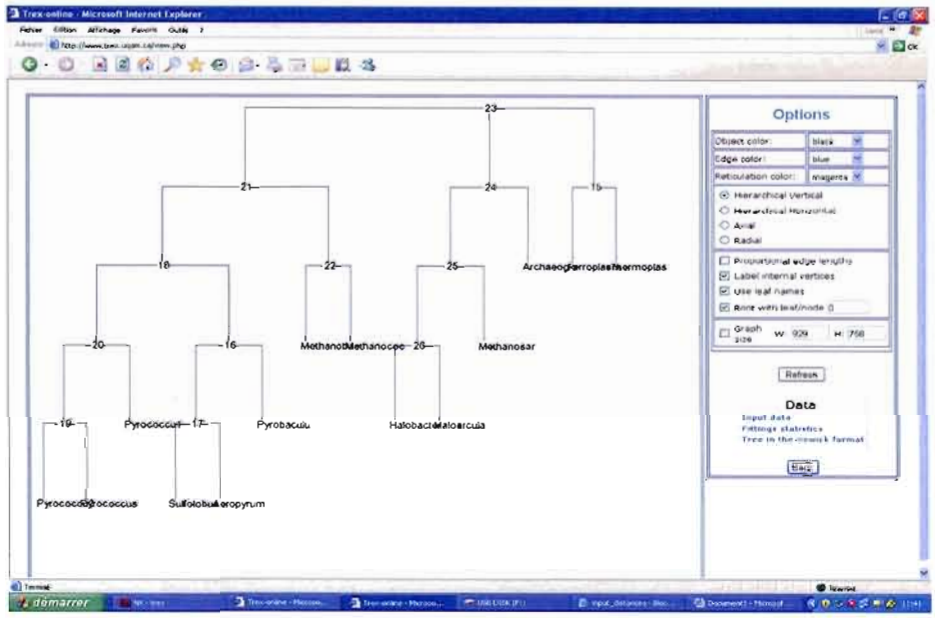


Figure 2.18 Représentation hiérarchique verticale de l'arbre (version Web)

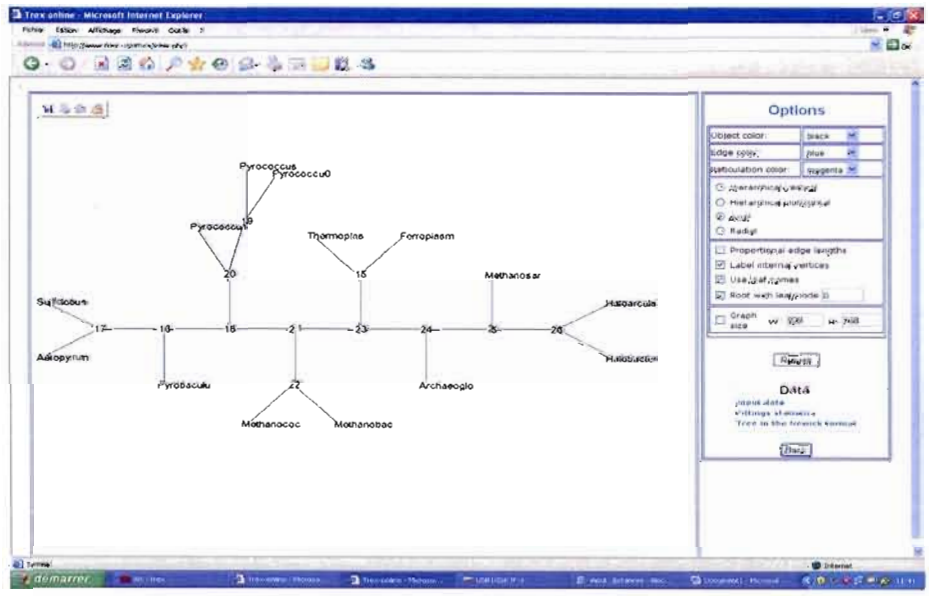


Figure 2.19 Représentation axiale de l'arbre (version Web)

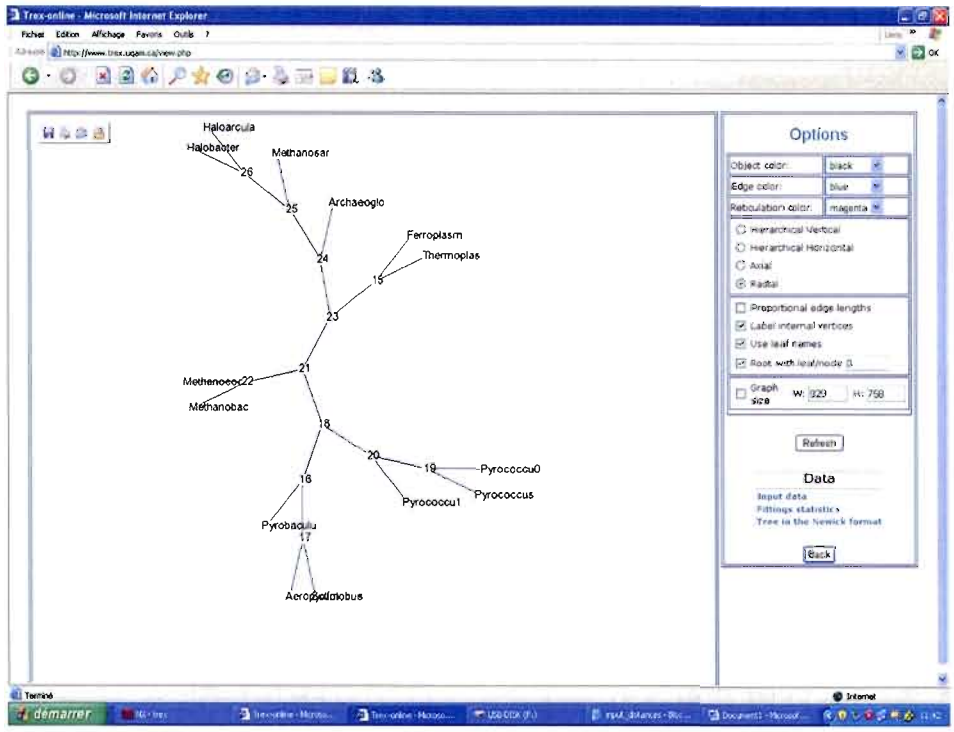


Figure 2.20 Représentation radiale de l'arbre (version Web)

CHAPITRE III

Reconstruction d'arbres et des réseaux dans T-Rex

Ce chapitre présente les méthodes pour la reconstruction des arbres phylogénétiques qui sont disponibles dans T-Rex Web et dans le développement desquelles j'ai été impliqué.

3.1 Tree inference (inférence d'arbres)

T-Rex offre trois options dans son menu pour l'inférence des arbres, inférence des arbres à partir d'une matrice complète, inférence à partir d'une matrice incomplète et inférence des réseaux réticulés.

3.1.1 Inférence d'arbres phylogénétiques

Ce programme permet la visualisation d'un arbre à partir d'une matrice complète en format Phylip, ou d'un ensemble des séquences en format Phylip. T-Rex inclut les méthodes d'inférences suivantes :

a- Méthodes de distance :

- Neighbor-Joining
- ADDTREE
- Unweighted Neighbor Joining
- Circular order reconstruction
- Weighted least-squares (MW)
- BIONJ
- FITCH (PHYLIP)
-

b- Méthodes de parcimonie :

- DNAPARS (PHYLIP)
- PROTPARS (PHYLIP)
- PARS (PHYLIP)
- DOLLOP (PHYLIP)

c- Maximum de vraisemblance :

- PHYML
- DNAML (PHYLIP)
- DNAMLK (PHYLIP)
- PROTML (PHYLIP)
- PROTMLK (PHYLIP)

Si on choisit la méthode de minimisation par les moindres carrés avec poids (Weighted Least-Squares), T-Rex offre des options supplémentaires telles qu'utiliser une matrice de poids particulière et de l'appliquer localement ou globalement (MW local optimizatoin, MW global optimization) pour plus de détails voir l'article de Makarenkov et Leclerc (1999), ainsi que le mode de calcul des écarts des moindres carrés (Weighted matrix W).

La matrice des poids (Weight matrix W) pour un modele de dimension *n*x*n* est définie comme matrice de nombres $W_{i,x}$ ou i est dans {1,2,3,.....,*n*}et x est dans {A,T,G,C} pour l'ADN.

Le score pour les $x_1, x_2, x_3, \ldots, x_n$ est donnée par $W_{1,x_1} + W_{2,x_2} + \ldots + W_{n,x_n}$.

Par exemple, la matrice de poids pour ces données se calcule comme suit :

	-2	-1	0	1	2	3	4	5
A	-0.43	-5.01	1.34	-1.06	0.02	-0.12	-0.51	-0.41
C	-0.77	1.08	-5.33	-0.24	-0.08	0.13	-0.19	-0.07
G	0.43	-4.50	-4.50	0.95	-4.50	0.02	0.48	0.22
T	0.54	-5.07	-1.67	0.01	0.50	-0.10	0.23	0.24
Concensus	T	C	A	G	T	n	G	t
	G						t	g

Exemple :	T	C	T	G	T	A	T	G
Score	0.54 +	1.08 +	-1.67 +	0.95 +	0.50 +	-0.12 +	0.23 +	0.22 = 1.73

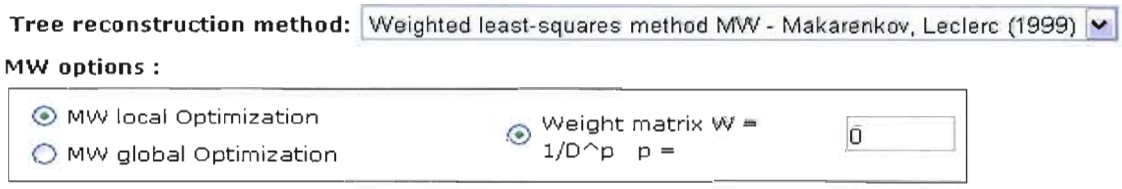


Figure 3.1 Options supplémentaires si le format d'entrée choisi est séquences

3.1.1.1 Model d'évolution des séquences (voir également le paragraphe 3.4)

Cette option offre la possibilité de choisir une méthode de calcul des distances à partir séquences observées. Les méthodes incluses dans T-Rex sont les suivantes :

- Jukes-Cantor (Jukes et Cantor 1969)
- Tajima-Nei (Tajima et Nei 1984)
- Kimura 2-Paramtres (Kimura 1980)
- Jin-Nei Gamma (JIN 90)
- Kimura Protéine (Kimura 1983)
- LogDet (Barry and Hartigan, 1987; Lake, 1994; Steel, 1994; Lockhart et. al., 1994).
- F84 (Kishino and Hasegawa, 1989; Felsenstein and Churchill, 1996)

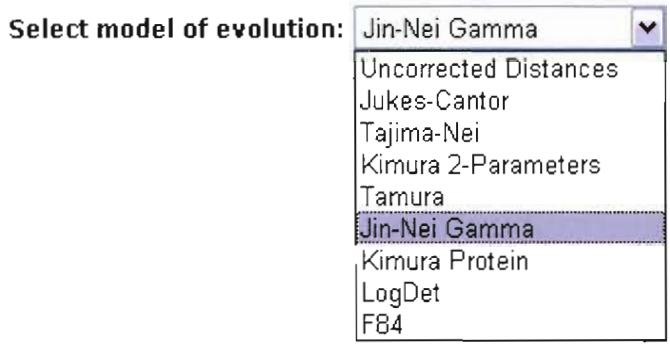


Figure 3.2 Models d'évolution dans T-Rex

Le programme SeqToDistance permet de transformer une séquence en format Phylip en une matrice de distance selon le modèle d'évolution choisi :

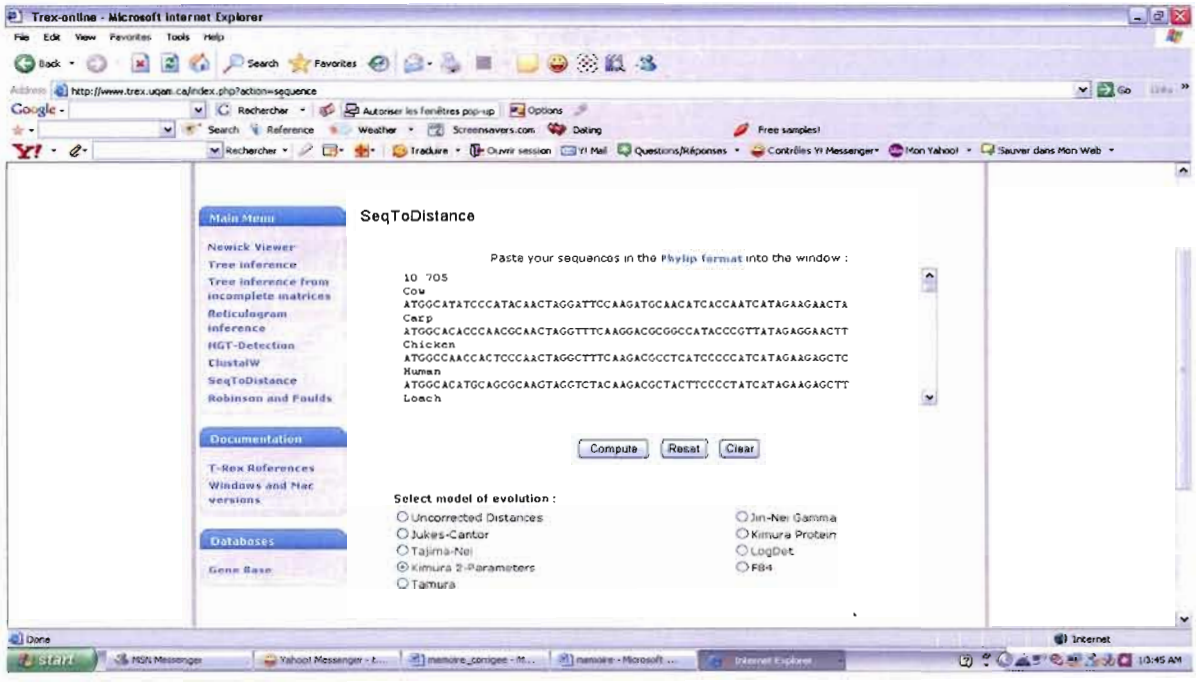


Figure 3.3 Option SeqToDistance dans (T-Rex WEB)

Exemple :

```
10
Cow 0.000000 0.389936 0.377150 0.372345 0.391747 0.264776 0.234731 0.222502 0.202384 0.358694
Carp 0.389936 0.000000 0.390024 0.464200 0.242591 0.370880 0.373650 0.374861 0.351609 0.297523
Chicken 0.377150 0.390024 0.000000 0.405968 0.419812 0.381943 0.370022 0.382495 0.369233 0.348172
Human 0.372345 0.464200 0.405968 0.000000 0.448174 0.347077 0.345616 0.382572 0.333142 0.426815
Loach 0.391747 0.242591 0.419812 0.448174 0.000000 0.394057 0.364578 0.393739 0.382714 0.315993
Mouse 0.264776 0.370880 0.381943 0.347077 0.394057 0.000000 0.168029 0.253449 0.259810 0.326221
Rat 0.234731 0.373650 0.370022 0.345616 0.364578 0.168029 0.000000 0.271403 0.253449 0.351622
Seal 0.222502 0.374861 0.382495 0.382572 0.393739 0.253449 0.271403 0.000000 0.225159 0.364816
Whale 0.202384 0.351609 0.369233 0.333142 0.382714 0.259810 0.253449 0.225159 0.000000 0.323372
Frog 0.358694 0.297523 0.348172 0.426815 0.315993 0.326221 0.351622 0.364816 0.323372 0.000000
```

3.1.1.2 Validation

a) Bootstrap

Cette méthode permet l'évaluation de la robustesse d'un arbre phylogénétique en attribuant une valeur à chacun de ses branches. Le bootstrap procède par un tirage aléatoire des caractères avec remise en jeu du caractère tiré. Il y a autant de tirages que la matrice contient de caractères. Le processus de tirage se répète, et à chaque fois on obtient un nouvel arbre qui peut être plus ou moins proche de l'arbre initial. Le résultat est généralement représenté sous la forme d'un arbre consensus majoritaire. A chaque nœud on attribue une valeur indiquant le pourcentage d'arbres issus du processus de tirage. Cette méthode part du principe que plus le nombre de caractères qui soutient un regroupement donné est grand, plus la probabilité qu'ils soient présents dans le tirage aléatoire est élevée et donc plus grande sera la proportion d'arbres qui contiendra ce regroupement. Si le pourcentage d'une branche est élevé elle est dite robuste. Cette méthode permet de mettre en évidence les groupes de taxa problématiques.

b) Jackknife

Cette méthode permet de créer une nouvelle matrice de données de taille inférieure à la matrice originale, en prenant de manière aléatoire les lignes ou les colonnes de la matrice initiale sans remise de l'élément après tirage. Un ou plusieurs caractères peuvent être ainsi soustraits de la matrice initiale. Cette méthode met également en évidence les noeuds les plus solides et les taxa problématiques.

Par défaut le nombre de tirage est 100.

3.1.1.3 Pénalité de gap

Ce champ est requis quand on choisit la méthode de Jukes-Cantor pour la correction des distances observées dans le champ model d'évolution. La pénalité de gap est une valeur numérique appliquée à des points de similitude pour l'introduction d'un espace d'insertion ou de suppression, la prolongation d'un espace, ou toutes les deux. Les pénalités de gap sont généralement soustraites des scores cumulatifs définis pour la comparaison de deux séquences ou plus par l'intermédiaire d'un algorithme d'optimisation qui essaye de maximiser ces scores.

3.1.1.4 PEMV estimation of missing bases values (méthode PEMV d'estimation des bases manquantes)

En choisissant la méthode Jukes-Cantor ou Kimura 2-paramètres comme méthodes de calcul des distances, un bouton radio s'affiche permettant d'ignorer les bases manquantes ou d'appliquer la méthode PEMV pour obtenir la distance corrigée en considérant les sites incomplets. Cette méthode permet de calculer la distance corrigée entre les paires de séquences nucléotidiques lors de l'inférence phylogénétique, en calculant préalablement une vraisemblance pour chacun des nucléotides manquants. Une fois ces vraisemblances obtenues, elles peuvent être utilisées pour calculer la distance corrigée entre les séquences en utilisant le modèle d'évolution voulu. Pour plus de détail voir l'article d'Abdoulaye Baniré Diallo, François-Joseph Lapointe et Vladimir Makarenkov 2006.

3.1.1.5 Calcul de la valeur du paramètre α ?

Si la méthode choisie pour le calcul des distances est Jin-Nei Gamma, un bouton radio s'affiche permettant de saisir le paramètre α ou de le calculer. Le paramètre α est évalué à partir des données et correspond à l'intervalle de variations. Plus α est petit, plus l'intervalle des variations est grand.

3.2 Réseaux réticulés

T-Rex permet également d'inférer les réseaux réticules, le programme calcule d'abord un arbre additif classique en utilisant un des cinq algorithmes disponibles pour la reconstruction d'arbres. Ensuite, à chaque étape de la procédure de reconstruction de réticulogrammes (i.e réseau réticule), on choisit une réticulation qui optimise la fonction des moindres carrés ou la fonction de pondération spécifique. On ajoute la branche optimale calculée au réticulogramme croissant. Deux critères statistiques (Q_1 et Q_2) sont proposés pour mesurer le gain en ajustement quand des réticulations (i.e branches supplémentaires) sont ajoutées. Le minimum de chacun de ces deux critères peut suggérer une règle d'arrêt pour l'addition des réticulations. Ainsi, l'utilisateur peut choisir un critère approprié pour arrêter le procédé d'ajout des réticulations ou pour indiquer un nombre exact de réticulations à placer dans un réticulogramme. Une description plus détaillée de la méthode de reconstruction des réticulogramme est disponible dans les articles de Legendre et Makarenkov (2002) et Makarenkov et Legendre (2004).

Les options disponibles permettent de modifier le mode de calcul des fonctions à optimiser et fixer le nombre de réticulations à ajouter. Dans l'exemple ci-dessous ce nombre est égal à 5.

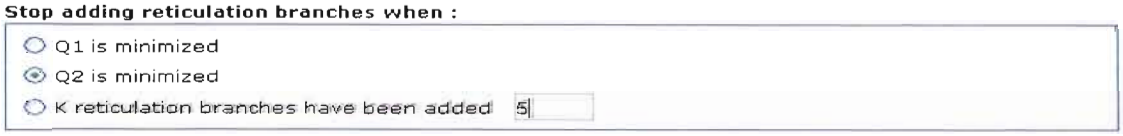


Figure 3.4 Mode de calcul des fonctions à optimiser dans T-Rex WEB

Exemple pratique (ensemble de 10 espèces représentées par les séquences d'ADN de 750 caractères)

```
10 705

Cow       ATGGCATATCCCATACAACTAGGATTCCAAGATGCAACATCACCAATCATAGAAGAACTA

Carp      ATGGCACACCCAACGCAACTAGGTTTCAAGGACGCGGCCATACCCGTTATAGAGGAACTT

Chicken   ATGGCCAACCACTCCCAACTAGGCTTTCAAGACGCCTCATCCCCCATCATAGAAGAGCTC

Human     ATGGCACATGCAGCGCAAGTAGGTCTACAAGACGCTACTTCCCCTATCATAGAAGAGCTT

Loach     ATGGCACATCCCACACAATTAGGATTCCAAGACGCGGCCTCACCCGTAATAGAAGAACTT

Mouse     ATGGCCTACCCATTCCAACTTGGTCTACAAGACGCCACATCCCCTATTATAGAAGAGCTA

Rat       ATGGCTTACCCATTTCAACTTGGCTTACAAGACGCTACATCACCTATCATAGAAGAACTT

Seal      ATGGCATACCCCCTACAAATAGGCCTACAAGATGCAACCTCTCCCATTATAGAGGAGTTA

Whale     ATGGCATATCCATTCCAACTAGGTTTCCAAGATGCAGCATCACCCATCATAGAAGAGCTC

Frog      ATGGCACACCCATCACAATTAGGTTTTCAAGACGCAGCCTCTCCAATTATAGAAGAATTA


CTTCACTTTCATGACCACACGCTAATAATTGTCTTCTTAATTAGCTCATTAGTACTTTAC

CTTCACTTCCACGACCACGCATTAATAATTGTGCTCCTAATTAGCACTTTAGTTTTATAT

GTTGAATTCCACGACCACGCCCTGATAGTCGCACTAGCAATTTGCAGCTTAGTACTCTAC

ATCACCTTTCATGATCACGCCCTCATAATCATTTTCCTTATCTGCTTCCTAGTCCTGTAT

CTTCACTTCCATGACCATGCCCTAATAATTGTATTTTTGATTAGCGCCCTAGTACTTTAT

ATAAATTTCCATGATCACACACTAATAATTGTTTTCCTAATTAGCTCCTTAGTCCTCTAT

ACAAACTTTCATGACCACACCCTAATAATTGTATTCCTCATCAGCTCCCTAGTACTTTAT

CTACACTTCCATGACCACACATTAATAATTGTGTTCCTAATTAGCTCATTAGTACTCTAC

CTACACTTTCACGATCATACACTAATAATCGTTTTTCTAATTAGCTCTTTAGTTCTCTAC

CTTCACTTCCACGACCATACCCTCATAGCCGTTTTTCTTATTAGTACGCTAGTTCTTTAC


ATTATTTCACTAATACTAACGACAAAGCTGACCCATACAAGCACGATAGATGCACAAGAA

ATTATTACTGCAATGGTATCAACTAAACTTACTAATAAATATATTCTAGACTCCCAAGAA

CTTCTAACTCTTATACTTATAGAAAAACTATCA---TCAAACACCGTAGATGCCCAAGAA
```

GCCCTTTTCCTAACACTCACAACAAAACTAACTAATACTAACATCTCAGACGCTCAGGAA

GTTATTATTACAACCGTCTCAACAAAACTCACTAACATATATATTTTGGACTCACAAGAA

ATCATCTCGCTAATATTAACAACAAAACTAACACATACAAGCACAATAGATGCACAAGAA

ATTATTTCACTAATACTAACAACAAAACTAACACACACAAGCACAATAGACGCCCAAGAA

ATTATCTCACTTATACTAACCACGAAACTCACCCACACAAGTACAATAGACGCACAAGAA

ATTATTACCCTAATGCTTACAACCAAATTAACACATACTAGTACAATAGACGCCCAAGAA

ATTATTACTATTATAATAACTACTAAACTAACTAATACAAACCTAATGGACGCACAAGAG

GTAGAGACAATCTGAACCATTCTGCCCGCCATCATCTTAATTCTAATTGCTCTTCCTTCT

ATCGAAATCGTATGAACCATTCTACCAGCCGTCATTTTAGTACTAATCGCCCTGCCCTCC

GTTGAACTAATCTGAACCATCCTACCCGCTATTGTCCTAGTCCTGCTTGCCCTCCCCTCC

ATAGAAACCGTCTGAACTATCCTGCCCGCCATCATCCTAGTCCTCATCGCCCTCCCATCC

ATTGAAATCGTATGAACTGTGCTCCCTGCCCTAATCCTCATTTTAATCGCCCTCCCCTCA

GTTGAAACCATTTGAACTATTCTACCAGCTGTAATCCTTATCATAATTGCTCTCCCCTCT

GTAGAAACAATTTGAACAATTCTCCCAGCTGTCATTCTTATTCTAATTGCCCTTCCCTCC

GTGGAAACGGTGTGAACGATCCTACCCGCTATCATTTTAATTCTCATTGCCCTACCATCA

GTAGAAACTGTCTGAACTATCCTCCCAGCCATTATCTTAATTTTAATTGCCTTGCCTTCA

ATCGAAATAGTGTGAACTATTATACCAGCTATTAGCCTCATCATAATTGCCCTTCCATCC

TTACGAATTCTATACATAATAGATGAAATCAATAACCCATCTCTTACAGTAAAAACCATA

CTACGCATCCTGTACCTTATAGACGAAATTAACGACCCTCACCTGACAATTAAAGCAATA

CTCCAAATCCTCTACATAATAGACGAAATCGACGAACCTGATCTCACCCTAAAAGCCATC

CTACGCATCCTTTACATAACAGACGAGGTCAACGATCCCTCCCTTACCATCAAATCAATT

CTACGAATTCTATATCTTATAGACGAGATTAATGACCCCCACCTAACAATTAAGGCCATG

CTACGCATTCTATATATAATAGACGAAATCAACAACCCCGTATTAACCGTTAAAACCATA

CTACGAATTCTATACATAATAGACGAGATTAATAACCCAGTTCTAACAGTAAAAACTATA

TTACGAATCCTCTACATAATGGACGAGATCAATAACCCTTCCTTGACCGTAAAAACTATA

TTACGGATCCTTTACATAATAGACGAAGTCAATAACCCCTCCCTCACTGTAAAAACAATA

CTTCGTATCCTATATTTAATAGATGAAGTTAATGATCCACACTTAACAATTAAAGCAATC

GGACATCAGTGATACTGAAGCTATGAGTATACAGATTATGAGGACTTAAGCTTCGACTCC

GGACACCAATGATACTGAAGTTACGAGTATACAGACTATGAAAATCTAGGATTCGACTCC

GGACACCAATGATACTGAACCTATGAATACACAGACTTCAAGGACCTCTCATTTGACTCC

GGCCACCAATGGTACTGAACCTACGAGTACACCGACTACGGCGGACTAATCTTCAACTCC

GGGCACCAATGATACTGAAGCTACGAGTATACTGATTATGAAAACTTAAGTTTTGACTCC

GGGCACCAATGATACTGAAGCTACGAATATACTGACTATGAAGACCTATGCTTTGATTCA

GGACACCAATGATACTGAAGCTATGAATATACTGACTATGAAGACCTATGCTTTGACTCC

GGACATCAGTGATACTGAAGCTATGAGTACACAGACTACGAAGACCTGAACTTTGACTCA

GGTCACCAATGATATTGAAGCTATGAGTATACCGACTACGAAGACCTAAGCTTCGACTCC

GGCCACCAATGATACTGAAGCTACGAATATACTAACTATGAGGATCTCTCATTTGACTCT

TACATAATTCCAACATCAGAATTAAAGCCAGGGGAGCTACGACTATTAGAAGTCGATAAT

TATATAGTACCAACCCAAGACCTTGCCCCCGGACAATTCCGACTTCTGGAAACAGACCAC

TACATAACCCCAACAACAGACCTCCCCCTAGGCCACTTCCGCCTACTAGAAGTCGACCAT

TACATACTTCCCCCATTATTCCTAGAACCAGGCGACCTGCGACTCCTTGACGTTGACAAT

TACATAATCCCCACCCAGGACCTAACCCCTGGACAATTCCGGCTACTAGAGACAGACCAC

TATATAATCCCAACAAACGACCTAAAACCTGGTGAACTACGACTGCTAGAAGTTGATAAC

TACATAATCCCAACCAATGACCTAAAACCAGGTGAACTTCGTCTATTAGAAGTTGATAAT

TATATGATCCCCACACAAGAACTAAAGCCCGGAGAACTACGACTGCTAGAAGTAGACAAT

TATATAATCCCAACATCAGACCTAAAGCCAGGAGAACTACGATTATTAGAAGTAGATAAC

TATATAATTCCAACTAATGACCTTACCCCTGGACAATTCCGGCTGCTAGAAGTTGATAAT

CGAGTTGTACTACCAATAGAAATAACAATCCGAATGTTAGTCTCCTCTGAAGACGTATTA

CGAATAGTTGTTCCAATAGAATCCCCAGTCCGTGTCCTAGTATCTGCTGAAGACGTGCTA

CGCATTGTAATCCCCATAGAATCCCCCATTCGAGTAATCATCACCGCTGATGACGTCCTC

CGAGTAGTACTCCCGATTGAAGCCCCCATTCGTATAATAATTACATCACAAGACGTCTTG

CGAATGGTTGTTCCCATAGAATCCCCTATTCGCATTCTTGTTTCCGCCGAAGATGTACTA

CGAGTCGTTCTGCCAATAGAACTTCCAATCCGTATATTAATTTCATCTGAAGACGTCCTC

CGGGTAGTCTTACCAATAGAACTTCCAATTCGTATACTAATCTCATCCGAAGACGTCCTG

CGAGTAGTCCTCCCAATAGAAATAACAATCCGCATACTAATCTCATCAGAAGATGTACTC

CGAGTTGTCTTACCTATAGAAATAACAATCCGAATATTAGTCTCATCAGAAGACGTACTC

CGAATAGTAGTCCCAATAGAATCTCCAACCCGACTTTTAGTTACAGCCGAAGACGTCCTC

CACTCATGAGCTGTGCCCTCTCTAGGACTAAAAACAGACGCAATCCCAGGCCGTCTAAAC

CATTCTTGAGCTGTTCCATCCCTTGGCGTAAAAATGGACGCAGTCCCAGGACGACTAAAT

CACTCATGAGCCGTACCCGCCCTCGGGGTAAAAACAGACGCAATCCCTGGACGACTAAAT

CACTCATGAGCTGTCCCCACATTAGGCTTAAAAACAGATGCAATTCCCGGACGTCTAAAC

CACTCCTGGGCCCTTCCAGCCATGGGGGTAAAGATAGACGCGGTCCCAGGACGCCTTAAC

CACTCATGAGCAGTCCCCTCCCTAGGACTTAAAACTGATGCCATCCCAGGCCGACTAAAT

CACTCATGAGCCATCCCTTCACTAGGGTTAAAAACCGACGCAATCCCCGGCCGCCTAAAC

CACTCATGAGCCGTACCGTCCCTAGGACTAAAAACTGATGCTATCCCAGGACGACTAAAC

CACTCATGGGCCGTACCCTCCTTGGGCCTAAAAACAGATGCAATCCCAGGACGCCTAAAC

CACTCGTGAGCTGTACCCTCCTTGGGTGTCAAAACAGATGCAATCCCAGGACGACTTCAT

CAAACAACCCTTATATCGTCCCGTCCAGGCTTATATTACGGTCAATGCTCAGAAATTTGC

CAAGCCGCCTTTATTGCCTCACGCCCAGGGGTCTTTTACGGACAATGCTCTGAAATTTGT

CAAACCTCCTTCATCACCACTCGACCAGGAGTGTTTTACGGACAATGCTCAGAAATCTGC

CAAACCACTTTCACCGCTACACGACCGGGGGTATACTACGGTCAATGCTCTGAAATCTGT

CAAACCGCCTTTATTGCCTCCCGCCCCGGGGTATTCTATGGGCAATGCTCAGAAATCTGT

CAAGCAACAGTAACATCAAACCGACCAGGGTTATTCTATGGCCAATGCTCTGAAATTTGT

CAAGCTACAGTCACATCAAACCGACCAGGTCTATTCTATGGCCAATGCTCTGAAATTTGC

CAAACAACCCTAATAACCATACGACCAGGACTGTACTACGGTCAATGCTCAGAAATCTGT

CAAACAACCTTAATATCAACACGACCAGGCCTATTTTATGGACAATGCTCAGAGATCTGC

CAAACATCATTTATTGCTACTCGTCCGGGAGTATTTTACGGACAATGTTCAGAAATTTGC

GGGTCAAACCACAGTTTCATACCCATTGTCCTTGAGTTAGTCCCACTAAAGTACTTTGAA

GGAGCTAATCACAGCTTTATACCAATTGTAGTTGAAGCAGTACCTCTCGAACACTTCGAA

GGAGCTAACCACAGCTACATACCCATTGTAGTAGAGTCTACCCCCCTAAAACACTTTGAA


GGAGCAAACCACAGTTTCATGCCCATCGTCCTAGAATTAATTCCCCTAAAAATCTTTGAA
GGAGCAAACCACAGCTTTATACCCATCGTAGTAGAAGCGGTCCCACTATCTCACTTCGAA
GGATCTAACCATAGCTTTATGCCCATTGTCCTAGAAATGGTTCCACTAAAATATTTCGAA
GGCTCAAATCACAGCTTCATACCCATTGTACTAGAAATAGTGCCTCTAAAATATTTCGAA
GGTTCAAACCACAGCTTCATACCTATTGTCCTCGAATTGGTCCCACTATCCCACTTCGAG
GGCTCAAACCACAGTTTCATACCAATTGTCCTAGAACTAGTACCCCTAGAAGTCTTTGAA
GGAGCAAACCACAGCTTTATACCAATTGTAGTTGAAGCAGTACCGCTAACCGACTTTGAA
AAATGATCTGCGTCAATATTA----------TAA
AACTGATCCTCATTAATACTAGAAGACGCCTCGCTAGGAAGCTAA
GCCTGATCCTCACTA----------CTGTCATCTTAA
ATA----------GGGCCCGTATTTACCCTATAG
AACTGGTCCACCCTTATACTAAAAGACGCCTCACTAGGAAGCTAA
AACTGATCTGCTTCAATAATT----------TAA
AACTGATCAGCTTCTATAATT----------TAA
AAATGATCTACCTCAATGCTT----------TAA
AAATGATCTGTATCAATACTA----------TAA
AACTGATCTTCATCAATACTA---GAAGCATCACTA------AGA

Si on choisit Q_1 pour mesurer le gain en ajustement on obtient le réticulogramme suivant :

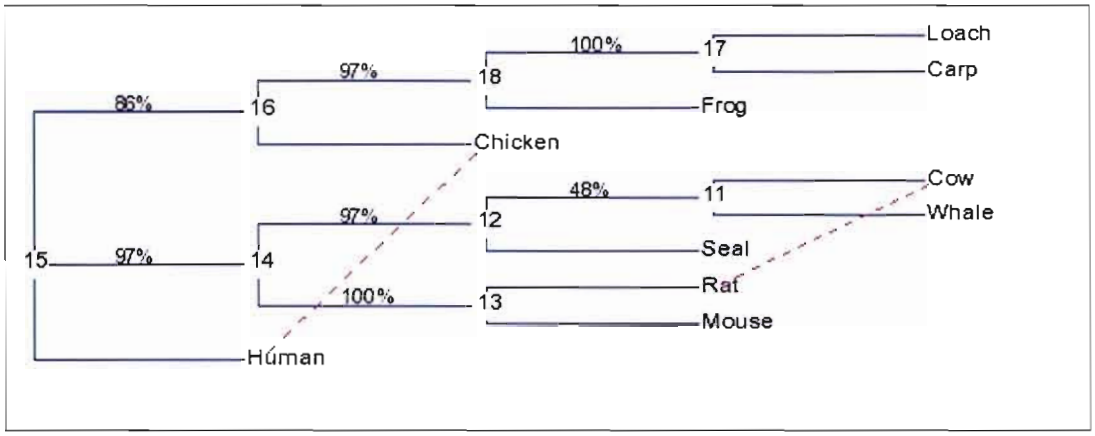


Figure 3.5 Réticulogramme obtenu avec le choix de Q_1 pour mesurer le gain en ajustement

Si on choisit Q_2 pour mesurer le gain en ajustement on obtient le réticulogramme suivant :

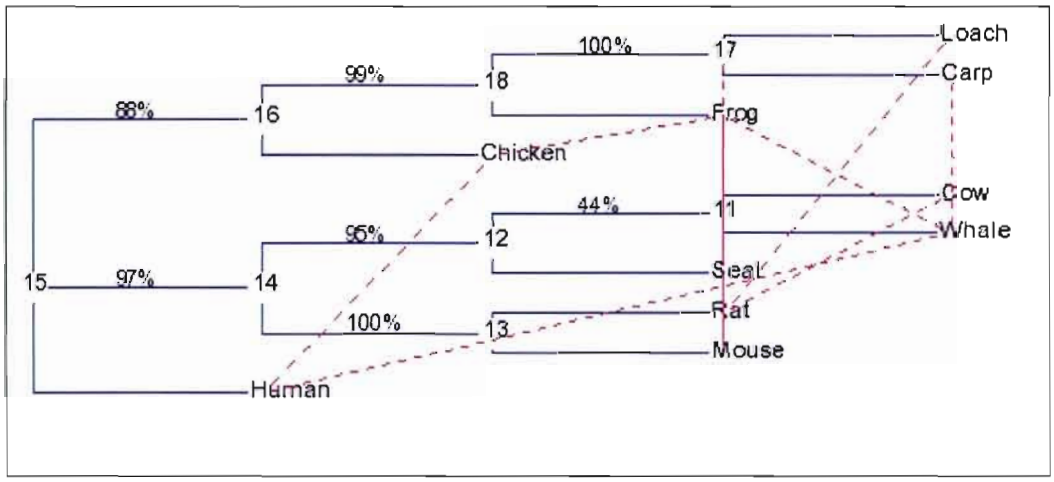


Figure 3.6 Réticulogramme obtenu avec le choix de Q_2 pour mesurer le gain en ajustement

Les traits rouges ajoutés entre les branches représentent des réticulations ajoutées à l'arbre original.

3.3 HGT-Detection - Détection des transferts horizontaux de gènes

3.3.1 Mécanisme

La conjugaison, la transformation et la transduction sont les trois principaux mécanismes qui expliquent les transferts horizontaux.

- Conjugaison : les organismes ont mis au point un système leur permettant de s'échanger du matériel génétique pour s'adapter à leur environnement.
- Des fragments d'ADN libres provenant d'organismes morts peuvent se trouver dans le milieu extérieur d'un organisme vivant, et peuvent être intégrés à l'intérieur d'une cellule particulière, puis intégrés au génome de cet organisme. Entre deux espèces différentes peut alors exister un transfert horizontal.
- Le fragment d'ADN est transféré d'une espèce à une autre via des virus ou des phages. Certains virus sont capables de transférer par erreur du matériel génétique d'une espèce à une autre.

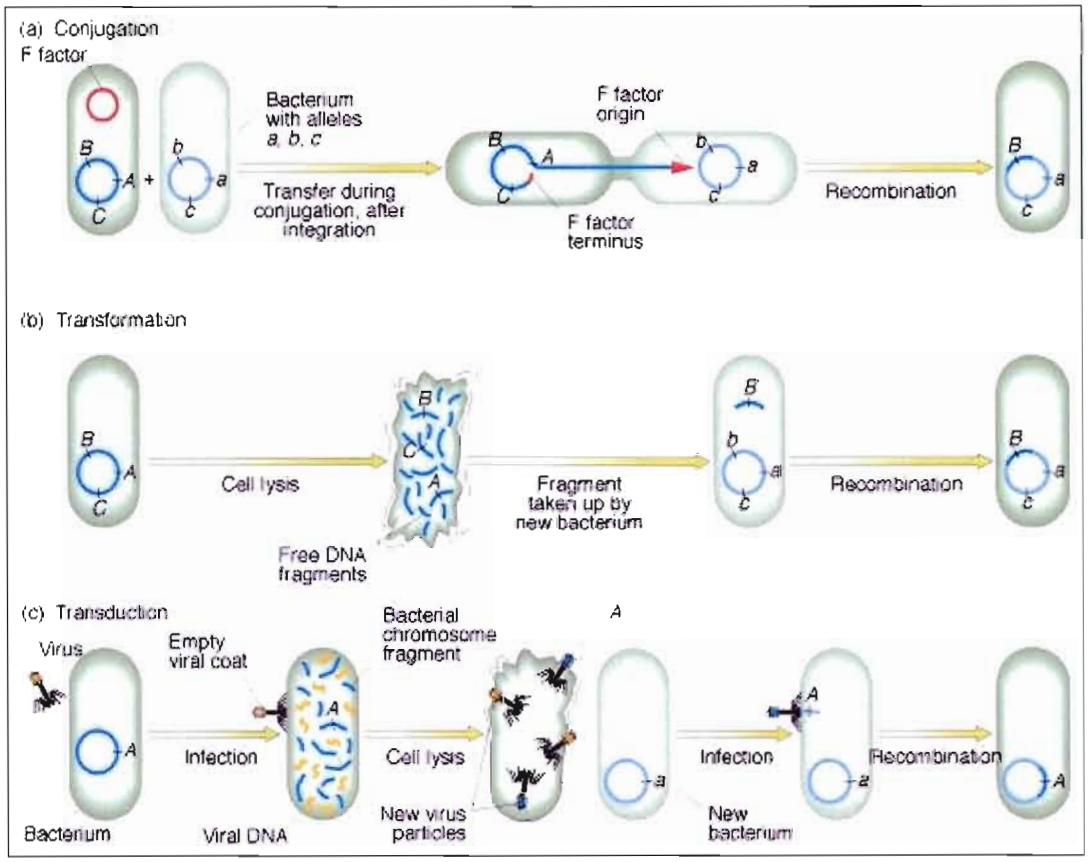


Figure 3.7 Trois mécanismes de transferts horizontaux de gènes.

3.3.2 Méthodes de détection des Transferts Horizontaux

Notre interface permet d'introduire l'arbre d'espèces et l'arbre de gène en format Newick ou en format Phylip, et de faire des choix parmi les paramètres suivants :

- Une case a cocher permettant d'enraciner l'arbre d'espèces sur le point médian.
- Une case a cocher permettant d'enraciner l'arbre de gène sur le point médian.
- Une case a cocher permettant d'utiliser les contraintes des sous-arbres.
- Un champ permettant de saisir le nombre maximal de transferts qui est fixé par défaut à 50.

- Un bouton radio permettant de choisir parmi deux méthodes de détection impliquant des scénarios unique et multiples des transferts horizontaux, pour plus de détails voir (Boc et Makarenkov 2003).
- Un bouton radio permettant de choisir entre Robinson et Foulds ou Least-Squares (moindres carrées) comme critère d'optimisation pour réconcilier les topologies de l'arbre de gène et celui d'espèces.

Exemple :

On soumet au programme les matrices suivantes :

- Matrice pour construire l'arbre d'espèces (i.e espèces de bactéries) :

9

Ferroplasm	0.000000	2.000000	7.000000	6.000000	7.000000	7.000000	7.000000	6.000000	5.000000
Thermoplas	2.000000	0.000000	7.000000	6.000000	7.000000	7.000000	7.000000	6.000000	5.000000
Aeropyrum_	7.000000	7.000000	0.000000	3.000000	2.000000	6.000000	6.000000	5.000000	6.000000
Pyrobaculu	6.000000	6.000000	3.000000	0.000000	3.000000	5.000000	5.000000	4.000000	5.000000
Sulfolobus	7.000000	7.000000	2.000000	3.000000	0.000000	6.000000	6.000000	5.000000	6.000000
Pyrococcus	7.000000	7.000000	6.000000	5.000000	6.000000	0.000000	2.000000	3.000000	6.000000
Pyrococcu0	7.000000	7.000000	6.000000	5.000000	6.000000	2.000000	0.000000	3.000000	6.000000
Methanococ	5.000000	5.000000	6.000000	5.000000	6.000000	6.000000	6.000000	5.000000	0.000000
Archaeoglo	4.000000	4.000000	7.000000	6.000000	7.000000	7.000000	7.000000	6.000000	5.000000

Tableau 3.1 Matrice d'espèces

- Matrice pour construire l'arbre du gène :

9

Ferroplasm	0.000000	2.800000	4.700000	3.600000	3.400000	2.600010	3.170000	3.100010	3.350000
Thermoplas	2.800000	0.000000	3.700000	2.600000	2.400000	2.800010	3.370000	3.300010	3.550000
Aeropyrum_	4.700000	3.700000	0.000000	4.100000	3.900000	4.700010	5.270000	5.200010	5.450000
Pyrobaculu	3.600000	2.600000	4.100000	0.000000	1.700000	3.600010	4.170000	4.100010	4.350000
Sulfolobus	3.400000	2.400000	3.900000	1.700000	0.000000	3.400010	3.970000	3.900010	4.150000
Pyrococcus	2.600010	2.800010	4.700010	3.600010	3.400010	0.000000	0.570010	0.500020	1.550010
Pyrococcu0	3.170000	3.370000	5.270000	4.170000	3.970000	0.570010	0.000000	0.070010	2.120000
Pyrococcu1	3.100010	3.300010	5.200010	4.100010	3.900010	0.500020	0.070010	0.000000	2.050010
Methanococ	3.350000	3.550000	5.450000	4.350000	4.150000	1.550010	2.120000	2.050010	0.000000

Tableau 3.2 Matrice du gène

Les résultats obtenus sont les suivants :

Arbre d'espèces :

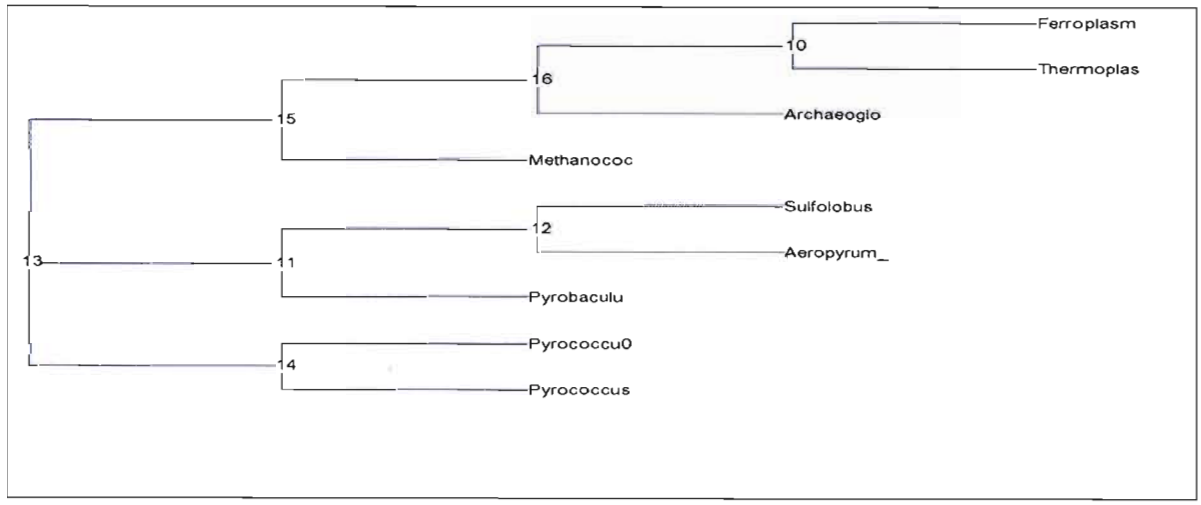


Figure 3.8 Arbre d'espèces

- Arbre du gène :

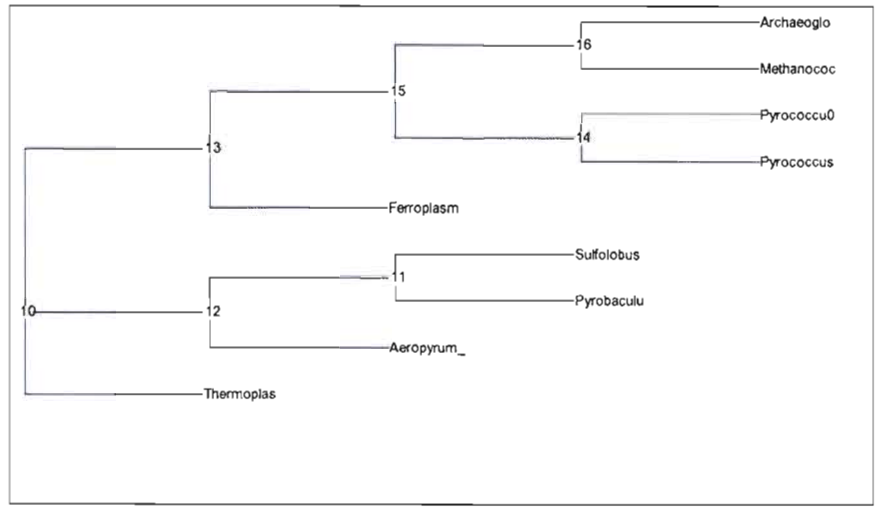


Figure 3.9 Arbre du gène

- Transferts horizontaux retrouvés pour réconcilier les deux topologies ci-dessus (l'option d'optimisation basée sur la distance topologique de Robinson et Foulds a été choisie) :

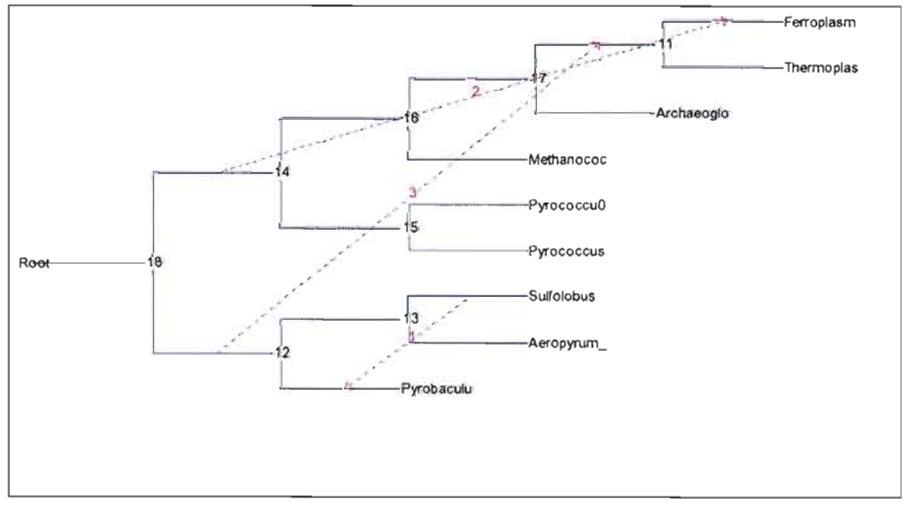


Figure 3.10 Transferts horizontaux retrouvés pour réconcilier les deux topologies

3.4 Clustal : L'alignement des séquences par ClustalW

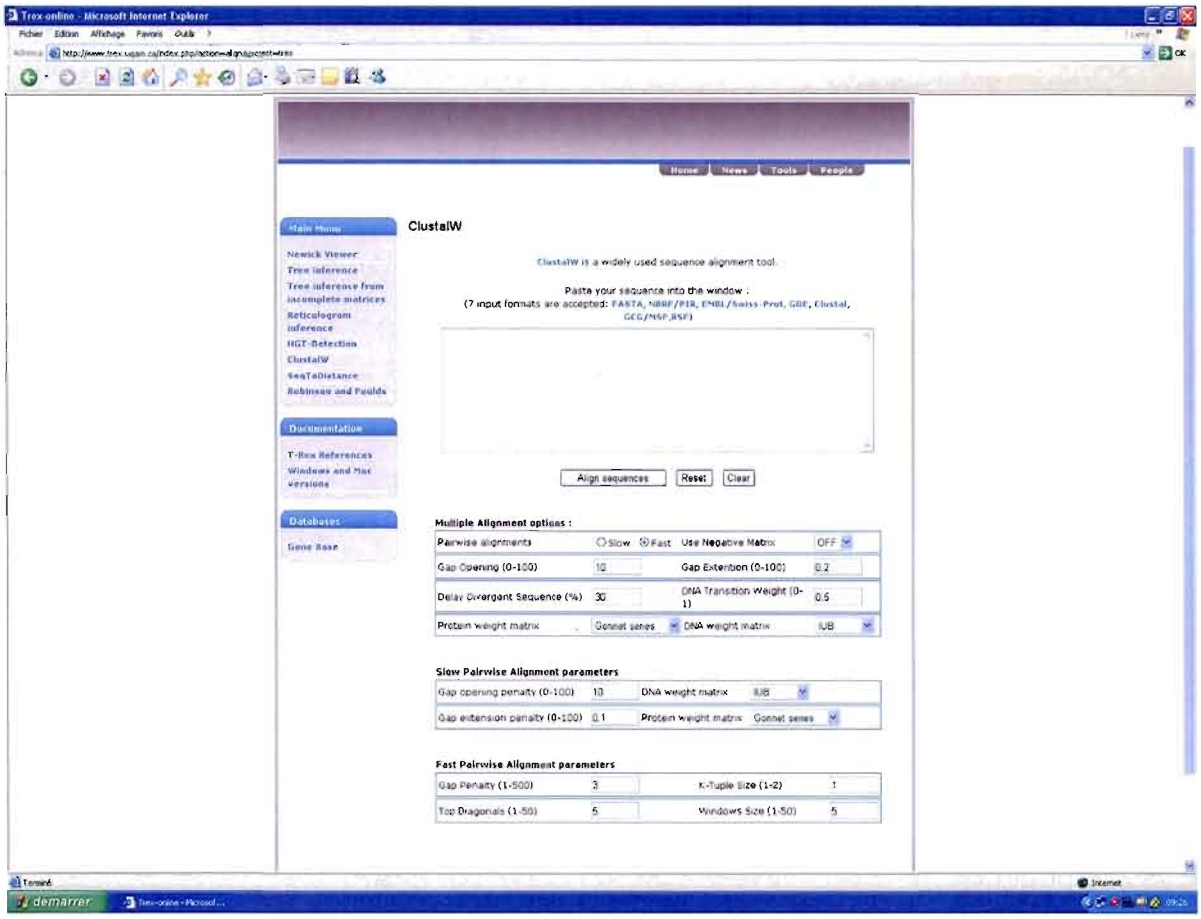


Figure 3.11 Interface du programme ClustalW

Le programme ClustalW permet d'établir l'alignement multiple des séquences nucléiques ou protéiques. Il génère également un arbre phylogénétique à partir d'un alignement donné selon la méthode NJ (Neighbour Joining). Il détermine le meilleur alignement de l'ensemble des séquences en entrée et les disposent de manière à distinguer les identités, similitudes et différences entre les bases homologues de chaque séquence.

L'alignement multiple permet de :

- Détecter des résidus identiques ou similaires pouvant jouer un rôle clé dans la fonction de la molécule ou dans sa structure tridimensionnelle.

- Caractériser de nouvelles familles de protéines.
- Détecter ou démontrer une homologie entre différentes séquences.
- Trouver un PRIMER consensus pour des PCR.
- Établir une phylogénie.
- Aider à la modélisation : les algorithmes de prédiction de structure secondaire exploitent très souvent les alignements multiples.

Le traitement est souvent très long et dépend de trois paramètres : le volume de données à traiter, la puissance de calcul des ordinateurs et les algorithmes utilisés. Pour calculer un alignement multiple, on se base sur deux grandes approches algorithmiques, l'algorithme de Neeedleman et Wunsch 1970, et les méthodes heuristiques.

- Algorithme de Neddleman et Wunsh : C'est un algorithme qui ne répond pas à tout les besoins, il est utilisé pour un petit nombre de séquences. Dans cet algorithme on recherche l'alignement multiple qui maximise la somme des scores de chaque alignement pour chaque paire. Cette technique le rend complexe, et sa complexité croit de façon exponentielle avec le nombre de séquences.
- Méthodes heuristiques : C'est la stratégie implémentée dans T-Rex. Elle est rapide et dans la plupart des cas donne de bons résultats. ClustalW aligne les séquences deux à deux et construit l'arbre des relations évolutives entre les séquences. Une fois cet arbre construit, le programme prend les deux séquences les plus proches et commence l'alignement multiple, Puis, il progresse vers les séquences plus distantes.

ClustalW supporte les séquences multiples en format :

- NBRF/PIR
- EMBL / UniProtKB/Swiss-Prot

- Pearson (Fasta)
- GDE
- ALN/ClustalW
- GCG/MSF
- RSF

Tous les caractères non alphabétiques sont ignorés, exceptés "-" ("." dans le format CG/MSF) pour désigner un gap. Minuscules et majuscules sont autorisées.

3.4.1 Étapes

Les différentes étapes d'un alignement multiple sont :

- Alignement 2 à 2 des séquences en utilisant une première série de paramètres et une méthode classique d'alignement de deux séquences.
- Élimination des séquences trop éloignées.
- Construction de groupes de séquences.
- Alignement multiple en utilisant une seconde série de paramètres et les groupes préalablement définis.

3.4.2 Paramètres

- Pairwise alignments_: C'est une option permettant l'algorithme lent ou rapide pour les alignements deux à deux.
- Use Negative Matrix : Cette option active ou désactive la permission des matrices négatives.
- Gap Opening : Pénalité pour ouvrir un 'gap', cette valeur est comprise entre 0 et 100.
- Gap Extention : Pénalité pour lorsqu'un 'gap' est étendu.

- Delay Divergent Sequence (%) : Permet de saisir le seuil au-dessus duquel l'alignement est retardé. Si la valeur est supérieur a 30%, l'alignement de la séquence est remis à plus tard.

- DNA Transition Weight : Cette option permet de saisir une valeur comprise entre 0 et 1 qui détermine le poids des transitions A <-> G, C <-> T. Plus la valeur est proche de 0 plus les séquences sont éloignées. Pour des séquences très proches elle est égale a 1.

- Protein weight matrix : Matrice de scores donnant la similarité des acides aminés par rapport aux autres. On peut choisir entre BLOSUM Series, PAM Series, Gonnet Series, et Identity Matrix (la matrice d'identité).

- DNA weight matrix : matrice de scores pour les acides nucléiques.

Le choix de l'algorithme lent nécessite les paramètres suivants :

- Gap opening penalty : Représente la pénalité sur l'introduction d'une première insertion c'est une valeur numérique comprise entre 1 et 100. Par défaut elle est égale a 10.

- Gap extension penalty : Représente la Pénalité sur l'élongation d'un gap avec une nouvelle insertion. C'est une valeur numérique comprise entre 1 et 100. Par défaut elle est égale a 0.1.

Le choix de l'algorithme rapide nécessite les paramètres suivants :

- Gap Penalty : Représente la pénalité sur l'insertion, c'est une valeur numérique comprise entre 1 et 500. Par défaut elle est égale a 3.

- Top Diagonals : Nombre de meilleures diagonales, par défaut il est égal a 5 et comprise entre 1 et 50.

- K-Tuple Size : Fenêtre autour de chaque meilleure diagonale. C'est une valeur numérique comprise entre 1 et 50. Par défaut elle est égale a 5.

- Windows Size : Taille des uplets de codification.

3.4.3 Résultats

- Arbre en format Newick.

- Appel de l'option SeqToDistance en générant automatiquement la séquence à saisir en format Phylip.

- Appel de l'option Newick Viewer en générant automatiquement la séquence en format Newick.

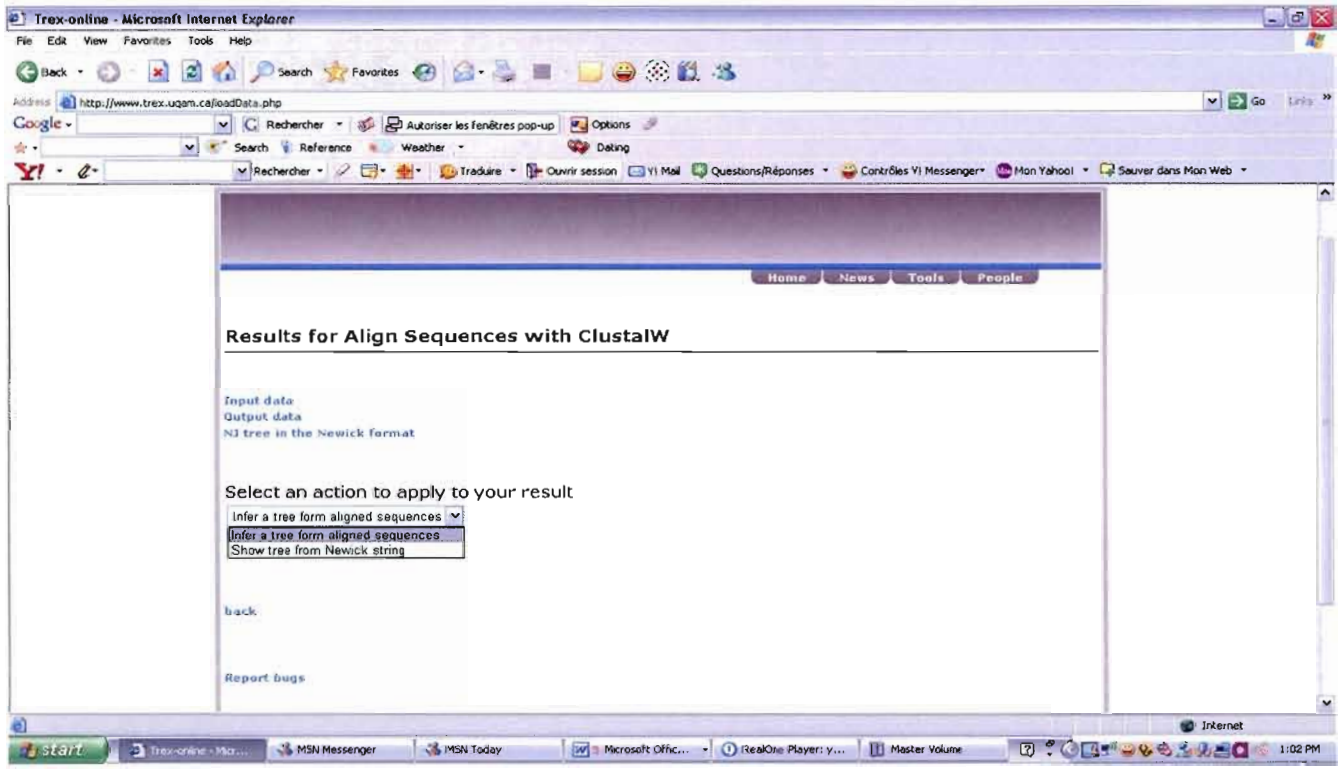


Figure 3.12 Résultats fournis par ClustalW

3.4.4 Exemple numérique

Cet exemple met en valeur l'alignement multiple de 5 séquences des mammifères représentées par les séquences protéiques de longueur 412.

L'entrée soumise a clustalW à travers notre interface :

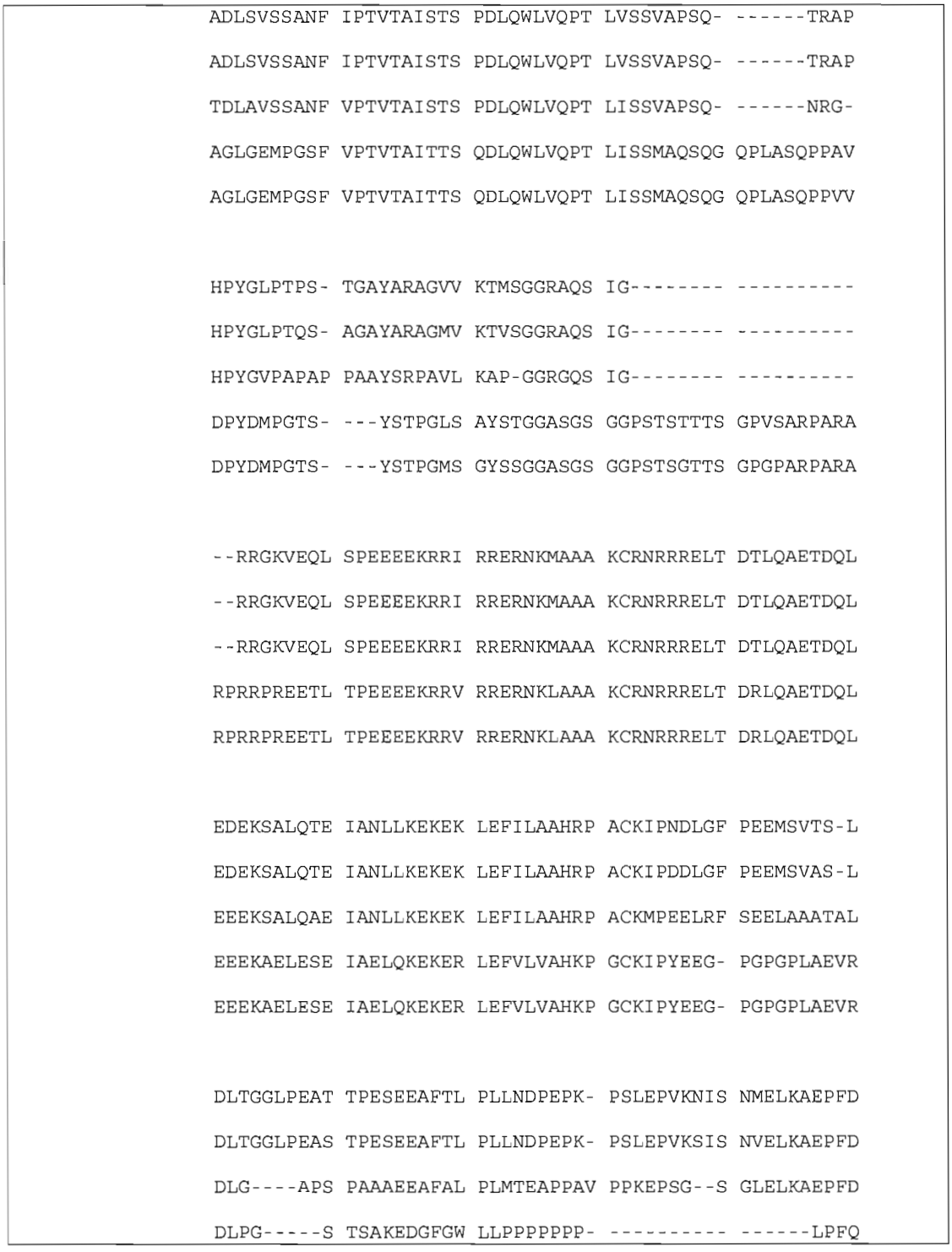
```
>FOSB_MOUSE Protein fosB
     MFQAFPGDYD SGSRCSSSPS AESQYLSSVD SFGSPPTAAA SQECAGLGEM PGSFVPTVTA
     ITTSQDLQWL VQPTLISSMA QSQGQPLASQ PPAVDPYDMP GTSYSTPGLS AYSTGGASGS
     GGPSTSTTTS GPVSARPARA RPRRPREETL TPEEEEKRRV RRERNKLAAA KCRNRRRELT
     DRLQAETDQL EEEKAELESE IAELQKEKER LEFVLVAHKP GCKIPYEEGP GPGPLAEVRD
     LPGSTSAKED GFGWLLPPPP PPPLPFQSSR DAPPNLTASL FTHSEVQVLG DPFPVVSPSY
     TSSFVLTCPE VSAFAGAQRT SGSEQPSDPL NSPSLLAL

>FOSB_HUMAN Protein fosB
     MFQAFPGDYD SGSRCSSSPS AESQYLSSVD SFGSPPTAAA SQECAGLGEM PGSFVPTVTA
     ITTSQDLQWL VQPTLISSMA QSQGQPLASQ PPVVDPYDMP GTSYSTPGMS GYSSGGASGS
     GGPSTSGTTS GPGPARPARA RPRRPREETL TPEEEEKRRV RRERNKLAAA KCRNRRRELT
     DRLQAETDQL EEEKAELESE IAELQKEKER LEFVLVAHKP GCKIPYEEGP GPGPLAEVRD
     LPGSAPAKED GFSWLLPPPP PPPLPFQTSQ DAPPNLTASL FTHSEVQVLG DPFPVVNPSY
     TSSFVLTCPE VSAFAGAQRT SGSDQPSDPL NSPSLLAL

>FOS_CHICK Proto-oncogene protein c-fos
     MMYQGFAGEY EAPSSRCSSA SPAGDSLTYY PSPADSFSSM GSPVNSQDFC TDLAVSSANF
     VPTVTAISTS PDLQWLVQPT LISSVAPSQN RGHPYGVPAP APPAAYSRPA VLKAPGGRGQ
     SIGRRGKVEQ LSPEEEEKRR IRRERNKMAA AKCRNRRREL TDTLQAETDQ LEEEKSALQA
     EIANLLKEKE KLEFILAAHR PACKMPEELR FSEELAAATA LDLGAPSPAA AEEAFALPLM
     TEAPPAVPPK EPSGSGLELK AEPFDELLFS AGPREASRSV PDMDLPGASS FYASDWEPLG
```

```
     AGSGGELEPL CTPVVTCTPC PSTYTSTFVF TYPEADAFPS CAAAHRKGSS SNEPSSDSLS
     SPTLLAL

>FOS_RAT Proto-oncogene protein c-fos
     MMFSGFNADY EASSSRCSSA SPAGDSLSYY HSPADSFSSM GSPVNTQDFC ADLSVSSANF
     IPTVTAISTS PDLQWLVQPT LVSSVAPSQT RAPHPYGLPT PSTGAYARAG VVKTMSGGRA
     QSIGRRGKVE QLSPEEEEKR RIRRERNKMA AAKCRNRRRE LTDTLQAETD QLEDEKSALQ
     TEIANLLKEK EKLEFILAAH RPACKIPNDL GFPEEMSVTS LDLTGGLPEA TTPESEEAFT
     LPLLNDPEPK PSLEPVKNIS NMELKAEPFD DFLFPASSRP SGSETARSVP DVDLSGSFYA
     ADWEPLHSSS LGMGPMVTEL EPLCTPVVTC TPSCTTYTSS FVFTYPEADS FPSCAAAHRK
     GSSSNEPSSD SLSSPTLLAL

>FOS_MOUSE Proto-oncogene protein c-fos
     MMFSGFNADY EASSSRCSSA SPAGDSLSYY HSPADSFSSM GSPVNTQDFC ADLSVSSANF
     IPTVTAISTS PDLQWLVQPT LVSSVAPSQT RAPHPYGLPT QSAGAYARAG MVKTVSGGRA
     QSIGRRGKVE QLSPEEEEKR RIRRERNKMA AAKCRNRRRE LTDTLQAETD QLEDEKSALQ
     TEIANLLKEK EKLEFILAAH RPACKIPDDL GFPEEMSVAS LDLTGGLPEA STPESEEAFT
     LPLLNDPEPK PSLEPVKSIS NVELKAEPFD DFLFPASSRP SGSETSRSVP DVDLSGSFYA
     ADWEPLHSNS LGMGPMVTEL EPLCTPVVTC TPGCTTYTSS FVFTYPEADS FPSCAAAHRK
     GSSSNEPSSD SLSSPTLLAL
```

ClustalW génère les résultats suivants :

- Les séquences alignées obtenues sont comme suit :

```
5    412
FOS_RAT     MMFSGFNADY EASSSRCSSA SPAGDSLSYY HSPADSFSSM GSPVNTQDFC
FOS_MOUSE   MMFSGFNADY EASSSRCSSA SPAGDSLSYY HSPADSFSSM GSPVNTQDFC
FOS_CHICK   MMYQGFAGEY EAPSSRCSSA SPAGDSLTYY PSPADSFSSM GSPVNSQDFC
FOSB_MOUSE -MFQAFPGDY DS-GSRCSS- SPSAESQ--Y LSSVDSFGSP PTAAASQE-C
FOSB_HUMAN -MFQAFPGDY DS-GSRCSS- SPSAESQ--Y LSSVDSFGSP PTAAASQE-C
```


ADLSVSSANF IPTVTAISTS PDLQWLVQPT LVSSVAPSQ- ------TRAP
ADLSVSSANF IPTVTAISTS PDLQWLVQPT LVSSVAPSQ- ------TRAP
TDLAVSSANF VPTVTAISTS PDLQWLVQPT LISSVAPSQ- ------NRG-
AGLGEMPGSF VPTVTAITTS QDLQWLVQPT LISSMAQSQG QPLASQPPAV
AGLGEMPGSF VPTVTAITTS QDLQWLVQPT LISSMAQSQG QPLASQPPVV
HPYGLPTPS- TGAYARAGVV KTMSGGRAQS IG-------- ----------
HPYGLPTQS- AGAYARAGMV KTVSGGRAQS IG-------- ----------
HPYGVPAPAP PAAYSRPAVL KAP-GGRGQS IG-------- ----------
DPYDMPGTS- ---YSTPGLS AYSTGGASGS GGPSTSTTTS GPVSARPARA
DPYDMPGTS- ---YSTPGMS GYSSGGASGS GGPSTSGTTS GPGPARPARA
--RRGKVEQL SPEEEEKRRI RRERNKMAAA KCRNRRRELT DTLQAETDQL
--RRGKVEQL SPEEEEKRRI RRERNKMAAA KCRNRRRELT DTLQAETDQL
--RRGKVEQL SPEEEEKRRI RRERNKMAAA KCRNRRRELT DTLQAETDQL
RPRRPREETL TPEEEEKRRV RRERNKLAAA KCRNRRRELT DRLQAETDQL
RPRRPREETL TPEEEEKRRV RRERNKLAAA KCRNRRRELT DRLQAETDQL
EDEKSALQTE IANLLKEKEK LEFILAAHRP ACKIPNDLGF PEEMSVTS-L
EDEKSALQTE IANLLKEKEK LEFILAAHRP ACKIPDDLGF PEEMSVAS-L
EEEKSALQAE IANLLKEKEK LEFILAAHRP ACKMPEELRF SEELAAATAL
EEEKAELESE IAELQKEKER LEFVLVAHKP GCKIPYEEG- PGPGPLAEVR
EEEKAELESE IAELQKEKER LEFVLVAHKP GCKIPYEEG- PGPGPLAEVR
DLTGGLPEAT TPESEEAFTL PLLNDPEPK- PSLEPVKNIS NMELKAEPFD
DLTGGLPEAS TPESEEAFTL PLLNDPEPK- PSLEPVKSIS NVELKAEPFD
DLG----APS PAAAEEAFAL PLMTEAPPAV PPKEPSG--S GLELKAEPFD
DLPG-----S TSAKEDGFGW LLPPPPPPP- ---------- ------LPFQ

```
DLPG-----S APAKEDGFSW LLPPPPPPP- ---------- ------LPFQ


DFLFPASSRP SGSETARSVP DVDLSG--SF YAADWEPLHS SSLGMGPMVT

DFLFPASSRP SGSETSRSVP DVDLSG--SF YAADWEPLHS NSLGMGPMVT

ELLFSAGPR- ---EASRSVP DMDLPGASSF YASDWEPLGA GSGG------

---------- ----SSRDAP -PNLTA--SL FTHS------ ----------

---------- ----TSQDAP -PNLTA--SL FTHS------ ----------


ELEPLCTPVV TCTPSCTTYT SSFVFTYPEA DSFPSCAAAH RKGSSSNEPS

ELEPLCTPVV TCTPGCTTYT SSFVFTYPEA DSFPSCAAAH RKGSSSNEPS

ELEPLCTPVV TCTPCPSTYT STFVFTYPEA DAFPSCAAAH RKGSSSNEPS

EVQVLGDPFP VVSP---SYT SSFVLTCPEV SAF---AGAQ R--TSGSEQP

EVQVLGDPFP VVNP---SYT SSFVLTCPEV SAF---AGAQ R--TSGSDQP


SDSLSSPTLL AL

SDSLSSPTLL AL

SDSLSSPTLL AL

SDPLNSPSLL AL

SDPLNSPSLL AL
```

- L'arbre phylogénétique correspondant a cet alignement est comme suit :

```
(((FOSB_MOUSE:0.01627,FOSB_HUMAN:0.02515):0.46708,FOS_CHICK:0.13
943):0.13769,FOS_RAT:0.01832,FOS_MOUSE:0.01326);
```

- Après avoir généré l'arbre en format Newick, ClustalW appelle le programme Newick Viewer en passant comme paramètre la séquence en format Newick. Ceci permet d'afficher l'arbre suivant :

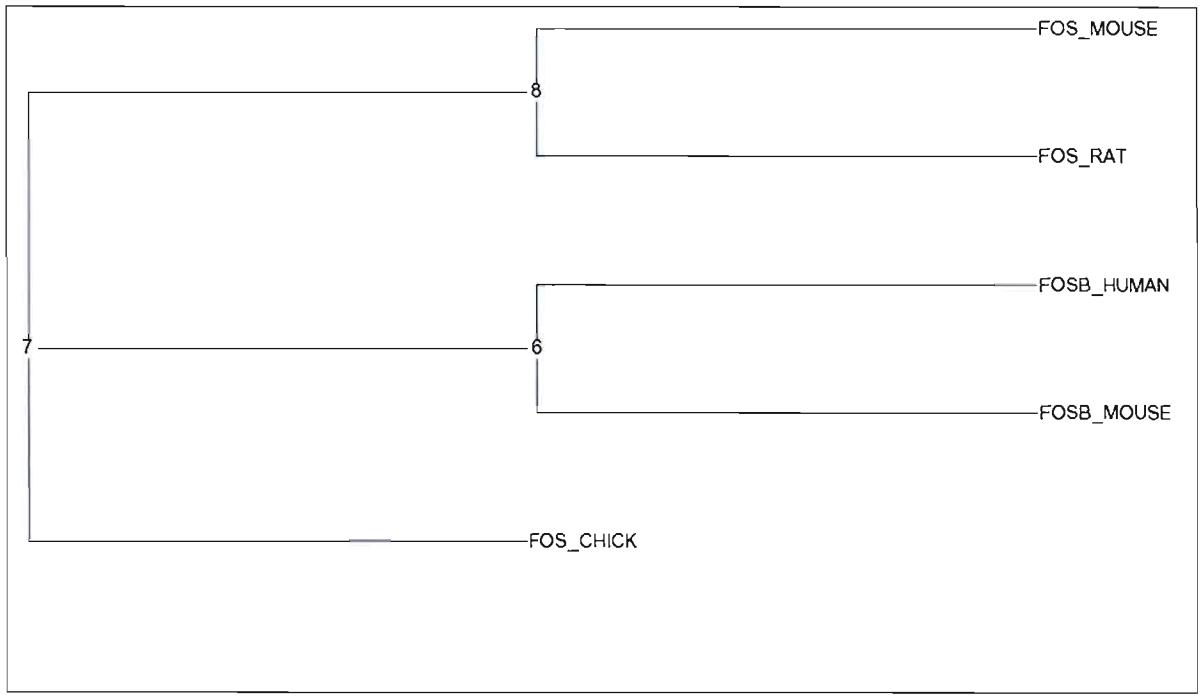


Figure 3.13 Arbre généré par ClustalW

- ClustalW permet de générer la matrice de distance suivante :

5

FOS_RAT	0.000000	0.032116	0.247949	0.637750	0.644083
FOS_MOUSE	0.032116	0.000000	0.244369	0.619000	0.619000
FOS_CHICK	0.247949	0.244369	0.000000	0.618437	0.630970
FOSB_MOUSE	0.637750	0.619000	0.618437	0.000000	0.042350
FOSB_HUMAN	0.644083	0.619000	0.630970	0.042350	0.000000

Tableau 3.3 Matrice de distances générée par ClustalW

- ClustalW permet d'inférer l'arbre directement à partir de la matrice ci-dessus.

3.5 Calcul de la distance topologique Robinson et Foulds

Ce programme permet de calculer, dans un temps optimal, la distance topologique de Robinson et de Foulds entre deux arbres phylogénétiques ou plus à partir de leur matrices de distance. Le programme est base sur l'algorithme proposé par Makarenkov et Leclerc (1999 b) pour le calcul de cette distance. L'algorithme implémenté dans ce programme emploie la notion des ordres circulaires pour comparer la topologie de deux arbres. L'algorithme décrit par Makarenkov et Leclerc (1999 b) a la complexité optimale, nécessitant un temps de $O(n^2)$ applique sur deux matrices de distances de taille *(nxn)*.

La distance topologique de Robinson et de Foulds est un outil important et fréquemment utilisé pour comparer les structures des arbres phylogénétiques, pour plus de détails voir Robinson et le Foulds 1981, ou Makarenkov et Leclerc (1999). Cette distance est égale au nombre minimum d'opérations élémentaires, telles que la fusion ou le dédoublement des noeuds, nécessaires pour transformer un arbre en autre. Dans l'article de Robinson et Foulds (1981), il a été prouvé que cette distance est également le nombre de bipartitions, ou des splits de Buneman (1971), qui appartiennent seulement à un des deux arbres.

Si on traite deux arbres non enracinés n'ayant aucun sommet interne marqué selon les éléments d'un l'ensemble X, la distance de Robinson et de Foulds de l'ensemble X a n éléments varie de 0 (quand les arbres sont isomorphes) à $2n$-6 (quand toutes les bipartitions non triviaux dans deux arbres sont différentes).

Exemple :

Cet exemple permet de calculer, la distance topologique de Robinson et de Foulds entre trois arbres phylogénétiques à partir de leur matrice de distance :

```
10
Ferroplasm 0.000000 2.000000 7.000000 6.000000 7.000000 7.000000 7.000000 6.000000 5.000000 4.000000
Thermoplas 2.000000 0.000000 7.000000 6.000000 7.000000 7.000000 7.000000 6.000000 5.000000 4.000000
Aeropyrum_ 7.000000 7.000000 0.000000 3.000000 2.000000 6.000000 6.000000 5.000000 6.000000 7.000000
Pyrobaculu 6.000000 6.000000 3.000000 0.000000 3.000000 5.000000 5.000000 4.000000 5.000000 6.000000
```

Sulfolobus 7.000000 7.000000 2.000000 3.000000 0.000000 6.000000 6.000000 5.000000 6.000000 7.000000

Pyrococcus 7.000000 7.000000 6.000000 5.000000 6.000000 0.000000 2.000000 3.000000 6.000000 7.000000

Pyrococcu0 7.000000 7.000000 6.000000 5.000000 6.000000 2.000000 0.000000 3.000000 6.000000 7.000000

Pyrococcu1 6.000000 6.000000 5.000000 4.000000 5.000000 3.000000 3.000000 0.000000 5.000000 6.000000

Methanococ 5.000000 5.000000 6.000000 5.000000 6.000000 6.000000 6.000000 5.000000 0.000000 5.000000

Archaeoglo 4.000000 4.000000 7.000000 6.000000 7.000000 7.000000 7.000000 6.000000 5.000000 0.000000

10

Ferroplasm 0.000000 2.800000 4.700000 3.600000 3.400000 2.600010 3.170000 3.100010 3.350000 3.400000

Thermoplas 2.800000 0.000000 3.700000 2.600000 2.400000 2.800010 3.370000 3.300010 3.550000 3.600000

Aeropyrum_ 4.700000 3.700000 0.000000 4.100000 3.900000 4.700010 5.270000 5.200010 5.450000 5.500000

Pyrobaculu 3.600000 2.600000 4.100000 0.000000 1.700000 3.600010 4.170000 4.100010 4.350000 4.400000

Sulfolobus 3.400000 2.400000 3.900000 1.700000 0.000000 3.400010 3.970000 3.900010 4.150000 4.200000

Pyrococcus 2.600010 2.800010 4.700010 3.600010 3.400010 0.000000 0.570010 0.500020 1.550010 1.600010

Pyrococcu0 3.170000 3.370000 5.270000 4.170000 3.970000 0.570010 0.000000 0.070010 2.120000 2.170000

Pyrococcu1 3.100010 3.300010 5.200010 4.100010 3.900010 0.500020 0.070010 0.000000 2.050010 2.100010

Methanococ 3.350000 3.550000 5.450000 4.350000 4.150000 1.550010 2.120000 2.050010 0.000000 1.750000

Archaeoglo 3.400000 3.600000 5.500000 4.400000 4.200000 1.600010 2.170000 2.100010 1.750000 0.000000

10

Pyrococcu1 0.000000 2.800000 4.700000 3.600000 3.400000 2.600010 3.170000 3.100010 3.350000 3.400000

Thermoplas 2.800000 0.000000 3.700000 2.600000 2.400000 2.800010 3.370000 3.300010 3.550000 3.600000

Aeropyrum_ 4.700000 3.700000 0.000000 4.100000 3.900000 4.700010 5.270000 5.200010 5.450000 5.500000

Pyrobaculu 3.600000 2.600000 4.100000 0.000000 1.700000 3.600010 4.170000 4.100010 4.350000 4.400000

Archaeoglo 3.400000 2.400000 3.900000 1.700000 0.000000 3.400010 3.970000 3.900010 4.150000 4.200000

Pyrococcus 2.600010 2.800010 4.700010 3.600010 3.400010 0.000000 0.570010 0.500020 1.550010 1.600010

Pyrococcu0 3.170000 3.370000 5.270000 4.170000 3.970000 0.570010 0.000000 0.070010 2.120000 2.170000

```
Ferroplasm 3.100010 3.300010 5.200010 4.100010 3.900010 0.500020 0.070010 0.000000 2.050010 2.100010

Methanococ 3.350000 3.550000 5.450000 4.350000 4.150000 1.550010 2.120000 2.050010 0.000000 1.750000

Sulfolobus 3.400000 3.600000 5.500000 4.400000 4.200000 1.600010 2.170000 2.100010 1.750000 0.000000
```

Le résultat obtenu est le suivant :

```
* ------------------------------------------------*
* Computation of the Robinson and Foulds          *
* topological distance between two (or more) trees. *
* ------------------------------------------------ *

RF Distance between Tree 1 and Tree 2 = 10
RF Distance between Tree 1 and Tree 3 = 14
```

CHAPITRE IV

Taxonomie d'espèces

4.1 Introduction

Species Taxonomy est une nouvelle option dans T-Rex qui permet de générer une matrice de distances additive et de reconstruire des arbres phylogénétiques à partir de lignées d'espèces. Nous avons développé deux algorithmes pour la construction de matrices de distances d'arbre à partir de la liste des lignées. La programmation de l'interface graphique se faisait en langage PHP (version 5.0).

4.2 Données d'entrée

Les données d'entrée pour le programme Species Taxonomy proviennent d'un fichier texte nommé NCBI.TXT. Chaque ligne de ce fichier débute par l'ancêtre le plus vieux suivi de ses descendants directs. Chaque espèce est séparée de son fils par un point virgule. Un point marque la fin de chaque ligne. Pour avoir la lignée d'une espèce donnée, il faut parcourir le fichier NCBI.TXT en lisant ses lignes de droite a gauche. La taille du fichier NCBI.TXT est importante, le nombre de lignes est environ 180 000 lignes.

cellular organisms ; Archaea; Crenarchaeota; Thermoprotei; Caldisphaerales; Caldisphaeraceae; Caldisphaera; Caldisphaera lagunensis.

cellular organisms ; Archaea; Crenarchaeota; Thermoprotei; Cenarchaeales; Cenarchaeaceae; Cenarchaeum; Cenarchaeum symbiosum.

cellular organisms ; Archaea; Crenarchaeota; Thermoprotei; Desulfurococcales; Desulfurococcaceae; Acidilobus; Acidilobus aceticus.

cellular organisms ; Archaea; Crenarchaeota; Thermoprotei; Desulfurococcales; Desulfurococcaceae; Acidilobus; Acidilobus sp. 124-87.

cellular organisms ; Archaea; Crenarchaeota; Thermoprotei; Desulfurococcales; Desulfurococcaceae; Acidilobus; Acidilobus sp. 345-15.
cellular organisms ; Archaea; Crenarchaeota; Thermoprotei; Desulfurococcales; Desulfurococcaceae; Acidilobus; Acidilobus sp. 405-16.
cellular organisms ; Archaea; Crenarchaeota; Thermoprotei; Desulfurococcales; Desulfurococcaceae; Acidilobus; Acidilobus sp. 722-67.

Figure 4.1 Exemple de lignées provenant du fichier NCBI.TXT

4.3 Traitement

Le traitement se fait en plusieurs étapes :

4.3.1 Formulaire de saisie des espèces recherchées et création d'un fichier des lignées des espèces données.

Ce formulaire permet d'interagir avec l'usager en temps réel. Il est composé des variables dynamiques de saisie et des boutons d'action pour traiter le formulaire.

La boite de saisie permet de saisir une liste d'espèces à aligner, le bouton « Clear » permet d'initialiser la boite de saisie, et le bouton « Submit » permet de lancer le traitement.

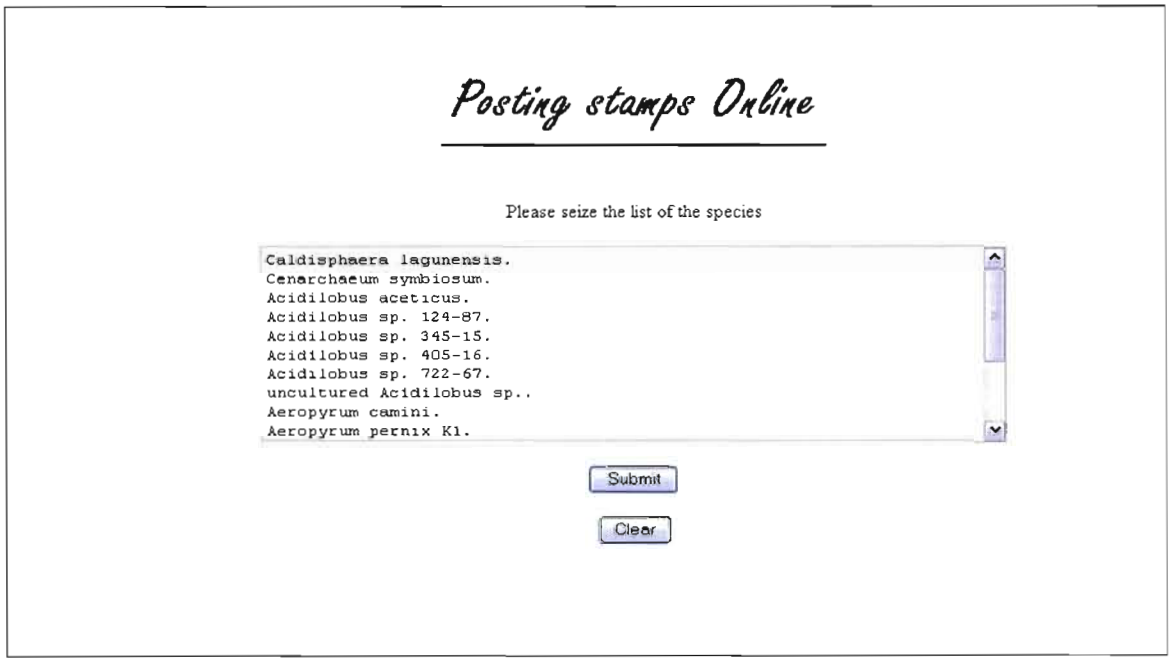


Figure 4.2 Formulaire de saisie des espèces

4.3.2 Création des lignées

Le traitement consiste à créer un fichier nommé Especе_ligneage à partir du fichier NCBI.TXT et de la liste d'espèces saisies par l'utilisateur. Chaque ligne du fichier Especе_lineage, est structurée comme suit :

- Le nom de l'espèce tel que saisi par l'utilisateur.
- Un caractère « @ », pour séparer l'espèce saisie du reste de la ligne.
- Le reste de la ligne se compose des ancêtres de l'espèce saisie, séparés par des « ; ».

Pour réaliser ce traitement plusieurs algorithmes sont possibles.

a) Premier algorithme :

C'est un algorithme naïf. Il consiste à :

- parcourir la liste des espèces saisies stockées dans un tableau E. Pour stocker les espèces saisies dans tableau, il existe une fonction split() du langage PHP qui retourne un tableau de chaînes : chacune d'entre elle est une sous-chaîne d'une chaîne délimitée par les occurrences trouvées de l'expression régulière recherchée. Si une erreur survient, la fonction retourne la valeur FALSE. Dans notre cas le séparateur est le retour de chariot.
- Pour chaque élément du tableau, on parcourt en lecture le fichier NCBI.TXT. On stocke chaque ligne lut dans un tableau nommé Fe. Puis, on compare le dernier élément du tableau Fe à l'espèce recherchée. Si le dernier élément du tableau Fe est identique au nom de l'espèce, on écrit dans le fichier lineage-especе.txt le dernier élément de Fe suivi du caractère « @ », suivit du reste des éléments du tableau Fe dans un ordre décroissant. Des points virgules jouent le rôle des séparateurs.

L'algorithme se présente comme suit :

Algorithme 1 : extraction des lignées des espèces saisies.

```
Matrice_distance (Tableau e)
Lexique
    - i : entier
    - emax : entier
    - E : tableau contenant les espèces saisies
    - Fe : Tableau contenant les parents de l'espèce saisie
    - Ligne_NCBI : Ligne dans le fichier NCBI
    - Ligne_lineage : Ligne dans le fichier NCBI
    - Ligne : caractère de taille 256.
Début
    Ouvrir en ajout Lineage_espece.Txt
    Pour i=0 à (longueur-1) faire
    Ouvrir en lecture NCBI.TXT
    Tant que non fin de fichier NCBI.TXT
        Fe ← split(";",ligne_ncbi)
        emax = count(Fe) // retourne le nombre d'élément dans Fe.
        Si E(i) = Fe(emax-1) // le dernier élément est l'espèce recherchée
            Alors
            Ligne_lineage ←(Fe(1)+ «@ » +

            substr(lig,1,strlen(lig)-strlen(Fe[count(Fe)-1])))
            // substr(lig,1,strlen(lig)-strlen(fe[count(Fe)-1]))) sont les ancêtres de
            l'espèce  trouvée
```

```
                Fin Si
        Fin Tant que
        Fermer NCBI.TXT
        Fin Pour
        Fermer Lineage_espece.Txt
Fin
```

Cet algorithme est efficace pour un nombre réduit d'espèces à aligner ou pour des espèces qui se trouvent sur les premières centaines d'enregistrements du fichier NCBI.TXT. Si on choisit un grand nombre d'espèces à aligner et si ces dernières se trouvent proches de la fin du fichier NCBI.TXT le programme devient lent et inefficace.

Exemple 1

Liste des espèces à aligner :

```
Caldisphaera lagunensis.
Cenarchaeum symbiosum.
Acidilobus aceticus.
Acidilobus sp. 124-87.
Acidilobus sp. 345-15.
Acidilobus sp. 405-16.
Acidilobus sp. 722-67.
uncultured Acidilobus sp..
Aeropyrum camini.
```

Aeropyrum pernix K1.

Desulfurococcus amylolyticus.

Desulfurococcus fermentans.

Desulfurococcus mobilis.

Desulfurococcus mucosus.

Desulfurococcus saccharovorans.

Le fichier Lineage_escpece.txt généré est le suivant :

Caldisphaera lagunensis.@cellular organisms ; Archaea; Crenarchaeota; Thermoprotei; Caldisphaerales; Caldisphaeraceae; Caldisphaera; Caldisphaera lagunensis.

Cenarchaeum symbiosum.@cellular organisms ; Archaea; Crenarchaeota; Thermoprotei; Cenarchaeales; Cenarchaeaceae; Cenarchaeum; Cenarchaeum symbiosum.

Acidilobus aceticus.@cellular organisms ; Archaea; Crenarchaeota; Thermoprotei; Desulfurococcales; Desulfurococcaceae; Acidilobus; Acidilobus aceticus.

Acidilobus sp. 124-87.@cellular organisms ; Archaea; Crenarchaeota; Thermoprotei; Desulfurococcales; Desulfurococcaceae; Acidilobus; Acidilobus sp. 124-87.

Acidilobus sp. 345-15.@cellular organisms ; Archaea; Crenarchaeota; Thermoprotei; Desulfurococcales; Desulfurococcaceae; Acidilobus; Acidilobus sp. 345-15.

Acidilobus sp. 405-16.@cellular organisms ; Archaea; Crenarchaeota; Thermoprotei; Desulfurococcales; Desulfurococcaceae; Acidilobus; Acidilobus sp. 405-16.

Acidilobus sp. 722-67.@cellular organisms ; Archaea; Crenarchaeota; Thermoprotei; Desulfurococcales; Desulfurococcaceae; Acidilobus; Acidilobus sp. 722-67.

uncultured Acidilobus sp..@cellular organisms ; Archaea; Crenarchaeota; Thermoprotei; Desulfurococcales; Desulfurococcaceae; Acidilobus; environmental samples; uncultured Acidilobus sp..

Aeropyrum camini.@cellular organisms ; Archaea; Crenarchaeota; Thermoprotei; Desulfurococcales; Desulfurococcaceae; Aeropyrum; Aeropyrum camini.

Aeropyrum pernix K1.@cellular organisms ; Archaea; Crenarchaeota; Thermoprotei; Desulfurococcales; Desulfurococcaceae; Aeropyrum; Aeropyrum pernix; Aeropyrum pernix K1.

Desulfurococcus amylolyticus.@cellular organisms ; Archaea; Crenarchaeota; Thermoprotei; Desulfurococcales; Desulfurococcaceae; Desulfurococcus; Desulfurococcus amylolyticus.

Desulfurococcus fermentans.@cellular organisms ; Archaea; Crenarchaeota; Thermoprotei; Desulfurococcales; Desulfurococcaceae; Desulfurococcus; Desulfurococcus fermentans.

Desulfurococcus mobilis.@cellular organisms ; Archaea; Crenarchaeota; Thermoprotei; Desulfurococcales; Desulfurococcaceae; Desulfurococcus; Desulfurococcus mobilis.

Desulfurococcus mucosus.@cellular organisms ; Archaea; Crenarchaeota; Thermoprotei; Desulfurococcales; Desulfurococcaceae; Desulfurococcus; Desulfurococcus mucosus.

Desulfurococcus saccharovorans.@cellular organisms ; Archaea; Crenarchaeota; Thermoprotei; Desulfurococcales; Desulfurococcaceae; Desulfurococcus; Desulfurococcus saccharovorans.

Deuxième algorithme :

Le deuxième algorithme parcourt en premier le fichier NCBI.TXT. Pour chaque lignée on extrait sa dernière chaîne, on compare cette chaîne à tous les éléments du tableau E contenant toutes les espèces saisies par l'utilisateur. Si cette chaîne est égale à l'un des éléments du tableau, alors on écrit dans le fichier Lineage_espece.TXT l'espèce suivi du caractère « @ » et du reste de ses ancêtre.

Pour éviter de parcourir tout le fichier NCBI.TXT nous faisons comme suit : si le nombre d'espèces saisies correspond au nombre de lignées ajoutées dans le fichier Lineage_espece.txt le traitement est arrêté.

Matrice_distance (Tableau e)

Lexique

- i : entier
- E : tableau contenant les espèces saisies
- Fe : Tableau contenant les parents de l'espèce saisie
- Ligne_NCBI : Ligne dans le fichier NCBI
- Ligne_lineage : Ligne dans le fichier NCBI
- Ligne : caractère de taille 256.
- substr (string string, int start [, int length]) : est une fonction PHP qui retourne le segment d'une chaîne définie par *start* et *length*. Si *start* est positif, la chaîne commencera au caractère numéro *start*, dans la chaîne *string*.
- Strpos() retourne la position numérique de la première occurrence de needle dans la chaîne de caractères haystack .

Début

Ouvrir en ajout Lineage_espece.Txt

```
Pour i=0 à (longueur-1) faire
        Ouvrir en lecture NCBI.TXT
                Tant que non fin de fichier NCBI.TXT
                        Fe ← split(";",ligne_ncbi)
                        Si E(i) = Fe(1) // le premier élément est l'espèce recherchée
                        alors
                        Ligne_lineage ←(Fe(1)+ «@ » +
                        substr(lig,1,strlen(lig)-strlen(fe[count(Fe)-1])))
                        // substr(lig,1,strlen(lig)-strlen(Fe[count(Fe)-1]))) est les
                        ancêtres de l'espèce trouvée
                        Fin Si
                Fin Tant que
        Fermer NCBI.TXT
Fin Pour
Fermer Lineage_espece.Txt
Fin
```

Cet algorithme est plus efficace que le précédent. On peut encore améliorer ce traitement en passant par les technologies des bases de données. Les données stockées dans le fichier texte NCBI.TXT peuvent être également stockées dans une table indexée par le nom de l'espèce.

Cette table sera composée d'un champ *espèce* et d'un champ *père*. Pour aligner une espèce il suffit d'accéder à la table avec le nom de l'espèce, récupérer le champ nommé *père* le stoker dans une variable, écrire sur le fichier lineage_espece.txt le nom de l'espèce, suivi

d'un point virgule et de son père. Ensuite on continue l'opération jusqu'à ce que on accède avec la variable et qu'aucun enregistrement ne corresponde à notre recherche.

Le nombre total d'enregistrement est au plus $2N$-3, où N est le nombre de toutes les espèces dans NCBI. Pour ajouter une nouvelle lignée, il suffit d'ajouter l'espèce comme clé de l'enregistrement et son père comme deuxième champ.

4.3.3 Matrice de distances

Ce programme permet de créer un fichier nommé matrice_distance.txt à partir du fichier lineage_especе.txt. Le traitement consiste à calculer la distance dans un arbre phylogénétique hypothétique entre chaque paire d'espèces saisies par l'utilisateur. Pour une liste de n espèces, le résultat sera une matrice carrée *(nxn)* avec des 0 sur la diagonale parce que la distance entre une espèce et elle-même est 0. La matrice résultat est une matrice symétrique parce que la distance entre une espèce A et une espèce B est égale à la distance entre B et A.

La première colonne de la matrice est composée des espèces saisies. La longueur de chaque espèce ne doit pas dépasser 20 caractères. Au début du traitement, on remplace toutes les espèces ayant une longueur qui dépasse 20 caractères par le même nom tronqué au 20-ème caractère. Certaines espèces peuvent avoir les 20 premiers caractères similaires. Dans ce cas, on ajoute le caractère « _ » suivie d'un numéro séquentiel et on génère un fichier lexique permettant d'interpréter tous les noms d'espèces.

Nous avons crée une fonction nommée occurrence_especе qui permet de calculer le nombre d'espèces ayant les mêmes premiers 20 caractères identiques a une espèce a donnée. Cette fonction renvoie une valeur entière x. Nous utilisons également les deux fonctions PHP suivantes :

Strpos(*ligne*,"@") : renvoie la position du caractère « @ » dans une lignée dans le fichier lineage_texte.txt.

substr(*ligne*,0,strpos(*ligne*,"@")) : permet d'extraire une chaîne de caractères d'une autre chaîne. Dans notre cas, elle permet d'extraire de la chaîne *ligne*, une chaîne de caractères du début de la ligne jusqu'à la position ou se trouve le caractère « @ », qui correspond à l'espèce donnée.

L'algorithme se présente comme suit :

```
Occurrence_espece (a,chaîne de caractères)
Lexique
        - a : caractère // espèce
        - x : entier     // nombre d'occurrences
        - ligne : caractère // une lignée dans le fichier lineage_espece.txt
Début
  Ouvrir (en lecture) Lineage_espece.txt
  Tant que la fin de fichier Lineage_espece.txt n'est pas atteinte
        a ← substr(ligne,0,strpos(ligne,"@"))
        Si (substr(a,0,20) == substr(espece,0,20)) Alors
              x ← x+ 1
        Fin Si
  Fin Tant que
  Fermer Lineage_espece.txt
Fin
```

Nous avons créé une fonction nommé distance(). Cette fonction permet de calculer pour deux espèces données *a* et *b*, la distance entre l'espèce *a* et le premier ancêtre commun

de *a* et de *b*. Étant donné deux espèces *a* et *b* ayant comme lignées *aa* et *bb* respectivement. Le traitement consiste à stocker la lignée de *a* dans un tableau, et la lignée de *b* dans un autre, de parcourir le premier tableau, et puis pour chaque élément de parcourir le deuxième jusqu'à ce qu'on trouve l'ancêtre commun.

La distance entre *a* et *b* sera égale à : distance(a,b)+distance(b,a).

L'algorithme de la fonction distance se présente donc comme suit :

```
Distance (aa : chaîne de caractères ; bb : chaîne de caractères)
Lexique
        a : tableau contenant la lignée de la premier espèce
        b : tableau contenant la lignée de la deuxième espèce
        i ← 0 : entier
        j ←0  : entier
        d ← 0 : entier
        racine ←  0 : entier
        imax ←  0 : entier //
Début
        a ← split(";",aa) // tableau contenant tout les ancêtres de la première espèce
        b ← split(";",bb) // tableau contenant tout les ancêtres de la deuxième espèce
        imax ← count(a) // Le nombre d'éléments dans le tableau a
        i← imax-1 // i = l'indice du dernier élément dans le tableau a
        j ← 0
  Tant que i >= 0
        J←0
        Tant que (a[i] <>  b[j] et j < count(b)-1)  // j < indice du dernier élément
```

```
                J← j+1
                d←d+1
            Fin Tant que
        Fin Tant que
        Retouner d
    Fin
```

On crée un fichier nommé matrice_distance.txt dans lequel on stocke la matrice générée. On copie le fichier lineage_espece.txt dans un autre fichier, on parcourt le fichier lineage_especе.txt et on stocke la chaîne de caractères à gauche du caractère « @ » dans une variable *a* et le reste de la lignée dans une variable *aa*. *a* est l'espèce et *aa* sa lignée.

Pour chaque ligne du fichier linage_espece.txt on écrit l'espèce dans le fichier distance_distance.txt, on stocke le nom de l'espèce dans une variable, on parcourt le fichier et on copie du fichier lineage_espece.txt. On calcule par la suite la distance entre la variable contenant l'espèce et chaque espèce se trouvant dans le fichier copie.

À la fin, on écrit dans le fichier Lexique.txt toutes les espèces dont les noms ont été modifiés.

L'algorithme se présente comme suit :

```
Matrice_distance
Lexique
        Ligne caractère // Lignée du fichier lineage_espece.txt
        a caractère // la variable a contient l'espèce
```

```
old_a caractère // la variable contient une copie de l'espèce avant de la
        tronquer
aa caractère // la variable aa contient la lignée de a
Ligne2 caractère // Ligne du fichier copie de lineage_espece.txt
b caractère // la variable a contient l'espèce
bb caractère // la variable aa contient la lignée de a
abrev caractère // variable contenant le nom de l'espèce tronquée
pref entier
s entier
Lexx caractère
str_replace() retourne une chaîne ou un tableau dont toutes les occurrences
de search dans subject ont été remplacées par replace.
Début
 Ouvrire en lecture lineage_espece.txt
 Ouvrire en écriture matrice_distance.txt
 Ouvrire en lecture copie de lineage_espece.txt
 Tant que non fin de fichier lineage_espece.txt
                a ← substr(ligne,0,strpos(ligne,"@"))
                aa ← substr(ligne,strpos(ligne,"@")+1,strlen(ligne))
                old_a ←a
                a ← str_replace(" ","_",a)
                Si strlen(old_a) > 20 Alors
                        Si ( occurence_espece(old_a) > 1 ) Alors
                                pref ← pref+1
```

```
                    v ← ""
                    Si (pref < 10 ) Alors
                        v ← "00" . pref
                    Fin Si
                    If (pref >9 et pref <100 ) Alors
                        v ← "0" . pref
                    Fin Si
                abrev ← abrev . substr(a,0,16) . "_" . v . " = " . a . " ; "
                a ← substr(a,0,16) . "_" . v ;
            Si ( occurence_espece(old_a) < 1 ) Alors
                abrev ← abrev . substr(a,0,20) . " = " . a . " ; "
                a ← substr(a,0,20)
            Fin Si
        Fin Si
        Si strlen(old_a) < 20 Alors
            Tant que strlen(a) < 20
                a ← a . " "
            Fin Tant que
        Fin Si
    ecrire a dans le fichier matrice_distance.txt
    ecrire « ; » dans le fichier matrice_distance.txt
    aller début de la copie du fichier lineage_distance.txt
    Tant que non fin du fichier copie lineage_espece.txt
```

```
                    b ← substr(ligne2,0,strpos(ligne2,"@")) ;
                    bb ← substr(ligne2,strpos(ligne2,"@")+1,strlen(ligne2))
                    Si a <> b Alors
                            d ← distance(aa,bb)+distance(bb,aa);
                        Sinon
                            d←0
                    Fin Si
            Fin Tant que
        Fin Tant que
        Fermer Lineage_espece.txt
        Fermer copie de Lineage_espece.txt
        Fermer matrice_distance.txt
        // Ecriture dans le fichier lexique.txt
        Créer fichier Lexique.txt
        lexx ← split(";" , abrev)
        s ← 0
        Tant que lexx[$s] <> "" et  s < count(lexx)-1)
            Ecrire (lex,trim(lexx[s]))
            Ecrire(retour chariot);
            s ← s + 1 ;
        Fin Tant que
        Fermer fichier Lexique.txt
    Fin
```

4.3.4 Arêtes

Pour tracer un arbre phylogénétique on a besoin de savoir toutes ses arêtes. Le programme *Arrête* permet de créer un fichier texte nomme aretes.txt qui contient toutes les arêtes d'un arbre phylogénétique correspondant a une liste de lignées données. Une arrête est représentée par les deux espèces entre deux parenthèses séparées par une virgule. Les arêtes sont séparées entre elles par des points virgules.

L'algorithme se présente comme suit :

```
Arêtes
Lexique
        a caractère (espèce)
        aa caractère (lignée de a)
        sa tableau (les ancêtres de a)
        i entier (indice du tableau sa)
        arete caractère (chaîne de carctere contenant les arêtes)
        ta caractère (chaîne de carctere contenant les arêtes sans éléments doubles)
        array_unique : fonction PHP qui prend un tableau et retourne un autre
        tableau complètement dédoublé.
Début
        Ouvrir en ecriture aretes.txt
        Ouvrir en lecture lineage_espece.txt
        Tant que non fichier lineage_espece.txt
                a← substr(ligne,0,strpos(ligne,"@"))
                aa ← substr(ligne,strpos(ligne,"@")+1,strlen(ligne))
                sa ← split(";",rtrim(aa));
```

```
                    i ← count($sa)-1;
                    arete ← arete . "( " .a . " , " . sa[i] . " ) ; "
                    Tant que i>o
                            arete ← arete . "( ". sa[i] . " , " . sa[i-1] . " ) ; "
                            i ← i - 1
                    Fin Tant que
            Fin Tant que
            ta ← array_unique(split(";",arete)) // enlever les éléments doubles
            i ← 0
            arete ← ""
            Tant que (i < count(ta)-1)
                            arete ← arete . ta[i] . " ; "
                            i ← i + 1
            Fin Tant que
            Ecrire arete dans le fichier arete.txt
            Fermer lineage_espece.txt
            Fermer aretes.txt
    Fin
```

4.3.5 Affichage des résultats

Une interface est développée pour permettre aux utilisateurs d'afficher les résultats suivants :

- distance matrix (matrice de distance)
- detail stamps (permet d'afficher les noms d'espèces)
- Lineage (affiche les lignées)
- Arêtes (affiche les arêtes de l'arbre)

Figure 4.3 Formulaire permettant l'affichage des résultats

Exemple :

Si on prend la liste des espèces suivantes :

Caldisphaera lagunensis.

Cenarchaeum symbiosum.

Acidilobus aceticus.

uncultured Acidilobus sp..

Aeropyrum camini.

Aeropyrum pernix K1.

Desulfurococcus amylolyticus.

Tribonia.

Ice core clone 23-4.

On obtient les résultats suivants :

- distance Matrix : La matrice de distance générée est comme suit :

```
9

Caldisphaera_lagunen    00.000 10.000 10.000 11.000 10.000 11.000 10.000 25.000 13.000

Cenarchaeum_symbiosu 10.000 00.000 10.000 11.000 10.000 11.000 10.000 25.000 13.000

Acidilobus_aceticus.    10.000 10.000 00.000 05.000 06.000 07.000 06.000 25.000 13.000

uncultured_Acidilobu    11.000 11.000 05.000 00.000 07.000 08.000 07.000 26.000 04.000

Aeropyrum_camini.       10.000 10.000 06.000 07.000 00.000 05.000 06.000 25.000 13.000

Aeropyrum_pernix_K1.    11.000 11.000 07.000 08.000 05.000 00.000 07.000 26.000 14.000

Desulfurococcus_amyl    10.000 10.000 06.000 07.000 06.000 07.000 00.000 25.000 13.000

Tribonia.               25.000 25.000 25.000 26.000 25.000 26.000 25.000 00.000 20.000

Ice_core_clone_23-4.    13.000 13.000 13.000 04.000 13.000 14.000 13.000 20.000 00.000
```

- detail stemps :

Dans la liste présentée seulement les quatre espèces suivantes ont les noms plus longs que 20 caractères. Les noms ont été remplacés par les mêmes noms tronqués à la 20-ème position. Le fichier suivant sert à déchiffrer les noms d'espèces tronquées :

```
Caldisphaera_lagunen = Caldisphaera_lagunensis.

Cenarchaeum_symbiosu = Cenarchaeum_symbiosum.

uncultured_Acidilobu = uncultured_Acidilobus_sp..

Desulfurococcus_amyl = Desulfurococcus_amylolyticus.
```

- Lineage : le lineage des espèces saisies est comme suit :

```
Caldisphaera lagunensis.@cellular organisms ; Archaea; Crenarchaeota;
Thermoprotei; Caldisphaerales; Caldisphaeraceae; Caldisphaera; Caldisphaera
lagunensis.
```

Cenarchaeum symbiosum.@cellular organisms ; Archaea; Crenarchaeota; Thermoprotei; Cenarchaeales; Cenarchaeaceae; Cenarchaeum; Cenarchaeum symbiosum.

Acidilobus aceticus.@cellular organisms ; Archaea; Crenarchaeota; Thermoprotei; Desulfurococcales; Desulfurococcaceae; Acidilobus; Acidilobus aceticus.

uncultured Acidilobus sp..@cellular organisms ; Archaea; Crenarchaeota; Thermoprotei; Desulfurococcales; Desulfurococcaceae; Acidilobus; environmental samples; uncultured Acidilobus sp..

Aeropyrum camini.@cellular organisms ; Archaea; Crenarchaeota; Thermoprotei; Desulfurococcales; Desulfurococcaceae; Aeropyrum; Aeropyrum camini.

Aeropyrum pernix K1.@cellular organisms ; Archaea; Crenarchaeota; Thermoprotei; Desulfurococcales; Desulfurococcaceae; Aeropyrum; Aeropyrum pernix; Aeropyrum pernix K1.

Desulfurococcus amylolyticus.@cellular organisms ; Archaea; Crenarchaeota; Thermoprotei; Desulfurococcales; Desulfurococcaceae; Desulfurococcus; Desulfurococcus amylolyticus.

winitii 99-33.@

Ice core clone 23-4.@cellular organisms ; Eukaryota; unclassified eukaryotes; environmental samples; Ice core clone 23-4.

- Le fichier arêtes contient les informations suivantes :

(Caldisphaera lagunensis. , Caldisphaera lagunensis.) ; (Caldisphaera lagunensis. , Caldisphaera) ; (Caldisphaera , Caldisphaeraceae) ; (Caldisphaeraceae , Caldisphaerales) ; (Caldisphaerales , Thermoprotei) ; (Thermoprotei , Crenarchaeota) ; (Crenarchaeota , Archaea) ; (Archaea , cellular organisms) ; (Cenarchaeum symbiosum. , Cenarchaeum symbiosum.) ; (Cenarchaeum symbiosum. , Cenarchaeum) ; (Cenarchaeum , Cenarchaeaceae) ; (Cenarchaeaceae , Cenarchaeales) ; (Cenarchaeales , Thermoprotei) ; (Acidilobus aceticus. , Acidilobus aceticus.) ; (Acidilobus aceticus. , Acidilobus) ; (Acidilobus , Desulfurococcaceae) ; (Desulfurococcaceae , Desulfurococcales) ; (Desulfurococcales , Thermoprotei) ; (uncultured Acidilobus sp.. , uncultured Acidilobus sp..) ; (uncultured Acidilobus sp.. , environmental samples) ; (environmental samples , Acidilobus) ; (Aeropyrum camini. , Aeropyrum camini.) ; (Aeropyrum camini. , Aeropyrum) ; (Aeropyrum , Desulfurococcaceae) ; (Aeropyrum pernix K1. , Aeropyrum pernix K1.) ; (Aeropyrum pernix K1. , Aeropyrum pernix) ; (Aeropyrum pernix , Aeropyrum) ; (Desulfurococcus amylolyticus. , Desulfurococcus amylolyticus.) ;

RÉFÉRENCES

Barry, D., et Hartigan, J. A. Asynchronous distance between homologous DNAsequences. Biometrics (1987). 43: 261–276.

Barthélémy Jean-Pierre et Guénoche Alain .Trees and proximity representations. With a preface by Michel Minoux (translated from the French by Gregor Lawden).Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons, Ltd., Chichester, xvi+238 pp. ISBN: 0-471-92263-3, (1991)

Boc, A. et Makarenkov, V. : New Efficient Algorithm for Detection of Horizontal gene Transfer Events. Algorithms in Bioinformatics, G. Benson and R. Page (Eds.), 3rd Annual WABI'03, Springer-Verlag, pp. (2003) 190-201.

Boc, A., Makarenkov, V. New Efficient Algorithm for Detection of Horizontal Gene Transfer Events, Algorithms in Bioinformatics, Springer, (2003) 190-201

Boc, A., Makarenkov, V. New Efficient Algorithm for Detection of Horizontal Gene Transfer Events, Algorithms in Bioinformatics, Springer, (2003), 190-201

Buneman, P. The recovery of trees from measures of dissimilarity. In: Kendall,D. and Tautu,P. (eds) Mathematics in Archaeological and Historical Sciences, Edinburg University Press, Edinburg, pp. (1971) 387–395

Delwiche, C.F., et J. D. Palmer.: Rampant Horizontal Transfer and Duplication of Rubisco Genes in Eubacteria and Plastids, Mol. Biol. Evol. (1996) 13, 873-882.

Doolittle, W. F.: Phylogenetic classification and the universal tree. Science (1999) 284:2124-2128

Felsenstein, J., et Churchill, G. A. A hidden Markov model approach to variation among sites in rate of evolution. Mol. Biol. Evol. (1996) 13: 93–104

Gascuel O. : BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data, Mol. Biol. Evol. (1997) 14(7):685-695

Guénoche, A., Leclerc B. The triangles method to build X-trees from incomplete distance matrices. RAIRO Operations Research, (2001) 35, 283--300

Hallet, M., et Lagergren, J.: Efficient algorithms for lateral gene transfer problems. RECOMB 2001, Montréal, ACM, (2001) 149-156

Hein, J. : A heuristic method to reconstructing the evolution of sequences subject to recombination using parsimony. Math. Biosci. (1990) 185-200.

Jerzy Neyman et Herbert B. Osborn Evidence of Widespread Effects of Cloud Seeding at Two Arizona Experiments PNAS (1971) 68: 649-652

JIN, L., et M. NEI. Limitations of evolutionary parsimony method of phylogenetic analysis. Mol. Biol. Evol. (1990) 7:82-102

Jukes, T. H., et C. R. Cantor. Evolution of protein molecules. Pp. 21–132 in H. N. Munro, ed. Mammalian protein metabolism. Academic Press, New York (1969)

Kimura, M. A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. (1980) 16, 111-120

Kimura, M. The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge, MA (1983)

Kishino, H. et M. Hasegawa. Evaiuation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. Journal of Molecular Evolution (1989) 29:170- 179

Lake, J. A. Reconstructing evolutionary trees from DNA and protein sequences: Paralinear distances. Proc. Natl. Acad. Sci. USA 91 (1994): 1455–1459

Landry, P.-A., Lapointe, F.-J. Estimation of Missing Distances in Path-Length Matrices: Problems and Solutions. Pp. 209-224, in Mathematical hierarchies and Biology (B. Mirkin, F.R. McMorris, F. Roberts, A. Rzhetsky, eds.), DIMACS Series in Discrete Mathematics and Theoretical Computer Science, Amer. Math. Soc., Providence, RI, (1997), 209-224

Legendre P., Makarenkov V., " Reconstruction of biogeographic and evolutionary networks using reticulograms ", Systematic Biology, vol. 2, n" 51, (2002), p. 199-216.

Legendre, P. et V. Makarenkov. Reconstruction of biogeographic and evolutionary networks using reticulograms. Systematic Biology (2002) 51(2): 199-216

Linnaeus, C. Systema Naturae. Ninth edition. Theodor Haak, Leiden (Lugdunum Batavorum) (1756)

Lockhart, P. J., Steel, M. A., Hendy, M. D., Penny, D. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. (1994) 11: 605–612

Makarenkov, V. : T-Rex: reconstructing and visualizing phylogenetic trees and reticulation

networks. Bioinformatics, (2001) 17:664-668

Makarenkov, V. Boc, A. et Diallo, A. B.: Representing lateral gene transfer in species classification. Unique scenario. Accepté pour publication à IFCS 2004, Chicago.

Makarenkov, V. et Leclerc, B. Comparison of additive trees using circular orders, Journal of Computational Biology, (2000) 7, 731-744.

Makarenkov, V. et Legendre, P. From a phylogenetic tree to a reticulated network, Journal of Computational Biology, (2004) 11 (1), 195-212

Makarenkov, V. et Legendre, P., Optimal Variable Weighting for Ultrametric and Additive Trees and K-means Partitioning : Methods and Software, Journal of Classification, (2001), 18, 245-271

Makarenkov, V., Lapointe, F.-J. A weighted least-squares approach for inferring phylogenies from incomplete distance matrices, Bioinformatics, (2004) 20, 2113-2121

Makarenkov, V., Leclerc, B. An algorithm for the fitting of a tree metric according to a weighted least-squares criterion, Journal of Classification (1999) 16 3-26

Makarenkov, V., Leclerc, B. An algorithm for the fitting of a tree metric according to a weighted least-squares criterion, Journal of Classification (1999) 16 3-26

Makarenkov, V., Leclerc, B.. Tree metrics and their circular orders: some uses for the reconstruction and fitting of phylogenetic trees, in Mathematical hierarchies and Biology (B. Mirkin, F.R. McMorris, F. Roberts, A. Rzhetsky, eds.), DIMACS Series in Discrete Mathematics and Theoretical Computer Science, Amer. Math. Soc., Providence, RI, (1997), 183-208

Makarenkov, V.: T-Rex: reconstructing and visualizing phylogenetic trees and reticulation networks. Bioinformatics (2001) 17, 664-668.

Needleman et Wunsch, Needleman, S. B. et Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. (1970) 48, 443-453

Page, R. D. M. et Charleston, M. A.: From gene to organismal phylogeny: Reconciled trees. Bioinformatics (1998) 14, 819-820.

Robinson D.R et Foulds L.R. : Comparison of phylogenetic trees, Mathematical Biosciences (1981) 53, 131-147

Saitou, N. et Nei, M.: The neighbour-joining method: a new method for reconstructing

phylogenetic trees. Mol. Biol. Evol. (1987) 4, 406-425

Sattah ,S. , et A. Tversky. Additive similarity trees.Psychometrika (1977) 42:319-345

Steel, M. A. Recovering a tree from a leaf colourations it generates under a Markov model. Appl. Math. Lett. (1994) 7: 19–23

Tajima, F. et M. Nei. Estimation of evolutionary distance between nucleotide sequences. Molecular Biology and Evolution (1984)1:269-285

The NCBI handbook [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information.The Reference Sequence (RefSeq) Project, (2002)

Yushmanov, S.V. Construction of a tree with p leaves from 2p-3 elements of its distance matrix (russian), Matematicheskie Zametki (1984) 35: 877-887

[Cette page a été laissée intentionnellement blanche]

ANNEXE : CODE SOURCE

```
?
/*
**=============================================================
** Function    : Création de la matrice de distances dans matrice_distance.txt à partir de
**                                          lineage_especе.txt.
** Date          : 15/02/2005
**=============================================================
*/

  Function matrice_Distance($x) {
        include "fonction.php";
        include "occurence_espece.php";
        if(@unlink("matrice_distance.txt"))   {   }
        if(@unlink("detail.txt"))   {   }
       $matrice=fopen("matrice_distance.txt", "a");
       $detail=fopen("detail.txt", "a");
       $lineage=fopen("espece_lineage.txt", "r");
       if(!$lineage){  print("Le fichier contenant les especes ne peut pas être ouvert
             !\n"); exit; }
       $lineage2=fopen("espece_lineage.txt", "r");
       $abrev = "";
       $pref = 0;
       $v ="";
       fwrite($matrice,$x);
       fwrite($matrice,"\r\n");
       while($ligne=fgets($lineage, 255)){
             $a = substr($ligne,0,strpos($ligne,"@")) ;
$aa = substr($ligne,strpos($ligne,"@")+1,strlen($ligne)) ;
             $old_a = $a;
             $a = str_replace(" ","_",$a);

             if ( strlen($old_a) > 20 ){
                   if ( occurence_espece($old_a) > 1 ) {
                         $pref = $pref+1 ;
                         $v= "" ;
                         if ($pref < 10 ) { $v = "00" . $pref ; }
                         if ($pref >9 && $pref <100 ) { $v = "0" . $pref ;}
                         $abrev = $abrev . substr($a,0,16) . "_" . $v . " = " . $a . " ; "
                               ;
                         $a = substr($a,0,16) . "_" . $v ; }

                         if ( occurence_espece($old_a) < 2 ) {
                               $abrev = $abrev . substr($a,0,20) . " = " . $a . " ; " ;
                               $a = substr($a,0,20); }
```

```
                                }


                                if ( strlen($a) < 20 ){
                                        while ( strlen($a) < 20 ) {
                                         $a = $a . " " ;
                                 }
                             }
          fwrite($matrice,$a);
          fwrite($matrice," ");
         rewind($lineage2);
         while($ligne2=fgets($lineage2, 255)  ){
                $b = substr($ligne2,0,strpos($ligne2,"@")) ;
                $bb = substr($ligne2,strpos($ligne2,"@")+1,strlen($ligne2)) ;

                  if ($old_a != $b ) {$d = distance($aa,$bb)+distance($bb,$aa); }
                        else{ $d=0;}
                     if ($d <10 ) { fwrite($matrice,0);
                                fwrite($matrice,number_format($d,3,".",","")); }
                     else { fwrite($matrice,number_format($d,3,".",","")); }
                                fwrite($matrice," ");
                              }
       fwrite($matrice,"\r\n");
  }
/* ecriture dans le fichier detail.txt */

 if(@unlink("lexique.txt"))  {   }
          $lex=fopen("lexique.txt", "a");
           fwrite($lex,"\r\n");
           fwrite($lex,"\r\n");
          $lexx = split(";" , $abrev);
          $s = 0;
          while ($lexx[$s] != "" && $s < count($lexx)-1) {

                fwrite($lex,trim($lexx[$s]));
                fwrite($lex,"\r\n");
                $s = $s + 1 ;
         }
fclose($lineage2);
fclose($lineage);
fclose($matrice);
fclose($lex);
}
?>
```

```
<?
**=============================================================
** Function     : Saisie des espèces
** Date         : 15/03/2005
**=============================================================

*/
$x = 0;
include "matrice_distance.php";
include "choix_options.php";
$e = split("\r",ltrim($especes));
      if(@unlink("espece_lineage.txt"))          {             }
$lineage=fopen("espece_lineage.txt", "a");
$ncbi=fopen("ncbi.txt", "r");
$i=0;
$trouve  = 0;
$lig="";
 while ($i < count($e)-1 && $e[$i] !="" ) {
            rewind($ncbi);
            while($ligne_ncbi=fgets($ncbi, 255)  ){
                  $fe = split(";",$ligne_ncbi);
                  $lig = trim($ligne_ncbi);
                   if (ltrim(rtrim($e[$i])) == trim($fe[count($fe)-1]) ) {
                              break;}
            }

            fwrite($lineage,$e[$i]);
            fwrite($lineage,"@");
            fwrite($lineage,$lig);
            fwrite($lineage,"\r");
            $i = $i + 1;
}
$x =count($e)-1;
fclose($lineage);
fclose($ncbi);
matrice_distance($x);
choix_options();
?>
```

```
<?

**=============================================================
** Function   : Calcule de distances entre la première espèce et la racine commune
** Input      : Les noms des deux espèces
** Output     : La distance entre les espèces (valeur numérique)
** Date       : 15/02/2005
*/
**=============================================================

Function Distance($aa,$bb) {
 $sa = split(";",rtrim($aa));
$sb = split(";",Rtrim($bb));
$i = 0 ;
$j = 0 ;
$d = 0 ;
$racine = 0;
$imax = count($sa);
$i= $imax-1;
$j = 0 ;
while ($i >= 0 && $sa[$i] != ""  ) {
        $j = 0 ;
        while ($sa[$i] != $sb[$j] && $j < count($sb)-1) {
                $j = $j+1 ;
         }
         $d = $d+1 ;
                if ($sa[$i] == $sb[$j]) {$racine = 1 ; break ;  }
         $i = $i - 1;
}
 if ($racine == 0) {$d = 0;}
  return $d;
}
?>
```

```
<?
```

```
**=============================================================
/*
** Function    : Création du fichier arretes.txt à partir de lineage_espece.txt
** Date        : 01/08/2005
**=============================================================
*/

 if(@unlink("arretes.txt"))  { }
$art=fopen("arretes.txt", "a");
$lineage=fopen("especе_lineage.txt", "r");
 if(!$lineage){  print("Le fichier contenant les especes ne peut pas être ouvert !\n"); exit;
}
 $arrete = "";
 while($ligne=fgets($lineage, 255)){
          $a = substr($ligne,0,strpos($ligne,"@")) ;
          $aa = substr($ligne,strpos($ligne,"@")+1,strlen($ligne)) ;
                        $sa = split(";",rtrim($aa));
                        $i = count($sa)-1;
                        $arrete = $arrete . "( " .$a . " , " . $sa[$i] . " ) ; "  ;
          while ($i > 0){
                              $arrete = $arrete . "( ". $sa[$i] . " , " . $sa[$i-1] . " ) ; " ;
                              $i = $i - 1 ;
                        }
          }
          $ta = array_unique(split(";",$arrete));

          $i = 0 ;
          $arrete = "";

    while ($i < count($ta)-1){
                        if (strlen($ta[$i]) > 0 ) {
                              $arrete = $arrete . $ta[$i] . " ; " ; }
                              $i = $i + 1 ;
                        }


   fwrite($art,$arrete);
   fclose($lineage);

?>
```

```
<?
/*
**=============================================================
** Function    : cherche le nombre d'occurrences des 20 premiers caractères des espèces
** Date        : 01/10/2005
**=============================================================
*/
Function occurence_espece($espece) {
$lineage=fopen("espece_lineage.txt", "r");
if(!$lineage){  print("Le fichier contenant les especes ne peut pas être ouvert !\n"); exit; }
 $x=0;
$a = "";
while($ligne=fgets($lineage, 255)){
          $a = substr($ligne,0,strpos($ligne,"@")) ;
           if (substr($a,0,20) == substr($espece,0,20)) { $x = $x + 1 ; }
 }
 return $x;
   }
   ?>
```

```
<?
** Function    : affichage des options
** Date        : 15/03/2005
Function choix_options() {
echo "<body bgcolor='#FFFFD8'>";
echo "<center>";
echo "<p> <a href= 'matrice_distance.txt'> distances matrix </a> </p>";
echo "<p> <a href= 'lexique.txt'> detail stamps </a> </p>";
echo "<p> <a href= 'espece_lineage.txt'> Lineage </a> </p>";
echo "<p> <a href= 'arretes.PHP'> Arretes</a> </p>";
echo "<p> <A HREF='http://www.trex.uqam.ca/index.php'>  Return to PHYLIP home
page</A> </p>";
$choix = array("save and return phylip home page");
echo "<select  name='$choix' >\n ";
echo "<option> $choix[0]\n";
echo "</select>\n";
echo "</center>";
echo "</body>";

}

?>
```

```
<html>
<head>
<title>Programme Saisie_especes.htm</title>
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
</head>
<body bgcolor='#FFFFD8'>
<p> </p>
<form method="post" action="saisie_especes.php">
<tr>
   <td>
        <p align="center"><b><font face="Freestyle Script"><font size="7">Posting
stamps</font><font size="7"> Online </font></font></b>
          <hr width="300" color="#000000"><br><br></p>
   </td>
</tr>
<center>
      <p> Please seize the list of the species </p>
      <tr>
       <td align='center' width="500" height="300"><textarea rows="10" name="especes"
cols="70"></textarea></td>
      </tr>
       <p>  <input type="submit" name="Submit" value=" Submit " > </p>
       <p>  <input type="reset" name="Clear" value=" Clear " > </p>

</center>
</form>
</body>
</html>
```

```
<?


/*
**  ================================================
** Function  : Méthode disponibles dabs T-Rex
** Date      : 03/02/2006
**  ================================================
*/
                echo "<script language='JavaScript'>\n";

                echo " function support(action){ \n";
                echo "     document.location.replace('index.php?support='+action);\n";
                echo "  }\n";

                echo "</script>\n";

                function afficherDocumentationTrex(){
                        echo "<div id='right' class='box'>
                                        <div class='colorbox'>
                                                <div
class='module'><div><div><div><h3>Documentation</h3>
                                                        <table width='100%' border='0'
cellpadding='0' cellspacing='0'>
                                                                <tr align='left'><td><a
href='http://www.info2.uqam.ca/~makarenv/trex.html#ref' class='mainlevel'
target='_blank' >T-Rex References</a></td></tr>
                                                                <tr align='left'><td><a
href='http://www.info2.uqam.ca/~makarenv/trex.html' class='mainlevel' target='_blank'
>Windows and Mac<br>versions</a></td></tr>
                                                        </table>
                                                </div></div></div></div>
                                        </div>
                                </div>";
                };

                function afficherFonctionTrex(){
                        echo "<div id='right' class='box'>
                                        <div class='colorbox'>
                                                <div
class='module'><div><div><div><h3>Main Menu</h3>
                                                        <table width='100%' border='0'
cellpadding='0' cellspacing='0'>
```

```
                                                    <tr align='left'><td><a
href='index.php?action=newick&project=trex' class='mainlevel' >Newick
Viewer</a></td></tr>
                                                    <tr align='left'><td><a
href='index.php?action=inference&project=trex' class='mainlevel' >Tree
inference</a></td></tr>
                                                    <tr align='left'><td><a
href='index.php?action=trex&menuD=3&project=trex' class='mainlevel' >Tree inference
from<br>incomplete matrices</a></td></tr>
                                                    <tr align='left'><td><a
href='index.php?action=trex&menuD=2&project=trex' class='mainlevel' >Reticulogram
inference</a></td></tr>
                                                    <tr align='left'><td><a
href='index.php?action=hgt&project=trex' class='mainlevel' >HGT-
Detection</a></td></tr>
                                                    <tr align='left'><td><a
href='index.php?action=align&project=trex' class='mainlevel' >ClustalW</a></td></tr>
                                                    <tr align='left'><td><a
href='index.php?action=sequence&project=trex' class='mainlevel'
>SeqToDistance</a></td></tr>
                                                    <tr align='left'><td><a
href='index.php?action=rf&project=trex' class='mainlevel' >Robinson and
Foulds</a></td></tr>";
                                                //      <tr align='left'><td><a
href='index.php?action=phylip&project=trex' class='mainlevel' >PHYLIP
(ML+Parsimony)</a></td></tr>
                                                echo "  </table>
                                                </div></div></div></div>
                                        </div>
                                </div>";

        }
        function afficherFonctionGbank(){
                echo "<div id='right' class='box'>
                                        <div class='colorbox'>
                                                <div
class='module'><div><div><div><h3>Databases</h3>
                                                <table width='100%' border='0'
cellpadding='0' cellspacing='0'>
                                                    <tr align='left'><td><a
href='http://www.trex.uqam.ca/~trex_dev/gbank' class='mainlevel' >Gene
Base</a></td></tr>";
                                                echo "  </table>
                                                </div></div></div></div>
                                        </div>
                                </div>";
```

```
        }


        function relatedPrograms($action){
            if($action =="phylip"){
                echo "<div id='right' class='box'>
                                <div class='colorbox'>
                                    <div
class='module'><div><div><div><h3>PHYLIP Package</h3>
                                                <table width='100%'
border='0' cellpadding='0' cellspacing='0'>
                                                    <tr
align='left'><td><a href='index.php?action=phylip&app=fitch' class='mainlevel'
>Fitch</a></td></tr>
                                                    <tr
align='left'><td><a href='index.php?action=phylip&app=pars' class='mainlevel'
>Pars</a></td></tr>
                                                    <tr
align='left'><td><a href='index.php?action=phylip&app=dnapars' class='mainlevel'
>Dnapars</a></td></tr>
                                                    <tr
align='left'><td><a href='index.php?action=phylip&app=protpars' class='mainlevel'
>Protpars</a></td></tr>
                                                    <tr
align='left'><td><a href='index.php?action=phylip&app=dollop' class='mainlevel'
>Dollop</a></td></tr>
                                                    <tr
align='left'><td><a href='index.php?action=phylip&app=dnaml' class='mainlevel'
>Dnaml</a></td></tr>
                                                    <tr
align='left'><td><a href='index.php?action=phylip&app=dnamlk' class='mainlevel'
>Dnamlk</a></td></tr>
                                                    <tr
align='left'><td><a href='index.php?action=phylip&app=proml' class='mainlevel'
>Proml</a></td></tr>
                                                    <tr
align='left'><td><a href='index.php?action=phylip&app=promlk' class='mainlevel'
>Promlk</a></td></tr>";
                                    echo "  </table>
                                    </div></div></div></div>
                            </div>
                        </div>";
            }
        }

function afficherFormAdmin(){
```

```
                    echo "<div id='right' class='box'>
                            <div class='colorbox'>
                                <div
class='module'><div><div><div><h3>Menu</h3>
                                    <table width='100%'
border='0' cellpadding='0' cellspacing='0'>
                                        <tr
align='left'><td><a href='index.php?action=pays' class='mainlevel' >View visits
stats</a></td></tr>
                                        <tr
align='left'><td><a href='index.php?action=prog' class='mainlevel' >View progs
stats</a></td></tr>
                                        <tr
align='left'><td><a href='index.php?action=all' class='mainlevel' >View all
stats</a></td></tr>
                                        <tr
align='left'><td><a href='index.php?action=delrep' class='mainlevel' >Delete
directories</a></td></tr>";
                                    echo "  </table>
                                    </div></div></div></div>
                                </div>
                            </div>";
                }

        function afficherMenuTools(){
                    echo "<div id='right' class='box'>
                            <div class='colorbox'>
                                <div
class='module'><div><div><div><h3>Tools list</h3>
                                    <table width='100%'
border='0' cellpadding='0' cellspacing='0'>
                                        <tr
align='left'><td><a href='index.php?tools=tools' class='mainlevel' >HTS-
Corrector</a></td></tr>
                                        <tr
align='left'><td><a href='index.php?tools=tools' class='mainlevel' >Human snoRNA
Database</a></td></tr>
                                        <tr
align='left'><td><a href='index.php?tools=tools' class='mainlevel' >Linear and
Polynomial RDA and CCA</a></td></tr>
                                        <tr
align='left'><td><a href='index.php?tools=tools' class='mainlevel' >Optimal Variable
Weighting</a></td></tr>
                                        <tr
align='left'><td><a href='index.php?tools=tools' class='mainlevel' >Robinson and
Foulds</a></td></tr>";
```

```
                                        echo "  </table>
                                        </div></div></div></div>
                                    </div>
                                </div>";
                    }

        function afficherMenuPeople(){
                                echo "<div id='right' class='box'>
                                        <div class='colorbox'>
                                            <div
class='module'><div><div><div><h3>People</h3>
                                                    <table width='100%'
border='0' cellpadding='0' cellspacing='0'>
                                                    <tr
align='left'><td><a href='index.php?tools=people&filtre=all' class='mainlevel' >View
all</a></td></tr>
                                                    <tr
align='left'><td><a href='index.php?tools=people&filtre=responsable' class='mainlevel'
>Responsable</a></td></tr>
                                                    <tr
align='left'><td><a href='index.php?tools=people&filtre=postdoc' class='mainlevel'
>Post-Doc</a></td></tr>
                                                    <tr
align='left'><td><a href='index.php?tools=people&filtre=phd' class='mainlevel' >Ph. D
degree</a></td></tr>
                                                    <tr
align='left'><td><a href='index.php?tools=people&filtre=master' class='mainlevel'
>Master degree</a></td></tr>
                                                    <tr
align='left'><td><a href='index.php?tools=people&filtre=training' class='mainlevel'
>Training</a></td></tr>";
                                        echo "  </table>
                                        </div></div></div></div>
                                    </div>
                                </div>";
                                echo "<div id='right' class='box'>
                                        <div class='colorbox'>
                                            <div
class='module'><div><div><div><h3>Past people</h3>
                                                    <table width='100%'
border='0' cellpadding='0' cellspacing='0'>
                                                    <tr
align='left'><td><a href='index.php?tools=people&filtre=p_all' class='mainlevel' >View
all</a></td></tr>
```

```
                                                                    <tr
align='left'><td><a href='index.php?tools=people&filtre=p_postdoc' class='mainlevel'
>Post-Doc</a></td></tr>
                                                                    <tr
align='left'><td><a href='index.php?tools=people&filtre=p_phd' class='mainlevel' >Ph.
D degree</a></td></tr>
                                                                    <tr
align='left'><td><a href='index.php?tools=people&filtre=p_master' class='mainlevel'
>Master degree</a></td></tr>
                                                                    <tr
align='left'><td><a href='index.php?tools=people&filtre=p_training' class='mainlevel'
>Training</a></td></tr>";
                                                echo "  </table>
                                                </div></div></div></div>
                                        </div>
                                </div>";
                        }
?>
```

```
<?php

/*
**  ================================================
** Function   : Utilitaires disponibles dans T-Rex
** Date       : 12/05/2006
**  ================================================
*/


$ERROR_SPECIES_TREE = 25;
$ERROR_GENE_TREE  = 26;
$FILE_NOT_EXIST = 27;
$ERROR_NB_SPECIES = 28;

function uploadFile($source,$destination){

        if (move_uploaded_file($_FILES[$source]['tmp_name'], $destination)) {
        } else {
                echo "Unable to upload ".$_FILES[$sources][$name];
        }
}


function phpMyVisites(){

        $REMOTE_ADDR = $_SERVER['REMOTE_ADDR'];
  if($REMOTE_ADDR != '132.208.135.229')
        echo "  <!-- phpmyvisites -->
                        <a href='http://www.phpmyvisites.net/' title='phpMyVisites | Open
source web analytics'
                        onclick='window.open(this.href);return(false);'>
                        <script type='text/javascript'>
                        <!--
                        var a_vars = Array();
                        var pagename='';

                        var phpmyvisitesSite = 2;
                        var phpmyvisitesURL =
'http://trex.labunix.uqam.ca/~alix/phpmv2/phpmyvisites.php';
                        //-->
                        </script>
```

```
                        <script language=javascript
src='http://trex.labunix.uqam.ca/~alix/phpmv2/phpmyvisites.js'
type='text/javascript'></script>
                        <noscript>
                        <p>phpMyVisites | Open source web analytics
                        <img
src='http://trex.labunix.uqam.ca/~alix/phpmv2/phpmyvisites.php' alt='phpMyVisites'
style='border:0' />
                        </p>
                        </noscript>
                        </a>
                        <!-- /phpmyvisites -->
                        ";
}

function AfficherBasPage(){
        echo"<br><p><hr align='left' width='45%' color='black'>";
        echo"Copyright &copy; 2005 Universit&eacute; du Qu&eacute;bec &agrave
Montr&eacute;al (UQAM)";
        echo"<br>Webmaster : <a href='mailto:dcarrey@gmail.com'>Alix
Boc</a></p>";
}

function AfficherOnglet($action){

                if($action=='home')   $home   = "class='active'"; else $home="";
                if($action=='admin')  $admin  = "class='active'"; else $admin="";
                if($action=='tools')  $tools  = "class='active'"; else $tools="";
                if($action=='people') $people = "class='active'"; else $people="";

                if($action=='admin') {$pos1='../'; $pos2="";} else {$pos1="";
$pos2='admin/';}

                echo "<div id='nav'>";
                echo " <ul>";
                echo "  <li $home><a
href='$pos1"."index.php?action=home&tools=trex'>Home</a></li>";
//          echo "  <li $news><a href='index.php?action=news&tools='>News</a></li>";
                echo "  <li $tools><a
href='$pos1"."index.php?action=tools&tools=tools'>Tools</a></li>";
                echo "  <li $people><a
href='$pos1"."index.php?action=people&tools=people&filtre=all'>People</a></li>";
                echo "  <li $admin><a
href='$pos2"."index.php?action=all'>Admin</a></li>";
                echo " </ul>";
                echo "</div>";
```

```
}

function FormatPhylipInput($dataInput){
        $inputdata = str_replace("\r","",$dataInput);
        return $inputdata;
}

function afficherMessage($etat,$message){
        if($etat==1)
                echo "$message";
}

function afficherEnteteResultat($chaine){
        echo "<h3>$chaine</h3>";
}

function gestionVisite(){

        $REMOTE_ADDR = $_SERVER ['REMOTE_ADDR'];

        if( (!isset($_SESSION['session_id']))&&($REMOTE_ADDR !=
"132.208.135.229") ){
                $_SESSION['session_id'] = session_id();
                $nbvisits = file_get_contents("compteur.txt");
                $_SESSION['nbvisits'] = $nbvisits = $nbvisits+1;
                $fp = fopen("compteur.txt","w+");
                 fputs($fp,$nbvisits);
                 fclose($fp);
                 $fp = fopen("visiteurs.txt","a+");
                 $fp2 = fopen("admin/ip.txt","a+");
                 fputs($fp2,"$REMOTE_ADDR\n");
                 fclose($fp2);
                 $jour = date("d-m-Y");
                 $heure = date("H:i");
                 $chaine = "$REMOTE_ADDR $jour $heure\n";
                 fputs($fp,$chaine);
                 fclose($fp);
        }
        else{
                $nbvisits = $nbvisits = file_get_contents("compteur.txt");    ;
        }

        return $nbvisits;
}
```

```
function VarPostSession($nomVariable, $valDefaut)
        {
        $resultat= $valDefaut;

        if (isset($_POST[$nomVariable]))
                {
                $resultat= $_POST[$nomVariable];
                $_SESSION[$nomVariable]= $resultat;
                }
        elseif (isset($_SESSION[$nomVariable]))
                $resultat= $_SESSION[$nomVariable];
        else
                $_SESSION[$nomVariable]= $resultat;

        return $resultat;
        }

  function VarGetSession($nomVariable, $valDefaut)
        {
        $resultat= $valDefaut;

        if (isset($_GET[$nomVariable]))
                {
                $resultat= $_GET[$nomVariable];
                $_SESSION[$nomVariable]= $resultat;
                }
        elseif (isset($_SESSION[$nomVariable]))
                $resultat= $_SESSION[$nomVariable];
        else
                $_SESSION[$nomVariable]= $resultat;

        return $resultat;
        }

  function VarRequestSession($nomVariable, $valDefaut)
        {
        $resultat= $valDefaut;

        if (isset($_REQUEST[$nomVariable]))
                {
                $resultat= $_REQUEST[$nomVariable];
```

```
            $_SESSION[$nomVariable]= $resultat;
            }
        elseif (isset($_SESSION[$nomVariable]))
            $resultat= $_SESSION[$nomVariable];
        else
            $_SESSION[$nomVariable]= $resultat;

        return $resultat;
        }



        function VarPost($nomVariable,$valDefaut)
        {
            if(isset($_POST[$nomVariable]))
                $resultat = $_POST[$nomVariable];
            else
                $resultat = $valDefaut;
            return $resultat;
        }

?>
```

```
<?
/*
**  ================================================
** Function   : Page Principale du site T-Rex
** Date       : 20/005/2006
**  ================================================
*/
        session_start();

        include("headerApp.php");
        include ("utils.php");
        include ("menu.php");

//      phpMyVisites();

        unset($_SESSION['actionTrex']);

        $trex_clusterState = "trex_clusterState.txt";
        if(isset($_REQUEST['tools']))
                $tools = $_SESSION['tools'] = $_REQUEST['tools'];
        else
                $tools = $_SESSION['tools'] = "trex";

        if(isset($_REQUEST['action']))  $action = $_SESSION['action'] =
$_REQUEST['action'];
        if(!isset($_SESSION['action'])) $action = $_SESSION['action'] = "inference";
else $action = $_SESSION['action'];

        /*if(isset($_GET['action']))
                $action = $_GET['action'];
        else
                $action='inference';
*/
        if(isset($_REQUEST['menuD']))  $menuD = $_SESSION['menuD'] =
$_REQUEST['menuD'];
        if(!isset($_SESSION['menuD'])) $menuD = $_SESSION['menuD'] = "1"; else
$menuD = $_SESSION['menuD'];
        $method = VarGetSession('method','1');
        $app = VarGetSession('app','dnaml');

        $_SESSION['CalculExecute'] = 0;
        $_SESSION['bootstrapFile'] = "";
        $_SESSION['viewOption'] = "";

        //echo "action = $action, tools = $tools";
```

```
        $progs = varRequestSession('progs','dnapars');
?>
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
        <meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1" />
        <title>Trex-online</title>
        <meta name="title" content="Trex-online" />
        <meta name="author" content="Alix Boc" />
        <meta name="description" content="Trex-online" />
        <META NAME="keywords" CONTENT="Makarenkov,TREX,T-
REX,Phylogeny,phylogenetic tree,inferring phylogenies,phylogenetic tree,phylogenetic
reconstruction,sequence alignment,ClustalX,ClustalW,HGT,LGT,horizontal gene
transfer,horisontal gene transfer,lateral gene
transfer,recombination,hybridization,hybridisation,reticulogram,reticulograms,
reticulations,reticulate evolution,evolutionary network,networks,trees,sequence to
distance,NJ,BIONJ,MW,ADDTREE,tree drawing,tree map,transfer
detection,bioinformatics software,tree inferring algorithms,tree inferring
methods,phylogenetic networks">

        <meta name="robots" content="index, follow" />
        <link href="css/default_css.css" rel="stylesheet" type="text/css" />
        <link href="css/blue_css.css" rel="stylesheet" type="text/css" />

</head>

<body class="white">

<table  align='center' class="encadrement" cellspacing="0" cellpading="0"><tr><td>

<a name="up" id="up"></a>
<div align="center">
 <div id="wrapper">
   <div id="header">

   <? if($tools =="trex")
        //echo "<img src='images/t-rex-t.gif' align='left'>";
   ?>
        <!--<table border=0 align=left><tr><td><br><br><font color = 'white' size='100'
face='freestyle script'><b>Trex Online</b></font></td></tr></table>-->
   </div>

   <div id="header_bottom">
                <script type="text/javascript">
        sfHover = function() {
                var sfEls =
document.getElementById("nav").getElementsByTagName("LI");
```

```
                        for (var i=0; i<sfEls.length; i++) {
                                sfEls[i].onmouseover=function() {
                                        this.className+=" sfhover";
                                }
                                sfEls[i].onmouseout=function() {
                                        this.className=this.className.replace(new
RegExp(" sfhover\\b"), "");
                                }
                        }
                }
                if (window.attachEvent) window.attachEvent("onload", sfHover);
        </script>
        <?
                AfficherOnglet($action);
        ?>
</div>
        <table><tr><td> </td></tr></table>
        <table width="100%" cellspacing="0" cellpadding="0" border="0"
><tr><td>
                <table width='100%'  border="0"><tr>
                        <!--<td width="19%" valign="top">-->

                        <?
                                if($tools == "trex"){

                                        echo "<td width='19%' valign='top'>";
                                        afficherFonctionTrex();
                                        relatedPrograms($action);
                                        afficherDocumentationTrex();
                                        afficherFonctionGbank();
                                        echo "</td>";
                                }

                                if($tools == "tools"){

                                        echo "<td width='19%' valign='top'>";
                                        afficherMenuTools();
                                        echo "</td>";
                                }
                                if($tools == "people"){

                                        echo "<td width='19%' valign='top'>";
                                        afficherMenuPeople();
                                        echo "</td>";
                                }
                        ?>
```

```
                        <!--</td>-->
                        <td width="1%"></td>
                        <td width="80%" valign="top" align='center'>
                            <table align='center' width='100%'
class="contentpaneopen" border="0">
                                <? if(($action != "test"
)&&($action!="home")&&($action!="tools")&&($action!="people")){?>
                                    <tr>
                                        <td class="contentheading"
width="100%" align="left">
                                            <?
                                              if ($action ==
"phylip") echo afficherTitre($app,0);
                                              else if ($action ==
"trex") echo afficherTitre($action,$menuD);
                                              else echo
afficherTitre($action,$menuD);
                        ?>
                                        </td>
                                    </tr>

                                    <tr>
                                        <td><br></td>
                                    </tr>
                                <?}?>
                            <tr>
                                <td>
                                    <?
                                        $clusterState =
@file_get_contents($trex_clusterState);

                                        switch($action){

                                            case "newick":
        include('interfaces/int_newick.php');

        $chaineNewick = $_SESSION['chaineNewick'];

        afficherFormNewick($chaineNewick,$support);

        break;

                                            case "trex" :
        include('interfaces/int_trex.php');
```

```
        AfficherFormTrex($menuD,$clusterState);
        break;
                                                                case
"sequence":include('interfaces/int_seqtodist.php');
        $dataInputSequence  = VarPostSession('dataInputSequence', "");
        $menuS      = VarPostSession('menuS', 32890);
        $validation  = VarPostSession('validation', "");
        $PEMV       = VarPostSession('PEMV', 0);
        $ComputeA    = VarPostSession('ComputeA', 1);
        $penaltyGap  = VarPostSession('penaltyGap',0);
        $a = VarPostSession('a','1.0');

        AfficherFormSequence($dataInputSequence,$validation,$menuS,$PEMV,$Comp
uteA,$penaltyGap,$a,$support);
        break;
                                                                case "hgt"    :
include('interfaces/int_hgt.php');
        $_SESSION['cluster'] = 0;
        $_SESSION['reload'] = 0;
        $_SESSION['hgt_root1'] = 0;
        $_SESSION['hgt_root2'] = 0;
        $_SESSION['ExecViewTree'] = 0;
        $_SESSION['hgt'] = "on";
        AfficherFormHGT($clusterState);
```

```
        break;

                                                        case "align"  :
        include('interfaces/int_clustal.php');

        AfficherFormAlign($clusterState);

        break;

                                                        case "lineage":
        $dataInputLineage = VarPostSession('dataInputLineage', "");

        AfficherFormLineage($dataInputLineage,$support);

        break;

                                                        case "home" :
include('interfaces/int_home_trex.php');

        home_trex();

        break;

                                                        case "phyml":
        include('interfaces/int_phyml.php');

        AfficherFormPhyML($clusterState);

        break;

                                                        case "people":
        include('interfaces/int_people.php');

        $filtre = $_REQUEST['filtre'];

        people($filtre);

        break;

                                                        case "tools":
        include('interfaces/int_tools.php');

        tools();
        break;
```

```
                                                    case "inference":
include('interfaces/int_inference.php');
        AfficherFormInference($clusterState);

        break;

                                                    case "rf"       :
        include('interfaces/int_rf.php');

        AfficherFormRF();

        break;

                                                    case "phylip":
include('interfaces/int_phylip.php');

        AfficherFormPhylip($app);

        break;

                                                    default   :   echo
"<p align='center'><img src='images/underconstruction.jpg'></p>";

        break;
                                                }

                                        ?>
                                        </td>
                                    </tr>
                                </table>

                        </td>
                        </tr>
                        </table>
                </td></tr>
                <tr><td>
                <?
                        $nbvisits = gestionVisite();
                        echo "<br><br><br><br>This site has been visited $nbvisits times
since Friday, November 25, 2005.";
                ?>
                </td></tr>
                </table>
  </div>
</div>
        <?afficherBasPage();?>
```

```
</td></tr></table>
</body>
</html>
<?
/*
**  ========================================================
** Function    : Ce fichier contient les fonctions utiles pouvant être employées
* *              à plusieurs endroits.
** Date         : 04/05/2005
**  ========================================================

*/

if (defined("INCLUDE_UTILITAIRES"))
        return;

define ("INCLUDE_UTILITAIRES", "INCLUDE_UTILITAIRES");

/*****
 * Rediriger : Permet de rediriger une page.
 *
 * $page -> page de destination
 */
function Rediriger($page)
        {
        echo "<script language='JavaScript'>\n
       document.location.replace('$page');
       </script>\n";
        }

function make_directory($repertoire){
        $cmd = "mkdir $repertoire";
        @mkdir($repertoire,0777);
        system("chmod 777 $repertoire");
}
```

```
?>
/*
**  ================================================
** Function   : Interface de differentes fonctions de T-Rex
** Date       : 04/08/2005
**  ================================================
*/


<html>
        <head>
                <link href="css/blue_css.css" rel="stylesheet" type="text/css" />
        </head>
</html>
                <script language='JavaScript'>

                function validerInputInference(){
                        if ( document.dataEntry.dataInput.value.length < 1) {
                                alert('You have to paste a distance matrix before
computation');
                                return false;
                        }
                                if ( document.dataEntry.menuD.value == '0') {
                                alert('You have to select a distance method before
computation');
                                return false;
                        }
                        var x = document.dataEntry;
                        x.action = 'loadData.php';
                        x.submit();
                }

                        function pasteSample1HGT(){
                        var x = document.FormHGT;
                        x.dataInputSpecies.value = '4 \nAlpha    0 2 3 3\nBeta     2 0 3
3\nGamma    3 3 0 2\nEpsilon   3 3 2 0';
                        x.dataInputGene.value    = '4 \nAlpha    0 3 3 2\nBeta     3 0 2
3\nGamma    3 2 0 3\nEpsilon   2 3 3 0';
                        }

                function pasteSample2HGT(){
                        var x = document.FormHGT;
                        x.dataInputSpecies.value = '(A:1.0,B:1.0,(C:1.0,D:1.0):1.0);';
```

```
                x.dataInputGene.value = '(A:1.0,C:1.0,(B:1.0,D:1.0):1.0);';
        }

        </script>
<?

    function AfficherFormTrex($menuD,$clusterState){


        $method     = VarREQUESTSession('method', "");
        $MWoptimization = VarPostSession('MWoptimization', 1);
        $k = VarPostSession('k',5);
        $p = VarPostSession('p',0);
        $WeigthMatrix = VarPostSession('WeigthMatrix',1);
        $reticulationOptimization = VarPostSession('reticulationOptimization',1);
        $menuS      = VarPostSession('menuS', 32890);
        $PEMV       = VarPostSession('PEMV', 0);
        $ComputeA   = VarPostSession('ComputeA', 1);
        $penaltyGap  = VarPostSession('penaltyGap',0);
        $a = VarPostSession('a','1.0');
        $bootstrap = VarPostSession('bootstrap',0);
        $nbRep = VarPostSession('nbRep',100);

        if($menuD==3){
            $dataType = $_SESSION['dataType'] = 1;
        }
        else
            $dataType = VarPostSession('dataType',1);
        $dataInput = VarPostSession('dataInput',"");

        if($support == "") $state = "";
        else $state = "disabled";

        if(($dataInput == "")&&($dataType==1))
            $dataInput = "4\nAlpha     0 2 3 3\nBeta      2 0 3 3\nGamma     3 3
0 2\nEpsilon   3 3 2 0";

        echo "<form enctype='multipart/form-data' name='dataEntry'
method='POST' action='index.php?action=trex#methodsChoice' >";
        echo "<table border=0 width='99%' align='center'>";
        echo "  <tr align='center'>";
        if($dataType == 1)
            echo "          <td valign='top' align='center'>Paste your distance
matrix  in the <a href='http://evolution.genetics.washington.edu/phylip/doc/distance.html'
target='_blank'>Phylip format</a>  into the window : <br></td>";
        else
```

```
                        echo "          <td valign='top' align='center'>Paste your sequences
in the <a
href='http://workshop.molecularevolution.org/resources/fileformats/phylip_dna.php'
target='_blank'>Phylip format</a>  into the window :<br></td>";
                echo "  </tr>";
                echo "  <input type='hidden' name='page' value='index.php'>";
                echo "  <input type='hidden' name='actionTrex' value='infer'>\n";
                echo "  <input type='hidden' name='project' value='trex'>\n";
                echo "  <input type='hidden' name='menuD' value='$menuD'>\n";

                if(($menuD == 1)||($menuD == 2)){
                        echo "<tr><td align='center'><table border=0><tr>";
                        echo "  <td ><b>Data type :</b></td>";
                        echo "  <td><input type='radio' value='1' name='dataType' "; echo
($dataType==1)?"checked":""; echo "
onClick='javascript:document.dataEntry.submit();'></td><td>Distance matrix</td>";
                        echo "  <td><input type='radio' value='2' name='dataType' "; echo
($dataType==2)?"checked":""; echo "
onClick='javascript:document.dataEntry.submit();'></td><td>Sequences</td>";
                        echo "</tr></table></td></tr>";
                }

                echo "  <tr><td><p align='center'><textarea rows='10' name='dataInput'
cols='70%' >"; echo $dataInput; echo "</textarea></p>";
                echo "  </td></tr>";
                echo "  <tr>";
                echo "  <td align='center' colspan='2'><br>";
                echo "          <input type='submit' value='Compute' onClick='return
validerInputInference();' $state>  ";
//              echo "          <input class='button' type='submit' value='Compute on
cluster' name='cluster' onClick = 'return validerInputInference();'
$clusterState>  ";
                echo "          <input class='button' type='reset' value='Reset'
name='RAZ'>  ";
                echo "          <input class='button' type='button' value='Clear'
name='clear' onClick ='javascript: dataInput.value = \"\";'></td>";
                echo "  </td>";
                echo " </tr>";
                if($clusterState != "")
                        echo "  <tr><td align='center'>Sorry, the cluster is temporarely
unavailable !!!</td></tr>";
                echo "<tr>";
                echo "  <td><br><hr align='center' width='300' color='#000000'></td>";
                echo "</tr>";

                /*if(($menuD != "")&&($menuD != 0)){
```

```
                    echo "<tr>";
                    echo "  <td width='98%' colspan='2'>";
                    echo "          <b><br>Parameters for distance methods</b><hr
align='left'><p>";
                    echo "  </td>";
                    echo "</tr>";
                }*/

                if(($menuD == "1")||($menuD == "2")||($menuD == "4")){

                    echo "<tr>";
                    echo "  <td width='98%' colspan='2'
align=left><br><table><tr><td><b>Tree reconstruction method: </b></td><td>";
                    echo "          <select name='method' onchange='javascript:
document.dataEntry.submit();'>";
                    echo "                  <option name=2 value=2 "; echo
($method==2)?"selected":""; echo ">Neighbor Joining - Saitou and Nei
(1987)</option>";
                    echo "                  <option name=1 value=1 "; echo
($method==1)?"selected":""; echo ">ADDTREE - Sattath and Tversky
(1977)</option>";
                    echo "                  <option name=3 value=3 "; echo
($method==3)?"selected":""; echo ">Unweighted Neighbor Joining - Gascuel
(1997)</option>";
                    echo "                  <option name=4 value=4 "; echo
($method==4)?"selected":""; echo ">Circular order reconstruction - Makarenkov, Leclerc
(1997)</option>";
                    echo "                  <option name=5 value=5 "; echo
($method==5)?"selected":""; echo ">Weighted least-squares method MW - Makarenkov,
Leclerc (1999)</option>";
                    echo "                  <option name=6 value=6 "; echo
($method==6)?"selected":""; echo ">BioNJ - Gascuel (1997)</option>";
                    echo "          </select>";
                    echo "  </td></tr></table></td>";
                    echo "</tr>";
                }

                if($menuD == "3"){

                    echo "<tr>";
                    echo "  <td width='98%' colspan='2'
align=left><br><table><tr><td><b>Tree reconstruction method: </b></td><td>";
                    echo "          <select name='method'>";
```

```
                    echo "                    <option name=1 value=1 "; echo
($method==1)?"selected":""; echo ">Triangles method - Guenoche, Leclerc
(2001)</option>";
                    echo "                    <option name=2 value=2 "; echo
($method==2)?"selected":""; echo ">Ultrametric procedure + MW - De Soete
(1984)</option>";
                    echo "                    <option name=3 value=3 "; echo
($method==3)?"selected":""; echo ">Additive procedure + MW - Landry et al.
(1996)</option>";
                    echo "                    <option name=4 value=4 "; echo
($method==4)?"selected":""; echo ">MW-modified - Makarenkov (2001)</option>";
                    echo "                    <option name=5 value=5 "; echo
($method==5)?"selected":""; echo ">MW* - Makarenkov, Lapointe (2004)</option>";
                    echo "          </select>";
                    echo "          </td></tr></table></td>";
                    echo "</tr>";
               }

               if(((($menuD == "1")||($menuD=="4")||($menuD == "2"))&&($method ==
"5")){

                    echo "<tr>";
                    echo "  <td colspan='2'><b>MW options :</b></td>";
                    echo "</tr>";
                    echo "<tr>";
                    echo "  <td>";
                    echo "    <table border=0 width='80%'><tr><td class='grey'
width='100%'>";
                    echo "          <table align=center border=0 bgcolor='#FFFFFF'
width='100%'>";
                    echo "                    <tr>";
                    echo "                              <td width='50%' ><b></b>";
                    echo "                                        <table><tr><td><input
type='radio' value='1' name='MWoptimization' "; echo
($MWoptimization==1)?"checked":""; echo "></td><td> MW local
Optimization<br></td></tr>";
                    echo "                                        <tr><td><input type='radio'
value='2' name='MWoptimization' "; echo ($MWoptimization==2)?"checked":""; echo
"></td><td> MW global Optimization  </td></tr></table>";
                    echo "                              </td>";
                    echo "                              <td width='50%'>";
                    echo "                                        <table><tr><td><input
type='radio' value='1' name='WeigthMatrix' "; echo ($WeigthMatrix==1)?"checked":"";
echo "></td><td>Weight matrix W = 1/D^p   p = </td><td><input
type='text' name='p' size='4' value='$p'></td></tr></table>";
```

```
                    //echo "                                                <input type='radio'
value='2' name='WeigthMatrix' "; echo ($WeigthMatrix==2)?"checked":""; echo ">in
file  <input name='MatrixW' type='file' size='15'>";
                    echo "                                </td>";
                    echo "                        </tr>";
                    echo "            </table>";
                    echo "    </td></tr></table>";
                    echo "  </td>";
                    echo "</tr>";
          }

          if($menuD == "2"){

                    echo "<tr>";
                    echo "  <td width='45%'><b><br>";
                    echo "            Stop adding reticulation branches when :</b>";
                    echo "  </td>";
                    echo "</tr>";
                    echo "<tr bgcolor='#336699'><td>";
                    echo "            <table align=center border=0 bgcolor='#FFFFFF'
width='100%'>";
                    echo "                        <tr>";
                    echo "                                <td width='50%'><table>";
                    echo "                                        <tr><td><input type='radio'
value='1' name='reticulationOptimization' "; echo
($reticulationOptimization==1)?"checked":""; echo"></td><td colspan=2>Q1 is
minimized<br></td></tr>";
                    echo "                                        <tr><td><input type='radio'
value='2' name='reticulationOptimization' "; echo
($reticulationOptimization==2)?"checked":""; echo"></td><td colspan=2>Q2 is
minimized<br></td></tr>";
                    echo "                                        <tr><td><input type='radio'
value='3' name='reticulationOptimization' "; echo
($reticulationOptimization==3)?"checked":""; echo"></td><td>K reticulation branches
have been added  </td>";
                    echo "                                        <td><input type='text'
name='k' size='2' value='$k' onClick ='javascript:
dataEntry.reticulationOptimization[2].checked = true';></td></tr>";
                    echo "                                </table></td>";
                    echo "                        </tr>";
                    echo "            </table>";
                    echo "</td></tr>";
          }

          //echo $menuD;
```

```
                if(($dataType==2)&&(($menuD==1)||($menuD==2))){
                        /*echo "          <tr>";
                        echo "          <td>";
                        echo "                  <br><b>Parameters for sequences
model</b>";
                        echo "                  <hr width='70%' align='left'>";
                        echo "          </td>";
                        echo "  </tr>";*/

                        echo "          <tr>";
                        echo "           <td width='50%' colspan='2'>";
                        echo "                  <br><table><tr><td><b>Select model of
evolution:</b></td><td>";
                        echo "                  <select name='menuS'
onChange='javascript:document.dataEntry.submit();'>";

                        echo "                          <option name='blosum'
value='32887' "; echo ($menuS==32887)?"selected":""; echo ">Uncorrected
Distances</option>";
                        echo "                          <option name='blosum'
value='32888' "; echo ($menuS==32888)?"selected":""; echo ">Jukes-Cantor</option>";
                        echo "                          <option name='blosum'
value='32889' "; echo ($menuS==32889)?"selected":""; echo ">Tajima-Nei</option>";
                        echo "                          <option name='blosum'
value='32890' "; echo ($menuS==32890)?"selected":""; echo ">Kimura 2-
Parameters</option>";
                        echo "                          <option name='blosum'
value='32891' "; echo ($menuS==32891)?"selected":""; echo ">Tamura</option>";
                        echo "                          <option name='blosum'
value='32892' "; echo ($menuS==32892)?"selected":""; echo ">Jin-Nei
Gamma</option>";
                        echo "                          <option name='blosum'
value='32893' "; echo ($menuS==32893)?"selected":""; echo ">Kimura
Protein</option>";
                        echo "                          <option name='blosum'
value='32894' "; echo ($menuS==32894)?"selected":""; echo ">LogDet</option>";
                        echo "                          <option name='blosum'
value='32895' "; echo ($menuS==32895)?"selected":""; echo ">F84</option>";
                        echo "                  </select>";
                        echo "           </td></tr></table></td>";
                        echo "          <tr>";

                        if(($menuS == 32887)||($menuS == 32888)){

                                echo "  <tr>";
                                echo "          <td>";
```

```
                        echo "          <table border=0 width='50%'><tr><td
class='grey' width='100%'>";
                        echo "              <table width='100%'>";
                        echo "                  <tr>";
                        echo "                      <td
width='100'>Penalty of gap : </td><td><input type='text' size='5' name='penaltyGap'
value='$penaltyGap'></td>";
                        echo "                  </tr>";
                        echo "              </table>";
                        echo "          </td></tr></table>";
                        echo "      </td>";
                        echo "  </tr>";
                }

                if(($menuS == 32888)||($menuS == 32890)){
                        echo "  <tr>";
                        echo "      <td>";
                        echo "          <table border=0 width='50%'><tr><td
class='grey' width='100%'>";
                        echo "              <table border=0 width='100%'>";
                        echo "                  <tr><td><input type='radio'
value='0' name='PEMV'"; echo ($PEMV==0)?"checked":""; echo "></td><td>Ignore
missing bases</td></tr>";
                        echo "                  <tr><td><input type='radio'
value='1' name='PEMV'"; echo ($PEMV==1)?"checked":""; echo "></td><td>PEMV
estimation of missing bases values</td></tr>";
                        echo "              </table>";
                        echo "          </td></tr></table>";
                        echo "      </td>";
                        echo "  </tr>";
                }

                if(($menuS == 32892)){
                        echo "  <tr>";
                        echo "      <td'>";
                        echo "          <table border=0
width='60%'><tr><td class='grey' width='100%'>";
                        echo "              <table><tr><td>Compute the
value of the parameter a ?</td> ";
                        echo "              <td><input type='radio'
value='1' name='ComputeA' "; echo ($ComputeA==1)?"checked":""; echo " onClick
='javascript:a.disabled = true;'></td><td>Yes</td>";
                        echo "              <td><input type='radio'
value='0' name='ComputeA' "; echo ($ComputeA==0)?"checked":""; echo " onClick
='javascript:a.disabled = false;'></td><td>No</td>";
```

```
                                echo "
        <td>    a : </td><td><input type='txt' name='a'
value='$a' size='5' "; echo ($ComputeA==1)?"disabled":""; echo"></td></tr></table>";
                                echo "                    </td></tr></table>";
                                echo "          </td>";
                                echo "  </tr>";
                        }

                        echo "          <tr><td>";
                        echo "                  <b><br>Validation :</b><br>";
                        echo "          </td></tr>";
                        echo "          <tr>";
                        echo "           <td valign='top'>";
                        echo "            <table border=0 width='50%'><tr><td class='grey'
width='100%'>";
                        echo "            <table border=0><tr><td>";
                        echo "                  <table><tr><td><input type='radio'
value='0' name='bootstrap'"; echo ($bootstrap==0)?"checked":""; echo " onClick
='javascript:nbRep.disabled = false;'></td><td>Bootstrap</td></tr>";
                        echo "                  <tr><td><input type='radio' value='1'
name='bootstrap'"; echo ($bootstrap==1)?"checked":""; echo " onClick
='javascript:nbRep.disabled = false;'></td><td>Jackknife</td></tr>";
                        echo "                  <tr><td><input type='radio' value='2'
name='bootstrap'"; echo ($bootstrap==2)?"checked":""; echo " onClick
='javascript:nbRep.disabled = true;'></td><td>No branch validation</td></tr></table>";
                        echo "            </td>";
                        echo "            <td align='center' valign='top'>";
                        echo "                  Number of replicates<br><input type='txt'
name='nbRep' value='$nbRep' size='5' "; echo ($bootstrap==2)?"disabled":""; echo">";
                        echo "            </td></tr></table>";
                        echo "            </td></tr></table>";
                        echo "           </td>";
                        echo "          </tr>";

                }

                echo "</form>";
                echo "</table>";
                echo "<a name='methodsChoice'></a>";
        }
?>
```

```
/*
**  ================================================
** Function    : Interface de ClustalW
** Date        : 17/04/2006
**  ================================================
*/

Interface ClustalW
<html>
        <head>
                <link href="css/blue_css.css" rel="stylesheet" type="text/css" />
        </head>
</html>

                <script language='JavaScript'>

                function validerInputAlign(){
                        var x = document.formAlign;
                        if ( document.formAlign.dataInputAlign.value.length < 1) {
                                alert('You have to paste your sequences in the text area');
                                return false;
                        }
                        if ( isNaN(x.gapopen.value) || x.gapopen.value < 0 ||
x.gapopen.value > 100) {
                                alert('The gap opening field must be a number between 0
and 100');
                                return false;
                        }
                        if ( isNaN(x.gapext.value) || x.gapext.value < 0 || x.gapext.value >
100) {
                                alert('The gap extention field must be a number between 0
and 100');
                                return false;
                        }
                        if ( isNaN(x.maxdiv.value) || x.maxdiv.value < 0 || x.maxdiv.value
> 100) {
                                alert('The delay divergent sequence field must be a
percentage (0-100)');
                                return false;
                        }
                        if ( isNaN(x.transweight.value) || x.transweight.value < 0 ||
x.transweight.value > 1) {
```

```
                    alert('The DNA transition weight field must be number
between 0 and 1');
                    return false;
                }
                if ( isNaN(x.pairgap.value) || x.pairgap.value < 0 || x.pairgap.value
> 100) {
                    alert('The gap penalty value for pairwise alignment must be
number between 0 and 100');
                    return false;
                }
                if ( isNaN(x.pwgapopen.value) || x.pwgapopen.value < 1 ||
x.pwgapopen.value > 100) {
                    alert('The gap penalty for pairwise alignment must be a
number between 1 and 100');
                    return false;
                }
                if ( isNaN(x.pwgapext.value) || x.pwgapext.value < 0 ||
x.pwgapext.value > 100) {
                    alert('The gap extention value for pairwise alignment must
be a number between 0 and 100');
                    return false;
                }
                if ( isNaN(x.ktuple.value) || x.ktuple.value < 1 || x.ktuple.value > 2)
{
                    alert('The K-Tuple size must be a number between 1 and
2');
                    return false;
                }
                if ( isNaN(x.topdiags.value) || x.topdiags.value < 1 ||
x.topdiags.value > 50) {
                    alert('The top diagonals must be a number between 1 and
50');
                    return false;
                }
                if ( isNaN(x.window.value) || x.window.value < 1 ||
x.window.value > 50) {
                    alert('The window size must be a number between 1 and
50');
                    return false;
                }
                x.action = 'loadData.php';
                x.submit();
            }

            </script>
```

```
<?

        function AfficherFormAlign($clusterState){

                $dataInputAlign = VarPostSession('dataInputAlign', "");
                $quicktree = VarPostSession('quicktree', "1");
                $matrix = VarPostSession('matrix', "gonnet");
                $dnamatrix = VarPostSession('dnamatrix', "iub");
                $gapopen = VarPostSession('gapopen', "10");
                $gapext = VarPostSession('gapext', "0.2");
                $negative = VarPostSession('negative', "OFF");
                $maxdiv = VarPostSession('maxdiv', "30");
                $transweight = VarPostSession('transweight', "0.5");
                $ktuple = VarPostSession('ktuple', "1");
                $topdiags = VarPostSession('topdiags', "5");
                $window = VarPostSession('window', "5");
                $pairgap = VarPostSession('pairgap', "3");
                $pwmatrix = VarPostSession('pwmatrix', "gonnet");
                $pwdnamatrix = VarPostSession('pwdnamatrix', "iub");

                $pwgapopen = VarPostSession('pwgapopen', "10");
                $pwgapext = VarPostSession('pwgapext', "0.1");

                $border = 0;
                $bgcolor = '#FFFFFF';
                if($support == "") $state = "";
                else $state = "disabled";
                echo "<form enctype='multipart/form-data' name='formAlign'
method='post' action='index.php#methodsChoice'>";
                echo "<table border='$border' width='90%' align='center'>";
                echo "  <tr align='center'>";
                echo "          <td><a
href='http://www.ebi.ac.uk/clustalw/clustalw_help.html#sequence' target
='_blank'>ClustalW</a>";
                echo "                  is a widely used sequence alignment

tool.<br><br>";
                echo "          </td>";
                echo "  </tr>";
                echo "  <tr>";
                echo "          <input type='hidden' name='project' value='trex'>\n";
                echo "          <td align='center'><span class='normalTexte'>Paste your
sequence into the window : <br>(7 input formats are accepted: <a
```

```
href='http://en.wikipedia.org/wiki/FASTA_format' target='_blank'>FASTA</a>, <a
href='http://www.ebi.ac.uk/help/pir_frame.html' target='_blank'>NBRF/PIR</a>, <a
href='http://www.ebi.ac.uk/help/embl_frame.html' target='_blank'>EMBL/Swiss-
Prot</a>, <a href='http://www.ebi.ac.uk/help/gde_frame.html' target='_blank'>GDE</a>,
<a href='http://www.ebi.ac.uk/help/ClustalW_frame.html' target='_blank'>Clustal</a>,
<a href='http://www.ebi.ac.uk/help/MSF_frame.html' target='_blank'>GCG/MSF</a>,<a
href='http://www.ebi.ac.uk/help/rsf_frame.html'
target='_blank'>RSF</a>)</span><br></td>";

                echo "  </tr>";
                echo "  <tr><td><input type='hidden' name='page' value='index.php'>";
                echo "  <input type='hidden' name='actionTrex' value='align'></td></tr>";
                echo "  <tr>";
                echo "          <td align='center'><textarea rows='10'

name='dataInputAlign' cols='70%' >"; echo $dataInputAlign; echo "</textarea></td>";
                echo " </tr>";

                echo "  <tr>";
                echo "          <td align='center' ><br>";
                echo "                  <input class='button' type='submit' value='Align
sequences' name='Load' onClick = 'return validerInputAlign();' $state>  ";
//              echo "                  <input class='button' type='submit' value='Align
(Cluster)' name='cluster' onClick = 'return validerInputAlign();'
$clusterState>  ";
                echo "                  <input class='button' type='reset' value='Reset'
name='RAZ'>  ";
                echo "                  <input class='button' type='button' value='Clear'
name='clear' onClick ='javascript: formAlign.dataInputAlign.value = \"\";'>";
                echo "          </td>";
                echo "  </tr>";
                if($clusterState != "")
                        echo "  <tr><td align='center'>Sorry, the cluster is temporarely
unavailable !!!</td></tr>";

                echo "  <tr>";
                echo "          <td><br><br><b>Multiple Alignment options :</b></td>";
                echo "  </tr>";

                echo "  <tr bgcolor='#336699'>";
                echo "  <td align = 'center'><table border='$border' width='100%'
bgcolor='$bgcolor'><tr>";
                echo "          <td width='35%'>Pairwise alignments</td>";
                echo "          <td  width='20%' align='left'>";
                echo "                  <input type='radio' value='0' name='quicktree'
checked>Slow ";
```

```
                echo "                  <input type='radio' value='1' name='quicktree' ";
echo ($quicktree=='1')?"checked":""; echo " >Fast ";
                echo "          </td>";
                echo "          <td width='30%'> Use Negative Matrix</td>";
                echo "          <td width='15%' align='left'>";
                echo "                  <select name='negative'>";
                echo "                          <option name=1 value='OFF' "; echo
($negative=='OFF')?"selected":""; echo ">OFF</option>";
                echo "                          <option name=2 value='ON' "; echo
($negative=='ON')?"selected":""; echo ">ON</option>";
                echo "                  </select>";
                echo "          </td>";
                echo "  </tr></table></td>";
                echo "  </tr>";
                echo "  <tr bgcolor='#336699'>";
                echo "  <td align = 'center'><table border='$border' width='100%'
bgcolor='$bgcolor'><tr>";
                echo "          <td width='35%'> Gap Opening (0-100)</td>";
                echo "          <td width='20%' align='left'>";
                echo "                  <input type='txt' name='gapopen' value='$gapopen'
size='5' >";
                echo "          </td>";
                echo "          <td width='30%'> Gap Extention (0-100)</td>";
                echo "          <td width='15%' align='left'>";
                echo "                  <input type='txt' name='gapext' value='$gapext'
size='5' >";
                echo "          </td>";
                echo "  </tr></table></td>";
                echo "  </tr>";
                echo "  <tr bgcolor='#336699'>";
                echo "  <td align = 'center'><table border='$border' width='100%'
bgcolor='$bgcolor'><tr>";
                echo "          <td width='35%'>Delay Divergent Sequence (%)</td>";
                echo "          <td width='20%' align='left'>";
                echo "                  <input type='txt' name='maxdiv' value='$maxdiv'
size='5' >";
                echo "          </td>";
                echo "          <td width='30%'>DNA Transition Weight (0-1)</td>";
                echo "          <td width='15%' align='left'>";
                echo "                  <input type='txt' name='transweight'
value='$transweight' size='5' >";

                echo "          </td>";
                echo "  </tr></table></td>";
                echo "  </tr>";
                echo "  <tr bgcolor='#336699'>";
```

```
                echo "  <td align = 'center'><table border='$border' width='100%'
bgcolor='$bgcolor'><tr>";
                echo "          <td width='35%'>Protein weight matrix</td>";
                echo "          <td width='20%'>";
                echo "                  <select name='matrix'>";

                echo "                          <option name='blosum' value='blosum' ";
echo ($matrix=='blosum')?"selected":""; echo ">BLOSUM series</option>";
                echo "                          <option name='pam' value='pam' "; echo
($matrix=='pam')?"selected":""; echo ">PAM series</option>";
                echo "                          <option name='gonnet' value='gonnet' ";
echo ($matrix=='gonnet')?"selected":""; echo ">Gonnet series</option>";
                echo "                          <option name='id' value='id' "; echo
($matrix=='id')?"selected":""; echo ">Identity matrix</option>";
                echo "                  </select>";
                echo "          </td>";
                echo "          <td width='30%'>DNA weight matrix</td>";
                echo "          <td width='15%'>";
                echo "                  <select name='dnamatrix'>";

                echo "                          <option name='iub' value='iub' "; echo
($dnamatrix=='iub')?"selected":""; echo ">IUB</option>";
                echo "                          <option name='clustalw' value='clustalw' ";
echo ($dnamatrix=='clustalw')?"selected":""; echo ">ClustalW</option>";
                echo "                  </select>";
                echo "          </td>";
                echo "  </tr></table></td>";
                echo "  </tr>";

                echo "  <tr>";
                echo "          <td><br><br><b>Slow Pairwise Alignment
parameters</b></td>";
                echo "  </tr>";
                echo " <tr bgcolor='#336699'>";
                echo "  <td align = 'center'><table border='$border' width='100%'
bgcolor='$bgcolor'><tr>";
                echo "          <td width='35%'>Gap opening penalty (0-100)</td>";
                echo "          <td width='10%'>";
                echo "                  <input type='txt' name='pwgapopen'
value='$pwgapopen' size='5' >";
                echo "          </td>";
                echo "          <td width='25%'>DNA weight matrix</td>";
                echo "          <td width='30%'>";
                echo "                  <select name='pwdnamatrix'>";
```

```
                echo "                          <option name='iub' value='iub' "; echo
($pwdnamatrix=='iub')?"selected":""; echo ">IUB</option>";
                echo "                          <option name='clustalw' value='clustalw' ";
echo ($pwdnamatrix=='clustalw')?"selected":""; echo ">ClustalW</option>";
                echo "                  </select>";
                echo "          </td>";
                echo "  </tr></table></td>";
                echo " </tr>";
                echo " <tr bgcolor='#336699'>";
                echo "  <td align = 'center'><table border='$border' width='100%'
bgcolor='$bgcolor'><tr>";
                echo "          <td width='35%'>Gap extension penalty (0-100)</td>";
                echo "          <td width='10%'>";
                echo "                  <input type='txt' name='pwgapext'
value='$pwgapext' size='5' >";
                echo "          </td>";
                echo "          <td width='25%'>Protein weight matrix</td>";
                echo "          <td width='30%'>";
                echo "                  <select name='pwmatrix'>";

                echo "                          <option name='blosum' value='blosum' ";
echo ($pwmatrix=='blosum')?"selected":""; echo ">BLOSUM series</option>";
                echo "                          <option name='pam' value='pam' "; echo
($pwmatrix=='pam')?"selected":""; echo ">PAM series</option>";
                echo "                          <option name='gonnet' value='gonnet' ";
echo ($pwmatrix=='gonnet')?"selected":""; echo ">Gonnet series</option>";
                echo "                          <option name='id' value='id' "; echo
($pwmatrix=='id')?"selected":""; echo ">Identity matrix</option>";
                echo "                  </select>";
                echo "          </td>";
                echo "  </tr></table></td>";
                echo " </tr>";

                echo "  <tr>";
                echo "          <td><br><br><b>Fast Pairwise Alignment
parameters</b></td>";
                echo "  </tr>";

                echo " <tr bgcolor='#336699'>";
                echo "  <td align = 'center'><table border='$border' width='100%'
bgcolor='$bgcolor'><tr>";
                echo "          <td width='35%'>Gap Penalty (1-500)</td>";
                echo "          <td width='25%'>";
                echo "                  <input type='txt' name='pairgap' value='$pairgap'
size='5' >";
```

```
                echo "          </td>";
                echo "          <td width='30%'>K-Tuple Size (1-2)</td>";
                echo "          <td width='25%'>";


                echo "                  <input type='txt' name='ktuple' value='$ktuple'
size='5' >";

                echo "          </td>";
                echo "  </tr></table></td>";
                echo " </tr>";
                echo " <tr bgcolor='#336699'>";
                echo "  <td align = 'center'><table border='$border' width='100%'
bgcolor='$bgcolor'><tr>";
                echo "          <td width='35%'>Top Diagonals (1-50)</td>";
                echo "          <td width='25%'>";
                echo "                  <input type='txt' name='topdiags' value='$topdiags'
size='5' >";
                echo "          </td>";
                echo "          <td width='30%'>Windows Size (1-50)</td>";
                echo "          <td width='25%'>";
                echo "                  <input type='txt' name='window' value='$window'
size='5' >";
                echo "          </td>";
                echo "  </tr></table></td>";
                echo " </tr>";

                echo "</table>";
                echo "<a name='methodsChoice'></a>";
                echo "</form>";

        }
```

```
<html>
	<head>
		<link href="css/blue_css.css" rel="stylesheet" type="text/css" />
	</head>
</html>

<?
/*
**  ==============================================
** Function   : Outils de T-Rex
** Creator    : Adel Younes
** Date       : 23/11/2005
**  ==============================================
*/



		function tools(){

			//echo "<table width='100%'><tr><td align='justify'>";
			//echo "<h3>NewickToDistance</h3>";
			//echo "<br>Transform a tree in the newick format in distance
matrix ";
			//echo "<br><br><a
href='usagers/sources/newicktodist/newicktodist.exe'>Download it
!</a></td></tr></table><br><br>";

			echo "<a name='hts'></a>";
			echo "<table width='100%'><tr><td align='justify'>";
			echo "<h3>HTS-Corrector</h3>";
			echo "<br>High-throughput screening (HTS) is a novel and
efficient technology for drug discovery. It allows for screening of more than 100,000
compounds a day per screen and requires effective procedures for quality control. We
have developed a method for evaluating a background surface of an HTS assay; it can be
used to correct raw HTS data. This correction is necessary to take into account systematic
errors that may affect the procedure of hit selection. The described method allows one to
analyze experimental HTS data and determine trends and local fluctuations of the
corresponding background surfaces. For a large amount of plates of the same assay, the
deviations of the background surface from a plane are caused by systematic errors. Their
influence can be minimized by the subtraction of the systematic background from the raw
data.";
```

```
                    echo "<br><br><a
href='http://www.labunix.uqam.ca/~makarenv/hts.html' target='_blank'>Go to HTS-
Corrector Home Page</a></td></tr></table><br><br>";

                    echo "<a name='snorna'></a>";
                    echo "<table width='100%'><tr><td align='justify'>";
                    echo "<h3>Human snoRNA Database</h3>";
                    echo "<br><br><a href='http://www.trex.uqam.ca/~snorna'
target='_blank'>Go to Human snoRNA Database Home
Page</a></td></tr></table><br><br>";

                    echo "<a name='rda'></a>";
                    echo "<table width='100%'><tr><td align='justify'>";
          echo "<h3>Linear and Polynomial RDA and CCA</h3>";
          echo "<br> In this program, classical linear redundancy analysis (Rao, 1964) and
canonical correspondence analysis (ter Braak, 1986, 1987) are computed using multiple
regressions followed by direct eigenanalysis of the matrix of fitted values. The method of
calculation is described in Chapter 11 of Legendre & Legendre (1998). Polynomial RDA
and CCA, which are generalizations of the linear forms, are implemented using a new
approach proposed by Makarenkov and Legendre (1999, 2002). The polynomial methods
are based on the use of polynomial multiple regression, during the first stage of RDA and
CCA, instead of the multiple linear regression used in the linear forms. The explanatory
variables are limited to their quadratic form in any term of the polynomial. The program
produces the output required to draw biplot diagrams for linear and polynomial RDA or
CCA. The explanatory variables can be represented in biplots in two different ways: (1)
the individual terms of the polynomial equation can be represented as separate variables,
or (2) one can choose to represent an explanatory variable using the multiple correlations
(rescaled as required by the selected scaling method) of the canonical ordination axes
against the linear and quadratic forms of the variable. A permutation procedure allows
one to test the significance of the two models (linear and polynomial) and of the
difference between them.";
          echo                                                    "<br><br><a
href='http://www.bio.umontreal.ca/Casgrain/en/labo/plrdacca.html' target='_blank'>Go to
Linear and Polynomial RDA and CCA Home Page</a></td></tr></table><br><br>";

                    echo "<a name='ovw'></a>";
                    echo "<table width='100%'><tr><td align='justify'>";
                    echo "<h3>Optimal Variable Weighting (OVW)</h3>";
                    echo "<br>This program performs optimal variable weighting for
ultrametric and additive tree clustering, following the method proposed by De Soete
(1986, 1988), as well as for K-means partitioning. It is described in Makarenkov &
Legendre.
                    The program, which is available free of charge to
academic users, provides some improvements and extra options, compared to De Soete's
(1988) program OVWTRE, which only implemented fitting to the first two families of
clustering methods mentioned above. Given a rectangular data matrix Y (n,m), containing
```

measurements of n objects on m variables, the procedure computes variable weights w(m) such that the resulting matrix of inter-object dissimilarities D(n,n) obtained from Y optimally satisfies either the ultrametric or the additive inequality, or optimally corresponds to a K-means partition with fixed number of groups K. The weights w are constrained to be nonnegative and their sum is equal to one. We used Polak-Ribiere's Conjugate Gradient Method (Numerical Recipes, Press et al., 1986) to carry out the optimization process. Once the optimal variable weights are obtained in the case of the ultrametric or additive tree clustering, the resulting inter-object dissimilarities D can be subjected to any of the existing ultrametric or additive tree fitting procedures. See De Soete (1986) or Makarenkov and Leclerc (1999) for an overview of these methods. It is worth noting that sometimes only a local minimum can be obtained as a final result. Hence a good choice of initial weights is essential. According to our investigations an initial guess consisting of all equal weights (as implemented in De Soete's software OVWTRE) cannot guarantee reaching the global minimum. An interesting feature of program OVW, compared to OVWTRE, is that it allows users to restart the optimization procedure any number of times, using different random initial guesses for the weights. As a consequence, OVW usually obtains better results than OVWTRE in the case of ultrametric and additive clustering; optimization for K-means partitionning is not available in OVWTRE. Moreover, the optimization in OVW is carried out in the way allowing to avoid a degenerate trivial solution in the case of the additive clustering. Such a solution consists of giving a weight of 1 to any one variable and weights of 0 to all other variables. Another important feature not mentioned by De Soete (1986, 1988) is that the global minimum of a function to minimize can be reached with several different sets of optimal weights w. This may lead to different inter-object dissimilarities matrices D, from which different hierarchies or additive trees can be inferred.";

echo "

Go to OVW Home Page</td></tr></table>

";

echo "";
echo "<table width='100%'><tr><td align='justify'>";
echo "<h3>Robinson and Foulds topological distance</h3>";
echo "
This program allows to compute in the optimal time the Robinson and Foulds topological distance between two or more additive trees given their distance matrices. The program is based on the algorithm proposed by Makarenkov and Leclerc (1999, b) for calculation of this distance. The algorithm implemented in this program uses the notion of circular orders to compare the topology of two trees. The algorithm described in Makarenkov and Leclerc (1999, b) has optimal time complexity, requiring O(n2) time when performed on two nxn distance matrices. The Robinson and Foulds topological distance is an important and frequently used tool to compare additive (phylogenetic) tree structures (see for instance Robinson and Foulds (1981) or Makarenkov and Leclerc (1999, a,b)). This distance is equal to the minimum number of elementary operations, consisting of merging or splitting nodes, necessary to transform one tree into the other. As proved in Robinson and Foulds (1981), it is also the number of bipartitions, or Buneman's splits (1971), which belong to exactly one of the two trees. If we deal with two unrooted trees having no internal vertices labeled according to the

```
elements of the set X, the Robinson and Foulds distance of on the set X of n elements varies between 0 (when the trees are isomorphic) and 2n-6 (when all non-trivial bipartitions in two trees are different; a trivial bipartition corresponds to an edge incident to a leaf).";
        echo "<br><br><a href='http://www.bio.umontreal.ca/Casgrain/en/labo/robinson_foulds.html' target='_blank'>Go to Robinson and Foulds topological distance Home Page</a></td></tr></table>";
        }
?>
```