Date le : 21 février 03

Démonstration numéro 6 :

Les Bases de Données

Un bon nombre de bases de données biologiques sont accessibles à travers le web. Leur utilisation dépendra de vos besoins.
La centralisation des ressources ainsi que leur diffusion a été fortement encouragée par l'évolution de l'internet et la forte demande des organisations mondiales.
En 1988, un réseau a été mis en pied pour relier les laboratoires européens œuvrant dans la recherche de la biologie moléculaire en se basant sur la bioinformatique.

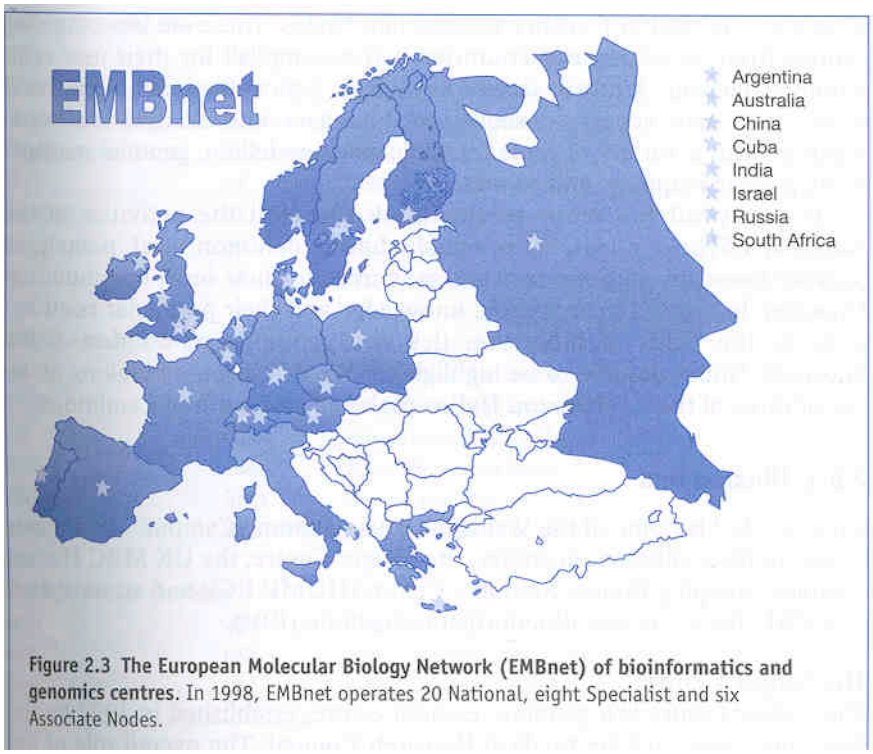


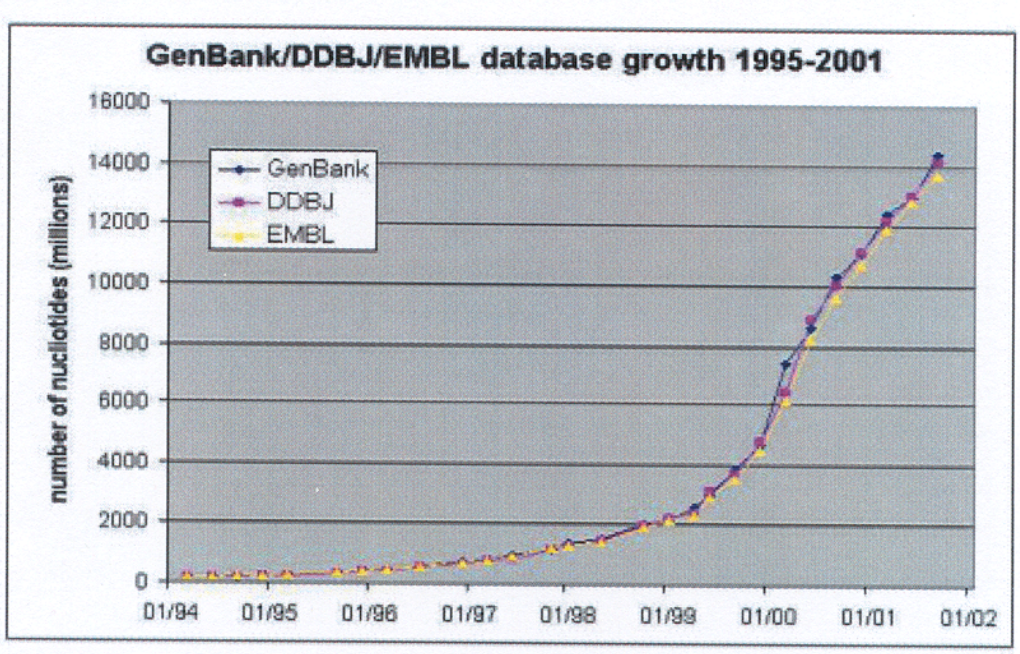
Figure 2.3 The European Molecular Biology Network (EMBnet) of bioinformatics and genomics centres. In 1998, EMBnet operates 20 National, eight Specialist and six Associate Nodes.

Les bases de données sont un moyen très efficace de stocker un grand volume d'informations. La différence entre elles va dépendre des informations à extraire et manipuler (séquences primaires ou structures 2D ou 3D).

Voici un résumé des bases de données les plus connues des séquences primaires que ça soit pour les acides nucléiques ou les protéines.

DNA (nucleotide)		Protein	
EMBL	UK	PIR	US
GenBank	US	MIPS	Germany
DDBJ	Japan	Swiss-Prot	Swiss
Celera	Celera	TrEMBL	Swiss
		NRL_3D	US
		GenPept	US

Les Bases de Données d'ADN



EMBL : **(European Molecular Biology Laboratory)**

C'est une base de donnée de l'institut européen de bioinformatique (EBI) en Angleterre.
Elle contient les séquences soumises directement par les auteurs, et les groupes de séquençage du génome, et de la littérature scientifique. La base de donnée est produite en collaboration avec celle du Japon (DDBJ) et celle des Etats-Unis (genbank), la mise à jour des données se fait de façon régulière.
En 1998, EMBL contenait plus de un million de données de plus de 15 500 espèces.

DDBJ : (DNA Data Bank of Japan)

Cette base de donnée d'ADN du Japon, débuta en 1986 en une collaboration entre EMBL et GenBank. Sa maintenance et sa distribution sont faite par l'institut national génétique.

GenBank :

La base de donnée de séquence d'ADN du centre national d'information en biotechnologie. Cette base de donnée est divisée en 17 divisions jusqu`à date :

En voici le tableau :

Division Code	Description
PRI	primate sequences
ROD	rodent sequences
MAM	other mammalian sequences
VRT	other vertebrate sequences
INV	invertebrate sequences
PLN	plant, fungal, and algal sequences
BCT	bacterial sequences
RNA	structural RNA sequences
VRL	viral sequences
PHG	bacteriophage sequences
SYN	synthetic sequences
UNA	unannotated sequences
EST	EST sequences (expressed sequence tags)
PAT	patent sequences
STS	STS sequences (sequence tagged sites)
GSS	GSS sequences (genome survey sequences)
HTG	HTGS sequences (high throughput genomic sequences)

ESTs (Expressed Sequence Tags) : de courts fragments d'ARN messager sont pris de plusieurs tissus et organismes. Ces échantillons sont amplifiés et séquencés. Le séquençage se fait en une seule passe de lecture, ce qui fait que les données ne sont pas très précises. Il y environ 6 million de séquences ESTS (plus du tiers sont clonées).

STSs (Sequence-Tagged Sites) : Echantillons courts du génome qui sert comme des marqueurs génomiques

HTGs : (High Throughput Genomic Sequences) : Séquences obtenues dans le cadre du séquencage humain complet. Les enregistrements de cette base de données sont classés en fonction du niveau d'avancement à travers l'achèvement complet de la séquence.

.

"Definition of a contig

In order to make it easier to talk about our data gained by the shotgun method of sequencing we have invented the word "contig". A contig is a set of gel readings that are related to one another by overlap of their sequences. All gel readings belong to one and only one contig, and each contig contains at least one gel reading. The gel readings in a contig can be summed to form a contiguous consensus sequence and the length of this sequence is the length of the contig."

Phase 0 : une ou quelque passes de lectures d'un seul clone (pas un contig).
Phase 1 : non complet, peut être non ordonnés, non orientés contigs avec des trous.
Phase 2 : des contigs non terminés, ordonnés et orientés avec ou sans trous.
Phase 3 : Finis, sans trous (avec ou sans annotation).

Les données dans GenBank, respectent un certain format reconnu par les logiciels d'extraction. Les formats les plus utilisés sont Genbank et Fasta.

Le format GenBank contient des mots clés et des sous clés et une table de caractéristique optionnelle, la fiche d'information se termine toujours par //.

A titre d'exemple le mot clé LOCUS introduit une étiquette, ainsi que des informations concernant la longueur de la séquence donnée en bp (base pair), le type de la séquence, la division, sa date d`émission.

Le format Fasta, ne contient qu'une brève description à la première ligne qui doit impérativement débuter avec le signe « > », et la ligne suivante contient la séquence proprement dite.

En voici un exemple du format Genbank:

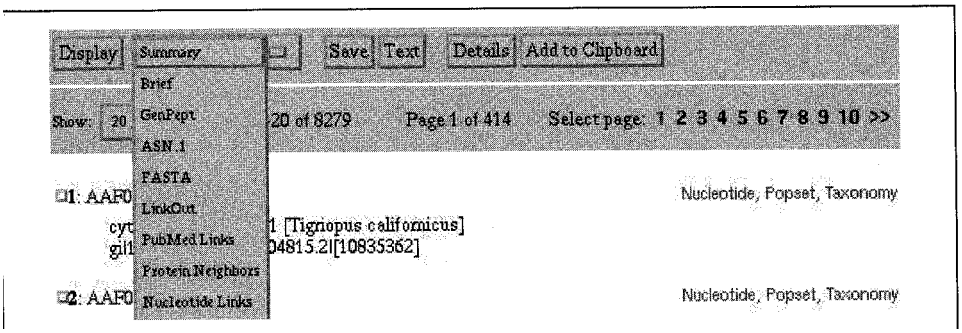
```
LOCUS       HSIGVH221               682 bp    DNA     linear   PRI 30-OCT-1995
DEFINITION  H.sapiens germline immunoglobulin heavy chain, variable region,
            (22-1).
ACCESSION   X92210
VERSION     X92210.1  GI:1045093
KEYWORDS    germ line; immunoglobulin.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 682)
  AUTHORS   Berman,J.E., Mellis,S.J., Pollock,R., Smith,C.L., Suh,H.,
            Heinke,B., Kowal,C., Surti,U., Chess,L., Cantor,C.R. and Alt,F.W.
  TITLE     Content and organization of the human Ig VH locus: definition of
            three new VH families and linkage to the Ig CH locus
  JOURNAL   EMBO J. 7 (3), 727-738 (1988)
  MEDLINE   88283641
   PUBMED   3396540
FEATURES             Location/Qualifiers
     source          1..682
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
BASE COUNT      178 a    156 c    147 g    198 t      3 others
ORIGIN
        1 cgaccgtctg catctcactc ttgttaggct gatgtgtcat ttatcttccc tttcttatca
       61 tggattgggc tttgagctaa gaaaggcttt gtctctatga atatgcaaat atactgatat
      121 ccactgaggt aaatatgttc tgtgccctga gagaatcacc tgagagaatc ccctgagagc
      181 acatctcctc atgggctgga cctgcaagnt cctcttcttg gtggcagcag ccacaggtaa
      241 gcagttccca ggtccaagta atgaggaggg gattgagtcc agtcaagggg gctttcatcc
      301 actcctgtgt cctccccaca ggtgcccact cccaggtgca gctggtgcaa tctggggctg
      361 aggtgaagaa gcctggggcc tcagtgaagg tctcctgcaa ggcttctgga tacaccttca
      421 cctactgcta cttgcactgg gtacgacagg cccttggana agggcttgaa tggacaggan
      481 tttagttatt tgagagattt ttcatacaac atttattctg taagcaaatt tcagggattg
      541 tagaatgaat cacattaaca aatctgacac agaacttcct ctgaatcaat ctttgtaaac
      601 atcaatttcc gaatcaatgt tgtaaatatt tcagaacaca agcacaaatt cacattttaa
      661 ctctactttt atctctattt aa
//
```

Exemple avec le format Fasta (plus abrégé, avec moins de détails) .

```
>gi|1045093|emb|X92210.1|HSIGVH221 H.sapiens germline immunoglobulin heavy
chain, variable region,  (22-1)
CGACCGTCTGCATCTCACTCTTGTTAGGCTGATGTGTCATTTATCTTCCCTTTCTTATCATGGATTGGGC
TTTGAGCTAAGAAAGGCTTTGTCTCTATGAATATGCAAATATACTGATATCCACTGAGGTAAATATGTTC
TGTGCCCTGAGAGAATCACCTGAGAGAATCCCCTGAGAGCACATCTCCTCATGGGCTGGACCTGCAAGNT
CCTCTTCTTGGTGGCAGCAGCCACAGGTAAGCAGTTCCCAGGTCCAAGTAATGAGGAGGGGATTGAGTCC
```

```
AGTCAAGGGGGCTTTCATCCACTCCTGTGTCCTCCCCACAGGTGCCCACTCCCAGGTGCAGCTGGTGCAA
TCTGGGGCTGAGGTGAAGAAGCCTGGGGCCTCAGTGAAGGTCTCCTGCAAGGCTTCTGGATACACCTTCA
CCTACTGCTACTTGCACTGGGTACGACAGGCCCTTGGANAAGGGCTTGAATGGACAGGANTTTAGTTATT
TGAGAGATTTTTCATACAACATTTATTCTGTAAGCAAATTTCAGGGATTGTAGAATGAATCACATTAACA
AATCTGACACAGAACTTCCTCTGAATCAATCTTTGTAAACATCAATTTCCGAATCAATGTTGTAAATATT
TCAGAACACAAGCACAAATTCACATTTTAACTCTACTTTTATCTCTATTTAA
```

Le passage d'un format à un autre se fait à travers l'interface, en le sélectionnant dans un menu.



Tester Genbank avec NM_022703 et XM_128610

A travers les sites hébergeant les bases de données permettent d'utiliser BLAST (Basic Local Alignment Search Tool), un outil de recherche standard.

Exemple avec EMBL

```
ID   HSIGVH221  standard; DNA; HUM; 682 BP.
XX
AC   X92210;
XX
SV   X92210.1
XX
DT   30-OCT-1995 (Rel. 45, Created)
DT   30-OCT-1995 (Rel. 45, Last updated, Version 2)
XX
DE   H.sapiens germline immunoglobulin heavy chain, variable region, (22-1)
XX
KW   germ line; immunoglobulin.
XX
OS   Homo sapiens (human)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia;
OC   Eutheria; Primates; Catarrhini; Hominidae; Homo.
XX
RN   [1]
RP   1-682
RX   MEDLINE; 88283641.
RA   Berman J.E., Mellis S.J., Pollock R., Smith C.L., Suh H., Heinke B.,
RA   Kowal C., Surti U., Chess L., Cantor C.R., Alt F.W.;
RT   "Content and organisation of the human Ig Vh locus: definition of
three new
RT   Vh families and linkage to the Ig Ch locus";
```

```
RL   EMBO J. 7:727-738(1988).
XX
DR   IMGT/LIGM; X92210; X92210.
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..682
FT                   /db_xref="taxon:9606"
FT                   /organism="Homo sapiens"
XX
SQ   Sequence 682 BP; 178 A; 156 C; 147 G; 198 T; 3 other;
    cgaccgtctg catctcactc ttgttaggct gatgtgtcat ttatcttccc tttcttatca     60
    tggattgggc tttgagctaa gaaaggcttt gtctctatga atatgcaaat atactgatat    120
    ccactgaggt aaatatgttc tgtgccctga gagaatcacc tgagagaatc ccctgagagc    180
    acatctcctc atgggctgga cctgcaagnt cctcttcttg gtggcagcag ccacaggtaa    240
    gcagttccca ggtccaagta atgaggaggg gattgagtcc agtcaagggg gctttcatcc    300
    actcctgtgt cctccccaca ggtgcccact cccaggtgca gctggtgcaa tctggggctg    360
    aggtgaagaa gcctggggcc tcagtgaagg tctcctgcaa ggcttctgga tacaccttca    420
    cctactgcta cttgcactgg gtacgacagg cccttggana agggcttgaa tggacaggan    480
    tttagttatt tgagagattt ttcatacaac atttattctg taagcaaatt tcagggattg    540
    tagaatgaat cacattaaca aatctgacac agaacttcct ctgaatcaat ctttgtaaac    600
    atcaatttcc gaatcaatgt tgtaaatatt tcagaacaca agcacaaatt cacattttaa    660
   ctctactttt atctctattt aa                                           682
//
```

Les adresses WEB

EMBL : http://www.ebi.ac.uk/ebi_docs/embl_db/ebi/topembl.html
DDBJ : http://www.ddbj.nig.ac.jp
GenBank: http://www.ncbi.nlm.nih.gov/Web/GenBank

D'autres bases de données:

dbEST: http://www.ncbi.nlm.nih.gov/dbEST
GSDB: http://genome-www.stanford.edu/Saccharomycesé/
UniGene: http://www.ncb.nlm.nih.gov/UniGene/
TDB: http://www.tigr.org/tdb/tdb.html
AceDB: http://www.sanger.ac.uk/Software/Acedb
Webace: http://webace.sanger.ac.uk

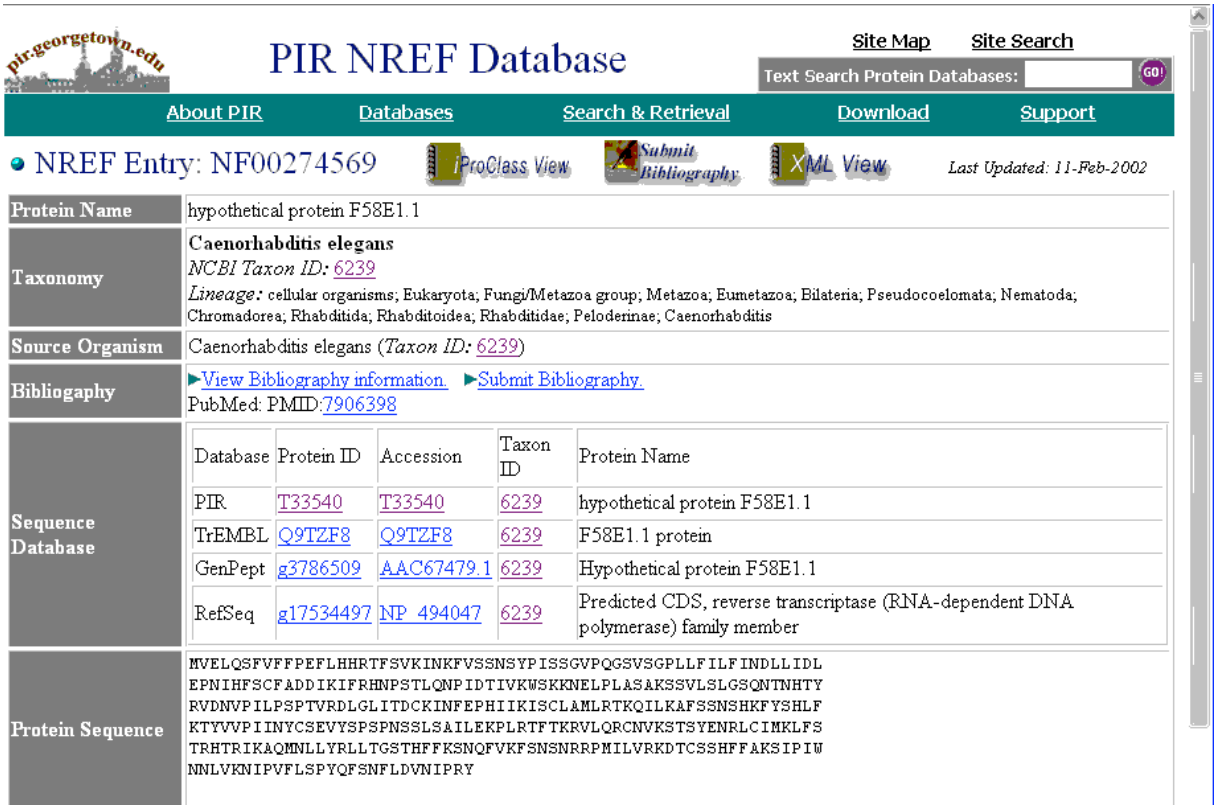
Les bases de données des séquences de protéines

PIR : International Protein Sequence Database (NBRF)

Développée au début des années 60 par Margaret DayHoff à la Fondation Nationale de Recherche Biomédicale.
Elle est divisée en différentes sections
PIR1 : Données complétement cklassées et annotées.
PIR2 : données préliminaires non révisées

PIR3 : données non vérifiées et non révisées
PIR4 : translation conceptuelles.

Adresse web de PIR : http://pir.georgetown.edu/



pir.georgetown.edu

PIR NREF Database

Site Map Site Search

Text Search Protein Databases: GO!

About PIR | Databases | Search & Retrieval | Download | Support

NREF Entry: NF00274569 iProClass View Submit Bibliography XML View *Last Updated: 11-Feb-2002*

Protein Name	hypothetical protein F58E1.1
Taxonomy	**Caenorhabditis elegans** *NCBI Taxon ID:* 6239 *Lineage:* cellular organisms; Eukaryota; Fungi/Metazoa group; Metazoa; Eumetazoa; Bilateria; Pseudocoelomata; Nematoda; Chromadorea; Rhabditida; Rhabditoidea; Rhabditidae; Peloderinae; Caenorhabditis
Source Organism	Caenorhabditis elegans (*Taxon ID:* 6239)
Bibliogaphy	▶View Bibliography information. ▶Submit Bibliography. PubMed: PMID:7906398

Sequence Database

Database	Protein ID	Accession	Taxon ID	Protein Name
PIR	T33540	T33540	6239	hypothetical protein F58E1.1
TrEMBL	Q9TZF8	Q9TZF8	6239	F58E1.1 protein
GenPept	g3786509	AAC67479.1	6239	Hypothetical protein F58E1.1
RefSeq	g17534497	NP_494047	6239	Predicted CDS, reverse transcriptase (RNA-dependent DNA polymerase) family member

Protein Sequence

```
MVELQSFVFFPEFLHHRTFSVKINKFVSSNSYPISSGVPQGSVSGPLLFILFINDLLIDL
EPNIHFSCFADDIKIFRHNPSTLQNPIDTIVKWSKKNELPLASAKSSVLSLGSQNTNHTY
RVDNVPILPSPTVRDLGLITDCKINFEPHIIKISCLAMLRTKQILKAFSSNSHKFYSHLF
KTYVVPIINYCSEVYSPSPNSSLSAILEKPLRTFTKRVLQRCNVKSTSYENRLCIMKLFS
TRHTRIKAQMNLLYRLLTGSTHFFKSNQFVKFSNSNRRPMILVRKDTCSSHFFAKSIPIW
NNLVKNIPVFLSPYQFSNFLDVNIPRY
```

Swiss-Prot :

Est maintenue en collaboration entre SIB et EBI/EMBL. Donnes de bonnes annotations en incluant une description de la fonction, la structure et les domaines des protéines.

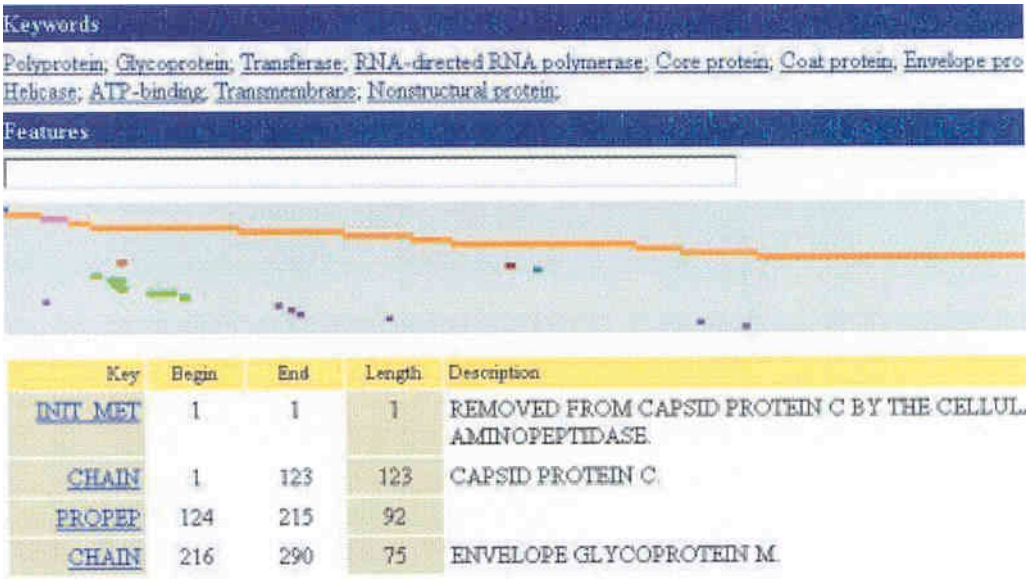
General information	
Entry name	POLG_WNV
Accession number	P06935
Created	Rel. 06, 1-JAN-1988
Sequence update	Rel. 06, 1-JAN-1988
Annotation update	Rel. 40, 16-OCT-2001

Description and origin of the Protein	
Description	GENOME POLYPROTEIN [CONTAINS: CAPSID PROTEIN C (CORE PROTEIN); MA (ENVELOPE PROTEIN M); MAJOR ENVELOPE PROTEIN E; NONSTRUCTURAL PR NS2B; NS4A AND NS4B; PROTEASE/HELICASE (EC 3.4.21.98) (NS3); RNA-DIRECTI POLYMERASE (EC 2.7.7.48) (NS5)]
Organism source	West Nile virus (WN).
Taxonomy	Viruses; ssRNA positive-strand viruses, no DNA stage; Flaviviridae; Flavivirus.
NCBI TaxID	11082

References		
[1]	Castle,E., Leidner,U., Nowak,T., Wengler,G.,	
	Primary structure of the West Nile flavivirus genome region coding for all nonstructural	
	(1986) *Virology* **149**:10	
	Position	SEQUENCE FROM N.A.
	Medline	86124703
	PubMed	3753811
[2]	Castle,E., Nowak,T., Leidner,U., Wengler,G.,	
	Sequence analysis of the viral core protein and the membrane- associated proteins V1 a	
	flavivirus West Nile virus and of the genome sequence for these proteins.	

Comments	
FUNCTION	THE SMALL PROTEINS NS2A, NS2B, NS4A AND NS4B ARE HYDROPHOB POSSIBLE MEMBRANE-RELATED FUNCTION. NS3 AND NS5 MAY PLAY VIRAL RNA REPLICATION.
CATALYTIC ACTIVITY	HYDROLYSIS OF FOUR PEPTIDE BONDS IN THE VIRAL PRECURSOR POI COMMONLY WITH ASP OR GLU IN THE P6 POSITION, CYS OR THR IN P ALA IN P1'.

Database cross-references	
EMBL	M12294;AAA48498.1;-.
PIR	A25256;GNWVWV.
HSSP	P14336;1SVB.
MEROPS	S07.001;-.
	IPR001410;DEAD.
	IPR001122;Flavi_capsid.
	IPR000336;Flavi_glycoprotE.
	IPR001069;T...



Keywords

Polyprotein; Glycoprotein; Transferase; RNA-directed RNA polymerase; Core protein; Coat protein; Envelope pro
Helicase; ATP-binding; Transmembrane; Nonstructural protein;

Features

Key	Begin	End	Length	Description
INIT_MET	1	1	1	REMOVED FROM CAPSID PROTEIN C BY THE CELLUL AMINOPEPTIDASE.
CHAIN	1	123	123	CAPSID PROTEIN C.
PROPEP	124	215	92	
CHAIN	216	290	75	ENVELOPE GLYCOPROTEIN M.

D'autres bases de données de proteines :

TrEMBL : Translated EMBL
GenPept
NRL_3D (séquences extraites de la base de donnée PDB)