

Licence 3 - DLSV316

Analyse des Séquences Génomiques

Matériel: <https://github.com/dgautheret/L3-ASG>

Objectifs du cours

- Rappels programmation R
- Extraction d'information de bases de données de séquences
- Faire et analyser sous R un alignement multiple de séquences ADN/protéines
- Manipuler des protéines en 3D
- Raisonner sur la relation entre conservation de séquence et fonction

Programme

	mar	mer	jeu	mar	mer	jeu	mar	mer	jeu	mar
	19-mars	20-mars	21-mars	26-mars	27-mars	28-mars	02-avr	03-avr	04-avr	09-avr
	Salle 315 au 336	Salle 315 au 336	Salle 315 au 336	Salle 224 au 336	Salle 315 au 336	Salle 223 au 336	Salle 315 au 336	Salle 315 au 336		Salle 315 au 336
8:15-10:15	DG		DG	DG			AL			AL
10:30-12:30		DG	AL		AL	AL		AL		
									Salle 223 au 336	
13:30-15:30						DG			DG	
15:45-17:45										
	Prise en main R	exercices R	Fin R . Banques de données génomiqu es	Lecture fasta R, calcul Entropie	Lecture fasta R, calcul Entropie	Entropie sur ali ADN	Traductio n ADN>prot. Entropie sur ali prot	Intro Pymol	Mapping entropie sur 3D. Visu 3D	

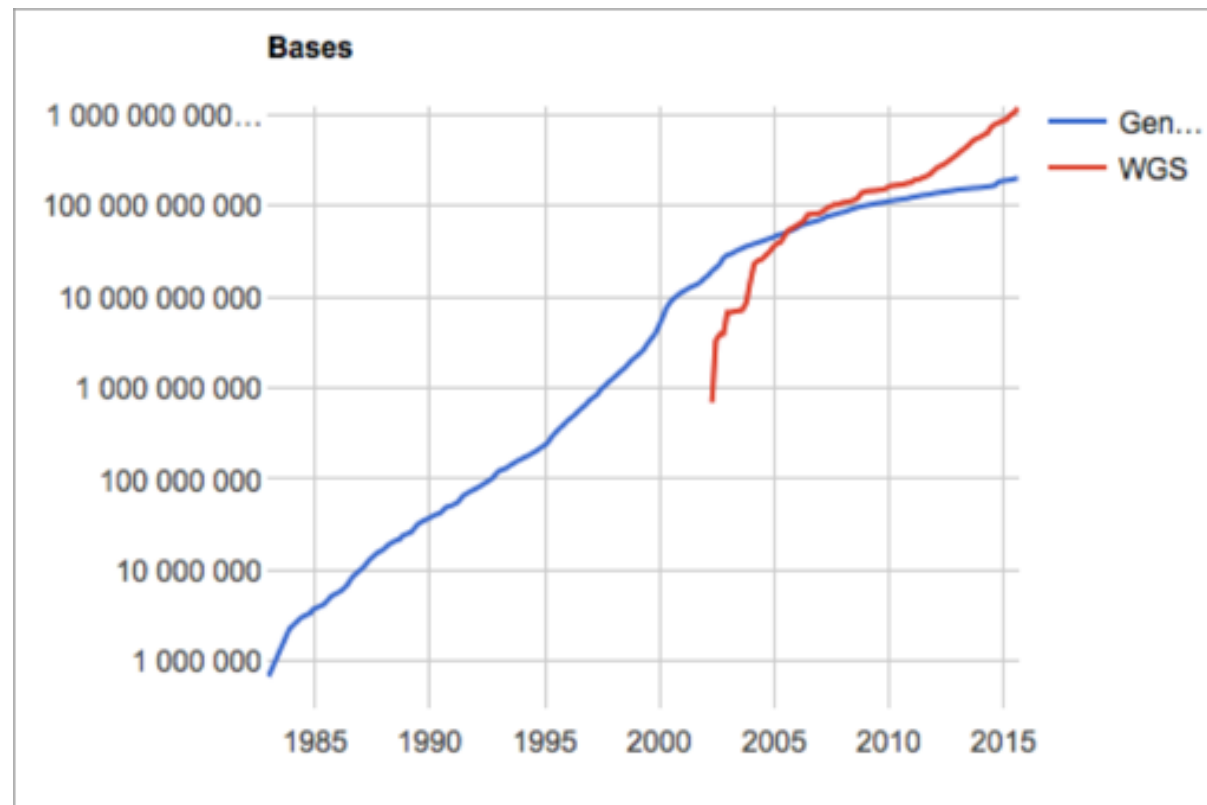
Les banques de séquences

Genbank: La banque d'ADN du NIH

- **Etat Genbank 2018**

- 253 Gbases
- 207M séquences
- Genbank double environ tous les 14 mois depuis ses débuts en 1982.
- Nouvelle version tous les 2 mois

(WGS: whole genome shotgun)



Identifiants Genbank

- Chaque enregistrement se voit attribuer un numéro d'accension, stable et unique, et chaque séquence un numéro de version (anciennement numéro GI.)
- Quand un changement est effectué dans un enregistrement Genbank, le num. d'accension reste, la version change.

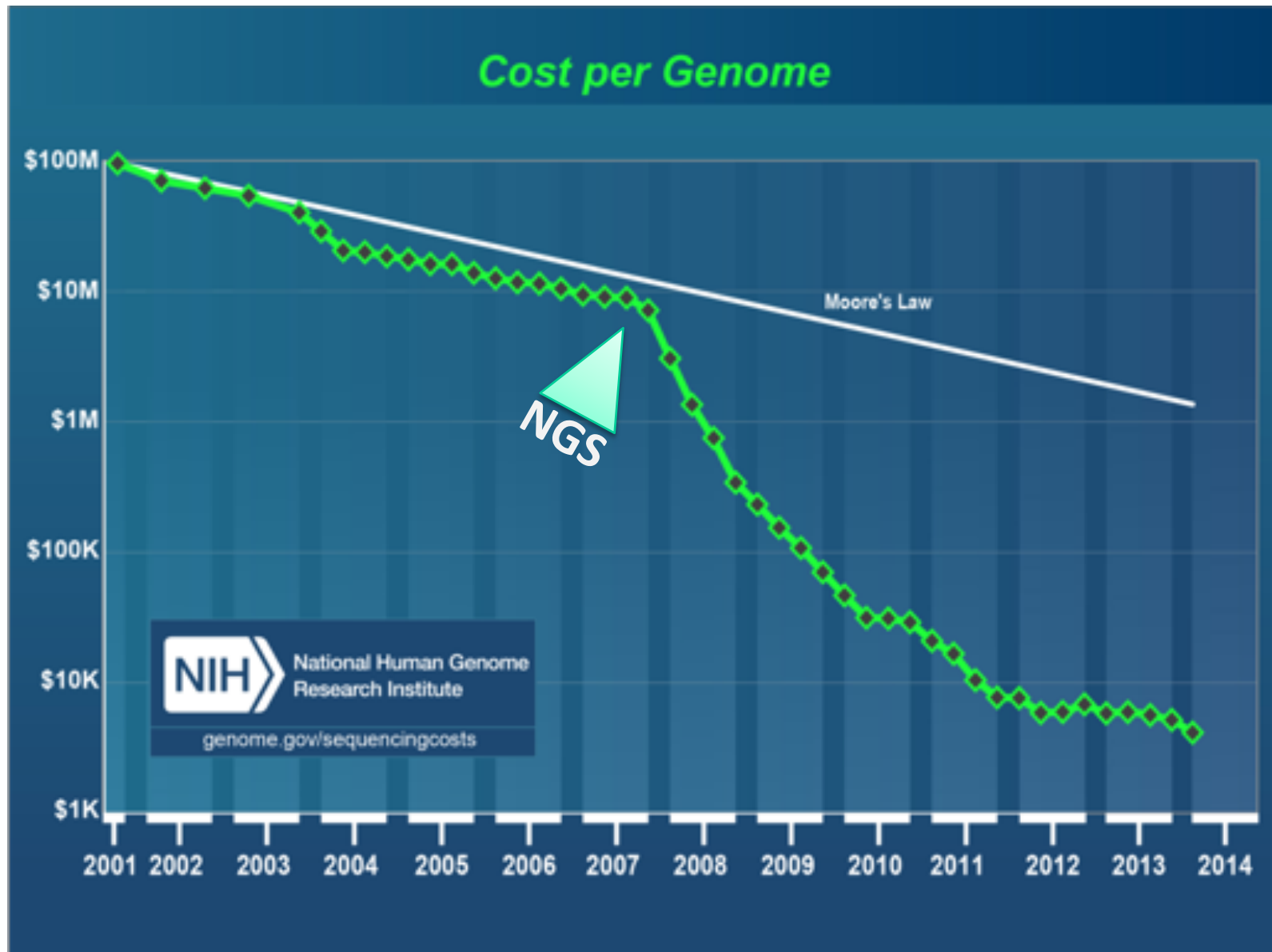
```
LOCUS NC_000913 4641652 bp DNA circular CON 08-AUG-2016
DEFINITION Escherichia coli str. K-12 substr. MG1655, complete genome.
ACCESSION NC_000913
VERSION NC_000913.3
```

Début de la fiche Genbank de E. coli

Autres banques nucléotidiques

- EMBL: Equivalent européen de Genbank. Format différent, contenu presque identique.
- DDBJ: équivalent au Japon
- Banques spécialisées Certaines collections de séquences, bien que généralement présentes dans Genbank, sont beaucoup plus utiles lorsqu'elles sont rassemblées dans des banques spécialisées, par ex:
 - Récepteurs des lymphocytes T (Réarrangements de l'ADN)
 - Génomes HIV, etc.
- SRA: short read archive...

Le bouleversement du «Next Generation Sequencing»



NGS sequencers



Nanopore
MinIon

50Mb



Lifetech Ion
torrent PGM

400 Mb



Illumina
MySeq

4 Gb



Lifetech Ion
proton

20 Gb



Illumina
Hi-Seq 2000

300 Gb

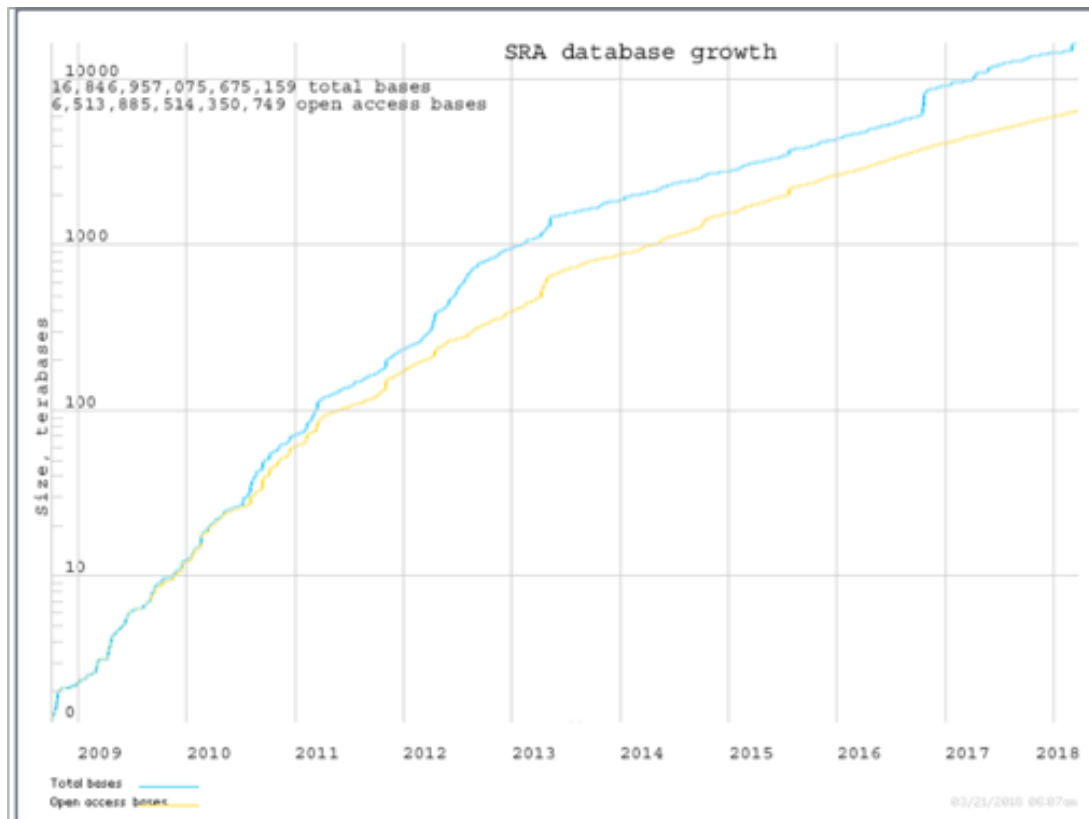


Illumina
NovaSeq

3Tb

The SRA (short read archive) database

- 3.200.000 entries (2018)
 - Each entry: ~50M sequences (DNA or cDNA fragments)



Banques protéiques

- Swissprot (UniProtKB/Swissprot).
 - La mieux annotée des banques protéiques. 2018: 550.000 entrées.
 - Curation par experts seulement (basé sur publis)
 - Attention: toutes les protéines connues n'y sont pas!
- TrEMBL (UniProtKB/TrEMBL):
 - banque protéique produite automatiquement par traduction banque EMBL. 2018: 108.000.000 entrées
- Uniprot=Swissprot+TrEMBL

Formats de données en bioinformatique

- La majorité des données de bioinfo sont de type texte:
 - FASTA, FASTQ, SAM, VCF, BED, GFF, GTF, TSV, CSV, WML, JSON
- Pour des raisons de performance et d'espace, certains sont en format binaire
 - BAM, VCF.GZ, FASTQ.GZ

Même donnée, différents formats

Format **JSON**

<https://tools.ietf.org/html/rfc4627>

```
users : {  
  first_name: "James", last_name:  
  "Watson", birthday: "1928-04-06"  
}
```

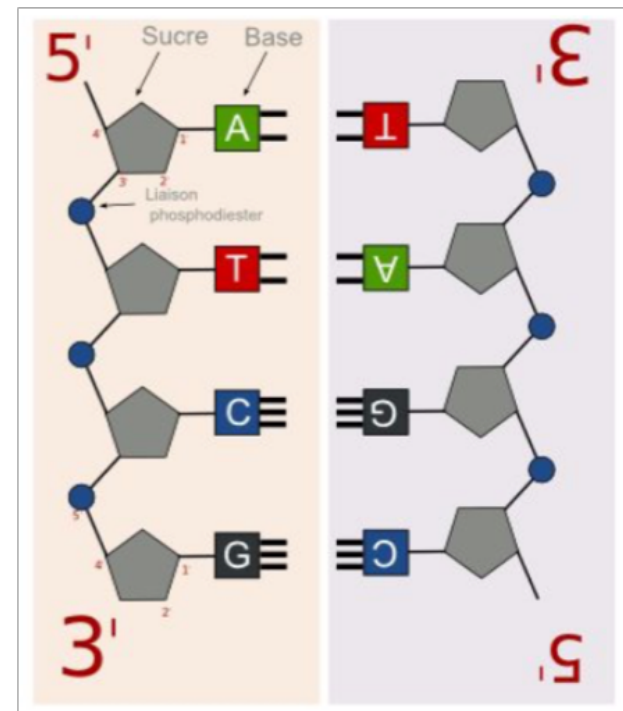
Format **XML**

<https://www.w3.org/TR/REC-xml/>

```
<users>  
  <first_name>James</first_name>  
  <last_name> Watson</last_name>  
  <birthday>19280406</birthday>  
</users>
```

Formats de séquences d'ADN

- Toujours dans le sens 5'→3'
- Sur quel brin?



Format fasta

*.fa , *.fasta

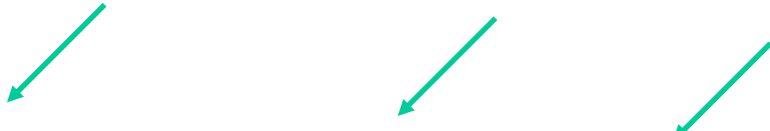
```
>identifiant1 commentaire libre  
CAGCATCGATCGTCGGCGATGCATGCGGATGCTAGCTGATCACGATGC  
CGCATGCTAGTCAGGCGGGAGGGATATTATTAGCGGCGTATCGGATGA  
CAGCATTTACGGCGGGAGTGCTATTATTTATGAGCGGCGAT  
>identifiant2 commentaire libre  
CAGGCGGAGGTTCTTTATTATATCGGCGGGCGGAGGCGGCGATGCATC  
CAGTGCACTGACGTAGTCAGCGATGCATTTTATGACTGACTCAGTTTT  
CCCGCTAGCTATGCTATGCTATTCGATCGATTTCGTGAGCTGATCTGGC  
CAGCTATGCTTTAGTA
```

Protein fasta (from Uniprot)

gene ID

species

full (english) name



```
>tr|Q4A489|Q4A489_BATTH Actin-1
CPESLFQPSFLGMESAGIHETTYNSIMKCDVDIRKDLYANTVLSGGTTMFPGIADRMQKE
ITALAPSTMKIKIIAPPERKYSVWIGGSILA
```


Format Genbank

```

LOCUS       L10986                47233 bp    DNA        linear    INV 21-SEP-2004
DEFINITION  Caenorhabditis elegans cosmid F10E9, complete sequence.
ACCESSION   L10986
VERSION     L10986.2  GI:38638818
KEYWORDS    HTG.
SOURCE      Caenorhabditis elegans
  ORGANISM  Caenorhabditis elegans
            Eukaryota; Metazoa; Nematoda; Chromadorea; Rhabditida;
            Rhabditoidea; Rhabditidae; Peloderinae; Caenorhabditis.
REFERENCE   1  (bases 1 to 47233)
AUTHORS     .
CONSRTM     WormBase Consortium
TITLE       Genome sequence of the nematode C. elegans: a platform for
            investigating biology. The C. elegans Sequencing Consortium
JOURNAL     Science 282 (5396), 2012-2018 (1998)
MEDLINE     99069613
PUBMED      9851916
FEATURES             Location/Qualifiers
     source          1..47233
                     /organism="Caenorhabditis elegans"
                     /mol_type="genomic DNA"
                     /strain="Bristol N2"
                     /db_xref="taxon:6239"
                     /chromosome="III"
                     /clone="F10E9"
     gene            265..26728
                     /gene="mig-10"
                     /locus_tag="F10E9.6"
     CDS             join(265..338,3266..3515,15194..15317,21507..21
                     21727..21887,23171..23335,24302..24472,24524..24608,
                     25012..25827,26284..26430,26478..26728)
                     /gene="mig-10"
     /translation="MDSCEECDLEVDSDEEDQLFGEKICISLLSLLPLSSSTLLSNA
                     INLELDEVERPPPLLNVLLEEQQFPKVCANIEEENELEADTEEDIAETADDEESKDPVE
                     KTFNFEPSPVMTDYDFDPYPVQIRARPQVPKPPIDTVRYSMNNIKESADWQLDELL
                     EELEALETQLNSSNGDQLLLGVSGIPASSSRENVKSISTLPPPPPALSYHQTPQQPQ
                     . . .
                     QVYTIGIGWEKKYKSPTPWCISIKLTALQMKRSQFIKYICAEDEMTFKKWLVALRIAKN
                     GAELLENYERACQIRRETLGPASSMSAASSSTAISEVPHSLSHHQRTSPSVASSIQLSS
                     HMMNNPTHPLSVNVRNQSPASFVNSCQQSHPSRTSAKLEIQYDEQPTGTIKRAPLDV
                     LRRVSRASTSSPTIPQEESSDSDEEFAPPPVAVSMRMPPVTPPKPCTPLTSKKAPPP
                     PPKRSDTTKLQSASPMAPAKNDLEAALARRREKMATMEC"
     . . .

```

```

BASE COUNT      2598 a    2024 c    1888 g    2449 t
ORIGIN
   1  ttctaaaaagt  cgaaaaaacga  gcaatttttg  atgctagatt  ttttgatttg  acgaattttt
  61  tcagttttttt  ttctttaaaa  aaggtttttg  accccttaaa  gttttccttt  cccttccaat
 121  tttttccttc  tttcttatac  gacttctcaa  gtttcaactc  taaaacaaag  ctacatgtac
 181  atttcoggta  aacttttgtg  ctccagaagt  ccattttctt  tttgttacat  ttattcaaga
 241  ttgaattcca  aaatttcagc  caatatggac  agttgcgaag  aggaatgcga  tctggaagtt
 301  gacagtgcgc  aagaagatca  actttttggt  gaaaagtggg  gagtctctat  tgtggtaacc
 361  aaagaaatgt  cagtggtccg  taaacacttg  actcccaaat  ggttctcgt  aattacctta
 421  tgcacacttt  tcaagtgttt  gccgtttgat  cttagccaat  ttgaaacggt  tagatgttaa
 481  atggaaaatg  ggtaaagtgt  tttattttat  agaaaaaagg  tttggaaaaa  aatcgagtca
 541  ctgaatagtt  tgaagaacgg  aaaaataaaa  ctttccaaaa  atcataaaac  atttagtggt
 601  tcgaaaatta  tagtggtttt  tttgttggtg  tgttttgaca  aaagctaaac  catctttatt
 661  gtagttttgt  aaaatgttca  caaagatgcg  ttttttttcc  aaatttgcca  ggctatcttt
 721  acattcacat  ttggataatt  caaatttttc  ttatcgctaa  caaattttcc  tatttttcca
 781  attattcggt  ttataaaagc  tttggtagta  tgttgtgtct  atcttttagt  gtcacagtt

```

Format fastq



Sequenceur NGS

Descriptif du read (position sur la piste de séquençage, taille,..)

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

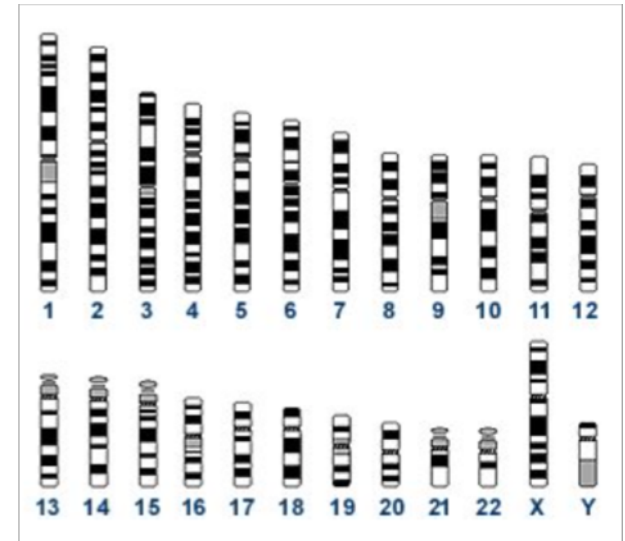
Qualité (probabilité que la base soit correcte) encodé par code ASCII

Les régions

Les coordonnées génomiques
permettent de définir une
région exacte du génome

<chromosome>:<start>-<end>

chr7:117465784-117715971



Accéder directement à une région

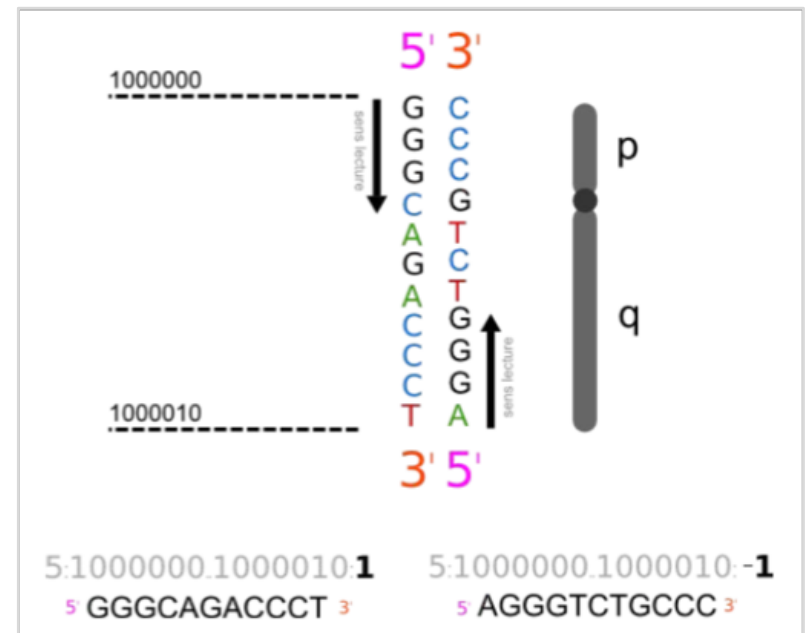
Via le browser Ensembl: <https://www.ensembl.org>

Puis choisir region du génome humain: 7:117465784..117715971

Via une URL:

<http://rest.ensembl.org/sequence/region/human/7:117465784..117715971:-1>

Attention aux versions
d'assemblage du
génom (Hg19, Hg38..)



Exercice Genbank

Récupérez sur le site du NCBI *dans la* section **nucleotides** l'accension NC_002549.1, au format gb

- De quelle séquence s'agit-il?
- Visualisez gènes, protéines, séquences régulatrices
- 10 premières bases du gène 1?

Exercices

Bioinformatics

vs.

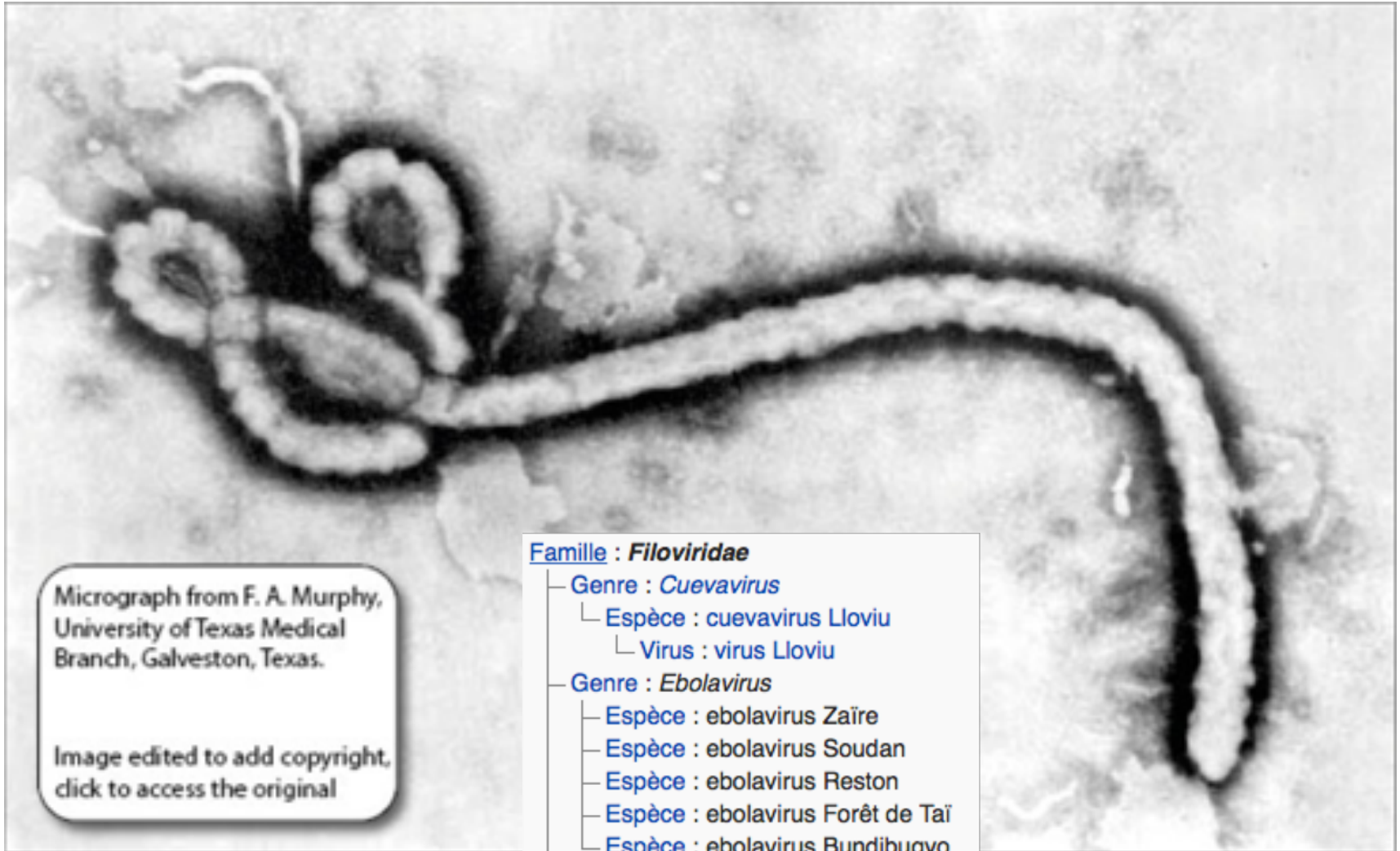
EbolaVirus

La maladie à virus Ebola

- Virus transmis à l'homme à partir d'animaux sauvages (chauve-souris, primates), puis entre humains
- Fièvre hémorragique
- Transmission: par contact fluides/muqueuses
- Taux de léthalité moyen de 50%
- Aucun vaccin ni traitement



EbolaVirus (EBOV): famille des filovirus



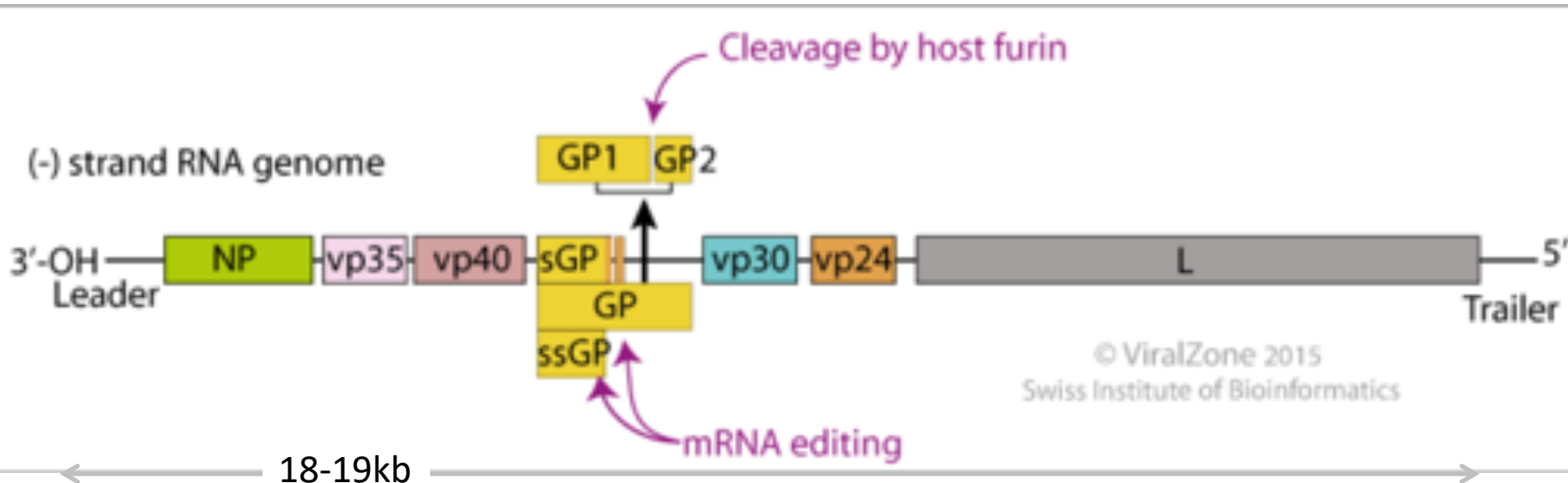
Micrograph from F. A. Murphy,
University of Texas Medical
Branch, Galveston, Texas.

Image edited to add copyright,
click to access the original

Famille : **Filoviridae**

- Genre : *Cuevavirus*
 - Espèce : cuevavirus Lloviu
 - Virus : virus Lloviu
- Genre : *Ebolavirus*
 - Espèce : ebolavirus Zaïre
 - Espèce : ebolavirus Soudan
 - Espèce : ebolavirus Reston
 - Espèce : ebolavirus Forêt de Taï
 - Espèce : ebolavirus Bundibugyo
- Genre : *Marburgvirus*
 - Espèce : marburgvirus Marburg
 - Virus : virus Marburg
 - Virus : virus Ravn

EBOV: un virus à ARN brin (-)



<http://viralzone.expasy.org/>

NP	nucléoprotéine
VP35	co-facteur RNA-pol
VP40	protéine de matrice
xGPx	glycoprotéines
VP30	activateur transcription
VP24	protéine de matrice
L	RNApol RNA-dépendante + coiffe et polyA

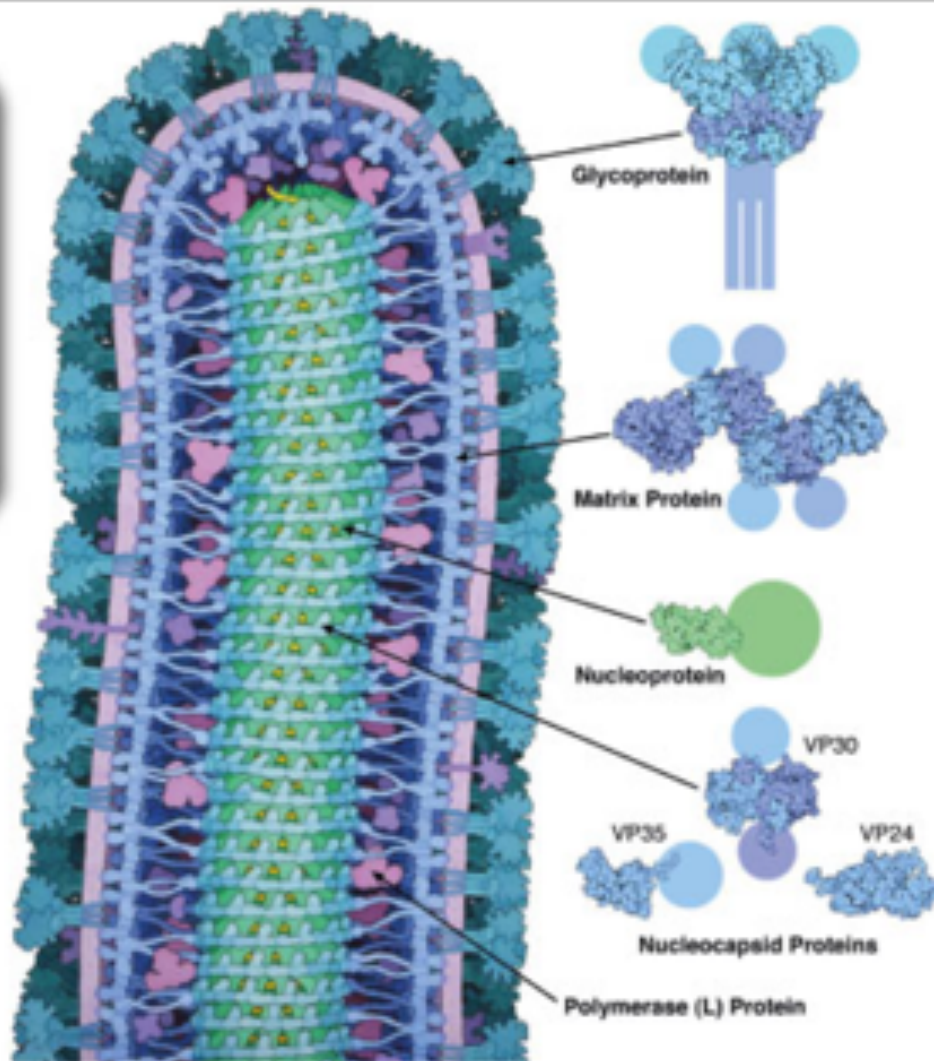
Composition d'EBOV

Ebolavirus Proteins

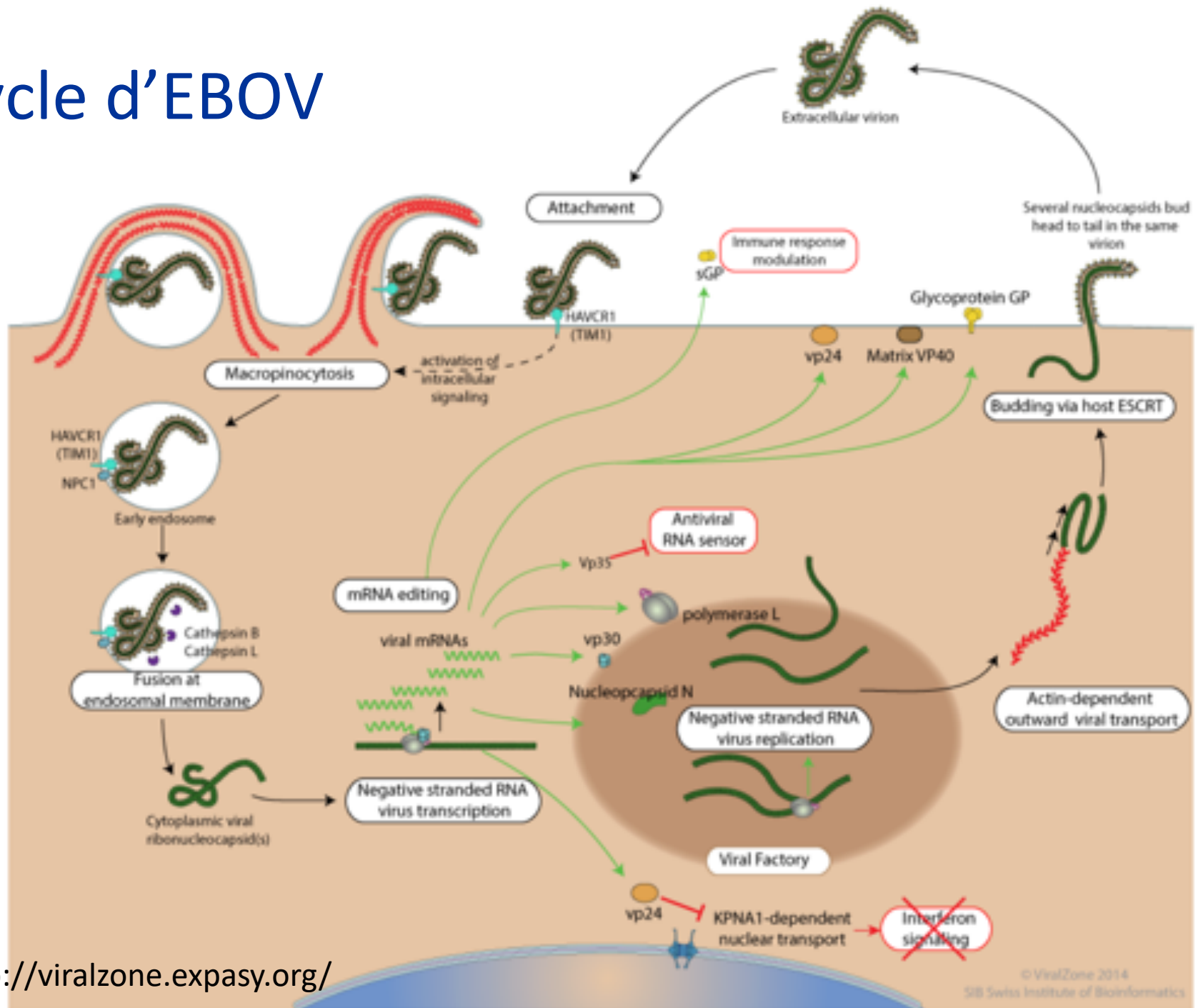
RCSB PDB-101

Credits: David S. Goodsell
and the RCSB PDB

Image edited to add copyright,
click to access the original

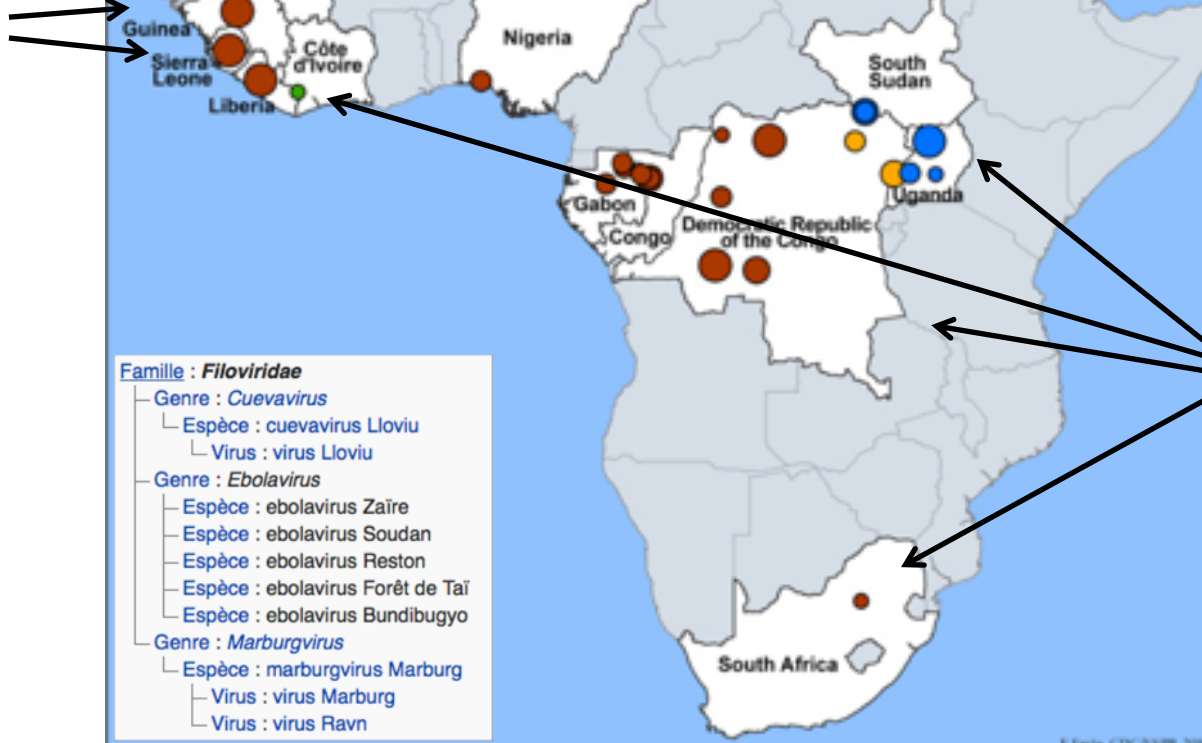


Cycle d'EBOV



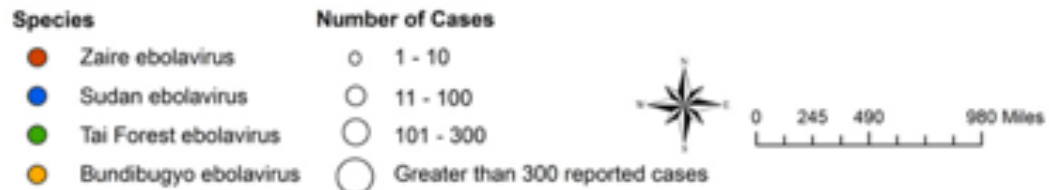
Les épidémies

Epidémie 2014
(Sierra Leone, Libéria,
Guinée):
28000 cas, 11000 décès



Epidémies
1976-2012

EBOLAVIRUS OUTBREAKS BY SPECIES AND SIZE, 1976 - 2014



Mini-Plateforme de séquençage EBOV

Déployée en 2015 en Guinée
Séquenceur MinION



Sequencing Ebolavirus in Guinea. A researcher prepares samples on the right, while MinIONs plugged into laptops are visible on the left.

IMAGE COURTESY OF EUROPEAN MOBILE LAB. PHOTOGRAPH BY TOMMY TRENCHARD

Un Séquenceur MinION



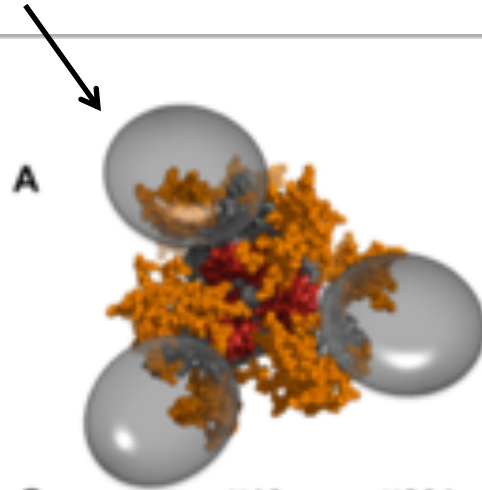
2016: 1693 génomes
complets ou partiels
séquencés

Les glycoprotéines GP

- Protéines d'enveloppe
- Permettent l'attachement du virus aux récepteurs de l'hôte
- Sont exprimées en surface de la cellule-hôte
- Induisent une réponse immunitaire
 - Développement de vaccin
 - Comprendre l'évolution/adaptation du virus

Structure de la GP

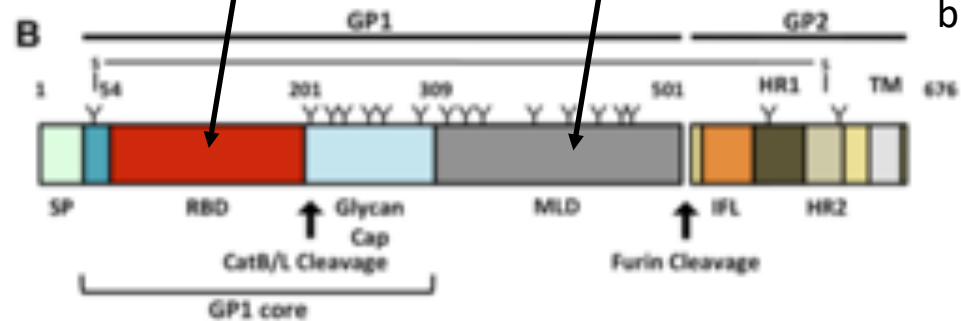
Trimère de GP1/GP2 en surface du virion



receptor-binding
domain (RBD)
très conservé

Glycan cap & mucin-like
domain (MLD): très
variables

GP2: dom.
transmem-
branaire



Panel C: Alignment of predicted N-linked glycan sites within the GP1 core of five Ebola virus species. N-X-S/T sequons are highlighted with a black background.

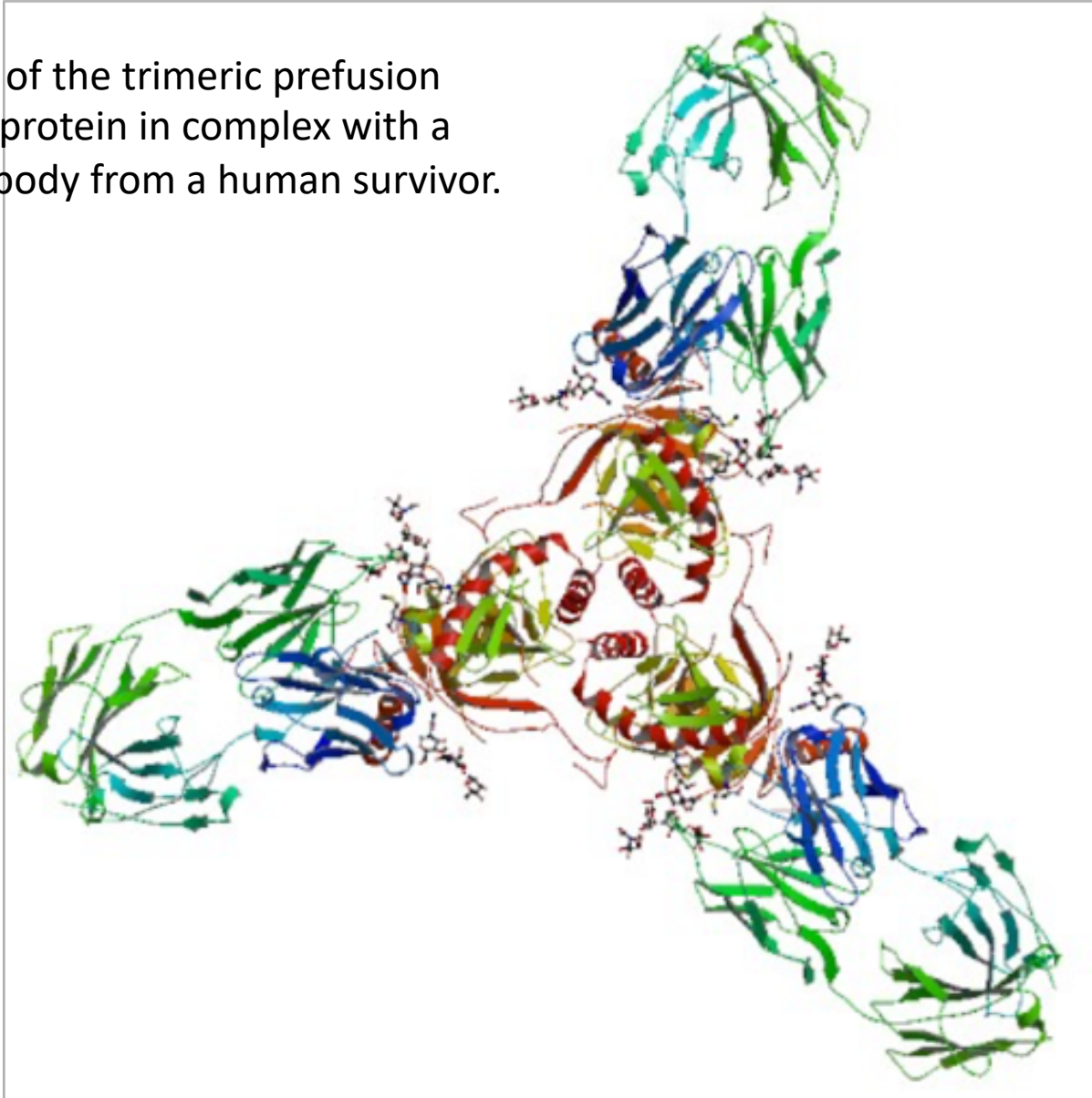
	N40	N204	N228	N238	N257	N268	N296
Ebola	IHNSTLQ	PVNATEDPSS	GTNETETLFEVDNLTIV	QINETIYTSGKRNTTGK	QINETIYTSGKRNTTGK	QINETIYTSGKRNTTGK	KKNLTRK
Sudan	VINSTLE	AVNYTENTSS	GAQHSTTLFKINNNTFV	QINDTIHLHQQLSNTTGK	QINDTIHLHQQLSNTTGK	QINDTIHLHQQLSNTTGK	KKNLSEQ
Tai Forest	VHNNTLQ	PANMTDPSS	GTNTTEFLFQVDHLTYV	LI NETIYSDNRRSNTTGK	LI NETIYSDNRRSNTTGK	LI NETIYSDNRRSNTTGK	KKNFTKT
Bundibugyo	VHNNTLQ	PANMTDPSS	GTNMTNPLFQVDHLTYV	QINETIYTNGRRSNTTGK	QINETIYTNGRRSNTTGK	QINETIYTNGRRSNTTGK	KKNFTKT
Reston	VINSTLK	PVNTTDDSTS	GGNESNTLFPVSNHTIV	QINETLRRNNRLSNSTGR	QINETLRRNNRLSNSTGR	QINETLRRNNRLSNSTGR	KKNFSQQ

Sites de
fixation
glycanes

FIG 1 Schematic diagrams of Ebola virus GP. (A) A molecular model of EBOV GP1/2 shown in a top-down view. Complex N-glycans are shown in orange, GP is shown in light gray, RBD is shown in red, and MLD structure that has not been solved is represented as a gray sphere. PDB ID [3CSY](#). (B) Linear model of EBOV GP. The disulfide bond between GP1 and GP2 is indicated, as well as the locations of N-linked glycans (marked with "Ys") in the GP1 and -2 domains, and the known protease cleavage sites are noted. SP, signal peptide; RBD, receptor-binding domain; MLD, mucin-like domain; IFL, internal fusion loop; HR1 and -2, heptad repeats 1 and 2; TM, transmembrane domain. (C) Alignment of predicted N-linked glycan sites within the GP1 core of the five Ebola virus species. N-X-S/T sequons are highlighted with a black background.

3CSY

Crystal structure of the trimeric prefusion Ebola virus glycoprotein in complex with a neutralizing antibody from a human survivor.
Lee et al. 2008



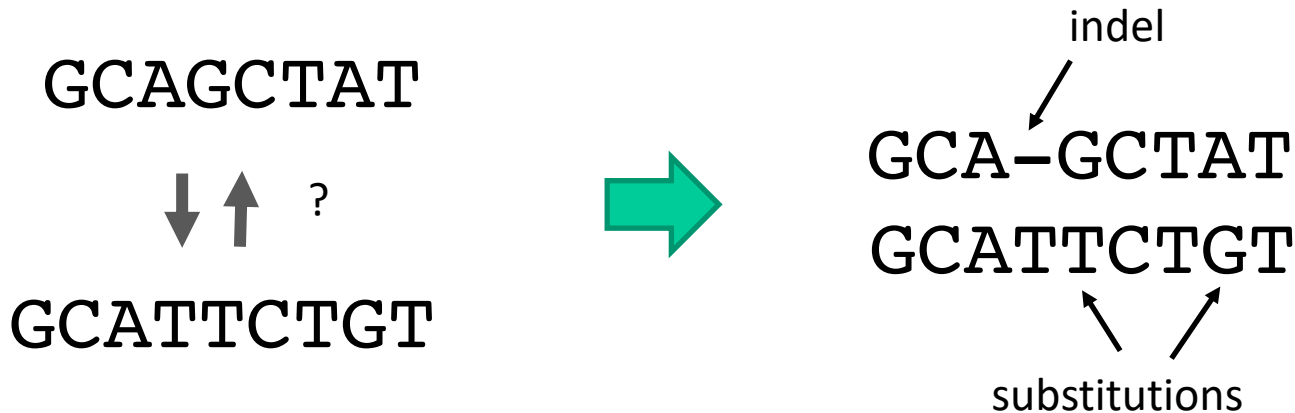
Questions

- Variation des GP au cours de l'évolution de EBOV
- Positions les plus variables dans la GP?
- Variations ADN vs. variations protéine
- A quoi correspondent ces positions sur la structure tridimensionnelle?

Notions de comparaison de séquences

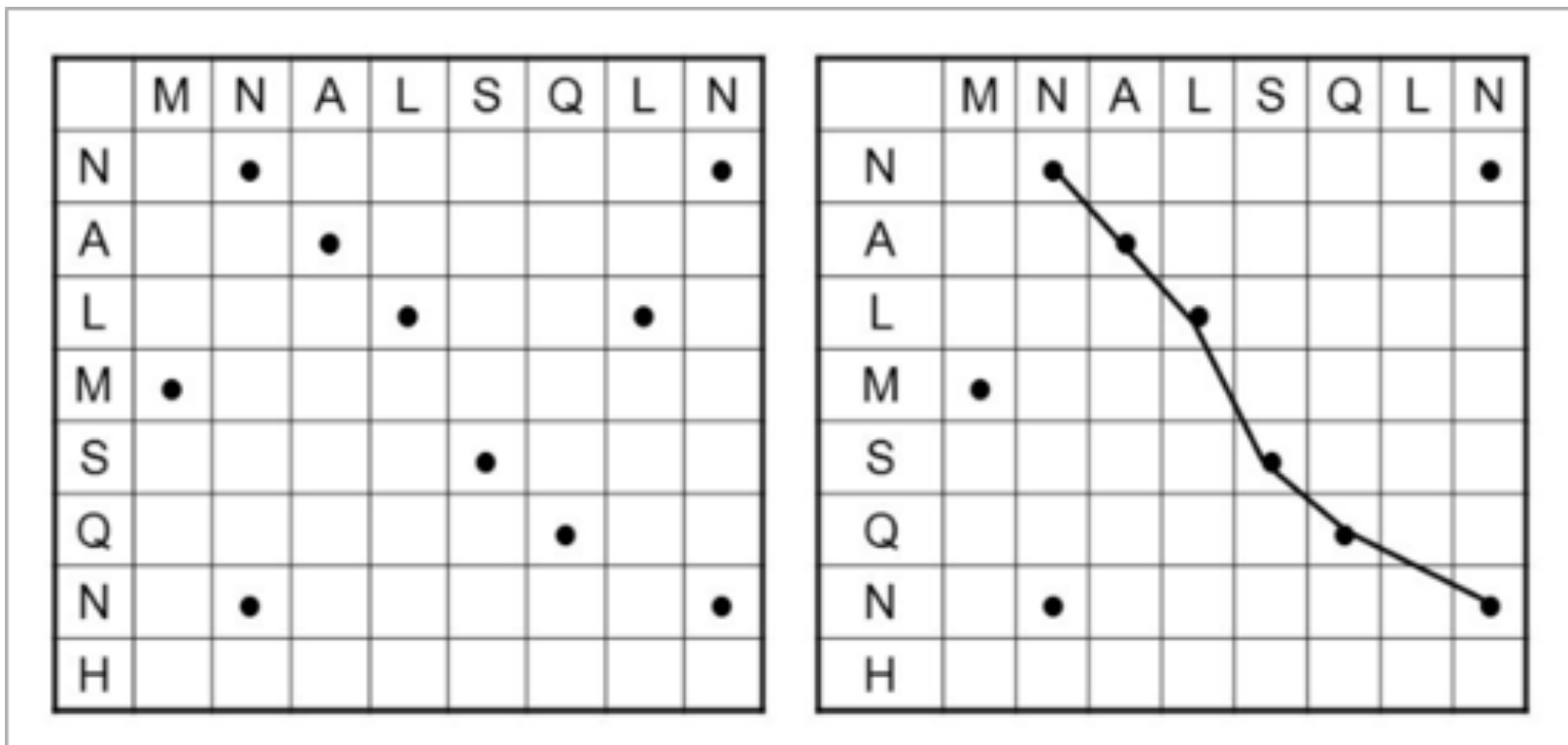
Comparaison de séquences

Quel est le taux de ressemblance entre 2 séquences?



Comparer = trouver l'alignement optimal

Un alignement entre 2 séquences



Trouver le chemin optimal maximisant les identités

Alignement global

- Algorithme de Needleman & Wunsch
 - Programmation dynamique
 - Trouve l'alignement de score optimal
- Score pour
 - Résidus identiques 😊
 - Substitution 😞
 - Indel 😞

Matrices de Substitution

- Matrice 4X4 (nt) ou 20x20 (aa) décrivant la distance ou la similitude entre résidus.
- Estiment le coût ou le taux de remplacement d'1 résidu par un autre (distance).
- Le choix d'une matrice affecte fortement le résultat de l'analyse. Chaque matrice de score représente implicitement une théorie évolutive donnée

Matrices DNA

	A	C	G	T
A	1	0	0	0
C	0	1	0	0
G	0	0	1	0
T	0	0	0	1

Matrice identité

	A	C	G	T
A	3	0	1	0
C	0	3	0	1
G	1	0	3	0
T	0	1	0	3

Matrice transition/transversion

Protéines: matrice de Dayoff (1979)

A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Z	
0.4	0.0	-0.4	0.0	0.0	-0.8	0.2	-0.2	-0.2	-0.2	-0.4	-0.2	0.0	0.2	0.0	-0.4	0.2	0.2	0.0	-1.2	-0.6	0.0	A
	0.5	-0.9	0.6	0.4	-1.0	0.1	0.3	-0.4	0.1	-0.7	-0.5	0.4	-0.2	0.3	-0.1	0.1	0.0	-0.4	-1.1	-0.6	0.4	B
		2.4	-1.0	-1.0	-0.8	-0.6	-0.6	-0.4	-1.0	-1.2	-1.0	-0.8	-0.6	-1.0	-0.8	0.0	-0.4	-0.4	-1.6	0.0	-1.0	C
			0.8	0.6	-1.2	0.2	0.2	-0.4	0.0	-0.8	-0.6	0.4	-0.2	0.4	-0.2	0.0	0.0	-0.4	-1.4	-0.8	0.5	D
				0.8	-1.0	0.0	0.2	-0.4	0.0	-0.6	-0.4	0.2	-0.2	0.4	-0.2	0.0	0.0	-0.4	-1.4	-0.8	0.6	E
					1.8	-1.0	-0.4	0.2	-1.0	0.4	0.0	-0.8	-1.0	-1.0	-0.8	-0.6	-0.6	-0.2	0.0	1.4	-1.0	F
						1.0	-0.4	-0.6	-0.4	-0.8	-0.6	0.0	-0.2	-0.2	-0.6	0.2	0.0	-0.2	-1.4	-1.0	-0.1	G
							1.2	-0.4	0.0	-0.4	-0.4	0.4	0.0	0.6	0.4	-0.2	-0.2	-0.4	-0.6	0.0	-0.4	H
								1.0	-0.4	0.4	0.4	-0.4	-0.4	-0.4	-0.4	-0.2	0.0	0.8	-1.0	-0.2	-0.4	I
									1.0	-0.6	0.0	0.2	-0.2	0.2	0.6	0.0	0.0	-0.4	-0.6	-0.8	0.1	K
										1.2	0.8	-0.6	-0.6	-0.4	-0.6	-0.6	-0.4	0.4	-0.4	-0.2	-0.5	L
											1.2	-0.4	-0.4	-0.2	0.0	-0.4	-0.2	0.4	-0.8	-0.4	-0.3	M
												0.4	-0.2	0.2	0.0	0.2	0.0	-0.4	-0.8	-0.4	0.2	N
													1.2	0.0	0.0	0.2	0.0	-0.2	-1.2	-1.0	-0.1	P
														0.8	0.2	-0.2	-0.2	-0.4	-1.0	-0.8	0.6	Q
															1.2	0.0	-0.2	-0.4	0.4	-0.8	0.6	R
																0.4	0.2	-0.2	-0.4	-0.6	-0.1	S
																	0.6	0.0	-1.0	-0.6	-0.1	T
																		0.8	-1.2	-0.4	-0.4	V
																			3.4	0.0	-1.2	W
																				2.0	-0.8	Y
																					0.6	Z

Matrice dérivée des substitutions observées dans les régions bien conservées des protéines

chaque case représente la probabilité de voir ces deux résidus remplacés l'un

Matrice dérivée des substitutions observées dans les régions bien conservées des protéines

Chaque case représente la probabilité de voir ces deux résidus remplacés l'un par l'autre dans un alignement. (matrice lod-score, de "log-odds" ou "log des chances").

$$S = \log (F_{ij} / (F_i \times F_j))$$

Alignement multiple

Q5E940_BOVIN	-----MPREDRATWKSNYFLKIIQLLDYPKCFIVGADNYGSKOMQDIEMSLRGK-AVVLMSGKNTMMRKAIRGHLENN--PALE	76
RLA0_HUMAN	-----MPREDRATWKSNYFLKIIQLLDYPKCFIVGADNYGSKOMQDIEMSLRGK-AVVLMSGKNTMMRKAIRGHLENN--PALE	76
RLA0_MOUSE	-----MPREDRATWKSNYFLKIIQLLDYPKCFIVGADNYGSKOMQDIEMSLRGK-AVVLMSGKNTMMRKAIRGHLENN--PALE	76
RLA0_RAT	-----MPREDRATWKSNYFLKIIQLLDYPKCFIVGADNYGSKOMQDIEMSLRGK-AVVLMSGKNTMMRKAIRGHLENN--PALE	76
RLA0_CHICK	-----MPREDRATWKSNYFMKIIQLLDYPKCFVVGADNYGSKOMQDIEMSLRGK-AVVLMSGKNTMMRKAIRGHLENN--PALE	76
RLA0_RANBY	-----MPREDRATWKSNYFLKIIQLLDYPKCFIVGADNYGSKOMQDIEMSLRGK-AVVLMSGKNTMMRKAIRGHLENN--PALE	76
Q7ZUG3_BHARE	-----MPREDRATWKSNYFLKIIQLLDYPKCFIVGADNYGSKOMQDIEMSLRGK-AVVLMSGKNTMMRKAIRGHLENN--PALE	76
RLA0 ICTPU	-----MPREDRATWKSNYFLKIIQLLDYPKCFIVGADNYGSKOMQDIEMSLRGK-AVVLMSGKNTMMRKAIRGHLENN--PALE	76
RLA0_DHOME	-----MYRENKAANKAQYFIKVVLFDDEPKCFIVGADNYGSKOMQDIEMSLRGK-AVVLMSGKNTMMRKAIRGHLENN--PALE	76
RLA0_DICDI	-----MSGAG-SKSKKLFIEKATKLTFTYDKMIVAEADPYGSSQLQKIKSINGI-GAVLMSGKNTMIRKIVIRDLADSK--PELD	75
Q54LP0_DICDI	-----MSGAG-SKSKNVFIEKATKLTFTYDKMIVAEADPYGSSQLQKIKSINGI-GAVLMSGKNTMIRKIVIRDLADSK--PELD	75
RLA0_PLAFB	-----MAKLSKQKQKQMYIEKLSSELQQYSKILIVHVDNYGSKOMASVEKSLGK-ATILMSGKNTMIRKIVIRDLADSK--PELD	76
RLA0_SULAC	-----NIGLAVTTTCKIAKWKYDEVAELTCKLTKTKTIIIANIEGFPADKLHEIEKKLRGK-ADIKVTKNNLFIKIALKNAG----YDCK	79
RLA0_SULTO	-----MRIMAVITQERKIAKWKIEEVKELECKLREHTIIIANIEGFPADKLHDIKKMREGM-AEIKVTKNTLFGIAAKNAG----LDVS	80
RLA0_SULSO	-----MKRLALALKQRKVASWKEEVKELTELKNSNTILIGNLEGFPADKLHEIEKKLRGK-ADIKVTKNTLFGIAAKNAG----IDIE	80
RLA0_AERPE	MSVYSLYGQMYKREKPIENKTLMLRELEELFSKHVVLPADLTGTPFVYGRVEKKLWKK-YDMMVAKKRIILHAKKAAGLE----LDDN	86
RLA0_PYRAE	NMLAIGKRRYVTRQYPAKVKIVSEATELLQKTPYVFLFDLGLSLERILHEIKYRLRRY-GYIKIIPPLFKIAFTKVYGG---IPAL	85
RLA0_METAC	-----MAERHNTENIPQKKDEIENIKELIQSHKVFQMGVIGILATKMKIARDLQDV-AVLKVSNTLIERALNQLG----ETIP	78
RLA0_METMA	-----MAERHNTENIPQKKDEIENIKELIQSHKVFQMGVIGILATKMKIARDLQDV-AVLKVSNTLIERALNQLG----ESIP	78
RLA0_ARCFU	-----MAAVRGS--PDTEVRAVEEIKRMISSEKPVVAIVSFRNYPAGOMKIEREFRGK-AEIKVTKNTLIERALDALG----GDYL	75
RLA0_METKA	MAYKAKGQPPSGYEPKVAEWKRREVKELKELMDEYENYGLVDLEGIPAFOLQEIKAKLREDTIIRMSNTLMRIALEEKLDER--PELE	80
RLA0_METTH	-----MAHVAEWKKKEVQELNDLIKQYEVVGIANLADIPAROLQKMQTLDS-ALIRMSKFLISLAKKAGREL--ENVQ	74
RLA0_METTL	-----MITAESENKIAPWKIEEVNKLKELLKNGQIVALVDVSMMPAYPLSQMRILIRENGLLLVSRNTLIELAIKKAAGELGKPELE	82
RLA0_METVA	-----MIDAKSENKIAPWKIEEVNALKELLKSANVIALIDVSMMPAYPLSQMRILIRENGLLLVSRNTLIELAIKKAAGELGKPELE	82
RLA0_METJA	-----METKVKAHVAPWKIEEVKTLKGLIKSKPVVAIVDMDVPAFOLQEIARDKIR-DEVKLRMSNTLIELAIKKAAGELGKPELE	81
RLA0_PYRAB	-----MAHVAEWKKKEVEELANLIKSTPVIALVDVSMMPAYPLSQMRILIRENGLLLVSRNTLIELAIKKAAGELGKPELE	77
RLA0_PYRNO	-----MAHVAEWKKKEVEELAKLIKSTPVIALVDVSMMPAYPLSQMRILIRENGLLLVSRNTLIELAIKKAAGELGKPELE	77
RLA0_PYRFU	-----MAHVAEWKKKEVEELANLIKSTPVIALVDVSMMPAYPLSQMRILIRENGLLLVSRNTLIELAIKKAAGELGKPELE	77
RLA0_PYRKO	-----MAHVAEWKKKEVEELANLIKSTPVIALVDVAGVPAYPLSKMDEKLRE-KALLVSRNTLIELAIKKAAGELGKPELE	76
RLA0_HALMA	-----MSAEERKTTETIPENKQEEVDIVMIESYESVGVVNIAGIPEROLQDMHDLNGT-AELVSRNTLIELALDDVD----DGLE	79
RLA0_HALVO	-----MSSEYRQTEYIPQKKREEVDLYDFIESYESVGVVGVAGIPEROLQDMHDLNGT-AELVSRNTLIELALDEVM----DGFE	79
RLA0_HALSA	-----MSAEQRTTEVPENKQEEVDLYDFIESYESVGVVGVAGIPEROLQDMHDLNGT-AELVSRNTLIELALDEVM----DGLE	79
RLA0_THEAC	-----MKEYSQQEKELYNEITIRIKASHVAIVDAGIREROLQDMHDLNGT-AELVSRNTLIELALDEVM----DGLE	72
RLA0_THEVO	-----MRKINPKKEIVSELAADITKSKAVAVDIXGVYIRQMDQIRAKNEDK-VKIKVVKELIFKALDSIND----EKLT	72
RLA0_PICTO	-----MTEPAQWKIDFVKNLENEINSEKYAAIVSIXGLRNNIFQKIKSINGIARDK-ARIKVSARLLRLAIENIGK---NNIV	72
ruler	1.....10.....20.....30.....40.....50.....60.....70.....80.....90	

Les usages des alignements multiples

- Phylogénie moléculaire
 - via calcul de distance
- Prédiction 3D
 - Par homologie avec autres protéines de repliement 3D connu
- Inférence fonctionnelle
 - Positions conservées=fonction conservée

Un arbre phylogénétique réalisé à partir d'un alignement de séquences de gènes GP35 de Ebola & Marburg V.

700-1300 ans

Suzuki & Gojobori 1997

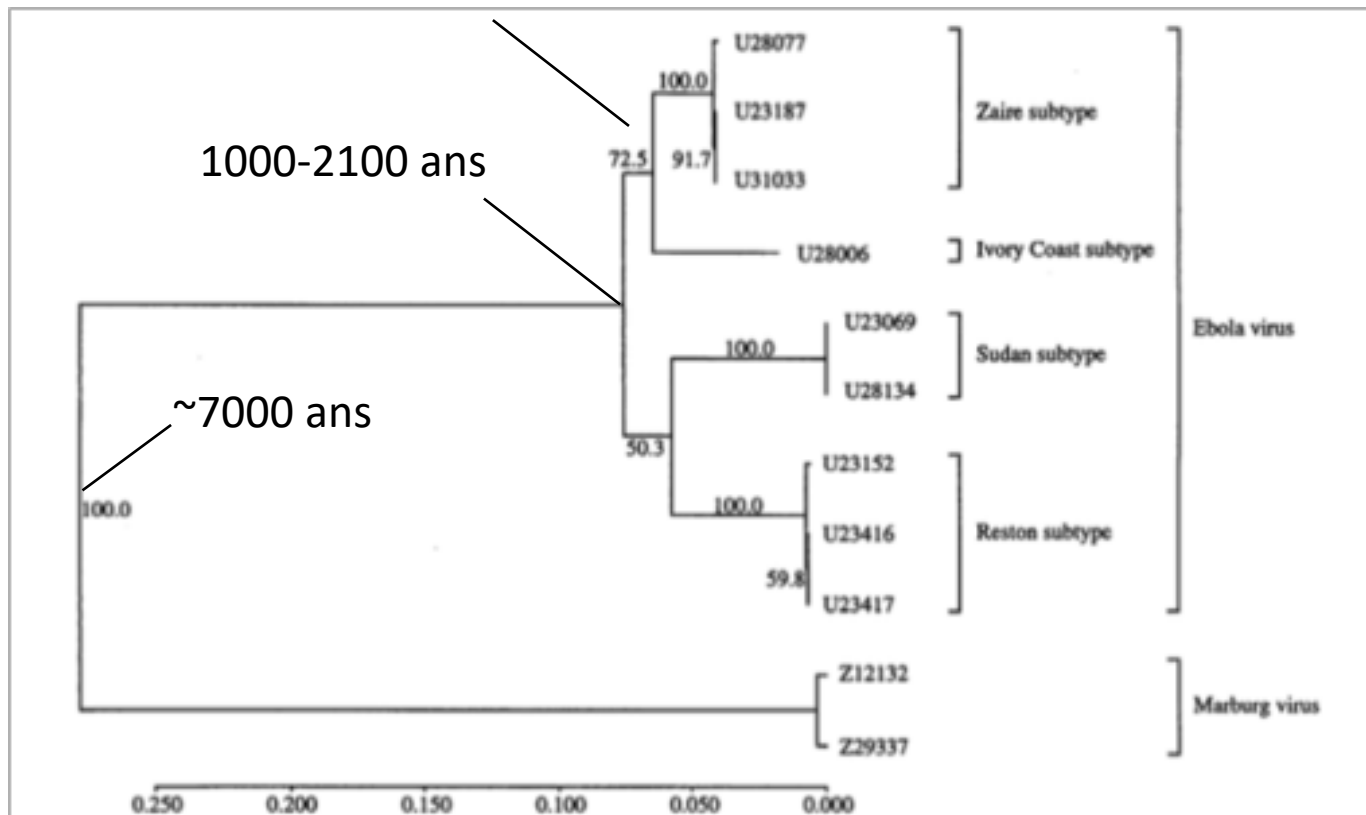


FIG. 1.—Phylogenetic tree constructed for the GP gene of Ebola and Marburg viruses by the neighbor-joining method (Saitou and Nei 1987), with distances for nonsynonymous sites estimated by the method of Nei and Gojobori (1986). The bootstrap probability for each node is also indicated (Felsenstein 1985). When we estimated the substitution rate of Ebola virus, we excluded the sequences of Marburg virus and constructed another phylogenetic tree (data not shown), in which the topology among Ebola virus strains was identical with that of the former one.

Entropie

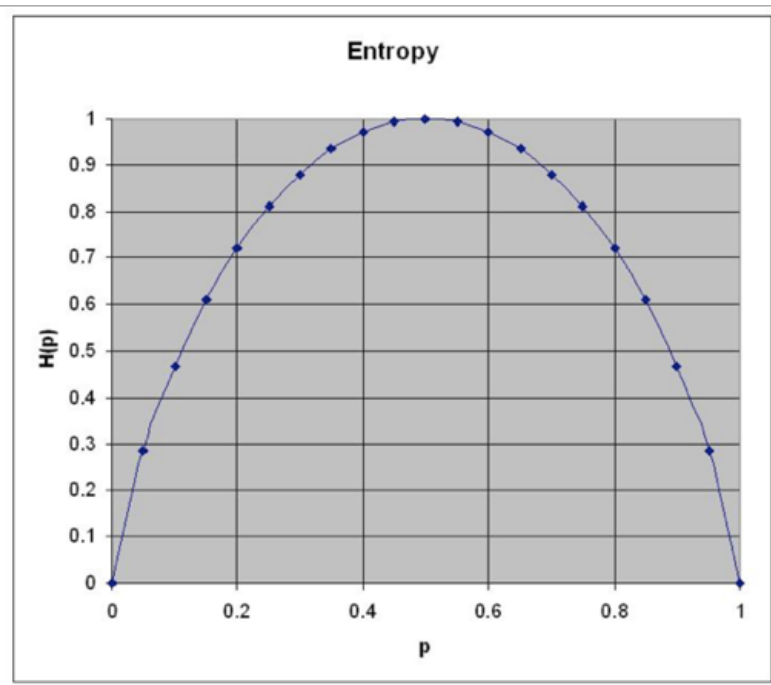
Entropie de Shannon

- Entropie de Shannon à la position i :

$$H_i = - \sum_{a=A,T,G,C} f_{a,i} \log_2(f_{a,i})$$

$f_{a,i}$: fréquence lettre a à la position i .

$f \cdot \log(f)$ tend vers 0 quand f tend vers 0 ou 1

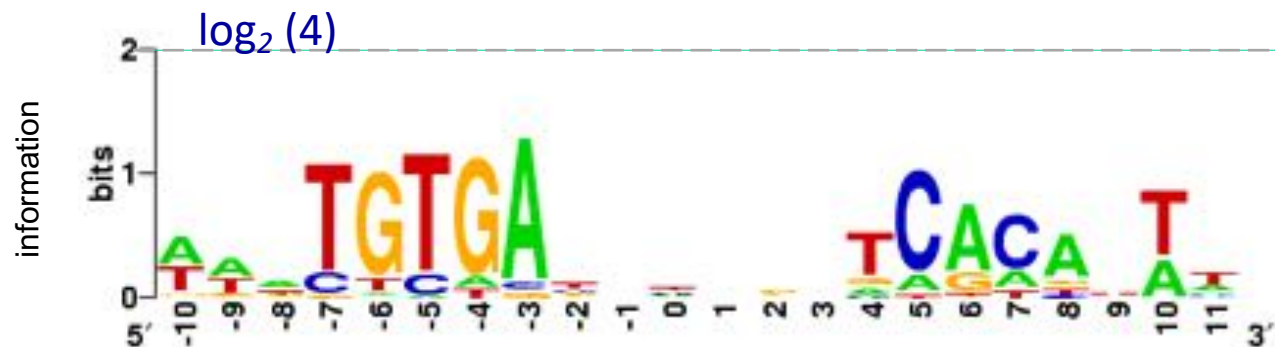


Variation de l'entropie en fonction de p pour un système à deux états

Entropie et contenu en information

- Le contenu en information est proportionnel à:

$$\log_2(4) - H_i = \text{information (pour ADN: } n=4\text{)}$$



Hauteur des lettres = $f \cdot \text{information}$

Sequence logos

(Schneider TD, Stephens RM. NAR. 1990)

Sites d'épissage

Représentation en fréquence:



Représentation en logo

*Beaucoup mieux
qu'une fréquence!
Fait ressortir
régions
conservées/
variables*

