UNIVERSITÉ PARIS SUD | FACULTÉ DES SCIENCES D'ORSAY

Analyse de Séquences

M1 BIBS

V. 2018.1

Contenu général de l'UE

- D.Gautheret
  - Les données de séquence: séquençage, structure des génomes, banques de données génomiques
- O. Lespinet
  - Alignement de séquence, phylogénie moléculaire
- D.Gautheret – A. Lopes
  - Homologie, annotation
  - TD Annotathon

Jeudi 20/09/2018	intro génomes / seq databases	DG
26-sept-18	Alignement séquences	OL
03-oct-18	Alignement séquences	OL
10-oct-18	Alignement séquences	OL
17-oct-18	Rech. Motifs	DG
24-oct-18	Arbres phylogénétiques	OL
	Vacances	
07-nov-18	exercices arbres	OL
14-nov-18	(libre)	
21-nov-18	Annotation+Annotathon 1	DG
28-nov-18	Annotathon	AL/DG
05-déc-18	Annotathon	AL
12-déc-18	Annotathon	AL
19-déc-18	Annotathon: prez par etudiants	DG / AL
	Vacances	
	Vacances	
09-janv-18	Annotathon	DG / AL
16-janv-18	(libre)	

Préface

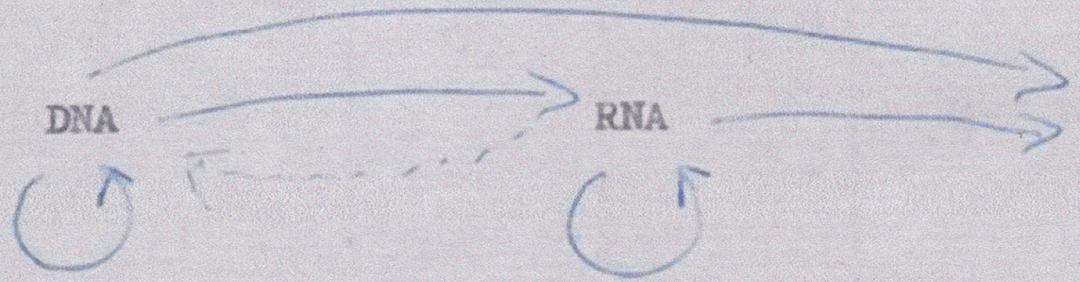
- Quelques notions d'information biologique

With slides from **Sacha Schutz**

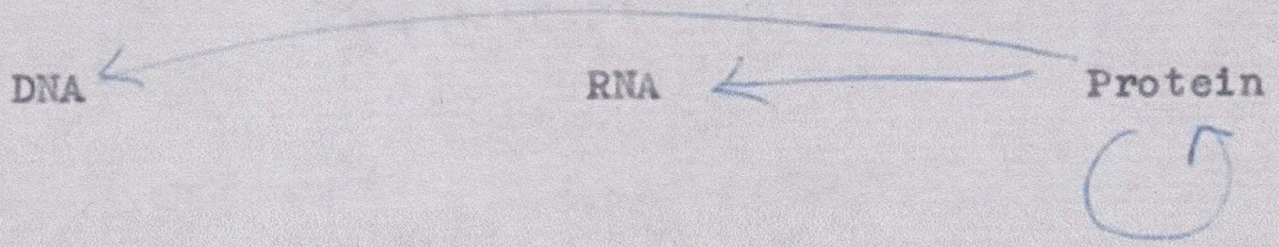
https://github.com/dridk

Biologie=science de l'information?

The Central Dogma: "Once information has got into a protein it can't get out again". Information here means the sequence of the amino acid residues, or other sequences related to it. That is, we may be able to have



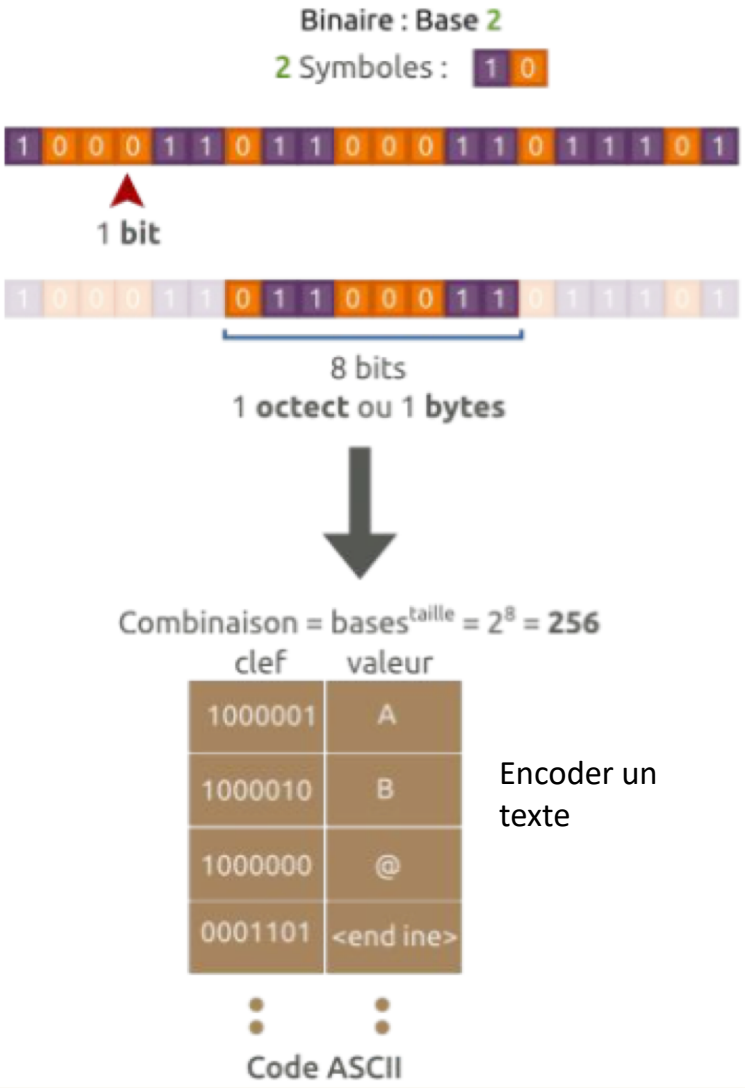
but never



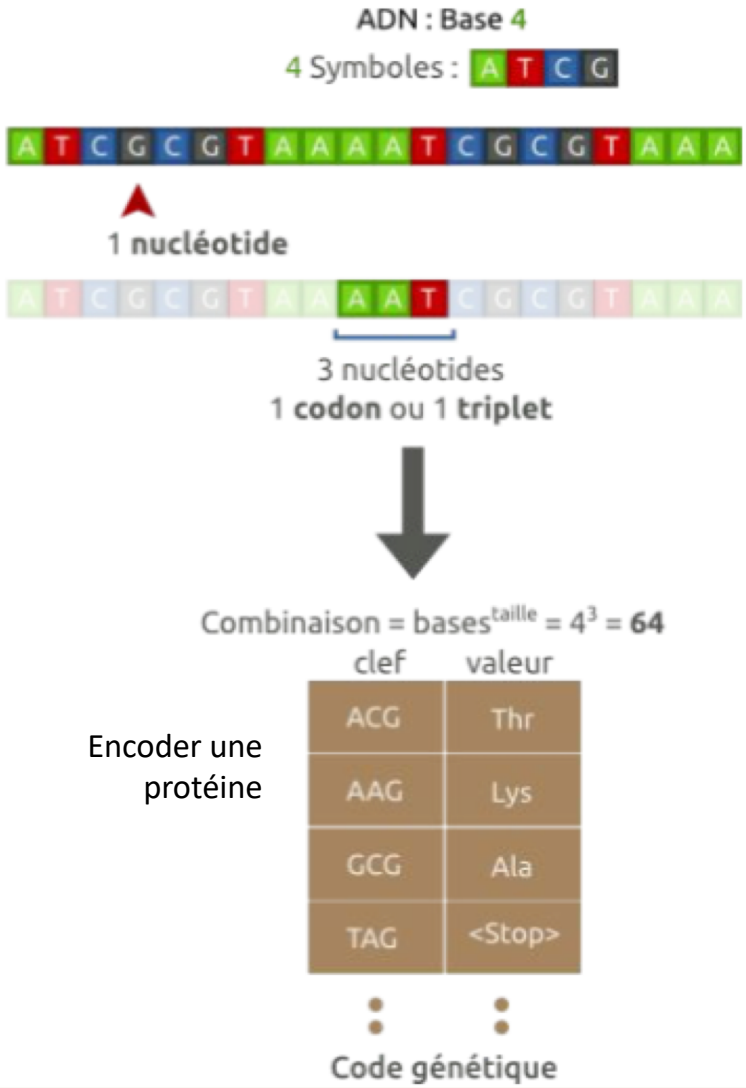
where the arrows show the transfer of information.

Cahier de labo de Francis Crick, 1956

Biologie=science de l'information?



Encoder un texte

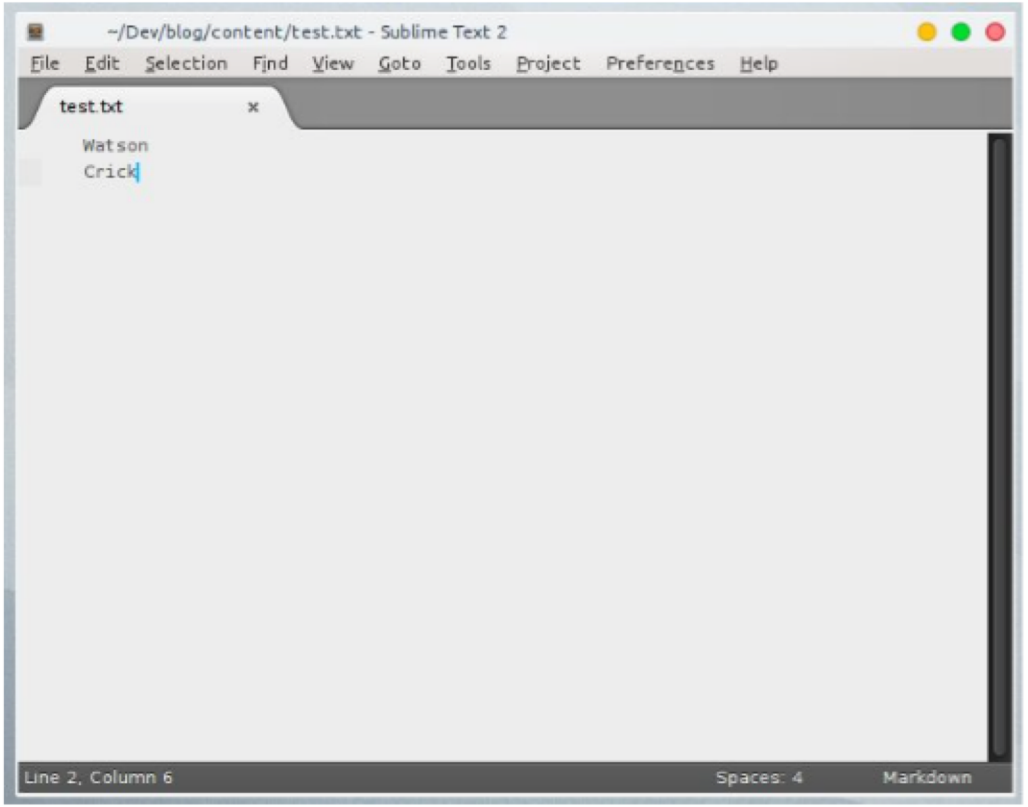


Encoder une protéine

Taille d’un fichier

Taille du fichier

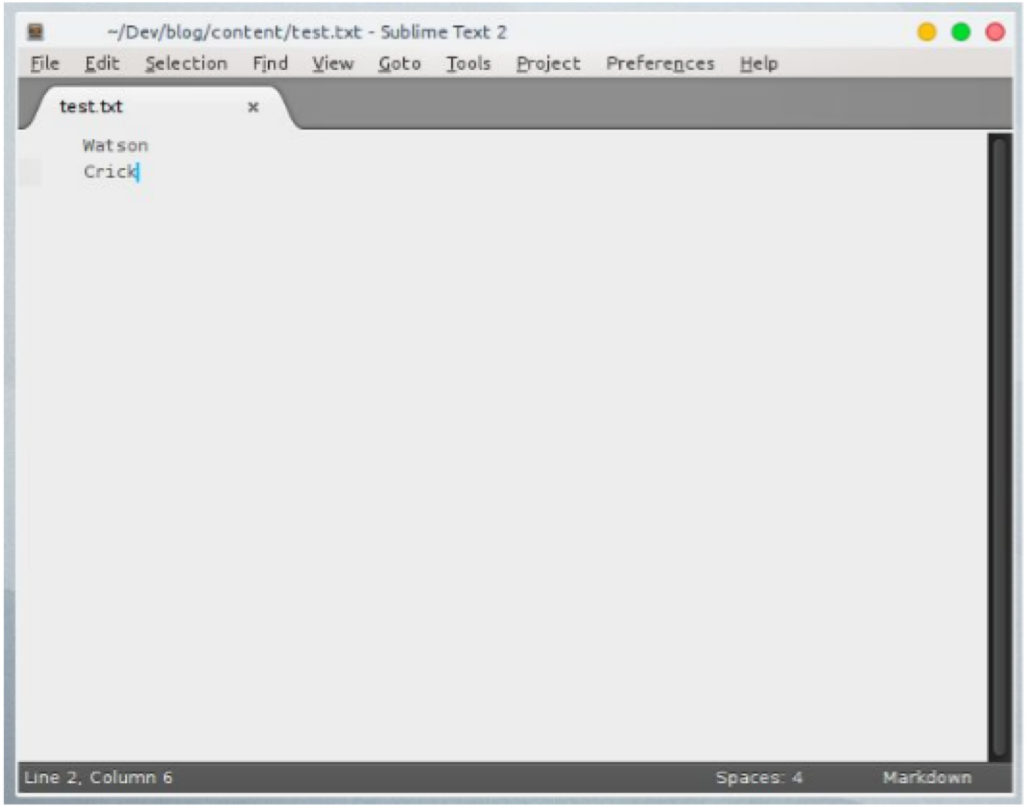
- en octets?
- En bits?



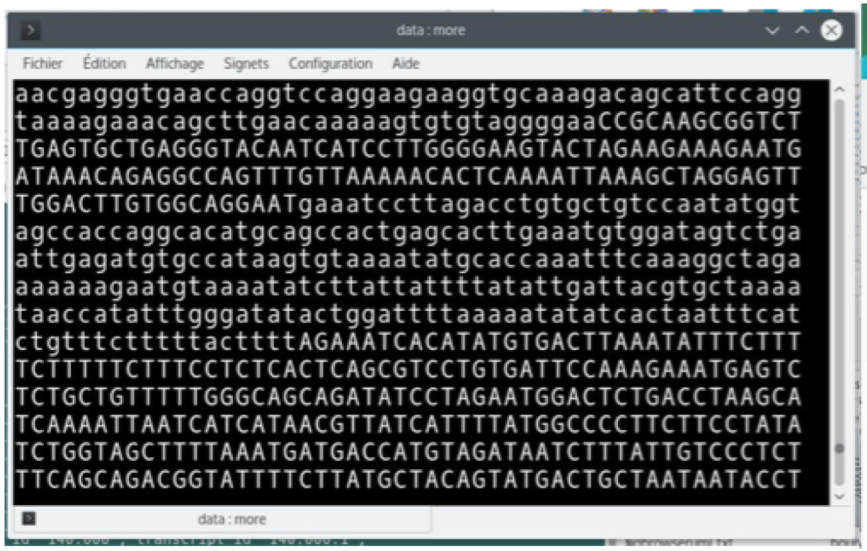
Taille d’un fichier

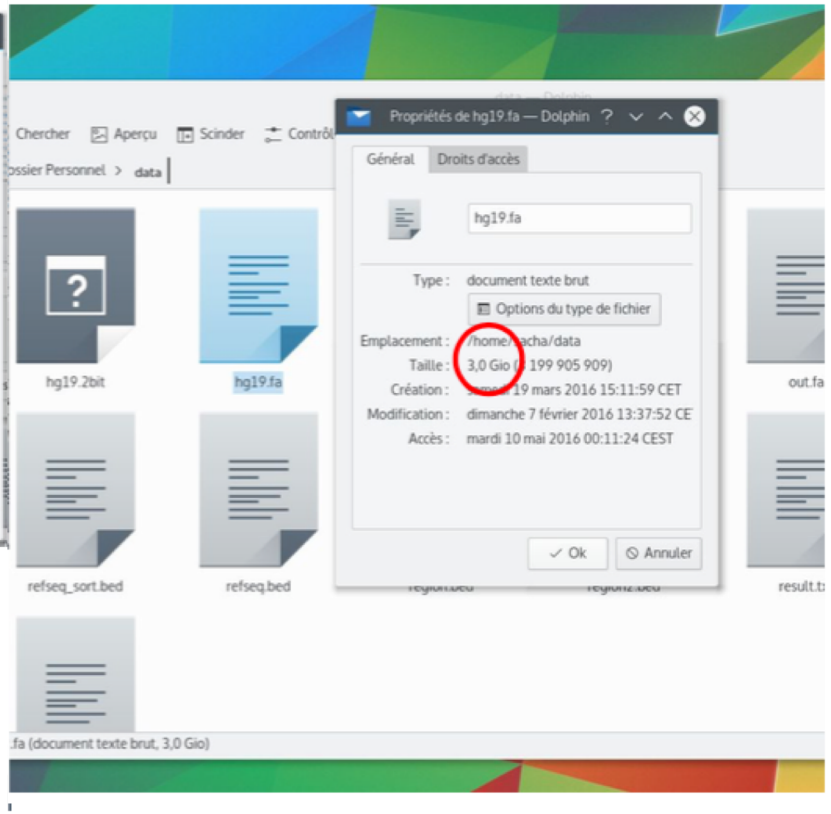
12 octets (bytes)=11+1

96 bits (12x8)



Taille d’un fichier





2 types de fichiers

Fichier texte

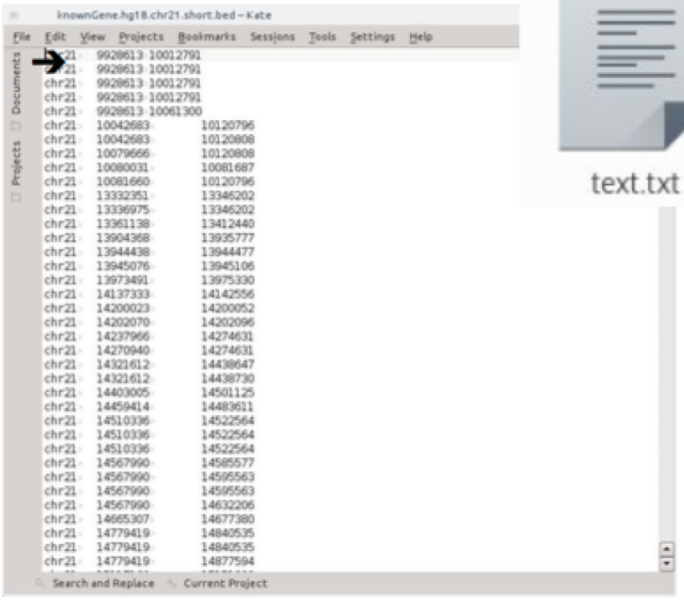
- Unité: 1 octet (byte)
- Lisible par un humain
- Dans un éditeur de texte
- Prend beaucoup d'espace
- Exemple:
  - HTML, Fasta, sam, csv, xml, vcf, fastq

Fichier binaire

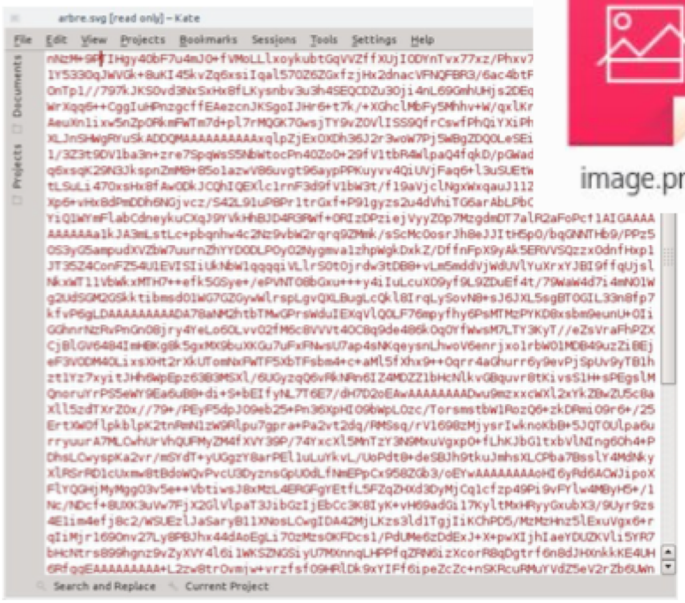
- Unité: 1 bit
- Non lisible par un humain
- Dans un éditeur hexadécimal
- Prend moins d'espace
- Exemple:
  - png, jpg, mp3, raw, bam, excel, word, fastq.gz

Exemples

Fichier texte



Fichier binaire



Espace texte vs. binaire

Exemple de fichier texte pour information Vrai/Faux (V/F):

VFVFFVFF

Total: 8 octets

Exemple de fichier binaire pour information Vrai/Faux (1/0):

10100100

Total: 1 octet

Formats de données en bioinformatique

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

Format et spécification

- Le format d'un fichier décrit comment les données sont représentées
- Cette description est fournie dans un document appelé **spécification**.

Même donnée, différents formats

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

Format **JSON**
https://tools.ietf.org/html/rfc4627

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

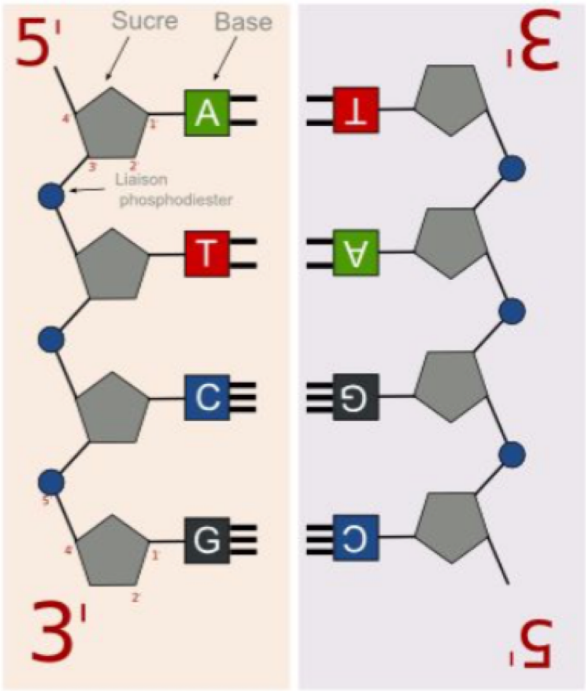
Format **XML**
https://www.w3.org/TR/REC-xml/

Séquences et régions

- En génomique on peut catégoriser les formats en:
  - Formats décrivant des **séquences**
  - Formats décrivant des **régions**

Séquences d'ADN

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



Format fasta

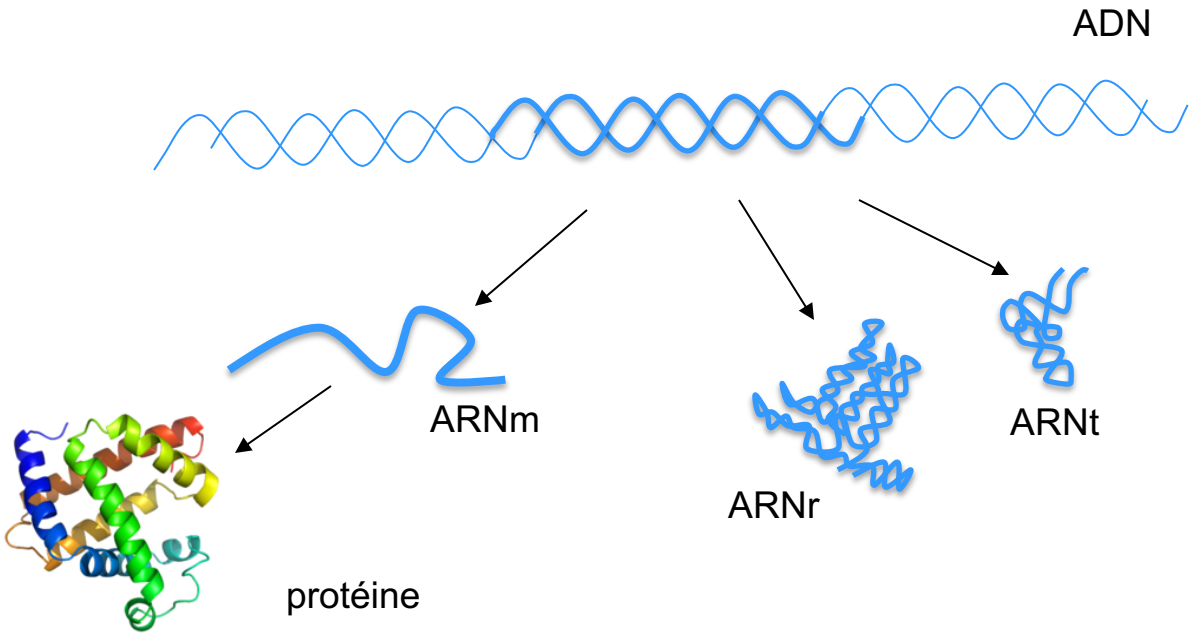
*.fa , *.fasta

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

1. Structure des génomes

Le gène

- Un gène est une séquence d'ADN qui spécifie la synthèse d'une protéine ou d'un ARN fonctionnel
- Un gène peut donc coder pour un ARN messager ou pour un ARN non-messager (ARNr, ARNt, …)



Caractéristiques des génomes procaryotes

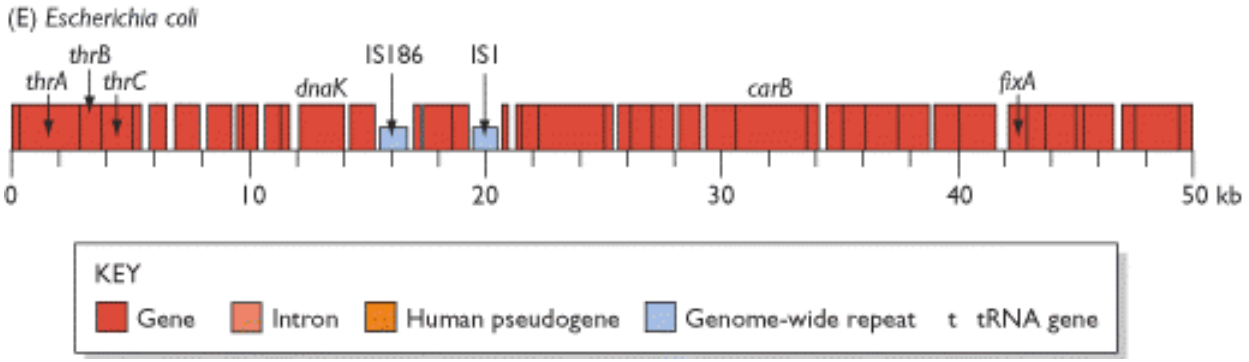
- Chromosome circulaire unique
- Présence possible de petites séquences d'ADN circulaires indépendantes : **les plasmides.**

Les gènes procaryotes

- Fraction codante des génomes élevée.
  - > 90% codant
  - Peu de séquences intergéniques
  - Génome « compact »
- Chez les procaryotes **la séquence des gènes est continue**. **Pas d'intron**
- Gènes organisés en opérons. 600 opérons dans le génome de *Escherichia coli*.

Densité des gènes procaryotes

- Longueur gène 950 nt. en moyenne (coli)
- Haute densité en gènes: 1gène / kb
- 95% du génome est transcrit chez E. coli.

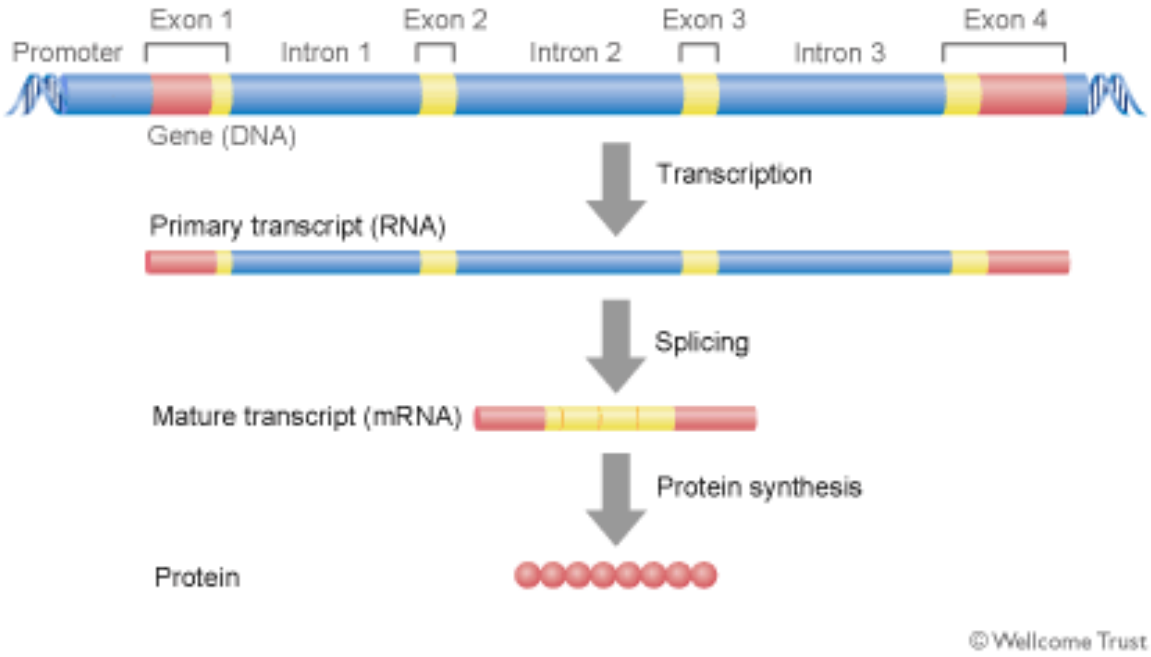


From « Genomes 2 », T.A. Brown

Caractéristiques des génomes eucaryotes

- Dans le noyau
- Taille >> procaryote
- Plusieurs chromosomes (homme 23, cheval 32, levure 16, drosophile 4…)
- Gènes « disloqués » (exons, introns)
- Grandes régions intergéniques de fonction inconnue

Le gène eucaryote



transcription

maturation

traduction

–Gène humain moyen: 27kb, 9 introns, codant:1,3kb , exon moyen: 145 bp, intron moyen:3365 bp.

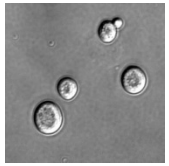
–Gènes "monstres": dystrophine: 2,4 Mb; Facteur de coagulation VIII: 186 kb, 26 exons; Tinine: codant: 80kb, 178 exons

Densité des gènes eucaryotes

Densité moyenne:

- S. Cerevisiae:

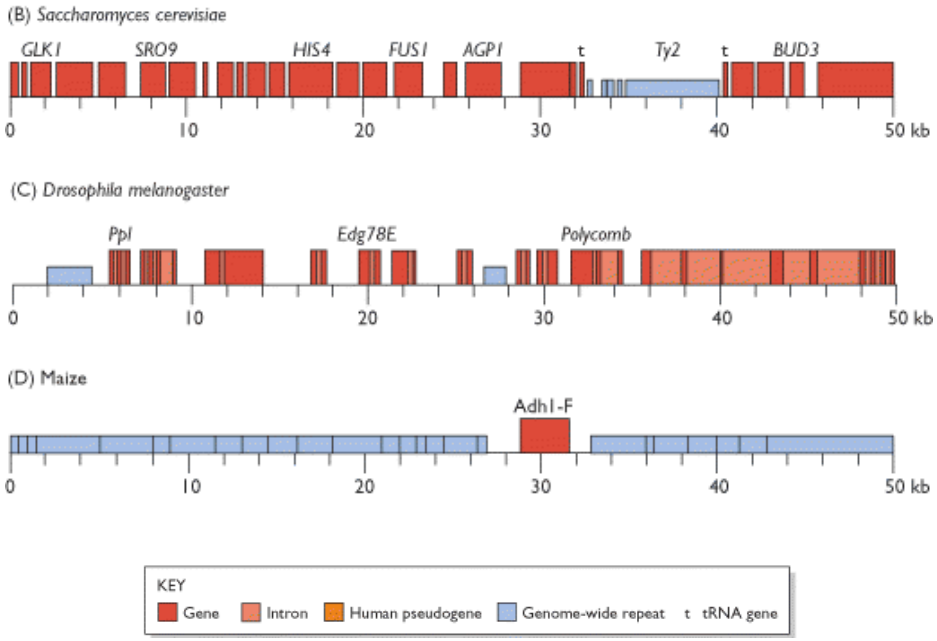
  1 gène/2kb.



- Drosophila:

  1gène/10kb

- Maïs: 1 gène tous les 70kb
- Humain: 1 gène tous les 100kb

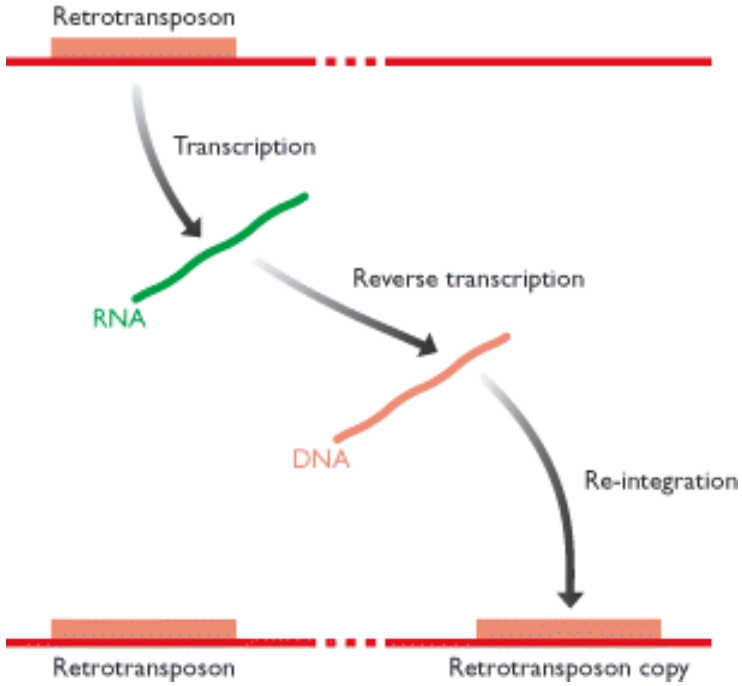


From « Genomes 2 », T.A. Brown

Junk DNA: les séquences répétées dans le génome humain

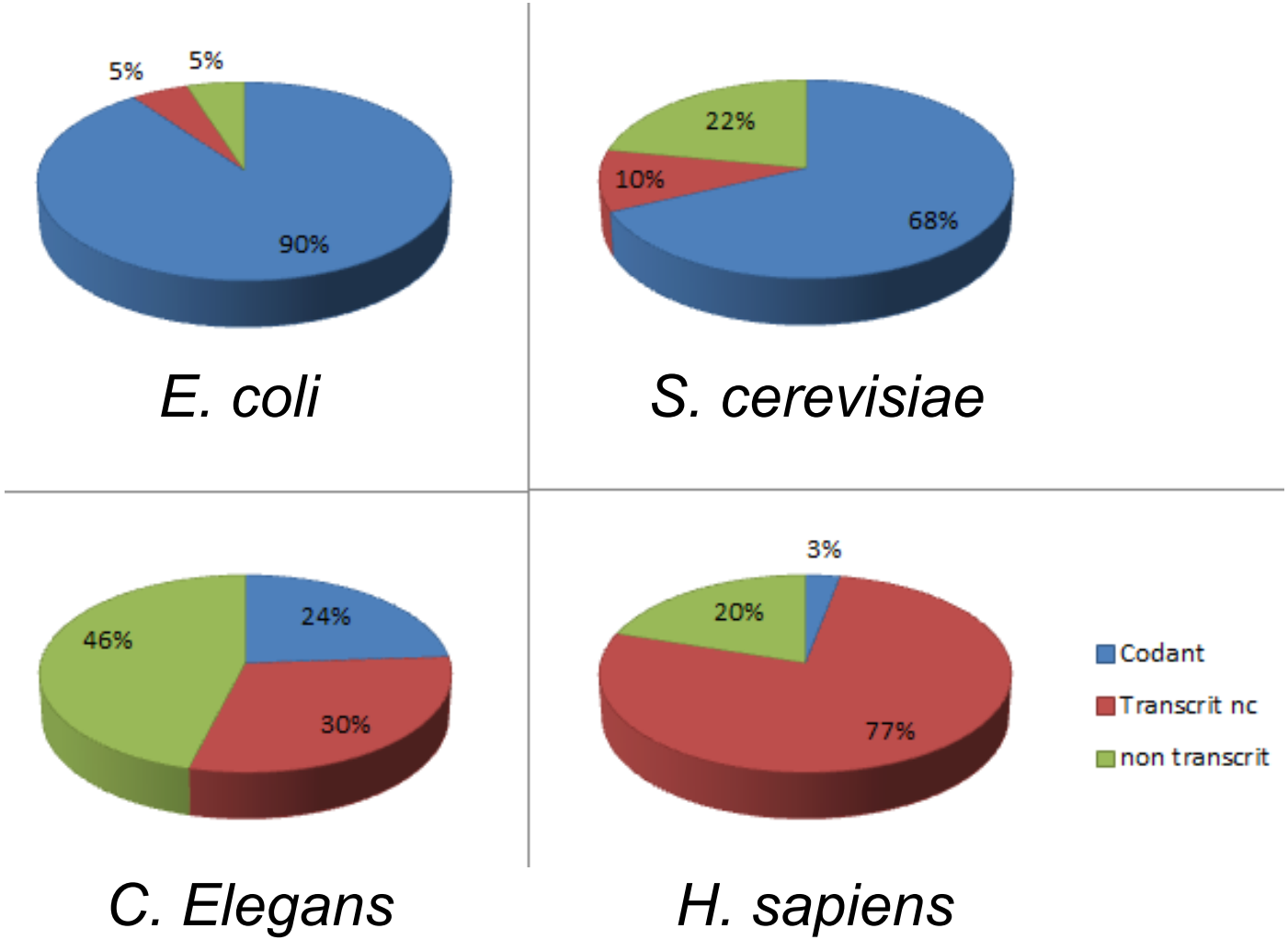
- 4 classes de séquences répétées
  - Répétition de type transposon (ou interspersed repeats)
  - copie inactives de gènes (processed pseudogenes)
  - Répétition simples de k-mères courts, p. ex. (A)n, (CA)n ou (CGG)n
  - Segments dupliqués: blocs de 10–300 kb copiés d'une région à l'autre ou en tandems

Retroposons: les principales séquences répétées chez l'homme



From « Genomes 2 », T.A. Brown

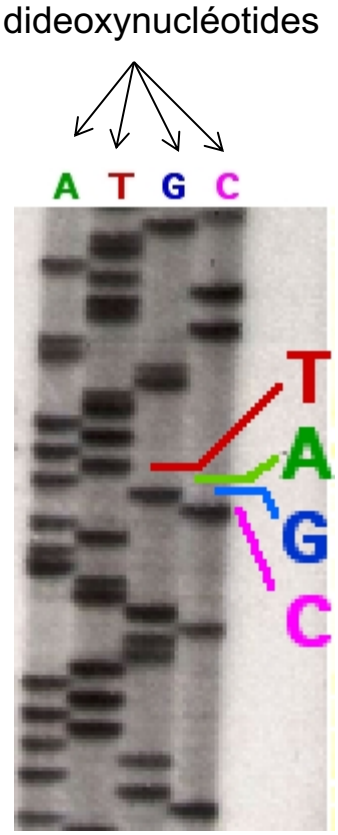
Fraction codante et non-codante des génomes



2. Le séquençage

Le séquençage de Sanger (1977)

- Séquençage par terminaison de chaine
  - Utilisation de dideoxynucléotides pour interrompre la synthèse à un certain type de base.
  - 4 réactions + marquage radioactif
- Amélioré en 1987 par l'introduction de marqueurs flurorescents (1 seule réaction) et l'automatisation.

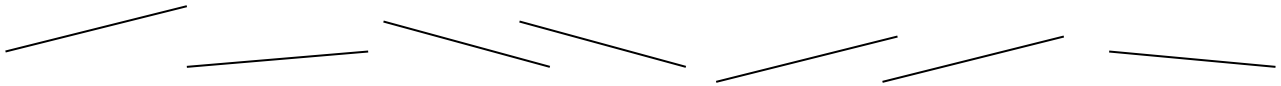


Wikipedia

Stratégie de séquençage Shotgun

Clone à séquencer

Fractionnement



Séquençage aléatoire

Assemblage

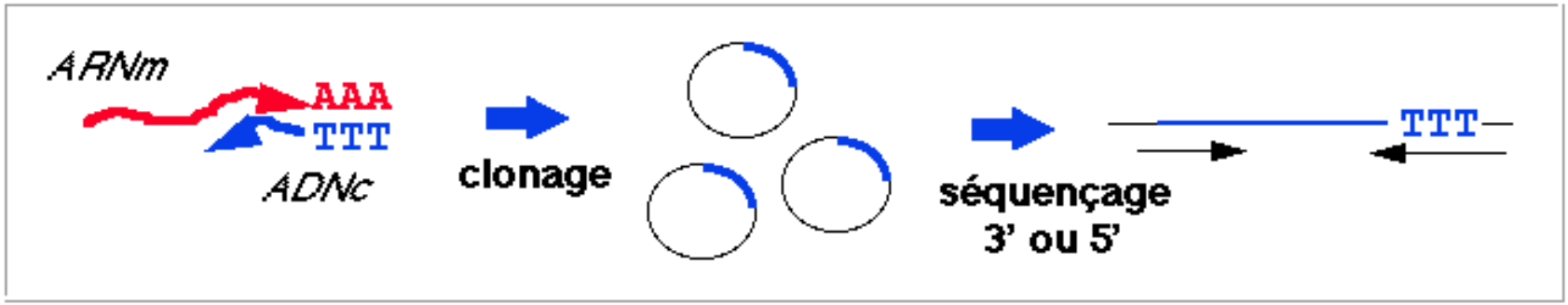
Scaffold (si information cartographie)

Problématique en présence de longues régions répétées)

Eric Lander, coordinateur: projet de séqueçage du génome humain

Les cDNA

- Idée initiale:
  - pourquoi vouloir tout séquencer (95% de junk DNA) si ce sont les gènes qui nous intéressent?
- Usage actuel:
  - Analyse du transcriptome



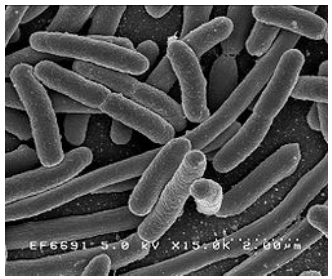
- EST (Expressed sequence Tag) = Séquences partielles d'ADNc clonés et prélevés aléatoirement.
- Full length cDNA: Séquences complète d'ADNc

Craig Venter a contribué à l'usage massif des EST dès 1991

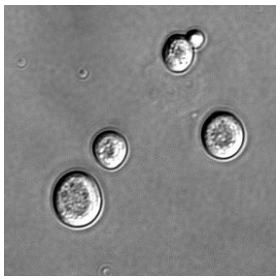
Le séquençage des organismes modèles

- Facilité expérimentale
  - Grosses cellules, corps transparent, temps de génération court, taux de reproduction élevé
- Représentant d'une fonction
  - Différenciation cellulaire, Système immunitaire, Photosynthèse, symétrie bilatérale…
- Génome dense
- Données accumulées
  - Mutants, génétique, biochimie…

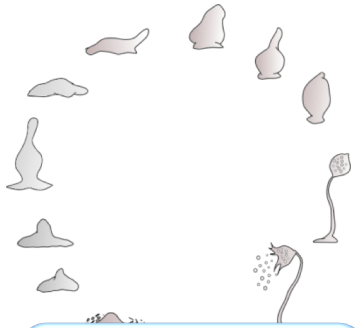
Les espèces-modèle



E. Coli
4.5 Mb
4200 gènes



S. cerevisiea
13 Mb
6000 gènes



Disctyostelium
34 Mb
12500 gènes

Arabidopsis thaliana
150 Mb
25000 gènes

C. elegans
100 Mb
15000 gènes

Drosophila melanogaster
150 Mb
15000 gènes

Danio rerio
1,5 Gb
45000 gènes

Mus musculus
3 Gb
20000 gènes

Et beaucoup d'autres …

Homo sapiens

- Phylum
  - Métazoaire, Mammifère, Primate
- Inconvénients
  - Expérimentation impossible sauf cellules somatiques en culture (Hela..)
  - Temps de génération lent

Raël

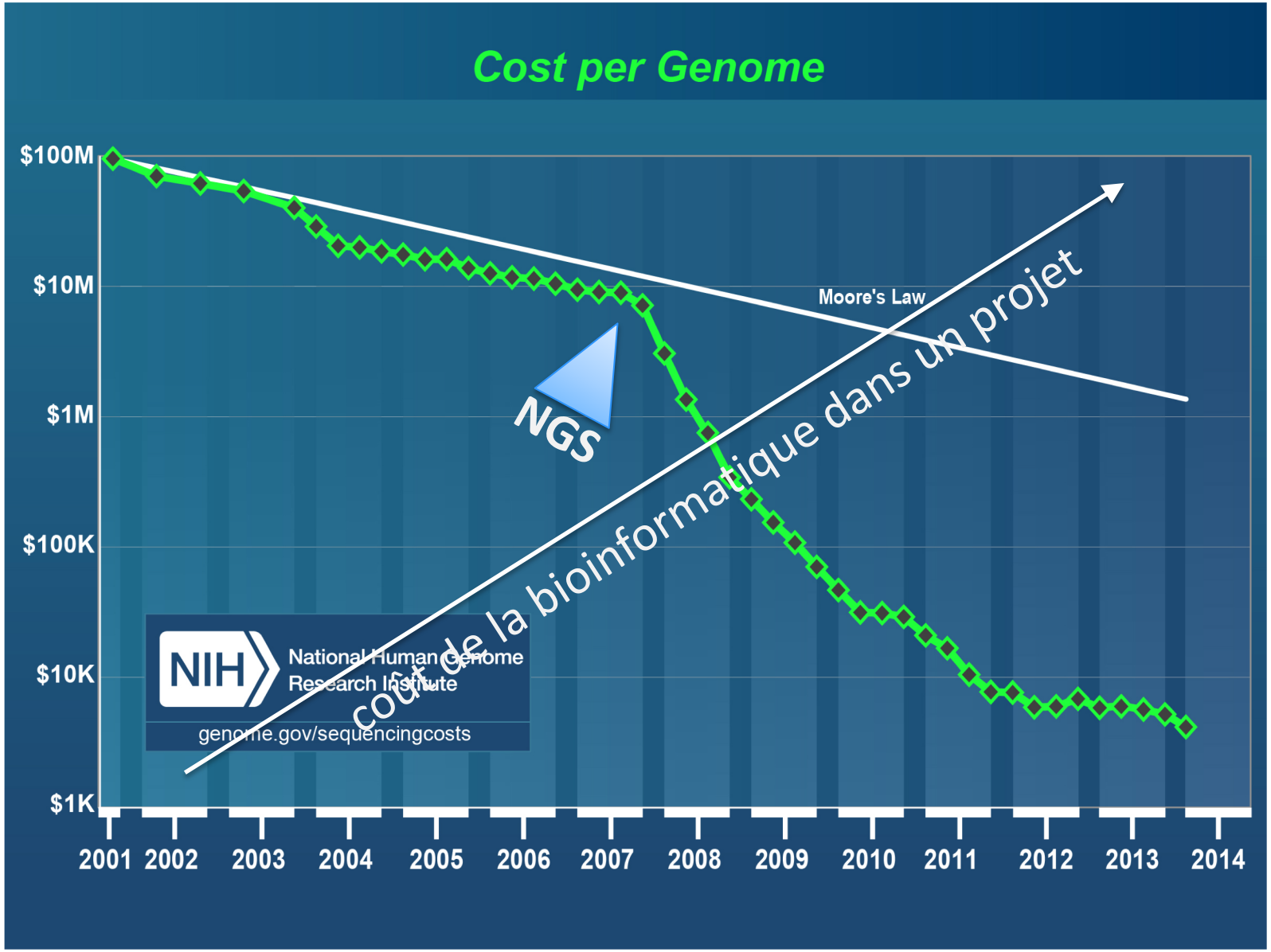
3 Gb

20000 gènes

Metagénomes

- Fragments d'ADN séquencés aléatoirement à partir d'un environnement donné
- Soit fragments ciblés (p. ex. ARNr 16S), soit ADN génomique
- Environnements étudiés:
  - Océan, intestin humain, drainage minier acide,
- Applications:
  - Populations microbiennes (structure, variations)
  - Découverte et séquençage de microorganismes
  - Découverte de gènes/fonctions

Le bouleversement des NGS

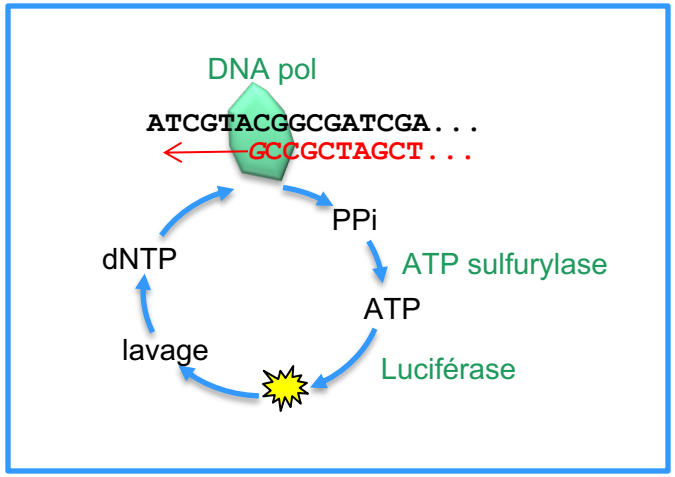


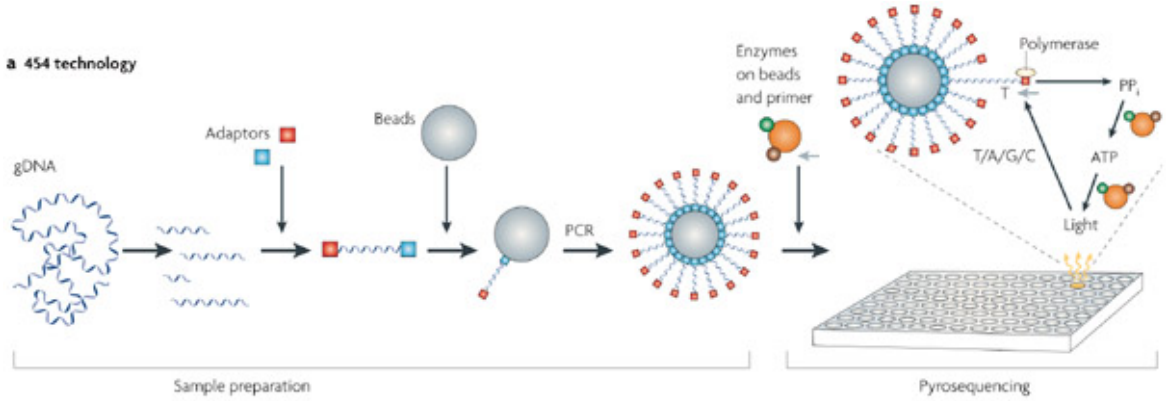
Début des NGS: pyroséquençage (2006)

454 LifeSciences/Roche
Biotage/Qiagen

1. ADN immobilisé (billes)
2. Synthèse brin complémentaire par ADN polymérase
3. Introduction des dNTP un par un
4. Si bon dNTP: libération PPi, synthèse ATP, et emission de lumière.

300-500nt à la fois x 400.000



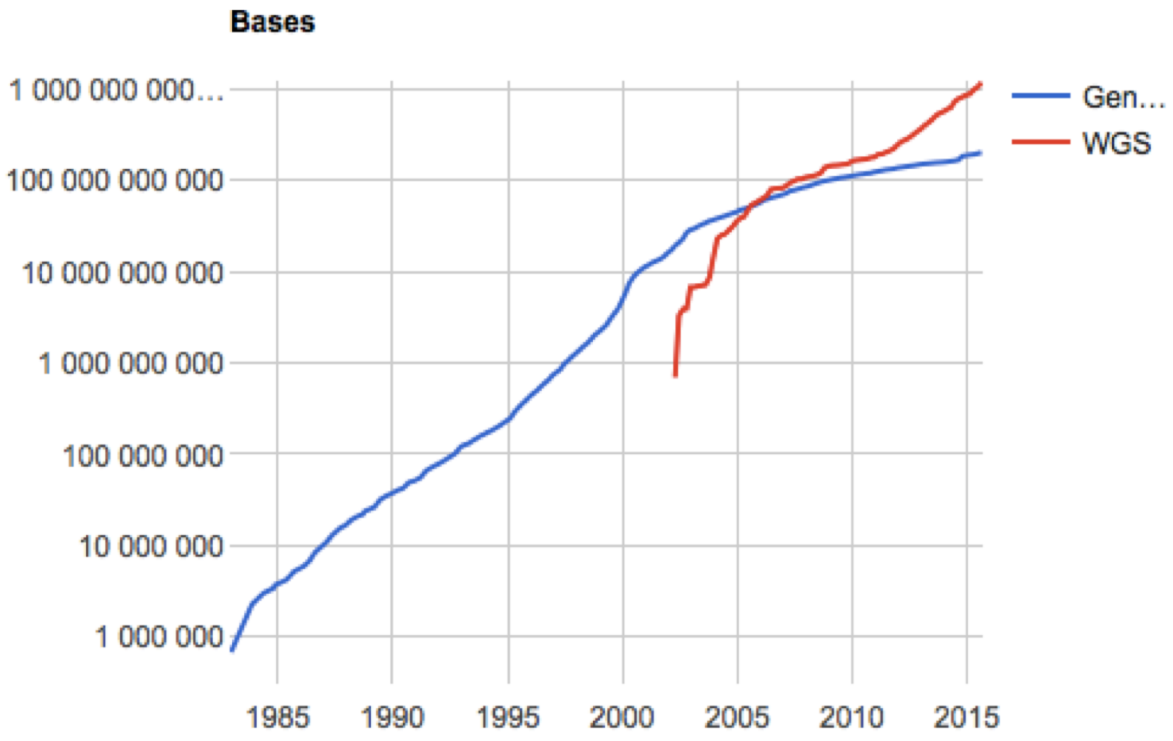


Nature Reviews Microbiology

3. Les banques de données génomiques

Genbank: La banque d'ADN du NIH

- **Etat 2018**
  - > 10^{11} bases
  - > 10^{8} séquences
  - Genbank double environ tous les 14 mois depuis ses débuts en 1982.
  - Nouvelle version tous les 2 mois



Enregistrement Genbank

- Chaque enregistrement se voit attribuer un numéro d'accession, stable et unique, et chaque séquence un numéro GI.
- Quand un changement est effectué dans un enregistrement Genbank, le num. d'accession reste, le GI change.

Séquence au Format Fasta

>U00096 Escherichia coli K-12 MG1655 complete genome.
agcttttcattctgactgcaacgggcaatatgtctctgtgtggattaaaaaaagagtgtc
tgatagcagcttctgaactggttacctgccgtgagtaaattaaaattttattgacttagg
tcactaaatactttaaccaatataggcatagcgcacagacagataaaaattacagagtac
acaacatccatgaaacgcattagcaccaccattaccaccaccatcaccattaccacaggt
aacggtgcgggctgacgcgtacaggaaacacagaaaaaagcccgcacctgacagtgcggg
ctttttttttcgaccaaaggtaacgaggtaacaaccatgcgagtgttgaagttcggcggt
acatcagtggcaaatgcagaacgttttctgcgtgttgccgatattctggaaagcaatgcc
aggcaggggcaggtggccaccgtcctctctgcccccgccaaaatcaccaaccacctggtg
gcgatgattgaaaaaaccattagcggccaggatgctttacccaatatcagcgatgccgaa
cgtatttttgccgaacttttgacgggactcgccgccgcccagccggggttcccgctggcg
caattgaaaactttcgtcgatcaggaatttgcccaaataaaacatgtcctgcatggcatt
agtttgttggggcagtgcccggatagcatcaacgctgcgctgatttgccgtggcgagaaa
atgtcgatcgccattatggccggcgtattagaagcgcgcggtcacaacgttactgttatc
gatccggtcgaaaaactgctggcagtggggcattacctcgaatctaccgtcgatattgct
gagtccacccgccgtattgcggcaagccgcattccggctgatcacatggtgctgatggca
ggtttcaccgccggtaatgaaaaaggcgaactggtggtgcttggacgcaacggttccgac
tactctgctgcggtgctggctgcctgtttacgcgccgattgttgcgagatttggacggac

Enregistrement Genbank avec annotation

```
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 f
            investigating biology. The C. elegans Sequencing Consort
  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="MDSCEEECDLEVDSDEEDQLFGEKCISLLSSLLPLSSSTLLSNA
                     INLELDEVERPPPLLNVLEEQQFPKVCANIEEENELEADTEEDIAETADDEESKDPVE
                     KTENFEPSVTMDTYDFPDPYPVQIRARPQVPPKPPIDTVRYSMNNIKESADWQLDELL
                     EELEALETQLNSSNGGDQLLLGVSGIPASSSRENVKSISTLPPPPPALSYHQTPQQPQ
. . .
                     QVYTGIGWEKKYKSPTPWCISIKLTALQMKRSQFIKYICAEDEMTFKKWLVALRIAKN
                     GAELLENYERACQIRRETLGPASSMSAASSSTAISEVPHSLSHHQRTPSVASSIQLSS
                     HMMNNPTHPLSVNVRNQSPASFSVNSCQQSHPSRTSAKLEIQYDEQPTGTIKRAPLDV
                     LRRVSRASTSSPTIPQEESDSDEEFPAPPPVASVMRMPPPVTPPKPCTPLTSKKAPPP
                     PPKRSDTTKLQSASPMAPAKNDLEAALARRREKMATMEC"
. . .
```

```
BASE COUNT     2598 a   2024 c   1888 g   2449 t
ORIGIN
        1 ttctaaaagt cgaaaaacga gcaatttttg atgctagatt ttttgatttg acgaattttt
       61 tcagtttttt ttctttaaaa aaggtttttg acccccttaaa gttttccttt cccttccaat
      121 ttttttccttc tttcttatac gacttctcaa gtttcaactc taaaacaaag ctacatgtac
      181 atttccggta aactttgtgt ctcagaagat ccattttctt tttgttacat ttattcaaga
      241 ttgaattcca aaatttcagc caatatggac agttgcgaag aggaatgcga tctggaagtt
      301 gacagtgacg aagaagatca acttttttggt gaaaagtggt gagttcttat tgtggtaacc
      361 aaagaaatgt cagtggtccg taaacacttg actcccaaat ggtttctcgt aattacctta
      421 tgcacacttt tcaagtgttt gccgtttgat cttagccaat ttgaaacgtt tagatgttaa
      481 atggaaaatg ggtaaagttt tttattttat agaaaaaagg tttggaaaaa aatcgagtca
      541 ctgaatagtt tgaagaacgg aaaaataaaa ctttccaaaa atcataaaac atttagtgtt
      601 tcgaaaatta tagtgttttt tttgttggta tgttttgaca aaagctaaac catctttatt
      661 gtagttttgt aaaatgttca caaagatgcg ttttttttttc aaatttggca ggctatcttt
      721 acattcacat ttggataatt caaatttttc ttatcgctaa caaattttcc tatttttcca
      781 attattcgtt tttataaagc tttggtagta tgttgtgtct atctttagtg gtcatcagtt
. . .
//
```

Mini TD

- Sur le site NCBI/genome, retrouvez le génome annoté de Candidatus Carsonella ruddii HT.
- De quel type d'espèce s'agit-il?
- Taille du génome?
- Que contiennent les champs…
  - CDS
  - Complement
  - Translation
- Sauvez le fichier au format Genbank « full » (annotations et séquences) et comptez les CDS avec la commande grep, option -c
- Récupérez la séquence au format fasta et comparez la taille des fichiers fasta et gb

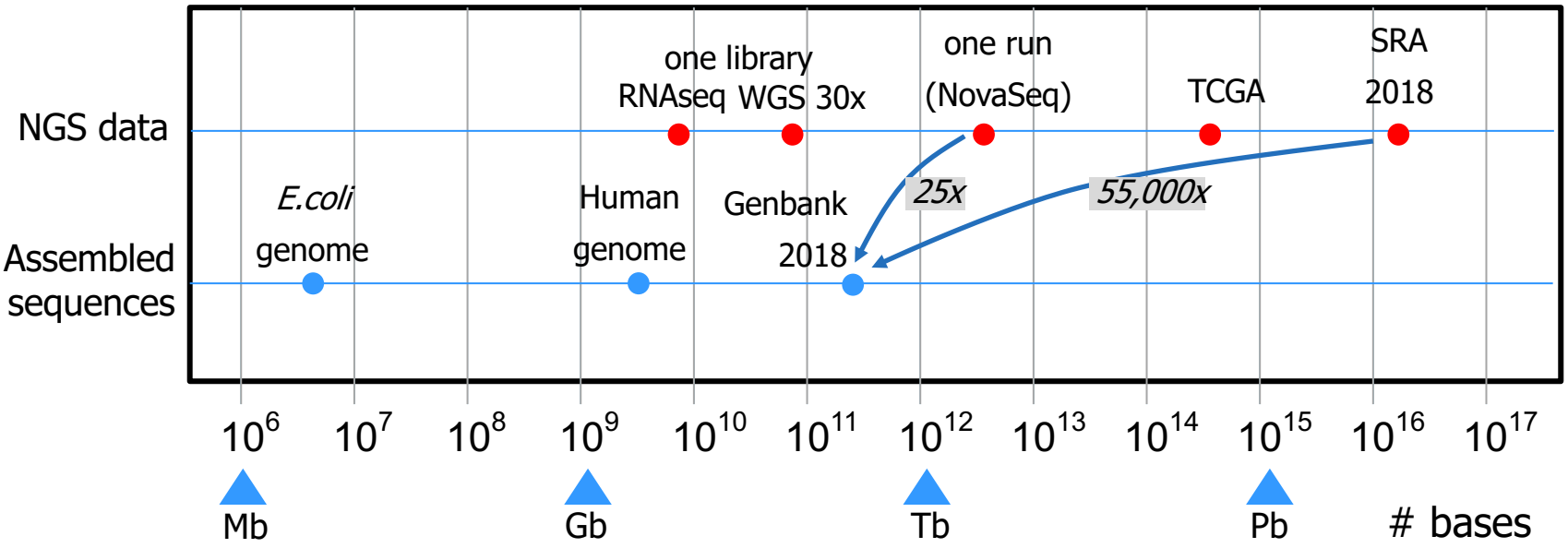
Autres banques nucléotidiques

- 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.
- Banques pour Blast
  - NR nucléique (« Non-redundant »). All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS, or phase 0, 1 or 2 HTGS sequences). (n'est plus "non-redondant")
  - DbEST: dbest Database of GenBank+EMBL+DDBJ sequences from EST Divisions

SRA: short read archive

- Partie de *European Nucleotide Archive* (sequences brutes issues de séquenceurs).
- Information sur chaque entrée:
  - Study
  - Sample
  - Experiment
  - Run
  - Organism
  - Instrument Platform
  - Library Name
  - Read Count
  - Base Count
  - File Name / File Size
- Séquences au format fastq

The incredible scale of NGS databases



Format fastq

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
IIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

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

Attention: plusieurs versions (Illumina, Sanger..)

$$Q_{\text{sanger}} = -10 \ \log_{10} p$$

Code ASCII et score de qualité

Score le plus bas (33=!): probabilité d'erreur élevée

Score le plus haut (126=~): probabilité d'erreur faible

Décimal	Octal	Hex	Binaire	Caractère	
000	000	00	00000000	NUL	(Null char.)
001	001	01	00000001	SOH	(Start of Header)
002	002	02	00000010	STX	(Start of Text)
003	003	03	00000011	ETX	(End of Text)
004	004	04	00000100	EOT	(End of Transmission)
005	005	05	00000101	ENQ	(Enquiry)
006	006	06	00000110	ACK	(Acknowledgment)
007	007	07	00000111	BEL	(Bell)
008	010	08	00001000	BS	(Backspace)
009	011	09	00001001	HT	(Horizontal Tab)
010	012	0A	00001010	LF	(Line Feed)
011	013	0B	00001011	VT	(Vertical Tab)
012	014	0C	00001100	FF	(Form Feed)
013	015	0D	00001101	CR	(Carriage Return)
014	016	0E	00001110	SO	(Shift Out)
015	017	0F	00001111	SI	(Shift In)
016	020	10	00010000	DLE	(Data Link Escape)
017	021	11	00010001	DC1	(XON)(Device Control 1)
018	022	12	00010010	DC2	(Device Control 2)
019	023	13	00010011	DC3	(XOFF)(Device Control 3)
020	024	14	00010100	DC4	(Device Control 4)
021	025	15	00010101	NAK	(Negative Acknowledgement)
022	026	16	00010110	SYN	(Synchronous Idle)
023	027	17	00010111	ETB	(End of Trans. Block)
024	030	18	00011000	CAN	(Cancel)
025	031	19	00011001	EM	(End of Medium)
026	032	1A	00011010	SUB	(Substitute)
027	033	1B	00011011	ESC	(Escape)
028	034	1C	00011100	FS	(File Separator)
029	035	1D	00011101	GS	(Group Separator)
030	036	1E	00011110	RS	(Request to Send)(Record Separator)
031	037	1F	00011111	US	(Unit Separator)
032	040	20	00100000	SP	(Space)
033	041	21	00100001	!	(exclamation mark)
034	042	22	00100010	"	(double quote)
035	043	23	00100011	#	(number sign)
036	044	24	00100100	$	(dollar sign)
037	045	25	00100101	%	(percent)
038	046	26	00100110	&	(ampersand)
039	047	27	00100111	'	(single quote)
040	050	28	00101000	(	(left opening parenthesis)
041	051	29	00101001	)	(right closing parenthesis)
042	052	2A	00101010	*	(asterisk)
043	053	2B	00101011	+	(plus)
044	054	2C	00101100	,	(comma)
045	055	2D	00101101	-	(minus or dash)
046	056	2E	00101110	.	(dot)
047	057	2F	00101111	/	(forward slash)
048	060	30	00110000	0	
049	061	31	00110001	1	
050	062	32	00110010	2	
051	063	33	00110011	3	

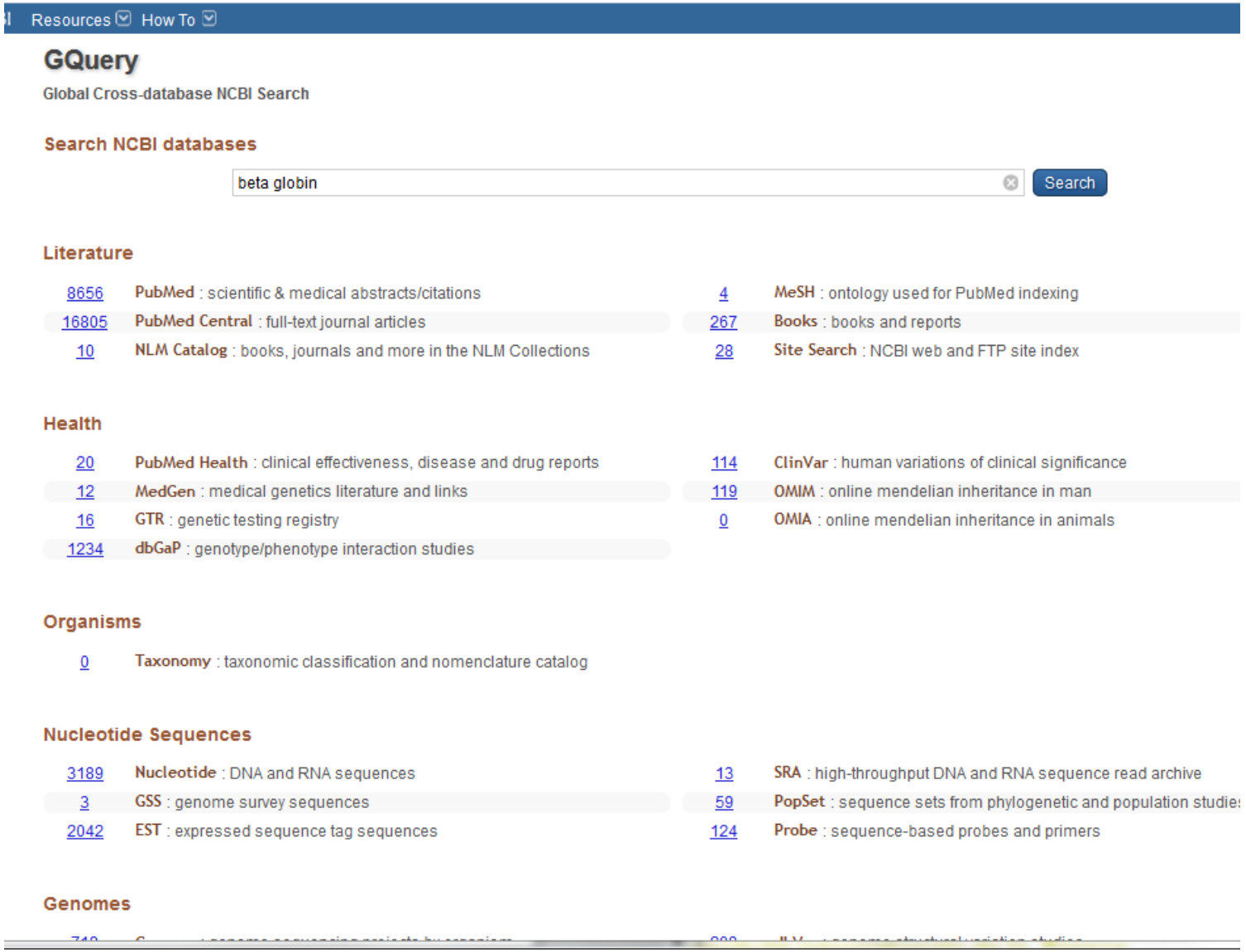
Banques protéiques

- Swissprot (UniProtKB/Swissprot).
  - La mieux annotée des banques protéiques. 2011: 530.000 entrées.
  - Curation par experts seulement (basé sur publis)
  - Attention: toutes les protéines connues n'y sont pas!
    1. Evidence at protein level 72765 13.7%
    2. Evidence at transcript level 69863 13.1%
    3. Inferred from homology 373177 70.1%
    4. Predicted 14474 2.7%
    5. Uncertain 1867 0.4%

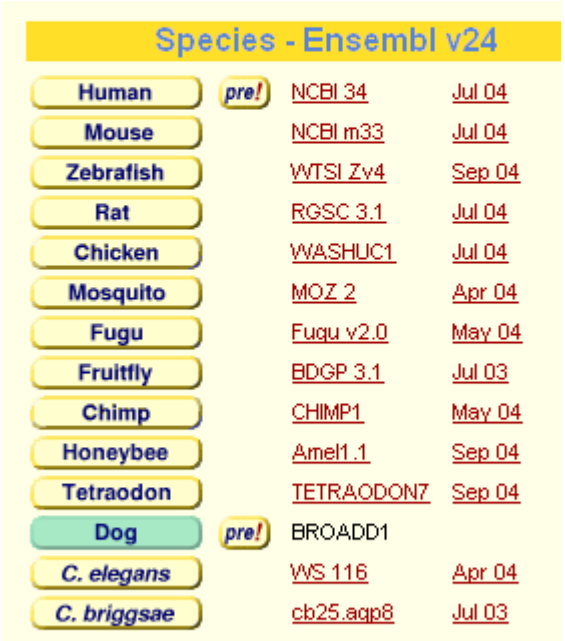
Banques protéiques

- TrEMBL (UniProtKB/TrEMBL):
  - banque protéique produite automatiquement par traduction banque EMBL. 2011: 17.000.000 entrées
- Uniprot=Swissprot+TrEMBL
- Dizaines de Banques spécialisées
  - Cazy (Carbohydrate Active Enzymes)
  - ABC transporters
  - etc.

NCBI Global query (recherche multi-bases)



Ensembl (www.ensembl.org)



- Plusieurs banques en une:
  - Génomes assemblés
  - Peptides confirmés
  - Transcrits confirmés
  - peptides prédits
  - Transcrits prédits
- + un genome browser

- Méthode de prédiction (système Genewise): Utilisent programme de prédiction Genscan, puis Blast contre: protéines, mRNA, EST
- Génome humain version Sep 2008 (NCBI 36) : Confirmed protein-coding genes: 21649; RNA genes: 4810; Predicted genes (Genscan): 49796; base pairs: 3,25 10^9.

Ensembl Species (2012 - partiel)


Alpaca
Vicugna pacos
vicPac1


Anole lizard
Anolis carolinensis
AnoCar2.0


Armadillo
Dasypus novemcinctus
dasNov2


Baboon (preview - assembly only)
Papio hamadryas
Pham


Budgerigar (preview - assembly only)
Melopsittacus undulatus
MelUnd6.3


Bushbaby
Otolemur Garnettii
OtoGar3


Ciona intestinalis
Ciona intestinalis
KH


Ciona savignyi
Ciona savignyi
CSAV2.0


Caenorhabditis elegans
Caenorhabditis elegans
WBcel215


Cat (preview new assembly Felis_catus-6.2)
Felis catus
CAT


Chicken (preview new assembly Galgal4)
Gallus gallus
WASHUC2


Gibbon
Nomascus leucogenys
Nleu1.0


Gorilla
Gorilla gorilla
gorGor3.1


Guinea Pig
Cavia porcellus
cavPor3


Hedgehog
Erinaceus europaeus
HEDGEHOG


Horse
Equus caballus
EquCab2


Human
Homo sapiens
GRCh37


Hyrax
Procavia capensis
proCap1


Kangaroo rat
Dipodomys ordii
dipOrd1


Lamprey
Petromyzon Marinus
Pmarinus_7.0


Lesser hedgehog tenrec
Echinops telfairi
TENREC


Macaque
Macaca mulatta
MMUL_1


Platyfish (preview - assembly only)
Xiphophorus maculatus
Xipmac4.4.2


Platypus
Ornithorhynchus anatinus
OANA5


Rabbit
Oryctolagus cuniculus
oryCun2


Rat (preview new assembly Rnor_5.0)
Rattus norvegicus
RGSC3.4


Saccharomyces cerevisiae
Saccharomyces cerevisiae
EF4


Sheep (preview - assembly only)
Ovis aries
oviAri1


Shrew
Sorex araneus
COMMON_SHREW1

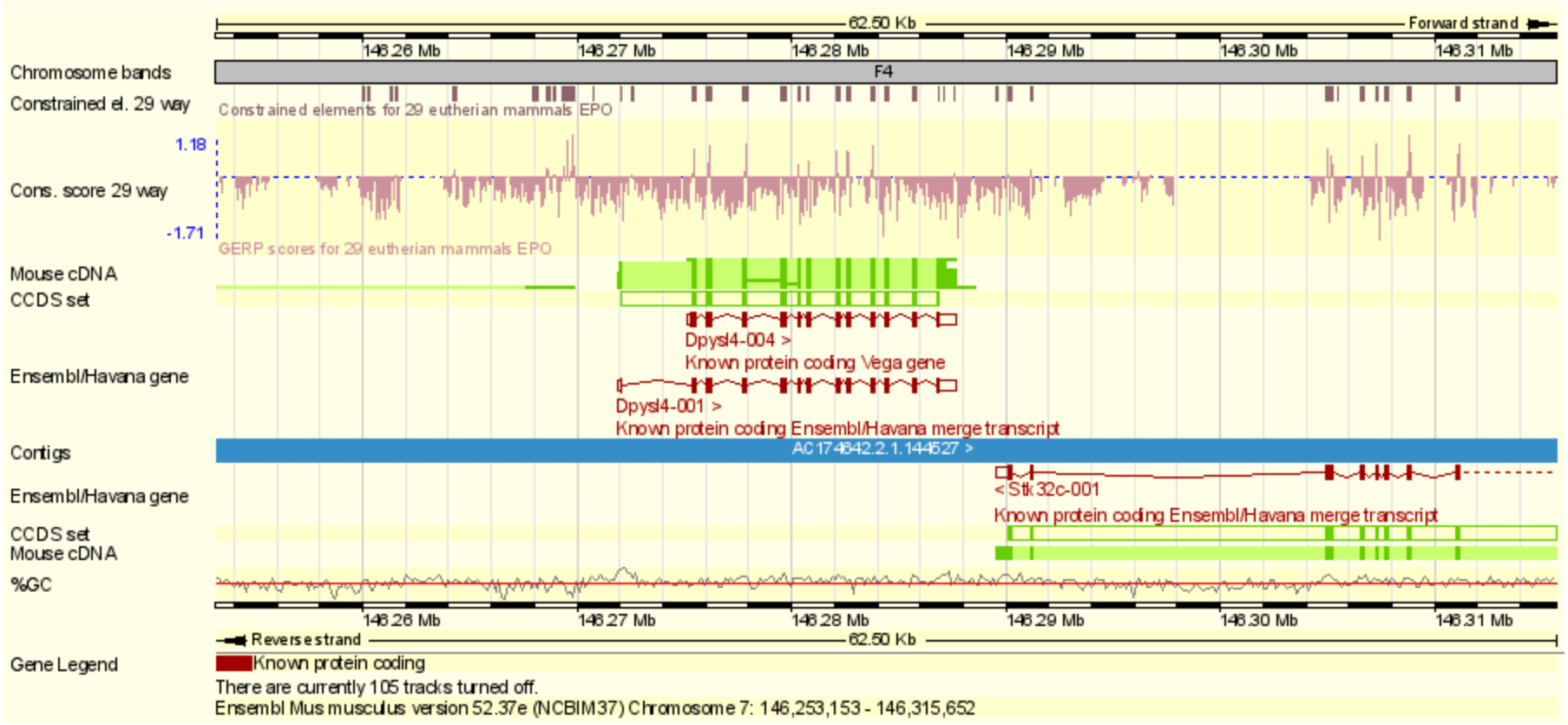

Sloth
Choloepus hoffmanni
choHof1


Spotted Gar (preview - assembly only)
Lepisosteus oculatus
LepOcu1


Squirrel
Ictidomys tridecemlineatus
spetri2

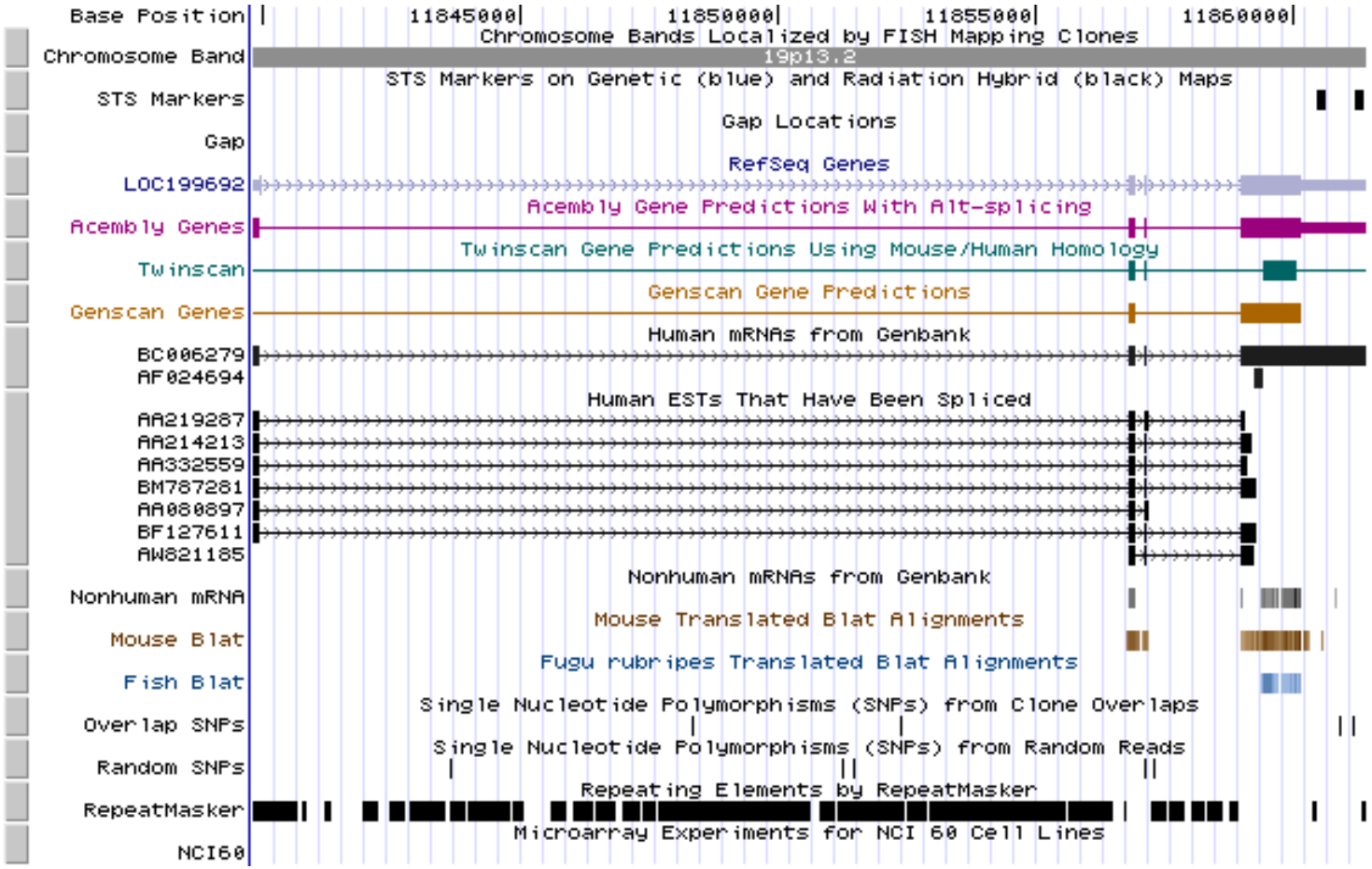

Squirrel monkey (preview - assembly only)
Saimiri boliviensis
SaiBol1.0

Ensembl: « chromosome view »



Genome browser UCSC

- http://www.genome.ucsc.edu/

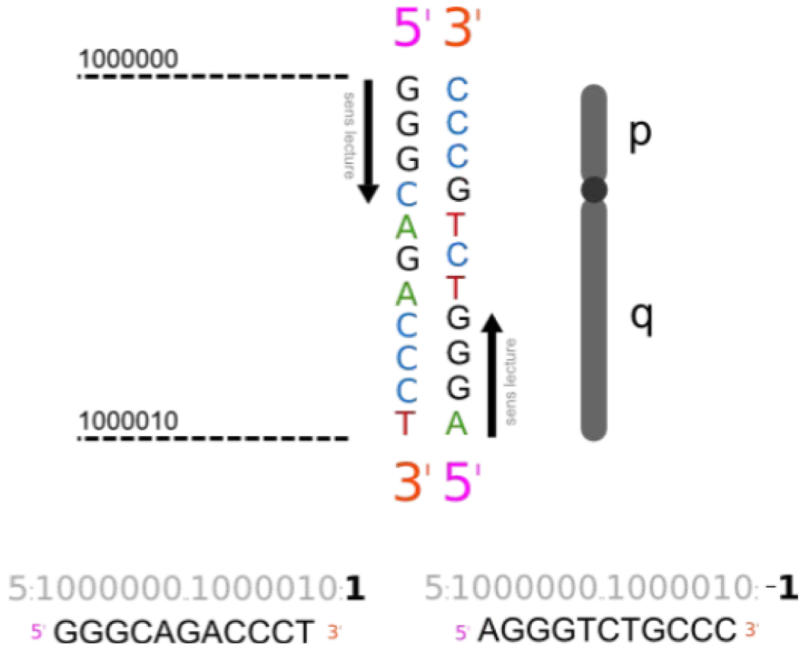


Accéder à une région génomique avec des API

API de la banque Ensembl:

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

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

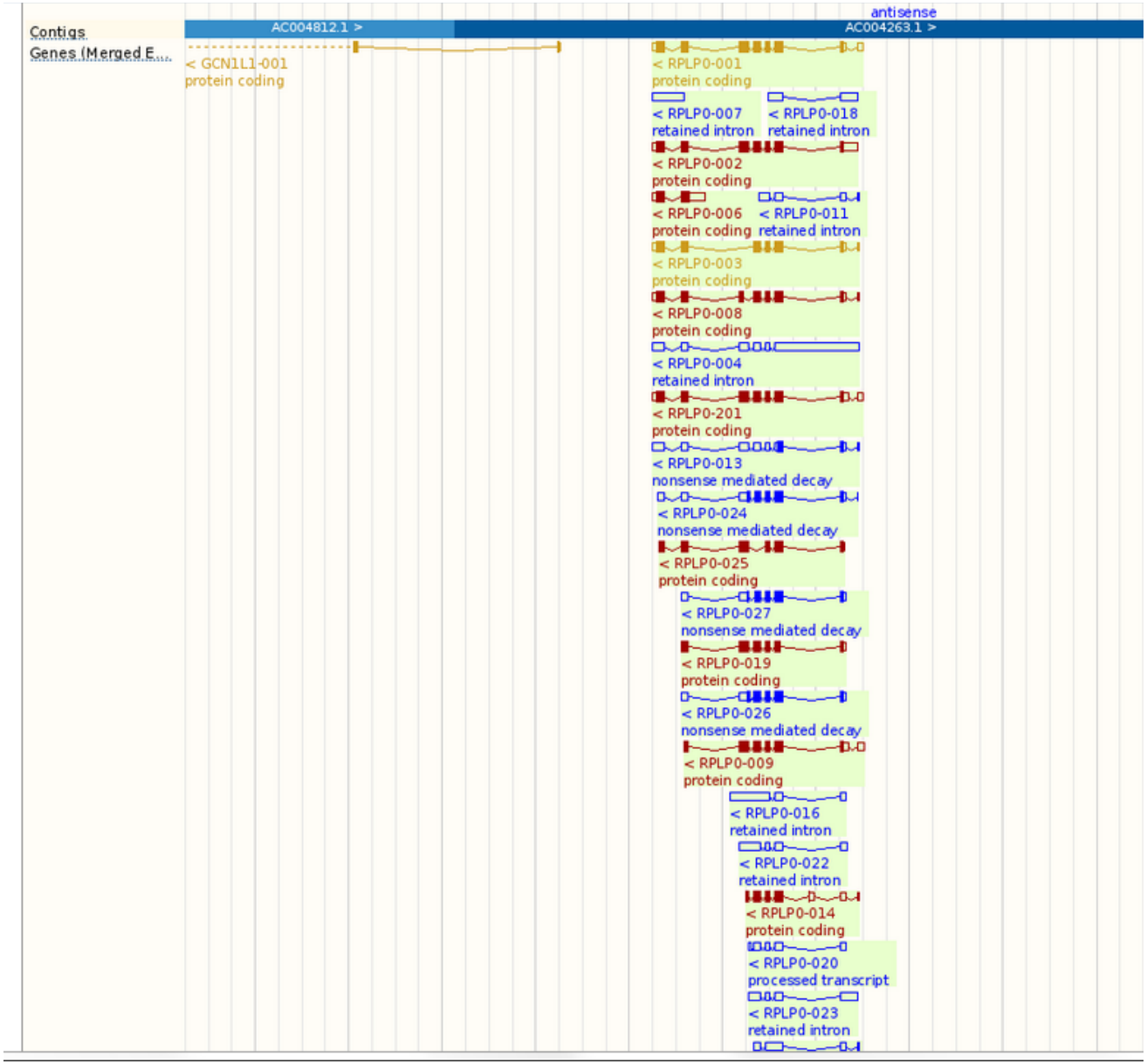


Mini TD

- Rendez-vous sur le site Ensembl
- Cherchez dans le génome humain le gène RPLP0
- Visualisez la région chromosomique, notez les coordonnées.
- Taille du gène? Combien d'introns, exons?
- Récupérer la séquence du mRNA
- Cherchez RPLP0 humain sur le site du NCBI
- Cherchez RPLP0 humain sur le site Uniprot
- Avec la commande wget (curl si vous êtes sur un Mac) et l'API Ensembl, récupérez la séquence du gène RPLP0 humain à partir de ses coordonnées.

  (ajouter options: « ?content-type=text/x-fasta » à l'URL)

Région chromosomique et annotation transcrits



Un transcrit (= mRNA)

Les variations du génome

En quoi nos génomes diffèrent-ils?

- Les individus d'espèces différentes ont des génomes différents par la taille, l'ordre et la nature des informations qu'ils contiennent.
- On considère souvent que 2 individus de la même espèce possèdent le « même » génome.
- En fait, le génome de chaque individu est unique. Chez l'homme le génome diffère de 0.1% entre 2 personnes non apparentées

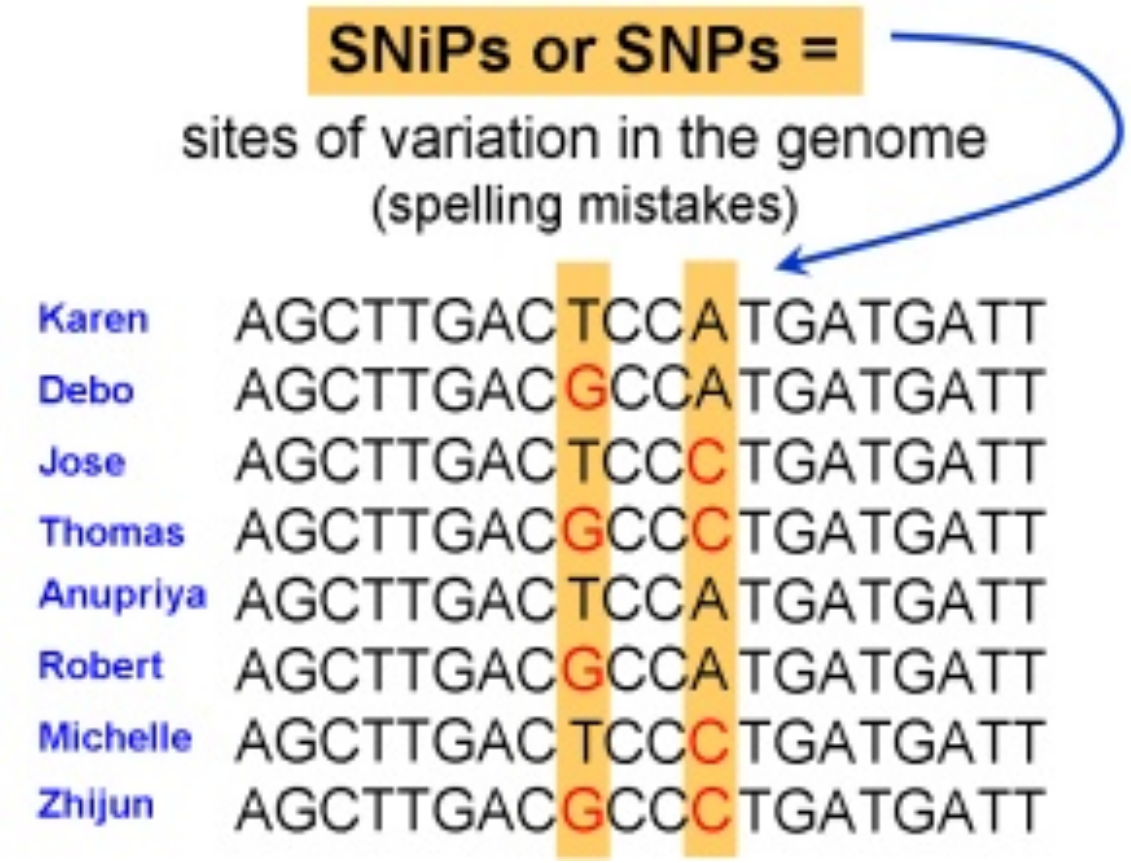
Les variations du génome dans une population

- **Très importantes médicalement**
  - Pharmacogénomique: comment chaque patient répond aux drogues
  - Marqueurs de susceptibilité aux maladies

- **Polymorphismes dans le génome humain**
  - Insertions, délétions, duplications, réarrangements
  - Microsatellites etc..
  - Single Nucleotide Polymorphism (SNP)
    - Les plus fréquents

Les SNP: Single Nucleotide Polymorphism

Les polymorphismes les plus fréquents dans la population



Dans le génome humain: **1 SNP chaque 500 bp** ⟶ **~6 million**

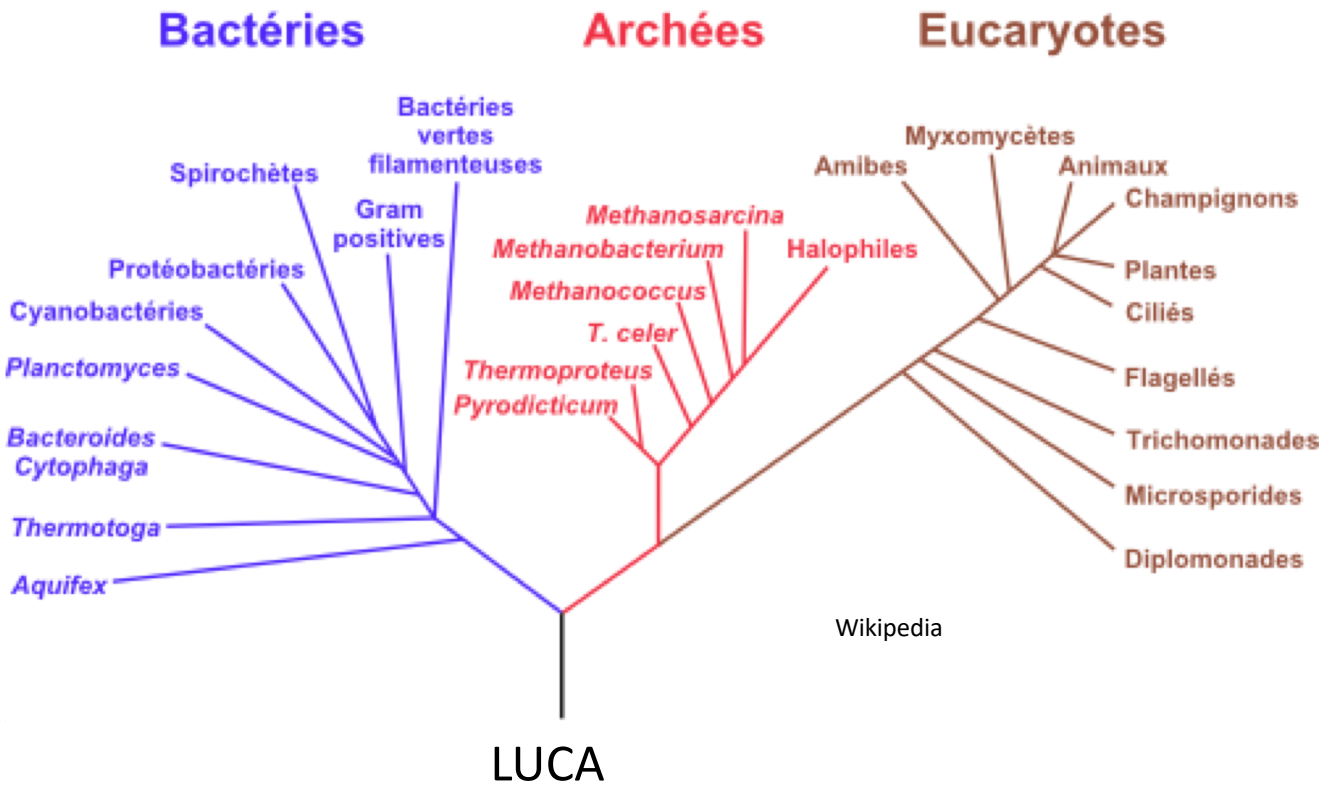
L'ADN de deux individus distincts est à 99,5% identique

cpmc.coriell.org/Sections/Medical/GeneInterac.

La ressemblance entre génomes

- Homme A / homme B
  - 99,9% identique (0,1% différence)
- Homme/chimpanzé
  - Codant: 98,5% identique
  - Non codant: ~96% identique
  - 90Mb d'insertions/délétions et 35 millions de différences ponctuelles
- Homme/souris
  - Codant: 90% identique
  - Non codant: la majorité est sans identité apparente, mais on trouve quand même de nombreux segments semblables (« conservés »)
- Homme/poulet
  - Codant: 80% identique
- Homme/poisson
  - Codant: 70% identique

Rappel: l'arbre du vivant



Wikipedia

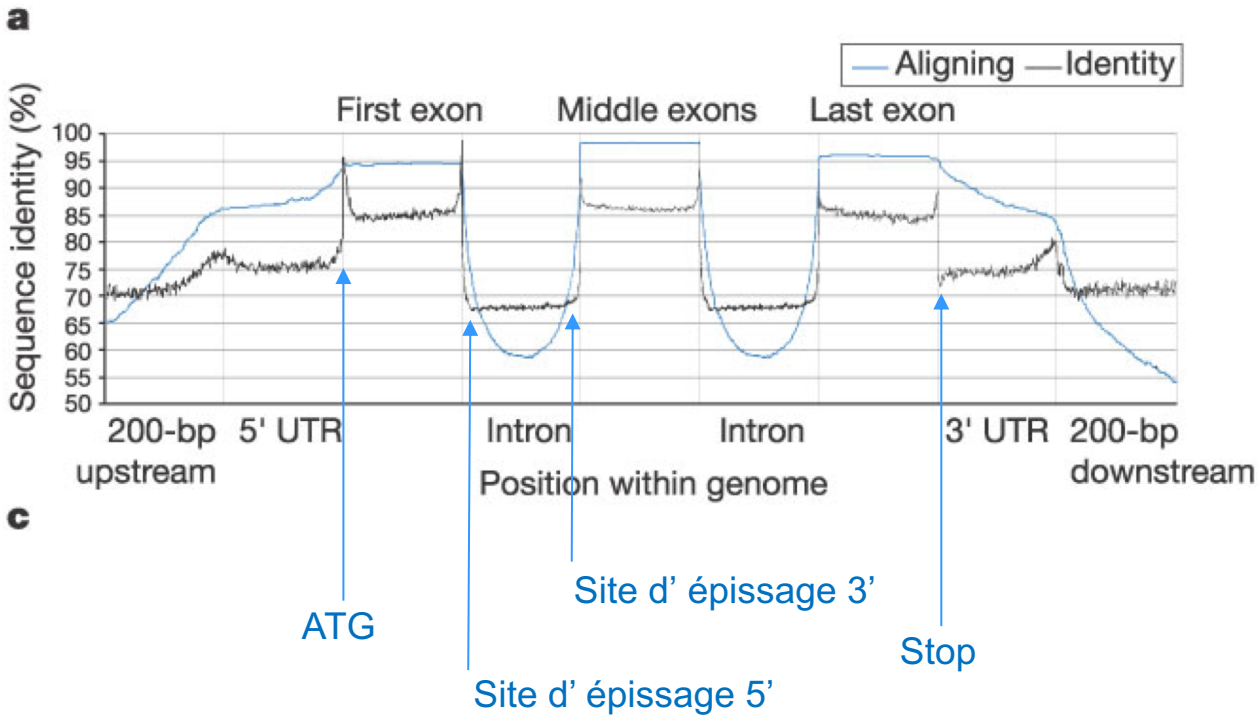
La conservation des séquences codantes ou non

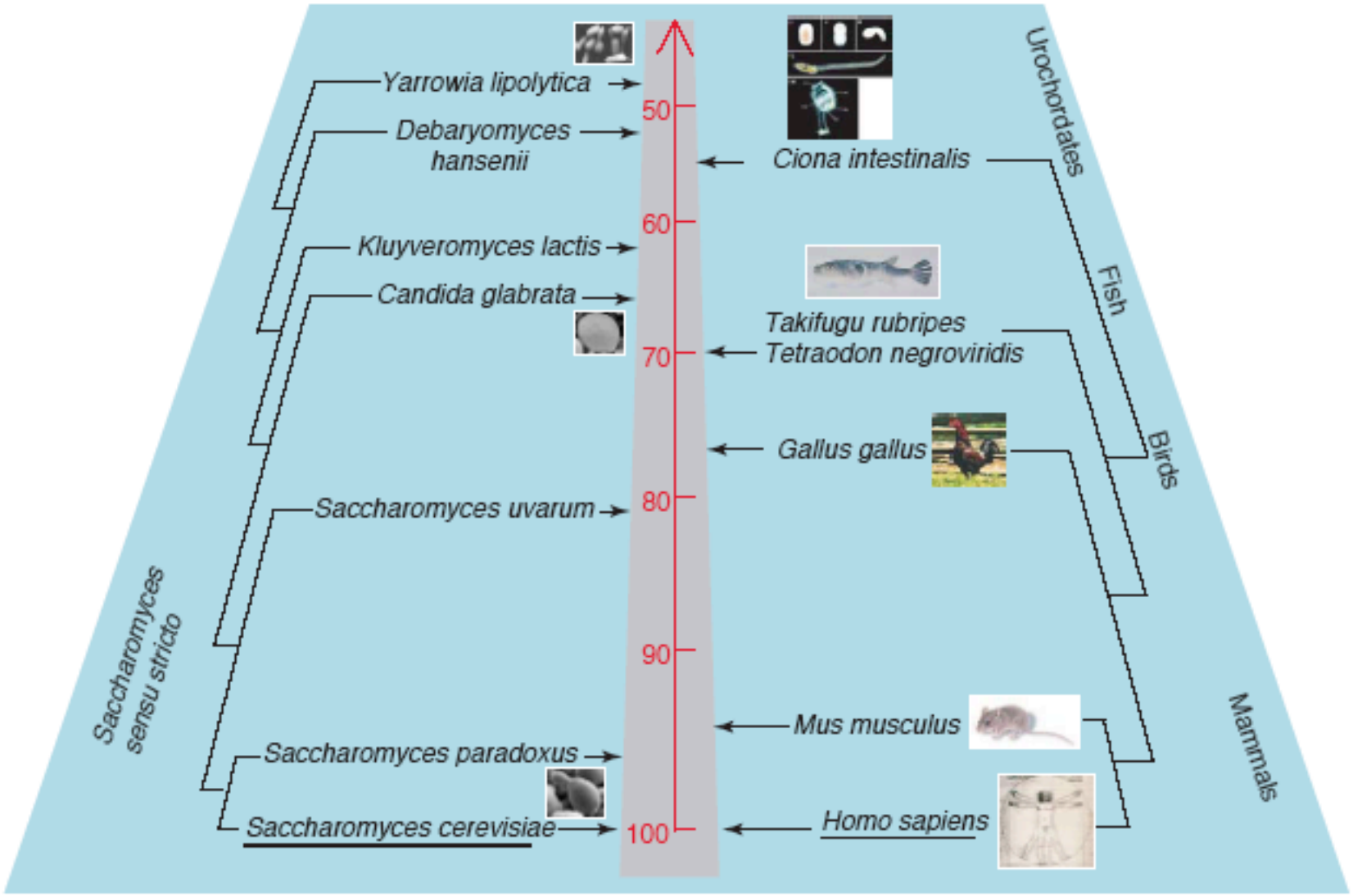
Figure de l'article de 2002 sur le génome de la souris.

"a wonderful visual guide to the most important features of mammalian genes" (Chris Pontig)

Identité homme-souris pour tous les couples de gènes orthologues

Divergence : 90M années





From B. Dujon, Trends in Genetics, 2006
Diatomées: Chris Bowler

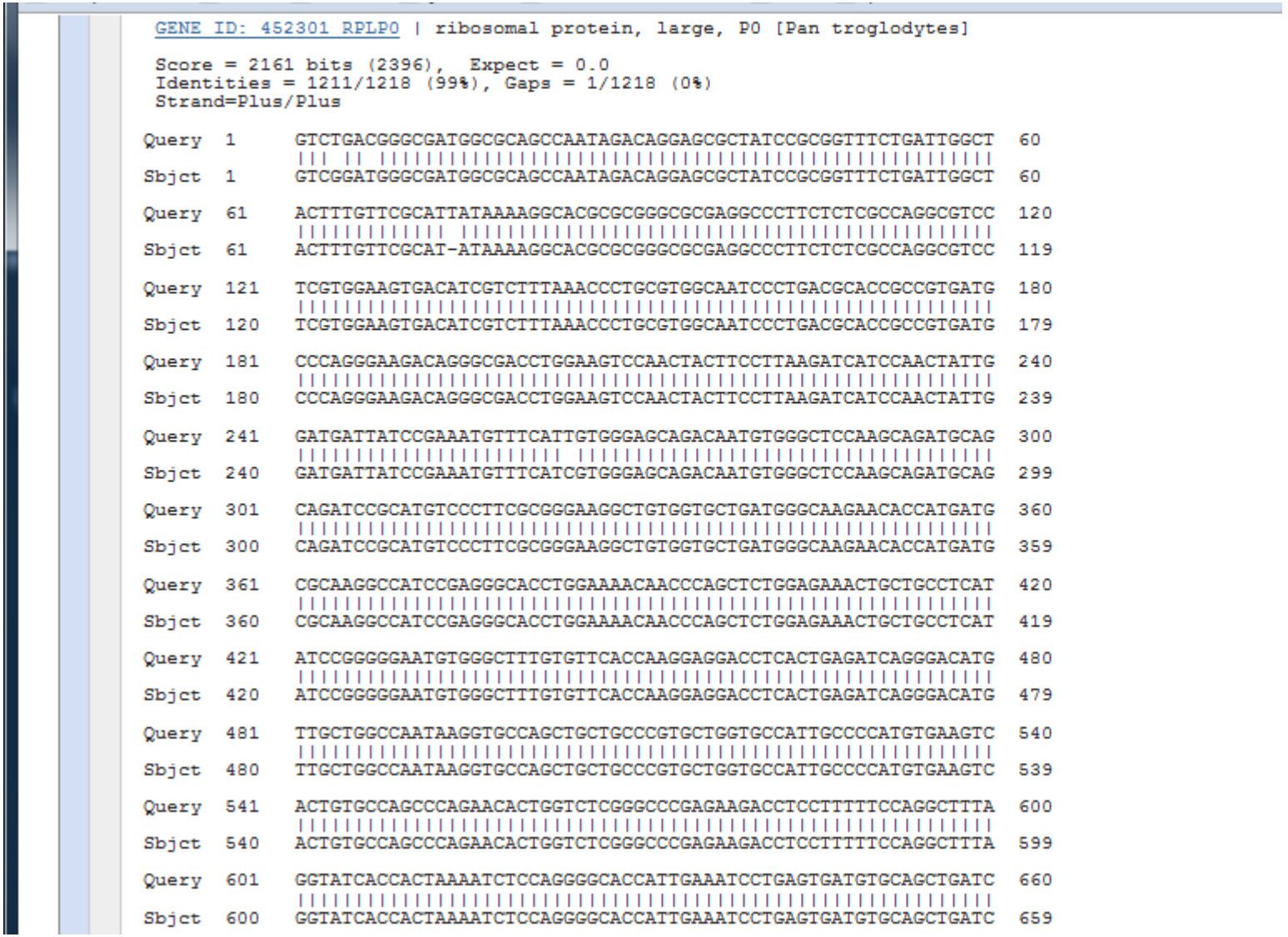
Comparer les séquences

RPLP0: un gène de protéine ribosomale universellement conservé.

mRNA humain de RPLP0 (exons seulement)

- >ENST00000392514 cdna:KNOWN_protein_coding
  GTCTGACGGGCGATGGCGCAGCCAATAGACAGGAGCGCTATCCGCGGTTTCTGATTGGCT
  ACTTTGTTCGCATTATAAAAGGCACGCGCGGGCGCGAGGCCCTTCTCTCGCCAGGCGTCC
  TCGTGGAAGTGACATCGTCTTTAAACCCTGCGTGGCAATCCCTGACGCACCGCCGTGATG
  CCCAGGGAAGACAGGGCGACCTGGAAGTCCAACTACTTCCTTAAGATCATCCAACTATTG
  GATGATTATCCGAAATGTTTCATTGTGGGAGCAGACAATGTGGGCTCCAAGCAGATGCAG
  CAGATCCGCATGTCCCTTCGCGGGAAGGCTGTGGTGCTGATGGGCAAGAACACCATGATG
  CGCAAGGCCATCCGAGGGCACCTGGAAAACAACCCAGCTCTGGAGAAACTGCTGCCTCAT
  ATCCGGGGGAATGTGGGCTTTGTGTTCACCAAGGAGGACCTCACTGAGATCAGGGACATG
  TTGCTGGCCAATAAGGTGCCAGCTGCTGCCCGTGCTGGTGCCATTGCCCCATGTGAAGTC
  ACTGTGCCAGCCCAGAACACTGGTCTCGGGCCCGAGAAGACCTCCTTTTTCCAGGCTTTA
  GGTATCACCACTAAAATCTCCAGGGGCACCATTGAAATCCTGAGTGATGTGCAGCTGATC
  AAGACTGGAGACAAAGTGGGAGCCAGCGAAGCCACGCTGCTGAACATGCTCAACATCTCC
  CCCTTCTCCTTTGGGCTGGTCATCCAGCAGGTGTTCGACAATGGCAGCATCTACAACCCT
  GAAGTGCTTGATATCACAGAGGAAACTCTGCATTCTCGCTTCCTGGAGGGTGTCCGCAAT
  GTTGCCAGTGTCTGTCTGCAGATTGGCTACCCAACTGTTGCATCAGTACCCCATTCTATC
  ATCAACGGGTACAAACGAGTCCTGGCCTTGTCTGTGGAGACGGATTACACCTTCCCACTT
  GCTGAAAAGGTCAAGGCCTTCTTGGCTGATCCATCTGCCTTTGTGGCTGCTGCCCCTGTG
  GCTGCTGCCACCACAGCTGCTCCTGCTGCTGCTGCAGCCCCAGCTAAGGTTGAAGCCAAG
  GAAGAGTCGGAGGAGTCGGACGAGGATATGGGATTTGGTCTCTTTGACTAATCACCAAAA
  AGCAACCAACTTAGCCAGTTTTATTTGCAAAACAAGGAAATAAAGGCTTACTTCTTTAAA
  AAGTCTCTGGACTCTTAA

Alignement avec mRNA RPLP0 de *Pan troglodytes*



GENE ID: 452301 RPLP0 | ribosomal protein, large, P0 [Pan troglodytes]

Score = 2161 bits (2396), Expect = 0.0
Identities = 1211/1218 (99%), Gaps = 1/1218 (0%)
Strand=Plus/Plus

```
Query  1     GTCTGACGGGCGATGGCGCAGCCAATAGACAGGAGCGCTATCCGCGGTTTCTGATTGGCT  60
             ||| || ||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1     GTCGGATGGGCGATGGCGCAGCCAATAGACAGGAGCGCTATCCGCGGTTTCTGATTGGCT  60

Query  61    ACTTTGTTCGCATTATAAAAGGCACGCGCGGGCGCGAGGCCCTTCTCTCGCCAGGCGTCC  120
             |||||||||||||| ||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61    ACTTTGTTCGCAT-ATAAAAGGCACGCGCGGGCGCGAGGCCCTTCTCTCGCCAGGCGTCC  119

Query  121   TCGTGGAAGTGACATCGTCTTTAAACCCTGCGTGGCAATCCCTGACGCACCGCCGTGATG  180
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  120   TCGTGGAAGTGACATCGTCTTTAAACCCTGCGTGGCAATCCCTGACGCACCGCCGTGATG  179

Query  181   CCCAGGGAAGACAGGGCGACCTGGAAGTCCAACTACTTCCTTAAGATCATCCAACTATTG  240
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  180   CCCAGGGAAGACAGGGCGACCTGGAAGTCCAACTACTTCCTTAAGATCATCCAACTATTG  239

Query  241   GATGATTATCCGAAATGTTTCATTGTGGGAGCAGACAATGTGGGCTCCAAGCAGATGCAG  300
             ||||||||||||||||||||||||| ||||||||||||||||||||||||||||||||||||
Sbjct  240   GATGATTATCCGAAATGTTTCATCGTGGGAGCAGACAATGTGGGCTCCAAGCAGATGCAG  299

Query  301   CAGATCCGCATGTCCCTTCGCGGGAAGGCTGTGGTGCTGATGGGCAAGAACACCATGATG  360
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  300   CAGATCCGCATGTCCCTTCGCGGGAAGGCTGTGGTGCTGATGGGCAAGAACACCATGATG  359

Query  361   CGCAAGGCCATCCGAGGGCACCTGGAAAACAACCCAGCTCTGGAGAAACTGCTGCCTCAT  420
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  360   CGCAAGGCCATCCGAGGGCACCTGGAAAACAACCCAGCTCTGGAGAAACTGCTGCCTCAT  419

Query  421   ATCCGGGGGAATGTGGGCTTTGTGTTCACCAAGGAGGACCTCACTGAGATCAGGGACATG  480
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  420   ATCCGGGGGAATGTGGGCTTTGTGTTCACCAAGGAGGACCTCACTGAGATCAGGGACATG  479

Query  481   TTGCTGGCCAATAAGGTGCCAGCTGCTGCCCGTGCTGGTGCCATTGCCCCATGTGAAGTC  540
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  480   TTGCTGGCCAATAAGGTGCCAGCTGCTGCCCGTGCTGGTGCCATTGCCCCATGTGAAGTC  539

Query  541   ACTGTGCCAGCCCAGAACACTGGTCTCGGGCCCGAGAAGACCTCCTTTTTCCAGGCTTTA  600
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  540   ACTGTGCCAGCCCAGAACACTGGTCTCGGGCCCGAGAAGACCTCCTTTTTCCAGGCTTTA  599

Query  601   GGTATCACCACTAAAATCTCCAGGGGCACCATTGAAATCCTGAGTGATGTGCAGCTGATC  660
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  600   GGTATCACCACTAAAATCTCCAGGGGCACCATTGAAATCCTGAGTGATGTGCAGCTGATC  659
```

Alignement avec mRNA RPLP0 de *Danio rerio*

```
>ref|NM_131580.2| Danio rerio ribosomal protein, large, P0 (rplp0), mRNA
 gb|BC062854.1| Danio rerio ribosomal protein, large, P0, mRNA (cDNA clone MGC:77791
IMAGE:7001273), complete cds
Length=1105

 GENE ID: 58101 rplp0 | ribosomal protein, large, P0 [Danio rerio]
(Over 10 PubMed links)

 Score =  888 bits (984),  Expect = 0.0
 Identities = 801/1000 (80%), Gaps = 10/1000 (1%)
 Strand=Plus/Plus

Query  137   GTCTTTAAACC--CTGCGTGGCAATCCCTG-ACGCACCGCCGTGATGCCCAGGGAAGACA  193
             ||||||||||   |||     | |||| || | |||| ||   ||||||||||||||||
Sbjct  31    GTCTTTAAACCGGCTGTTCACCGATCCTTGGAAGCACTGCAAAGATGCCCAGGGAAGACA  90

Query  194   GGGCGACCTGGAAGTCCAACTACTTCCTTAAGATCATCCAACTATTGGATGATTATCCGA  253
             |||| || |||||||||||||||||| || || ||||||||||  ||||||| || || |
Sbjct  91    GGGCCACGTGGAAGTCCAACTACTTTCTGAAAATCATCCAACTGCTGGATGACTACCCCA  150

Query  254   AATGTTTCATTGTGGGAGCAGACAATGTGGGCTCCAAGCAGATGCAGCAGATCCGCATGT  313
             | |||||||| ||||| |||||||||| ||||||||||||||||||   |||||  |||
Sbjct  151   AGTGTTTCATCGTGGGCGCAGACAATGTCGGCTCCAAGCAGATGCAGACCATCCGTCTGT  210

Query  314   CCCTTCGCGGGAAGGCTGTGGTGCTGATGGGCAAGAACACCATGATGCGCAAGGCCATCC  373
             |||| || || ||||| || ||||| ||||| || |||||||||||| | ||||||||||
Sbjct  211   CCCTGCGGGGCAAGGCCGTCGTGCTCATGGGGAAAAACACCATGATGAGGAAGGCCATCC  270

Query  374   GAGGGCACCTGGAAAACAACCCAGCTCTGGAGAAACTGCTGCCTCATATCCGGGGGAATG  433
             | || |||||||||||||||||||||||||||||  ||||| || || ||||| ||||| |
Sbjct  271   GTGGCCACCTGGAAAACAACCCAGCTCTGGAGAGGCTGCTTCCCCACATCCGCGGGAACG  330

Query  434   TGGGCTTTGTGTTCACCAAGGAGGACCTCACTGAGATCAGGGACATGTTGCTGGCCAATA  493
             ||||||| || |||||||||||||| || |||||| || | ||| || |||||||| | |
Sbjct  331   TGGGCTTCGTCTTCACCAAGGAGGATCTGACTGAGGTCCGAGACCTGCTGCTGGCAAACA  390

Query  494   AGGTGCCAGCTGCTGCCCGTGCTGGTGCCATTGCCCCATGTGAAGTCACTGTGCCAGCCC  553
             | ||||| ||||||||||||||||||||| ||||| ||||| || || ||||| ||||
Sbjct  391   AAGTGCCCGCTGCTGCCCGTGCTGGTGCCATCGCCCCCTGTGAGGTGACCGTGCCGGCCC  450

Query  554   AGAACACTGGTCTCGGGCCCGAGAAGACCTCCTTTTTCCAGGCTTTAGGTATCACCACTA  613
             ||||||| || ||||| || |||||||||| || |||||||||| || |||||||| |
Sbjct  451   AGAACACCGGGCTCGGTCCTGAGAAGACCTCTTTCTTCCAGGCTTTGGGAATCACCACCA  510

Query  614   AAATCTCCAGGGGCACCATTGAAATCCTGAGTGATGTGCAGCTGATCAAGACTGGAGACA  673
             | |||||||| || |||||||||||| ||||||| || ||||||||||  |||||||||
```

Alignement avec mRNA RPLP0 de *Drosophia melanogaster*

>ref|NM_079487.3| Drosophila melanogaster ribosomal protein LP0 (RpLP0), mRNA
Length=1261

GENE ID: 40451 RpLP0 | Ribosomal protein LP0 [Drosophila melanogaster]
(Over 10 PubMed links)

Score = 340 bits (376), Expect = 2e-90
Identities = 423/577 (73%), Gaps = 2/577 (0%)
Strand=Plus/Plus

```
Query  184  AGGGAAGACAGGGCGACCTGGAAGTCCAACTACTTCCTTAAGATCATCCAACTATTGGAT  243
            |||||  ||| |||  | ||||||  |  | ||||||  | ||| |  |  |||| || |||
Sbjct  142  AGGGAGAACAAGGCAGCGTGGAAGGCTCAGTACTTCATCAAGGTTGTGGAACTGTTCGAT  201

Query  244  GATTATCCGAAATGTTTCATTGTGGGAGCAGACAATGTGGGCTCCAAGCAGATGCAGCAG  303
            || |  || || || ||||| ||||| || ||||| ||||||||||||||||||||||  |
Sbjct  202  GAGTTCCCAAAGTGCTTCATCGTGGGCGCCGACAACGTGGGCTCCAAGCAGATGCAGAAC  261

Query  304  ATCCGCATGTCCCTTCGCGGGAAGGCTGTGGTGCTGATGGGCAAGAACACCATGATGCGC  363
            |||||  |    ||| || ||    ||| || ||||| ||||||||||||||||||||||||
Sbjct  262  ATCCGTACCAGCCTGCGTGGACTGGCCGTCGTGCTTATGGGCAAGAACACCATGATGCGC  321

Query  364  AAGGCCATCCGAGGGCACCTGGAAAACAACCCAGCTCTGGAGAAACTGCTGCCTCATATC  423
            |||||||||| || || ||||| ||||||||     |||||||| ||||| || || |||
Sbjct  322  AAGGCCATCCGCGGTCATCTGGAGAACAACCCGCAGCTGGAGAAGCTGCTACCCCACATC  381

Query  424  CGGGGGAATGTGGGCTTTGTGTTCACCAAGGAGGACCTCACTGAGATCAGGGACATGTTG  483
              ||| || ||||| || ||||||||||||||  || ||| | ||| |   | |||| | ||
Sbjct  382  AAGGGCAACGTGGGATTCGTGTTCACCAAGGGCGATCTCGCCGAGGTGCGCGACAAGCTG  441

Query  484  CTGGCCAATAAGGTGCCAGCTGCTGCCCGTGCTGGTGCCATTGCCCCATGTG-AAGTCAC  542
            ||||      ||||||||   ||   | ||||||  | ||  || ||||||||  | ||  | ||||
Sbjct  442  CTGGAGTCCAAGGTGCGCGCCCCCGCCCGTCCCGGCGCTATTGCCCC-TCTGCACGTCAT  500

Query  543  TGTGCCAGCCCAGAACACTGGTCTCGGGCCCGAGAAGACCTCCTTTTTCCAGGCTTTAGG  602
              | || || |||||||| ||   | || ||||||||||||    || |||||||   |
Sbjct  501  CATCCCGGCGCAGAACACCGGCTTGGGACCCGAGAAGACCAGTTTCTTCCAGGCCCTGTC  560

Query  603  TATCACCACTAAAATCTCCAGGGGCACCATTGAAATCCTGAGTGATGTGCAGCTGATCAA  662
             ||| | || ||||| |||| ||| || |||||||| | |  |||||||   |  | ||
Sbjct  561  CATCCCGACCAAAATTTCCAAGGGAACAATTGAAATCATCAACGATGTGCCCATCCTGAA  620

Query  663  GACTGGAGACAAAGTGGGAGCCAGCGAAGCCACGCTGCTGAACATGCTCAACATCTCCCC  722
            | |||| ||||| || || |||   ||| || || ||||| ||||||| | ||||||| ||
Sbjct  621  GCCTGGCGACAAGGTCGGCGCCTCCGAGGCGACACTGCTCAACATGTTGAACATCTCGCC  680
```

Alignement avec mRNA RPLP0 de *Arabidopsis thaliana*

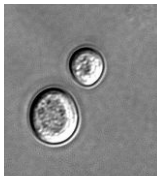
>ref|NM_129559.2| UEGM Arabidopsis thaliana 60S acidic ribosomal protein P0-1 (AT2G400
mRNA, complete cds
Length=1153

GENE ID: 818589 AT2G40010 | 60S acidic ribosomal protein P0-1
[Arabidopsis thaliana] (10 or fewer PubMed links)

Score = 66.2 bits (72), Expect = 6e-08
Identities = 261/402 (65%), Gaps = 14/402 (3%)
Strand=Plus/Plus

```
Query  409  CTGCTGCCTCATATCCGGGGGAATGTGGGCTTTGTGTTCACCAAGGAGGACCTCACTGAG  468
            ||  ||  ||||  |  |  |  ||||||||||  ||  ||   |  |||||  ||||   |||  |  |   ||
Sbjct  241  CTCCTTCCTCTTCTTCAGGGGAATGTTGGGTTGATCTTCACTAAGGGTGACTTGAAGGAA  300

Query  469  ATCAGGGAC-ATGTTGCTGGCCAATAAGGTGCCAGCTGCTGCCCGTGCTGGTGCCATTGC  527
             ||||  ||   |  ||||||     |   |||||    ||||  ||||  ||||   |||      |  ||
Sbjct  301  GTCAGTGAAGAGGTTGCTAAGTAC-AAGGTTGGAGCTCCTGCTCGTGTAGGTTTAGTCGC  359

Query  528  CCCATGTGAAGTCACTGTGCCAGCCCAG--AACACTGGTCTCGGGCCCGAGAAGACCTCC  585
             |||   |||  ||     ||||  |   ||||   ||||||||||||  |    ||       |||||||||
Sbjct  360  TCCAATTGATGTGGTCGTGCAA--CCAGGCAACACTGGTCTTGACCCTTCACAGACCTCC  417

Query  586  TTTTTCCAG--GCTTTAGGTATCACCACTAAAATCTCCAGGGGCACCATTGAAATCCTGA  643
            ||  ||||||   ||||  |    ||   |  ||  ||||||   ||   ||  ||   ||||  |||  |  |
Sbjct  418  TTCTTCCAGGTGCTTAA--CATTCCAACCAAAATCAACAAAGGTACGGTTGAGATCATAA  475

Query  644  GTGATGTGCAGCTGATCAAGACTGGAGACAAAGTGGGAGCCAGCGAAGCCACGCTGCTGA  703
                ||||  ||||  |||||||   ||  ||||||||  ||   |    |||  ||   ||||  ||
Sbjct  476  CCCCTGTGGAGCTCATCAAGAAAGGCGACAAAGTCGGTTCATCCGAGGCTGCGCTTCTTG  535

Query  704  ACATGCTCAACATCTCCCCCTTCTCCTTTGGGCTGGTCATCCAGCAGGTGTTCGACAATG  763
             ||  |||     |||    |||||  ||  |  |||  ||  ||   |   ||    ||  |  |||  ||||
Sbjct  536  CCAAGCTTGGAATCAGGCCCTTTTCGTATGGTCTCGTTGTTGAGTCAGTCTACGATAATG  595

Query  764  G--CAGCATCTACAACCCTGAAGTGCTTGATATCACAGAGGA  803
             |   |||   |  |   ||||||||||||||||  |   ||||  ||  ||
Sbjct  596  GGTCAG--TGTTTAACCCTGAAGTGCTTAACCTCACTGAAGA  635
```

Alignement avec mRNA RPLP0 de *Schizosaccharomyces pombe*



>ref|NM_001023549.1| Schizosaccharomyces pombe 972h- 60S acidic ribosomal protein Rpp1-3 (rpp103), mRNA
Length=330

GENE ID: 2539598 rpp103 | 60S acidic ribosomal protein Rpp1-3
[Schizosaccharomyces pombe 972h-] (10 or fewer PubMed links)

Score = 53.6 bits (58), Expect = 4e-04
Identities = 77/108 (71%), Gaps = 3/108 (3%)
Strand=Plus/Plus

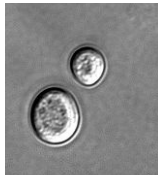
```
Query  1023  tgctgccaccacagctgctcctgctgctgctgcagccccagctAAGGTTGAAGCCAAGGA  1082
             ||||| | || | ||| || |||  ||||| | ||    |   ||||  ||||||||||
Sbjct  225   TGCTGGCGCCGCCGCTCCTGCTGAAGCTGCCGAAGAAGAAAAGAAGGAAGAAGCCAAGGA  284

Query  1083  AGAGTCGGAGGAGTCGGACGAGGATATGGGATTTGGTCTCTTTGACTA  1130
               ||   ||| ||||| || ||||| ||||| |||||   | ||||||||
Sbjct  285   GGA---GGAAGAGTCTGATGAGGACATGGGTTTTGGCTTGTTTGACTA  329
```

RPLP0: la protéine

>sp|P05388|RLA0_HUMAN 60S acidic ribosomal protein P0 OS=Homo sapiens GN=RPLP0 PE=1 SV=1
MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGKAVVLMGKNTM
MRKAIRGHLENNPALEKLLPHIRGNVGFVFTKEDLTEIRDMLLANKVPAAARAGAIAPCE
VTVPAQNTGLGPEKTSFFQALGITTKISRGTIEILSDVQLIKTGDKVGASEATLLNMLNI
SPFSFGLVIQQVFDNGSIYNPEVLDITEETLHSRFLEGVRNVASVCLQIGYPTVASVPHS
IINGYKRVLALSVETDYTFPLAEKVKAFLADPSAFVAAAPVAAATTAAPAAAAAPAKVEA KEESEESDEDMGFGLFD

```
>sp|P19889.1|RLA0_DROME RecName: Full=60S acidic ribosomal protein P0; AltName: Full=Apurinic
endonuclease; AltName: Full=DNA-(apurinic
or apyrimidinic site) lyase
Length=317

 Score =  417 bits (1072),  Expect = 2e-146, Method: Compositional matrix adjust.
 Identities = 210/317 (66%), Positives = 255/317 (80%), Gaps = 0/317 (0%)

Query  1    MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGKAVVLMGKNTM  60
            M RE++A WK+ YF+K+++L D++PKCFIVGADNVGSKQMQ IR SLRG AVVLMGKNTM
Sbjct  1    MVRENKAAWKAQYFIKVVELFDEFPKCFIVGADNVGSKQMQNIRTSLRGLAVVLMGKNTM  60

Query  61   MRKAIRGHLENNPALEKLLPHIRGNVGFVFTKEDLTEIRDMLLANKVPAAARAGAIAPCE  120
            MRKAIRGHLENNP LEKLLPHI+GNVGFVFTK DL E+RD LL +KV A AR GAIAP
Sbjct  61   MRKAIRGHLENNPQLEKLLPHIKGNVGFVFTKGDLAEVRDKLLESKVRAPARPGAIAPLH  120

Query  121  VTVPAQNTGLGPEKTSFFQALGITTKISRGTIEILSDVQLIKTGDKVGASEATLLNMLNI  180
            V +PAQNTGLGPEKTSFFQAL I TKIS+GTIEI++DV ++K GDKVGASEATLLNMLNI
Sbjct  121  VIIPAQNTGLGPEKTSFFQALSIPTKISKGTIEIINDVPILKPGDKVGASEATLLNMLNI  180

Query  181  SPFSFGLVIQQVFDNGSIYNPEVLDITEETLHSRFLEGVRNVASVCLQIGYPTVASVPHS  240
            SPFS+GL++ QV+D+GSI++PE+LDI  E L ++F +GV N+A+VCL +GYPT+AS PHS
Sbjct  181  SPFSYGLIVNQVYDSGSIFSPEILDIKPEDLRAKFQQGVANLAAVCLSVGYPTIASAPHS  240

Query  241  IINGYKRVLALSVETDYTFPLAEKVKAFLADPSAFVAAAPVAAATTAAPAAAAAPAKVEA  300
            I NG+K +LA++  T+  F  A  +K ++ DPS F AAA  +AA  A  A         +
Sbjct  241  IANGFKNLLAIAATTEVEFKEATTIKEYIKDPSKFAAAASASAAPAAGGATEKKEEAKKP  300

Query  301  KEESEESDEDMGFGLFD  317
            + ESEE D+DMGFGLFD
Sbjct  301  ESESEEEDDDMGFGLFD  317
```

```
>sp|P57691.1|RLA03_ARATH RecName: Full=60S acidic ribosomal protein P0-3
Length=323

 GENE ID: 820296 AT3G11250 | 60S acidic ribosomal protein P0-3
[Arabidopsis thaliana] (10 or fewer PubMed links)

 Score =  311 bits (798),  Expect = 1e-104, Method: Compositional matrix adjust.
 Identities = 168/321 (52%), Positives = 220/321 (69%), Gaps = 4/321 (1%)

Query  1    MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGKAVVLMGKNTM  60
            M +  +A  K  Y  K+ QL+D+Y +  +V ADNVGS Q+Q IR  LRG +VVLMGKNTM
Sbjct  1    MVKATKAEKKIAYDTKLCQLIDEYTQILVVAADNVGSTQLQNIRKGLRGDSVVLMGKNTM  60

Query  61   MRKAIRGHLEN--NPALEKLLPHIRGNVGFVFTKEDLTEIRDMLLANKVPAAARAGAIAP  118
            M++++R H EN  N A+  LLP ++GNVG +FTK DL E+ + +   KV A AR G +AP
Sbjct  61   MKRSVRIHSENSGNTAILNLLPLLQGNVGLIFTKGDLKEVSEEVAKYKVGAPARVGLVAP  120

Query  119  CEVTVPAQNTGLGPEKTSFFQALGITTKISRGTIEILSDVQLIKTGDKVGASEATLLNML  178
             +V V   NTGL P +TSFFQ L I TKI++GT+EI++ V+LIK GDKVG+SEA LL  L
Sbjct  121  IDVVVQPGNTGLDPSQTSFFQVLNIPTKINKGTVEIITPVELIKQGDKVGSSEAALLAKL  180

Query  179  NISPFSFGLVIQQVFDNGSIYNPEVLDITEETLHSRFLEGVRNVASVCLQIGYPTVASVP  238
             I PFS+GLV+Q V+DNGS+++PEVLD+TE+ L  +F  G+  V S+ L + YPT+A+ P
Sbjct  181  GIRPFSYGLVVQSVYDNGSVFSPEVLDLTEDQLVEKFASGISMVTSLALAVSYPTLAAAP  240

Query  239  HSIINGYKRVLALSVETDYTFPLAEKVKAFLADPSAFVAAAPVAAATTAAPAAAAAPAKV  298
            H  IN YK  LA++V TDYTFP AEKVK FL DPS FV AA   +A     +A A  A
Sbjct  241  HMFINAYKNALAIAVATDYTFPQAEKVKEFLKDPSKFVVAAAAVSADAGGGSAQAGAAAK  300

Query  299  EAKEESEESDEDM--GFGLFD  317
              +++ E  +ED   GFGLFD
Sbjct  301  VEEKKEESDEEDYEGGFGLFD  321
```

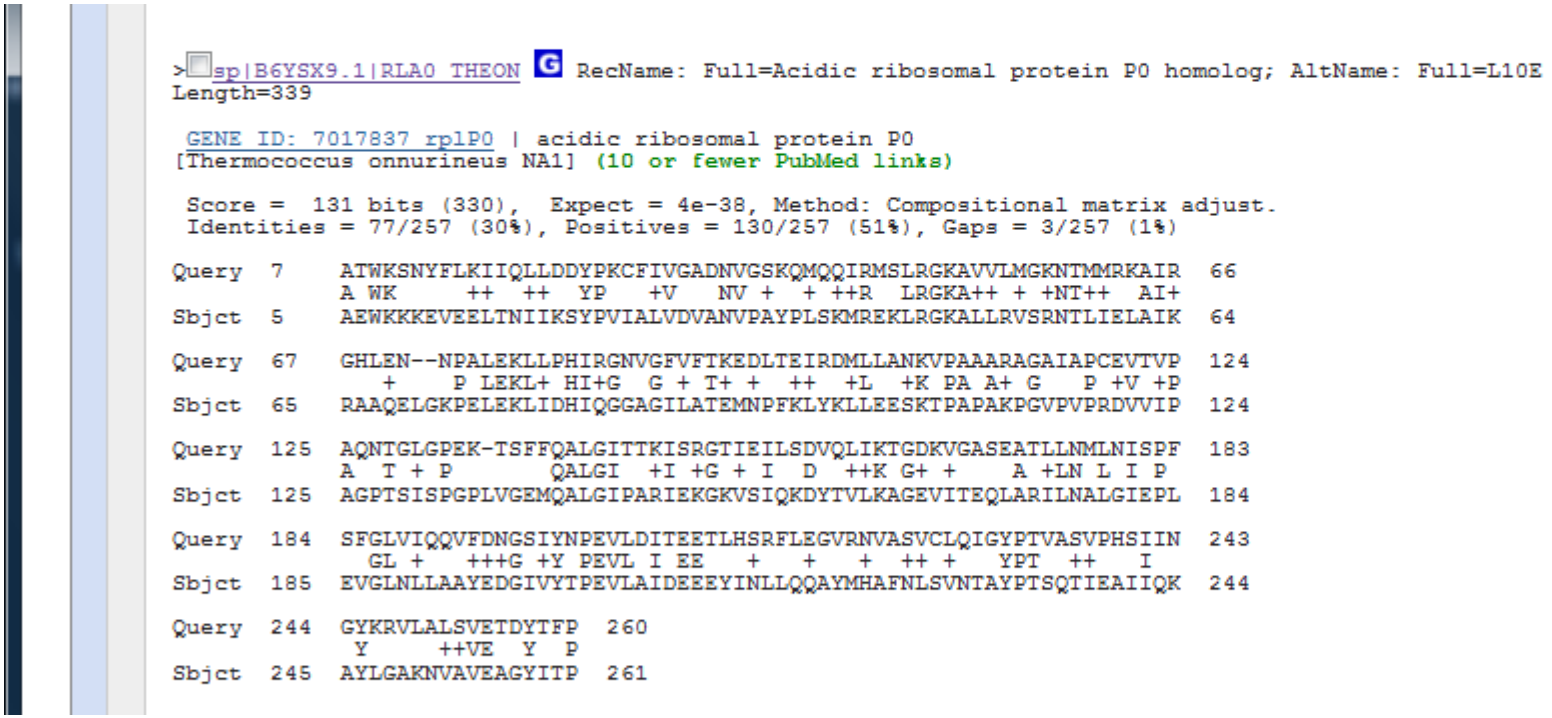


>sp|O74864.1|RLA0_SCHPO RecName: Full=60S acidic ribosomal protein P0
Length=312

GENE ID: 2538893 rpp0 | 60S acidic ribosomal protein Rpp0 (predicted)
[Schizosaccharomyces pombe 972h-] (10 or fewer PubMed links)

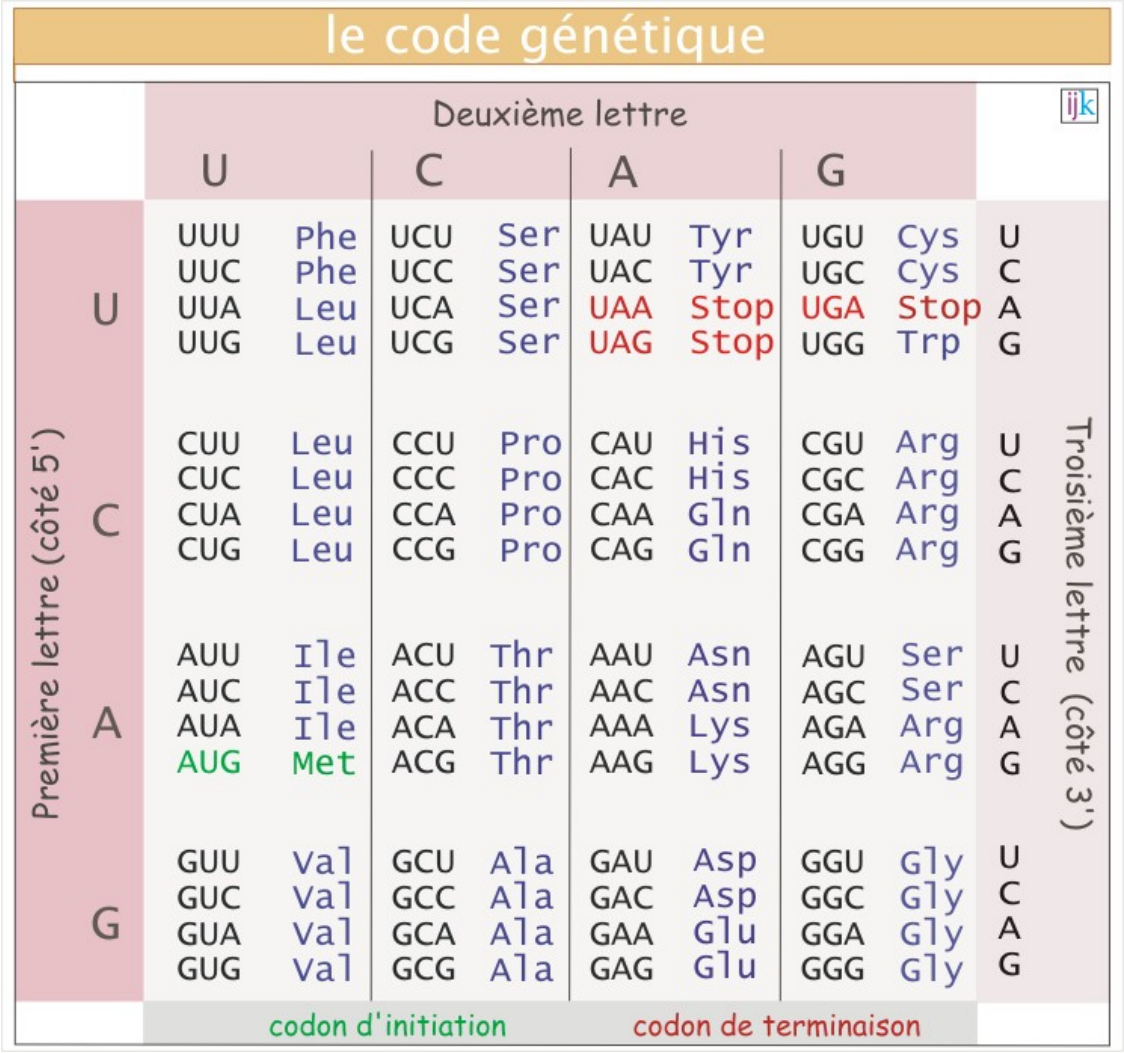
Score = 318 bits (816), Expect = 1e-107, Method: Compositional matrix adjust.
Identities = 170/308 (55%), Positives = 219/308 (71%), Gaps = 3/308 (1%)

```
Query  10   KSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGKAVVLMGKNTMMRKAIRGHL  69
            K+ YF K+  L + Y   F+V  DNV S+QM  +R  LRG A ++MGKNTM+R+A+RG +
Sbjct  8    KAQYFEKLRSLFEKYNSLFVVNIDNVSSQQMHTVRKQLRGTAELIMGKNTMIRRAMRGII  67

Query  70   ENNPALEKLLPHIRGNVGFVFTKEDLTEIRDMLLANKVPAAARAGAIAPCEVTVPAQNTG  129
             + P LE+LLP +RGNVGFVFT  DL E+R+ ++AN + A AR  AIAP +V VPA NTG
Sbjct  68   NDMPELERLLPVVRGNVGFVFTNADLKEVRETIIANVIAAPARPNAIAPLDVFVPAGNTG  127

Query  130  LGPEKTSFFQALGITTKISRGTIEILSDVQLIKTGDKVGASEATLLNMLNISPFSFGLVI  189
            + P KTSFFQALGI TKI+RGTIEI SDV L+    KVG SEATLLNMLNISPF++G+ +
Sbjct  128  MEPGKTSFFQALGIPTKITRGTIEITSDVHLVSKDAKVGPSEATLLNMLNISPFTYGMDV  187

Query  190  QQVFDNGSIYNPEVLDITEETLHSRFLEGVRNVASVCLQIGYPTVASVPHSIINGYKRVL  249
              ++D G++++PE+LD++EE L    L     + ++ L   YPT+ SV HS++N YK ++
Sbjct  188  LTIYDQGNVFSPEILDVSEEDLIGHLLSAASIITAISLGANYPTILSVMHSVVNAYKNLV  247

Query  250  ALSVETDYTFPLAEKVKAFLADPSAFVAAAPVAAATTAAPAAAAAAPAKVEAKEESEESDE  309
            A+S+ T+YTF   E+ KAFLADPSAFV A   AA   AA   A APA   A EE EESDE
Sbjct  248  AVSLATEYTFEGTEQTKAFLADPSAFVVA---AAPAAAAGGEAEAPAAEAAAEEEEESDE  304

Query  310  DMGFGLFD  317
            DMGFGLFD
Sbjct  305  DMGFGLFD  312
```



```
>sp|B6YSX9.1|RLA0_THEON RecName: Full=Acidic ribosomal protein P0 homolog; AltName: Full=L10E
Length=339

 GENE ID: 7017837 rplP0 | acidic ribosomal protein P0
[Thermococcus onnurineus NA1] (10 or fewer PubMed links)

 Score =  131 bits (330),  Expect = 4e-38, Method: Compositional matrix adjust.
 Identities = 77/257 (30%), Positives = 130/257 (51%), Gaps = 3/257 (1%)

Query  7    ATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGKAVVLMGKNTMMRKAIR  66
            A WK     ++  ++  YP   +V   NV +  + ++R  LRGKA++ + +NT++  AI+
Sbjct  5    AEWKKKEVEELTNIIKSYPVIALVDVANVPAYPLSKMREKLRGKALLRVSRNTLIELAIK  64

Query  67   GHLEN--NPALEKLLPHIRGNVGFVFTKEDLTEIRDMLLANKVPAAARAGAIAPCEVTVP  124
               +    P LEKL+ HI+G  G + T+ +  ++  +L  +K PA A+ G   P +V +P
Sbjct  65   RAAQELGKPELEKLIDHIQGGAGILATEMNPFKLYKLLEESKTPAPAKPGVPVPRDVVIP  124

Query  125  AQNTGLGPEK-TSFFQALGITTKISRGTIEILSDVQLIKTGDKVGASEATLLNMLNISPF  183
            A  T + P       QALGI  +I +G + I  D  ++K G+ +    A +LN L I P
Sbjct  125  AGPTSISPGPLVGEMQALGIPARIEKGKVSIQKDYTVLKAGEVITEQLARILNALGIEPL  184

Query  184  SFGLVIQQVFDNGSIYNPEVLDITEETLHSRFLEGVRNVASVCLQIGYPTVASVPHSIIN  243
              GL +   +++G +Y PEVL I EE   +   +   +  ++ +   YPT  ++   I
Sbjct  185  EVGLNLLAAYEDGIVYTPEVLAIDEEEYINLLQQAYMHAFNLSVNTAYPTSQTIEAIIQK  244

Query  244  GYKRVLALSVETDYTFP  260
             Y     ++VE  Y  P
Sbjct  245  AYLGAKNVAVEAGYITP  261
```

Pourquoi la comparaison de protéines est-elle plus sensible que la comparaison d'ADN?

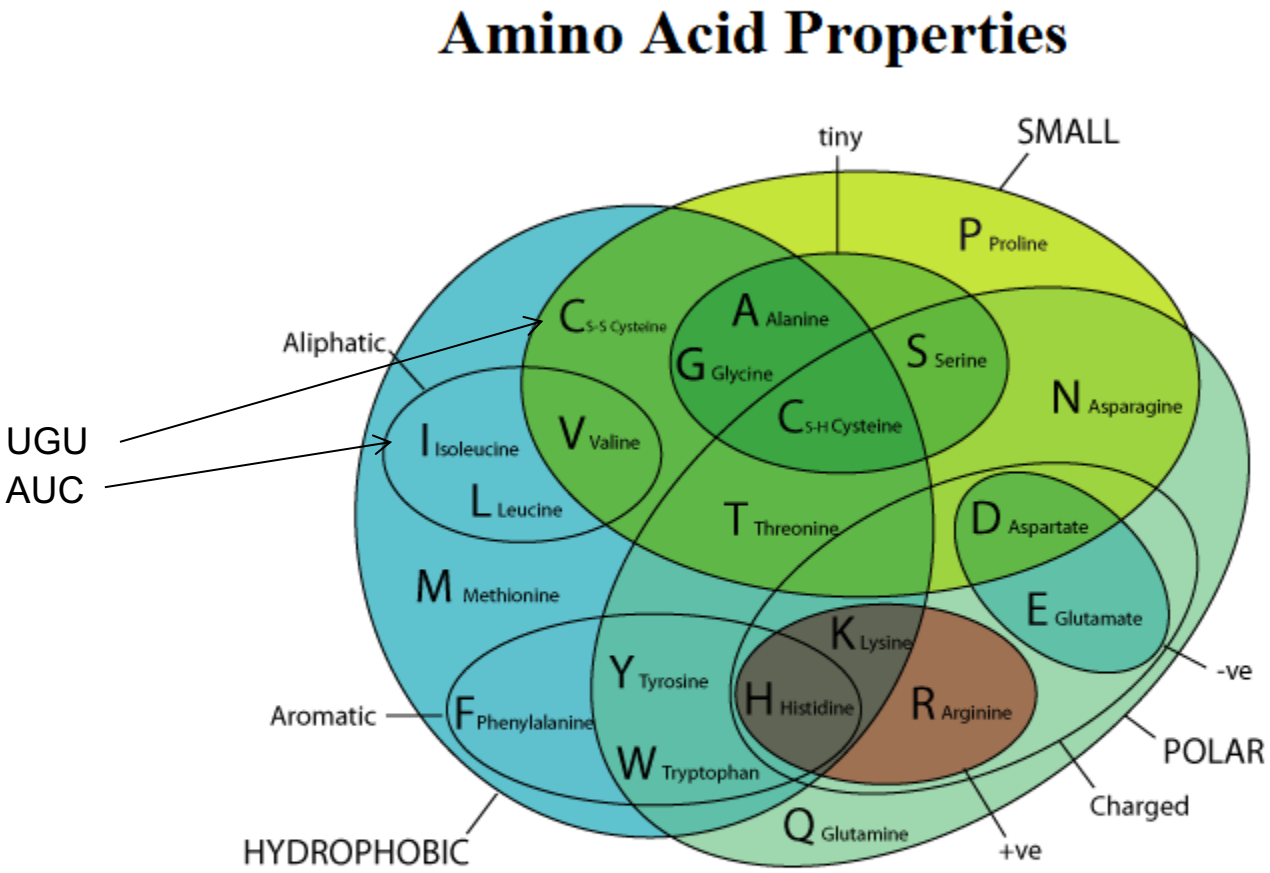


le code génétique

Première lettre (côté 5')	Deuxième lettre U		Deuxième lettre C		Deuxième lettre A		Deuxième lettre G		Troisième lettre (côté 3')
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys	C
	UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop	A
	UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC	Leu	CCC	Pro	CAC	His	CGC	Arg	C
	CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg	A
	CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg	G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser	C
	AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg	A
	AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg	G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly	C
	GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly	A
	GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly	G

codon d'initiation — codon de terminaison

Deuxième raison (plus importante):



From Livingstone, C. D. and Barton, G. J. (1993), "Protein Sequence Alignments: A Strategy for the Hierarchical Analysis of Residue Conservation", Comp. Appl. Bio. Sci., 9, 745-756.

-> nécessité de capturer ces propriétés dans un score

Bioinformatics Wisdom

Bio-informatique et bioinformatique

Biologiste + Informaticien =

Inf Biologiste =

Informaticien Biologiste =