

Notions de bioinformatique

Souvent les avancées des biotechnologies ont été possibles du fait d'avancées technologiques relevant d'autres domaines que la biologie.

- En juillet 1995 le séquençage d'*Haemophilus Influenza* rd KW 20 est achevé
 - Depuis plus de 5000 génomes entiers ont été séquencés et de nombreux autres séquençages sont en cours.
- (sources ncbi 2013/03)

Procaryotes	plus de 1000
Virus	3210
Eucaryotes	247 dont <i>l'Homosapiens sapiens</i> !!! (24/03/2008)

↳ Les données biologiques sont donc très nombreuses.

Les biologistes ont donc eu besoin d'un outil pour gérer ce grand nombre de données :

La bioinformatique est née

Définition

La bio informatique correspond à la conception et à l'utilisation d'outil informatique pour permettre le recueil, le stockage, l'analyse et l'exploitation de données biologiques telles que des séquences nucléiques et des séquences protéiques.

Les bio-informaticiens peuvent donc être

- des spécialistes de l'informatique (gestion de l'informations en réseaux, programmation,) qui travaillent de concert avec les biologistes
ce qui ne nous intéresse pas ici !!
- des utilisateurs de l'outil bio-infomatique pour pouvoir stocker des résultats expérimentaux et analyser des données..... ce qui nous intéresse.

Ainsi en ce qui nous concerne on retiendra que la bioinformatique permet de réaliser des analyses « *in silico* » (Qui complètent les analyses « in vivo » et « in vitro »)

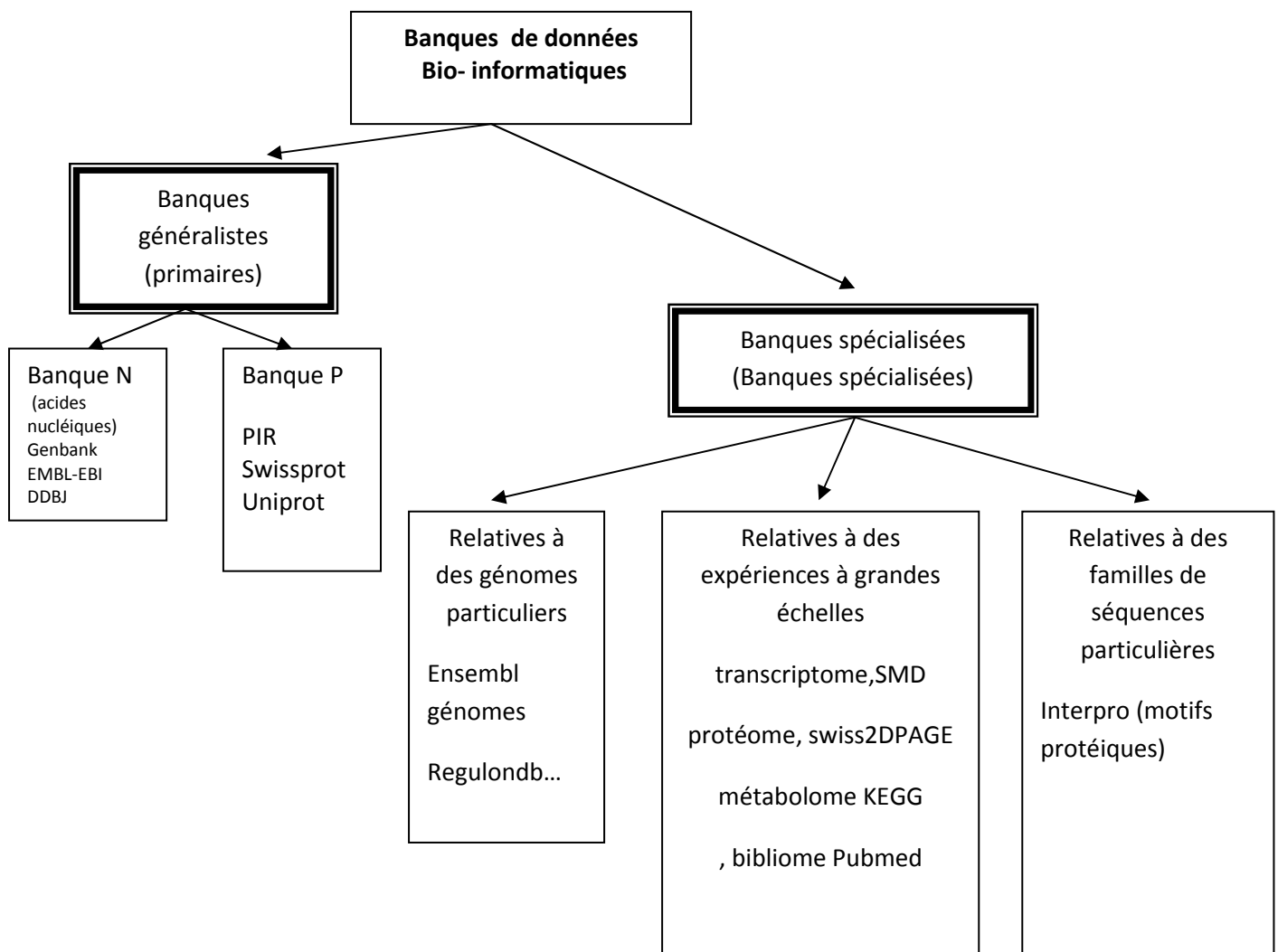
Ces analyses s'effectuent à partir de banques de données qui renferment des fichiers dans lequel se trouvent entre autres des séquences nucléotidiques (ADN) et protéiques assorties de données appelées annotations

I- bases ou banques de données et contenu

I-1- Les bases ou banques de données bio-informatiques

Les Banques (ou bases) de données sont des espaces virtuels de stockage de données sous forme de fichiers (sans répétitions)

L'Enjeu majeur de ces bases ou banques de données est de donner accès à de nombreuses données même si elles ne proviennent pas d'un même laboratoire ou d'une même source (à un premier niveau par des liens hypertextes, avec un vocabulaire contrôlé)



I-2 Contenu des fichiers présents dans les banques de données

Les banques de données généralistes renferment leurs données dans des fichiers dont le contenu est très proche. Les objectifs essentiels étant

- la traçabilité
- Le classement des fichiers en vue de recherches selon des critères.

Ainsi on retrouve constamment :

ID un identificateur
AC un numéro d'accèsion

DE une description de la séquence
 SQ la séquence dans le sens 5'...3'

Format de Gen bank

Champ	Définition
LOCUS	Identificateur (nom et taille)
DEFINITION	Description de la séquence
ACCESSION	Numéro d'accension
VERSION	Numéro de la version
KEY WORD	Mot(s)-clef(s)
SOURCE	Organisme d'où provient la sequence
ORGANISME	Classification taxonomique de l'organisme
REFERENCE	Reference bibliographique de l'entrée
AUTHORS	Auteurs des articles
TITLE	Titre de l'article
JOURNAL	Reference du journal
FEATURES	Caractéristiques de la séquence
ORIGINE	Sequence (6 blocs de 10 caractères par lignes
//	Fin de l'entrée

LOCUS A04646 2220 bp DNA linear PAT 03-JUL-2002

DEFINITION C.diphtheriae gene for diphtheria toxin protein.

ACCESSION A04646

VERSION A04646.1 GI:21694064

KEYWORDS .

SOURCE Corynebacterium diphtheriae

ORGANISM [Corynebacterium diphtheriae](#)

Bacteria; Actinobacteria; Actinobacteridae; Actinomycetales;
 Corynebacterineae; Corynebacteriaceae; Corynebacterium.

REFERENCE 1 (bases 1 to 2220)

AUTHORS Kaczorek,M., Tiollais,P., Chenciner,N., Streeck,R.E. and Boquet,P.

TITLE Nucleotide sequence coding for the signal peptide of diphtheria
 toxine, vector containing this nucleotide sequence, its use in the
 transformation of microorganisms and peptide compositions obtained

JOURNAL Patent: EP 0133403-A2 2 20-FEB-1985;

INSTITUT PASTEUR; INSERM

FEATURES Location/Qualifiers

source 1..2220

/organism="Corynebacterium diphtheriae"

/mol_type="unassigned DNA"

/db_xref="taxon:[1717](#)"

CDS 301..1983

/codon_start=1

/transl_table=[11](#)

/product="diphtheria toxin protein"

/protein_id="[CAA00374.1](#)"

/db_xref="GI:21694065"

/translation="MSRKLFASXLIGALLGIGAPPSAHAGADDVVDSSKSFVMENFSS
 YHGTPGYVDSIQKGIQPKSGTQRNYDDDWKGFYSTDNKYDAAGYSVDNENPLSGKA
 GDVVKVTYPGLTKVLALKVDNAETIKKELGLSLTEPXMEQVGTEEFXKRFGDGASRVV
 LSLPFAEGSSSVEYINNWEQAKALSVKLEINFETRGRGQDAMYEYMAQACAGNRVRR
 SVGRSLSCINLDWDVIRDKTKTKIESLKEHGPIKNKMSESPNKTVSQEKAKQYLEEFH
 QTALEHPQLSELKTVTGTNPLFAGANYAAWVNVAQVIDSETADNLEKTTAALSILPG
 IGSVMGIADGAVHHNTEEIQAQSIALLSSLMVAQAIPLVGELVDIGFAPYNFVESIINL
 FQVVHNSYNRSAYSPGHKTQPFLLHDGYAVSWNTVEDSIIRTGFQGESGHDIKITAENT
 PLPIASVLLPTIPGKLDVNKSKTHISVNGRKIRMRCRAIDGDVTFCRPKSPVYVGNV
 HXNLHVAFHRSSSEKIXSNEISSDSIGVLGYQKTVDHDKVNSKLSLFFEXKS"

ORIGIN

1 atcttttgcgg tgtggtacac ctgatctggt cgggttcattg ttgtggtggt caacgctggg
 61 gtaaccggcg ttgcgtatcc agtggctaca ctcaggttgt aatgattggg atgatgtacc ...

[...]

```

2041 gcacaggtag agcagaattn gaatntgact acggatcaga aggttggggg ttcgaatccc
2101 tccgggcgca caantgaaac cccagctcat agcatgtttg agctgggggt tctcatggcg
2161 tgtnngttgt ctgactgttg gctgtnttg cgggtggttg tgctngtacc gaaccgaacg

```

//

I-3 le format des séquences nucléotidiques et protéiques

Le format des séquences N ou P est uniforme :

- universel (compatible avec les logiciels de traitements de texte)
- le point de départ de la séquence est toujours identique

Le Format FASTA

```

>N°, lettre et description du genome ↵
Séquence nucléotidique

```

Exemple

```

>gi|9629357|ref|NC_001802.1| Human immunodeficiency virus 1, complete genome
GGTCTCTCTGGTTAGACCAGATCTGAGCCTGGGAGCTCTCTGGCTAACTAGGGAACCCACTGCTTAAGCC
TCAATAAAGCTTGCCTTGAGTGCTTCAAGTAGTGTGTGCCCGTCTGTTGTGTGACTCTGGTAAGTAGAGA
TCCCTCAGACCCTTTTAGTCAGTGTGGAAAATCTCTAGCAGTGGCGCCCGAACAGGGACCTGAAAGC  A
.....

```

Remarques

★ Pour les séquences nucléotidiques

- Il y a 15 écritures possibles pour chaque nucléotide

BASES	SYMBOLE	BASES	SYMBOLE	BASES	SYMBOLE
Adénosine	A	puRiques (A ou G)	R	non A (C, G ou T)	B
Cytosine	C	pYrimidiques (C ou T)	Y	non G (A, C ou T)	H
Guanosine	G	bases quelconques (aNy =A, G, C, T)	N	non C (A, G ou T)	D
Thymidine	T	bases céto (Kéto) (G ou T)	K	non T (A, C ou G)	V
bases à 3 liaisons H "Strong" (G ou C)	S	bases aMino (A ou C)	M	bases à 2 liaisons H "Weak" (A ou T)	W

- Seul 1 brin de nucléotide d'ADN est répertorié dans les banques de données !
 - Même si l'ADN est une molécule double brin !
 - même si génome est à ARN pour les virus !
- La séquence est toujours dans le sens **5'P3'OH**

Exercices

Ecrire le complémentaire de la séquence -1- .

Comment cette séquence serait écrite alors écrite ?

★ Pour les séquences protéiques

- Une lettre correspond à un AA (voir tableau)
- La séquence est écrite dans le sens $\text{NH}_2 \dots \text{COOH}$

SYMBOLE	CODE TROIS LETTRES	ACIDE AMINES
A	Ala	Alanine
B	Asp, Asn	acide aspartique ou asparagine
C	Cys	cystéine
D	Asp	acide aspartique
E	Glu	acide glutamique
F	Phe	phénylalanine
G	Gly	glycine
H	His	histidine
I	Ile	isoleucine
K	Lys	lysine
L	Leu	leucine
M	Met	methionine
N	Asn	asparagine
P	Pro	proline
Q	Gln	glutamine
R	Arg	arginine
S	Ser	sérine
T	Thr	thréonine
V	Val	valine
W	Trp	tryptophane
X	xxx	inconnu
Y	Tyr	tyrosine
Z	Glu, Gln	acide glutamique ou glutamine
★	fin	stop

I-4 Les annotations

Une annotation est une caractéristique localisée sur une séquence nucléique ou protéique

Exemple d'obtention d'une annotation permettant de localiser un gène dans une séquence nucléotidique permettant la synthèse d'une protéine

-L'annotation comprend deux étapes

- l'annotation automatique (ou intrinsèque) : Un algorithme recherche sur la séquence un site de fixation de l'ARN polymérase suivie d'une séquence qui permettra l'obtention d'un codon initiateur et plus loin d'un codon stop sur l'ARNm permettant

-**L'annotation manuelle** (ou extrinsèque) : Des biologistes parcourent une à une les annotations automatiques et décide de valider ou non le gène repéré dans l'étape précédente au moyen de données bibliographiques ou référencées dans des banques de données.

Les annotations ne sont donc jamais vraiment terminées....

D'autres annotations sont possibles :

Annotations	Objectifs
A l'échelle nucléique	
- usage du code génétique - statistique de « mots » (répétitions...) - comparaisons avec des banques de données	- identification de gènes - identifications de signaux particuliers
A l'échelle protéique	
- comparaison avec des banques de séquences protéiques - prédictions de structures secondaires et tertiaires	- identification de la fonction biologique des gènes identifiés
Pour un ensemble de gènes d'un même génome	
- relation de voies métaboliques - définitions de classe de gènes	- identification d'opéron - identification de fonctions biologiques de gènes
Comparaison de génomes	
- Etude de familles de gènes	-Etudes des interactions entre des gènes voisins -Identification de fonctions absentes ou spécifiques d'un génome donné

II- Les Outils de recherche, de visualisation et d'analyse des données

II-1 Les outils de recherche et de visualisation de données

Ces outils permettent d'interroger, de recouper des données pour que les utilisateurs comparent leurs données expérimentales aux données scientifiques archivées.

Ces outils sont d'accès aisé (via internet), didactique (intuitif), exhaustifs (complet pour que l'utilisateur n'ait pas à aller chercher ailleurs d'autres données)

Il existe deux grands types d'outils :

-les data browser qui permettent d'interroger les bases à partir de mots clés, de séquence, d'identifiant, de bibliographie ou de propriétés de séquence.

-les génomes browser qui permettent de parcourir le génome complet pour localiser une séquence donnée sur un génome

II-2 Exemples d'utilisation d'outils d'analyses relevant de la bioinformatique

On se limitera aux exigences du programme de terminale. Le but de ce paragraphe est de ne présenter que quelques exemples simples d'analyses bioinformatiques de données biologiques.

II-2-1 Etudier une séquence

Il est possible d'analyser statistiquement une séquence nucléotidique (par exemple en déterminant un nombre de nucléotides, en calculant le GC%)

II-2-2 Réaliser *in silico* des phénomènes biologiques qui s'effectuent normalement *in vivo*

Il est possible de repérer sur une séquence nucléotidique :

- une séquence codante (reading frame : cadre de lecture à partir d'un triplet donnant un codon initiateur sur l'ARNm et se terminant par un triplet donnant un codon stop sur l'ARNm) ,
- un gène.

Il est possible de traduire « *in silico* » cette séquence codante en protéine (= translation)

Pour les organismes procaryotes, le logiciel réalise « *in silico* » à partir d'un nucléotide donné la transcription de l'ADN en ARNm puis la traduction de l'ARN en protéines au moyen du code génétique.

Pour les organismes eucaryotes, il faudra veiller à ce que le logiciel traduit en protéine uniquement les exons. (ADN ... (transcription (noyau)... ARNprém(maturation (coiffage,ajout queue poly A, épissage= élimination des introns)(noyau)) ...ARNm.....(traduction).... Protéine)

II-2-3 Réaliser *in silico* des expérimentations qui relèvent du laboratoire

Il est possible de faire agir *in silico* des enzymes de restrictions sur une séquence d'ADN. Ces enzymes hydrolysent l'ADN en des sites de coupures spécifiques. Virtuellement, on obtiendra donc des fragments d'ADN de tailles différentes.

Il est possible de prévoir *in silico* le résultat d'une séparation électrophorétique des fragments d'ADN de tailles différentes.

Il est possible de déterminer *in silico* les couples d'amorces à utiliser pour réaliser l'amplification d'une séquence données pour une manipulation de PCR.

II-2-4 Comparer les séquences avec des séquences répertoriées dans des banques de données

Il est possible de rechercher des similitudes entre séquence afin de pouvoir éventuellement les identifier. Pour cela on réalise ce que l'on nomme alignement.

La qualité de l'alignement peut être appréciée par des calculs avec par exemple

- Des pourcentages de similitudes (même position pour des nucléotides identiques)
- Par un score d'alignement
- Par une valeur statistique qui calcule la probabilité d'avoir du fait du hasard le même alignement (e-value : plus la e value est proche de 1 moins l'alignement est bon)

De nombreux algorithmes et programmes peuvent permettre de faire ces alignements (FASTA, BLAST sont les plus connus).

Exercise

Exercice 1
Soit deux séquences à aligner.

Séquence 1 : ATTGTTTC

Séquence 2 : ACCGTTG

Calculer le % de similitude (nombre de nucléotides identiques/ nombre de nucléotides totales)

Calculer le score d'alignement avec la règle suivante si il ya identité +2, non identité -1

Dc

II-2-5 Etudier la taxonomie (classification) et lire des arbres phylogéniques (en lien avec CBSV de première)

La taxonomie et les arbres phylogénique dépendent du contenu génétique pris en compte.

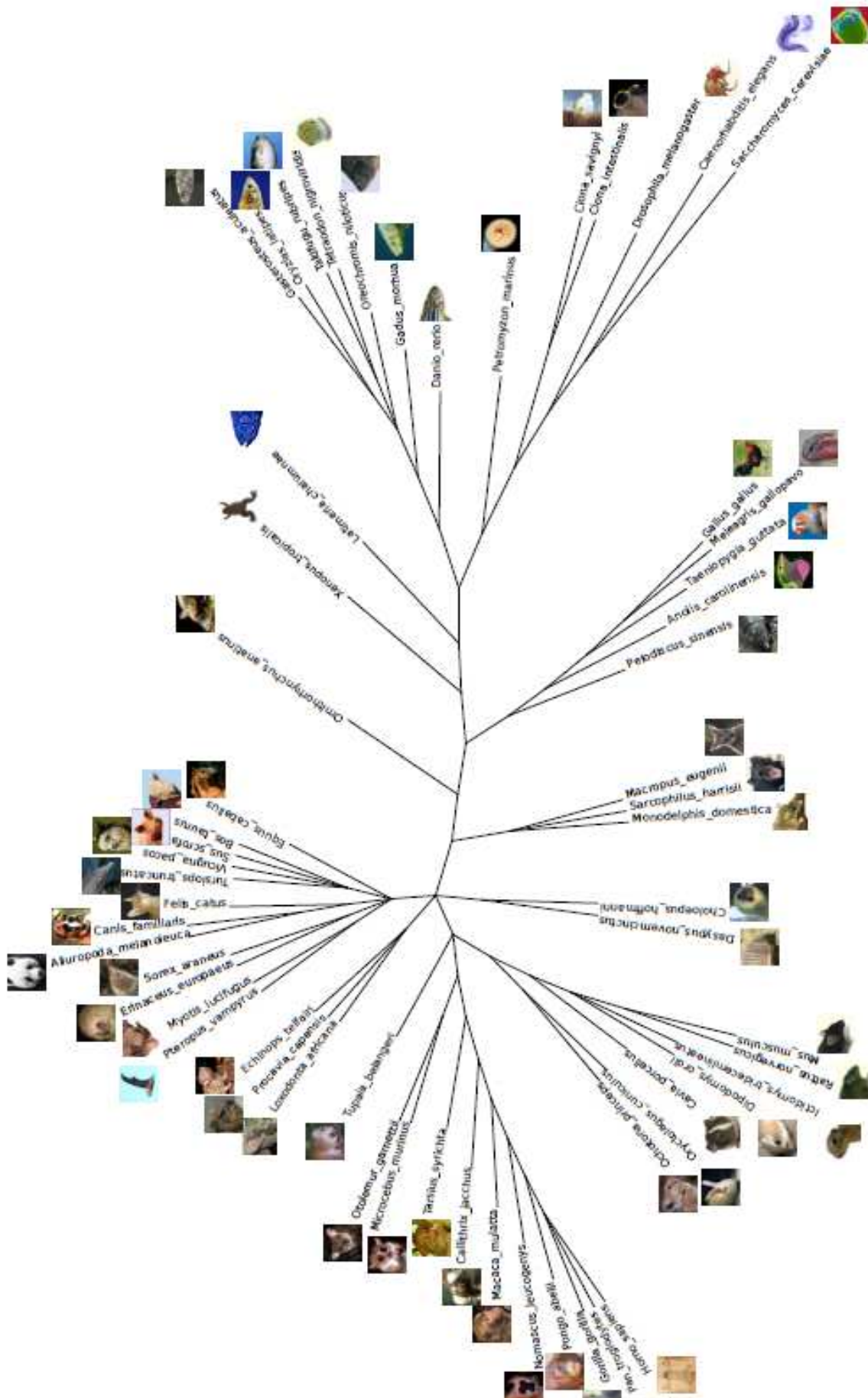


Image obtained using Dendroscope (D.H. Huson et al. "Dendroscope: An Interactive viewer for large phylogenetic trees", *BMC Bioinformatics* 8:460, 2007)

(bacille de Pfeiffer, gram – capsulé ou non, vaccination non obligatoire en fce mais svt dans vaccins associés : pentacoq, infanrix quinta, voies respiratoires ménigite voire septicémie)

le complémentaire de	5'- ACGT RY KM SW BDHV N-3'
est	3'- TGCA YR MK SW VHDB N-5'

% de similitude : $4/7 =$ plus proche de 1 mieux c'est

Score d'alignement $+2-1-1+2+2+2-1=5$ plus grand mieux c'est