

HMSN204

TD Banques de données biologiques

Formats de fichiers de séquences

1. Formats de fichiers de séquences : FASTA et Genbank

1.1 Exercice 1 : En-tête FASTA

Décrire de manière détaillée les informations qui sont retournées par les en-têtes listées ci-dessous (banque de données de provenance (aide en annexe), type de séquence, organisme, définition, ...). Vous trouverez en annexe des indications sur les banques de données de provenance dans les en-tête FASTA.

1. >NM_003140.1| Homo sapiens sex determining region Y (SRY), mRNA
2. >sp|Q05066|SRY_HUMAN Sex-determining region Y protein OS=Homo sapiens GN=SRY
3. >NP_035694.1| sex-determining region Y protein [Mus musculus]
4. >HM757931.1| Homo sapiens isolate HSA-34 SRY gene, partial sequence
5. >Q28783.1|SRY_PONPY RecName : Full=Sex-determining region Y protein
6. >tr|U3LIR9|U3LIR9_HIV1 Envelope glycoprotein gp160 OS=Human immunodeficiency virus 1 GN=env PE=4 SV=1

1.2 Exercice 2 : Conversion Genbank - FASTA

Une portion d'en-tête de fichier au format Genbank vous est fourni. Construire l'en-tête FASTA associée.

```
LOCUS      NM_001008988          910 bp    mRNA    linear    PRI 02-JUL-2011
DEFINITION Pan troglodytes sex determining region Y (SRY), mRNA.
ACCESSION  NM_001008988 XM_521375
VERSION    NM_001008988.1 GI:57113870
KEYWORDS   .
SOURCE     Pan troglodytes (chimpanzee)
```

2. Formats de fichiers de séquences : GFF et Genbank

2.1 Rappel

GFF est un format simplifié qui facilite la maintenance et la mise à jour d'annotations biologiques à partir d'un simple éditeur de texte. GFF simplifie également l'échange des annotations entre différentes applications.

GFF comprend 9 colonnes séparées par des tabulations (8 tabulations + retour chariot en fin de ligne) :

1. identifiant de séquence
2. source de données d'origine (ex. GenBank)
3. type de la sous-région, contraint à un terme du vocabulaire SOFA (ou identifiant SOFA)
4. start : position de début
5. end : position de fin

6. score : réel : valeur calculée à partir d'un éventuel outil (sinon .)
7. strand : brin positif +, ou négatif -, sinon . (pas de brin) ou ? pour incertain
8. phase : feature de type CDS (obligatoire) : ORF : 0, 1 et 2
9. attributes : couples propriété-valeur tag=valeur séparés par des ; domaine de valeurs des tags (ID, name, Alias, Parent (partof relationship), Target, Gap, Derives_from, Note, Dbxref, Ontology_term)

Des exemples (fantaisistes) de fichier au format GFF sont donnés en annexe.

2.2 Exercice : Conversion Genbank - GFF

A partir de la table des caractéristiques simplifiée présentée ci-dessous, construire un fichier tabulé au format GFF qui permettra de rendre compte des principaux éléments structuraux de la séquence étudiée.

FEATURES	Location/Qualifiers
gene	1..2117 /gene="ATF2" /gene_synonym="CRE-BP1; CREB2; HB16; TREB7" /note="activating transcription factor 2" /db_xref="GeneID:1386" /db_xref="HGNC:784"
exon	1..120 /gene="ATF2"
	/number=1
exon	121..219 /gene="ATF2"
	/number=2
exon	220..294 /gene="ATF2"
	/number=3
misc_feature	248..250 /note="upstream in-frame stop codon"
CDS	263..1780 /gene="ATF2" /gene_synonym="CRE-BP1; CREB2; HB16; TREB7" /note="cAMP responsive element binding protein 2" /codon_start=1 /product="cyclic AMP-dependent transcription factor ATF-2" /protein_id="NP_001871.2" /db_xref="GI:22538422" /db_xref="CCDS:CCDS2262.1" /db_xref="MIM:123811"
exon	295..364 /gene="ATF2"
	/number=4
exon	365..461 /gene="ATF2"
	/number=5
exon	462..580 /gene="ATF2"
	/number=6
exon	581..709 /gene="ATF2"
	/number=7
exon	710..888 /gene="ATF2"
	/number=8
exon	889..1003 /gene="ATF2"
	/number=9
exon	1004..1090 /gene="ATF2"
	/number=10
exon	1091..1240 /gene="ATF2"
	/number=11
STS	1232..1361 /gene="ATF2" /db_xref="UniSTS:4379"
exon	1241..1447 /gene="ATF2"
	/number=12
exon	1448..1553 /gene="ATF2"

```

        /number=13
exon      1554..2111
        /gene="ATF2"
        /number=14
polyA_signal 2090..2095
        /gene="ATF2"
polyA_site  2111
        /gene="ATF2"
        /experiment="experimental evidence"

```

3. Annexes

3.1 Indication banque de données de provenance pour les en-têtes FASTA

Quelques champs sur lesquels la consultation est permise pour les rubriques Sequence (Nucleotide et Protein) sont listés dans le tableau ci-dessous.

GenBank (comme GenPept)	gi gi-number gb accession locus
EMBL Data Library	gi gi-number emb accession locus
DDBJ, DNA Database of Japan	gi gi-number dbj accession locus
NBRF PIR	pir entry
Protein Research Foundation	prf name
SWISS-PROT	sp accession name
Brookhaven Protein Data Bank (1)	pdb entry chain
Brookhaven Protein Data Bank (2)	entry:chain PDBID CHAIN SEQUENCE
Patents	pat country number
GenInfo Backbone Id	bbs number
General database identifier	gnl database identifier
NCBI Reference Sequence	ref accession locus
Local Sequence identifier	lcl identifier

3.2 Exemple 1 de fichier au format GFF et son rendu visuel

```

##gff-version 3
ctg123 GenBank exon 1 1500 . + . ID=exon1;note=chromobox homolog 8
ctg123 GenBank gene 1 4500 . + . ID=gene;Dbxref=GeneID:779897
ctg123 GenBank exon 3000 4500 . + . ID=exon2
ctg123 GenBank intron 1501 2999 . + . ID=intron1
ctg123 GenBank snp 3080 3080 . + . ID=snp1

```

Visualisation du fichier GFF ci-dessus :

3.3 Exemple 2 de fichier au format GFF et son rendu visuel

```

##gff-version 3
NM_00314e GenBank gene 1 4500 . + . ID=gene;name=AZQ;Dbxref=GeneID:779897,HGNC:11311;gene_synonym=SRX
NM_00314e GenBank exon 1 1500 . + . ID=exon1;Name=AZQ.1;Parent=gene;note=chromobox homolog 8
NM_00314e GenBank exon 3000 4500 . + . ID=exon2;Parent=gene;Name=AZQ.2
NM_00314e GenBank intron 1501 2999 . + . ID=intron1;Parent=gene
NM_00314e GenBank snp 3080 3080 . + . ID=snp1;Parent=gene

```

Visualisation du fichier GFF ci-dessus :