Bioinformatics Data Formats

TIGR Plant Genome Annotation Workshop
May 2007

Biological Data and Bioinformatics

- The amount of biological data being generated and stored continues to increase.
- The data is composed of many different types: sequence (genome, ESTs), annotation of features, protein structural information, gene expression data, and alignment data.
- Another valuable resource for bioinformatics is web-based computational tools.

Popular Bioinformatics Resources

- NCBI- http://www.ncbi.nlm.nih.gov/
- EMBL- http://www.ebi.ac.uk/embl/
- PIR- http://pir.georgetown.edu/
- PDB- http://www.pdb.org/
- Google- http://www.google.com/

Early Data Formats

- These early databases stored sequence data in a file. The file held the sequence in ASCII (plain) text and had a descriptive filename.

- This method became limiting when researchers wanted to include annotations and information about the source of the sequence.

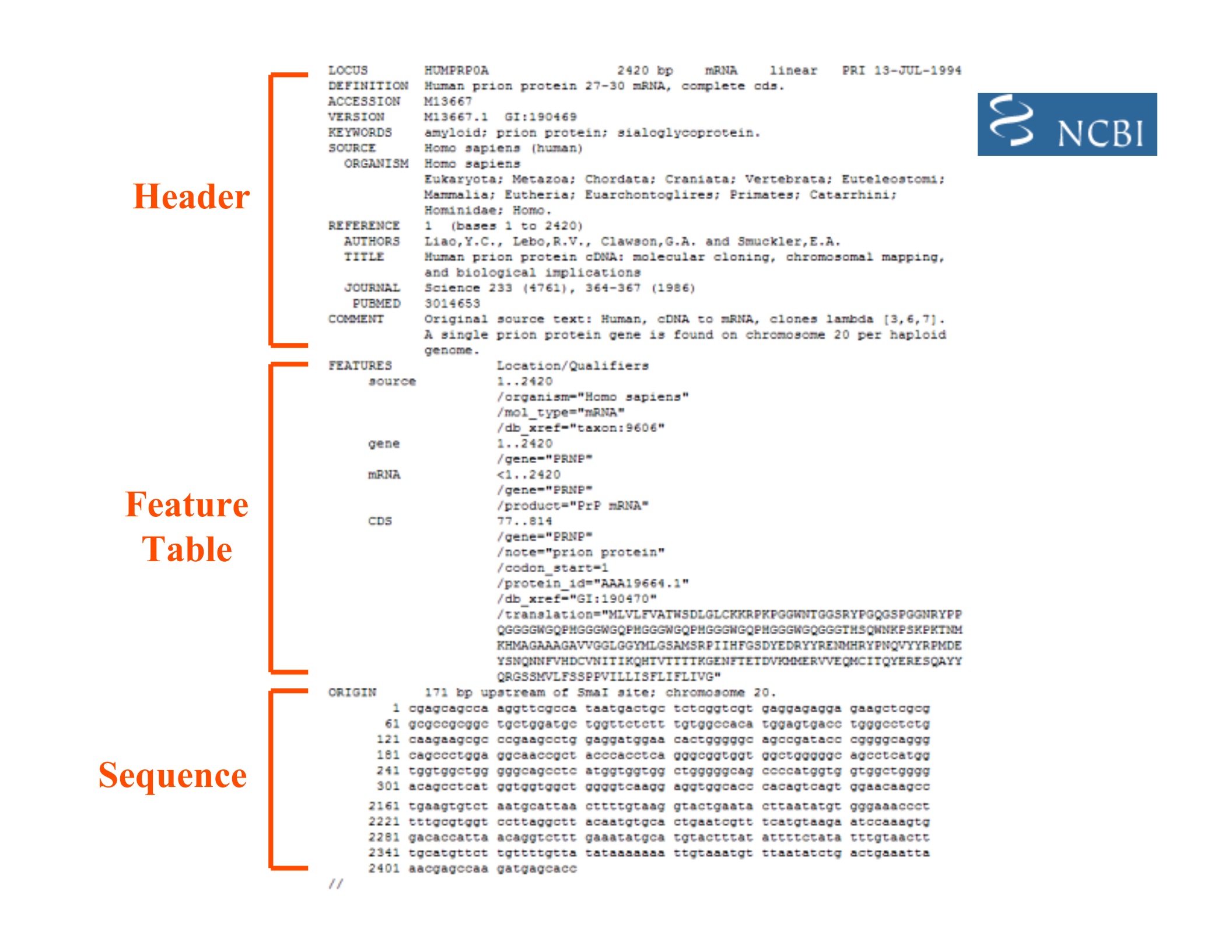
- Difficulty in searching for sequences was also an issue.

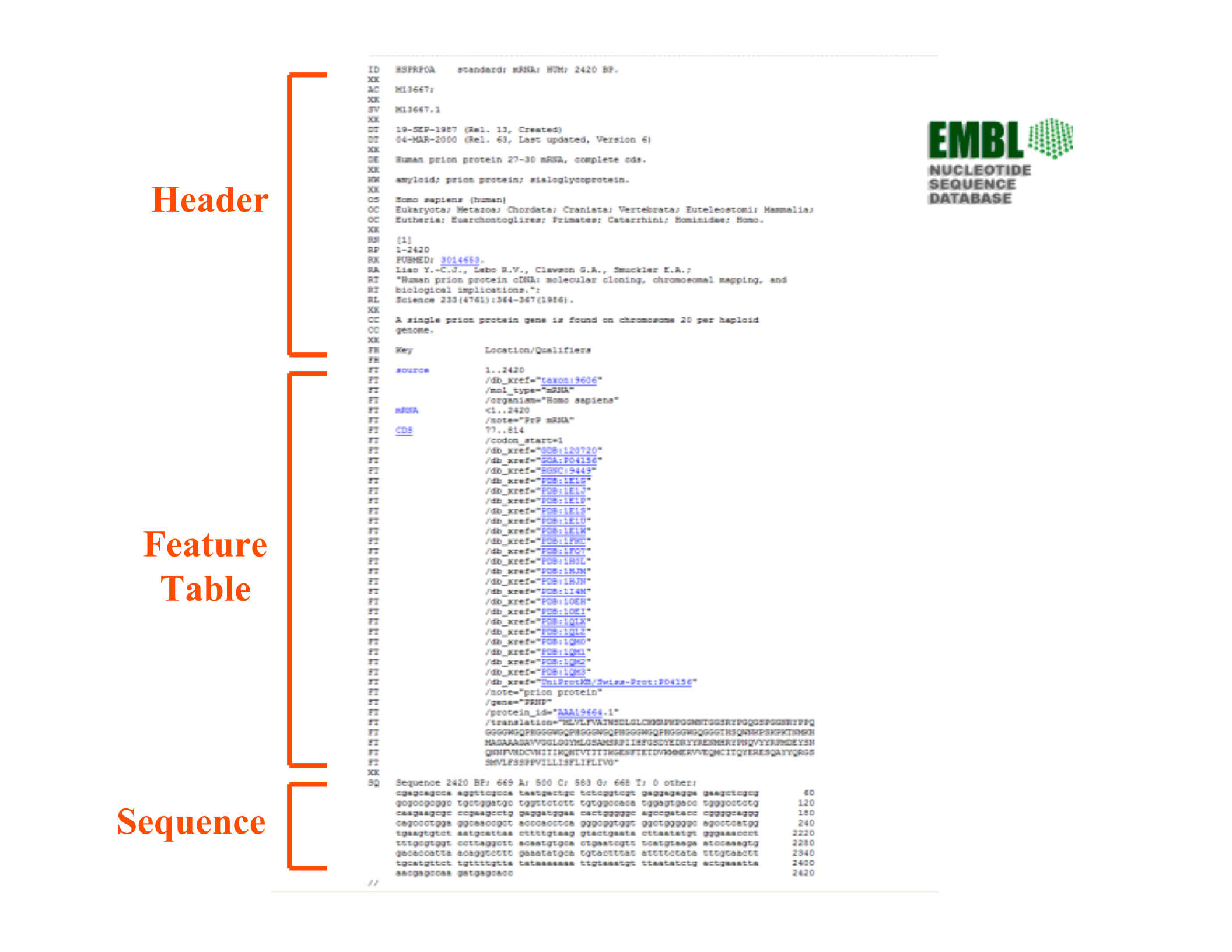
Flat File Storage Data Formats

- When GenBank, EMBL and DDBJ formed a collaboration (1986), sequence databases had moved to a defined flat file format with a shared feature table format and annotation standards.

- The PIR also adopted a similar format for protein sequences (http://www.molecularevolution.org/resources/fileformats/)

- The flat file formats from the sequence databases are still used to access and display sequence and annotation. They are also convenient for storage of local copies.

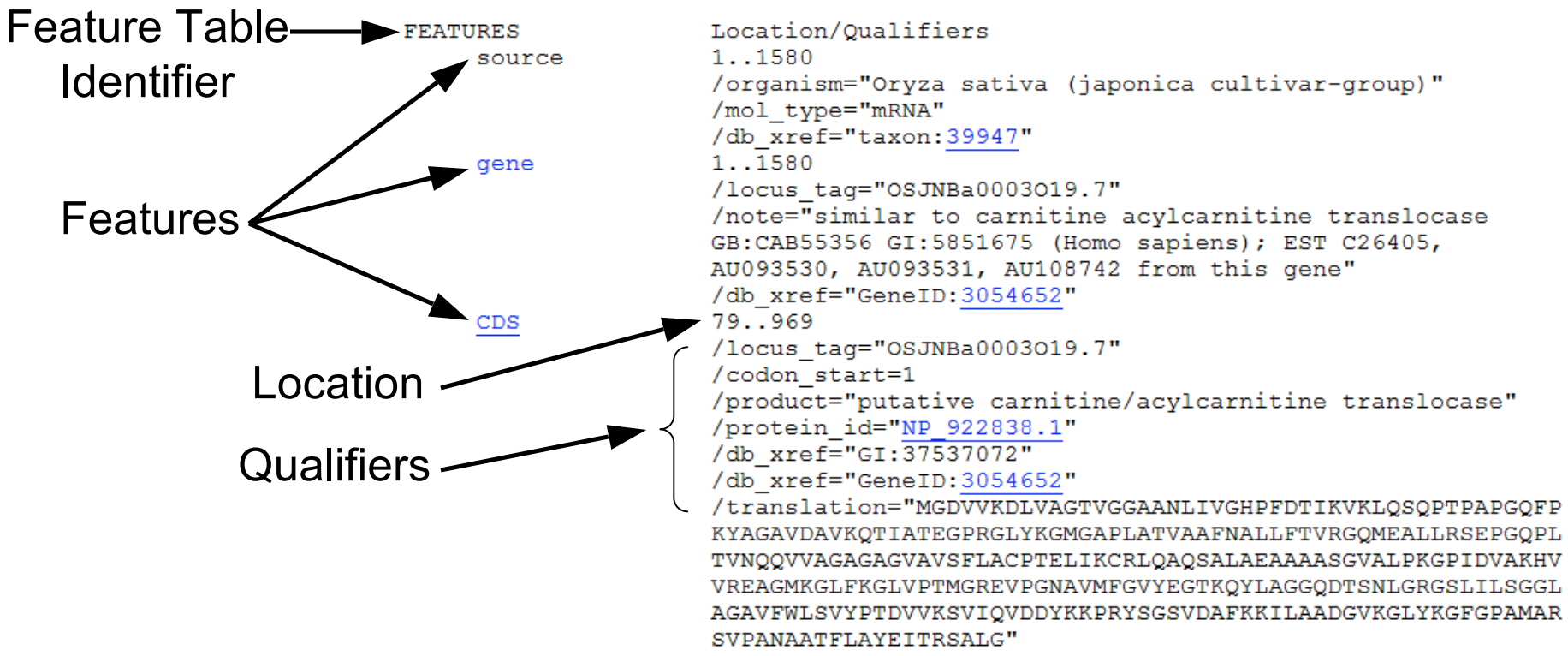

Header
Feature Table
Sequence
LOCUS HUMPRP0A 2420 bp mRNA linear PRI 13-JUL-1994
DEFINITION Human prion protein 27-30 mRNA, complete cds.
ACCESSION M13667
VERSION M13667.1 GI:190469
KEYWORDS amyloid; prion protein; sialoglycoprotein.
SOURCE Homo sapiens (human)
ORGANISM Homo sapiens
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Euarchontoglires; Primates; Catarrhini;
Hominidae; Homo.
REFERENCE 1 (bases 1 to 2420)
AUTHORS Liao,Y.C., Lebo,R.V., Clawson,G.A. and Smuckler,E.A.
TITLE Human prion protein cDNA: molecular cloning, chromosomal mapping,
and biological implications
JOURNAL Science 233 (4761), 364-367 (1986)
PUBMED 3014653
COMMENT Original source text: Human, cDNA to mRNA, clones lambda [3,6,7].
A single prion protein gene is found on chromosome 20 per haploid
genome.
FEATURES Location/Qualifiers
source 1..2420
/organism="Homo sapiens"
/mol_type="mRNA"
/db_xref="taxon:9606"
gene 1..2420
/gene="PRNP"
mRNA <1..2420
/gene="PRNP"
/product="PrP mRNA"
CDS 77..814
/gene="PRNP"
/note="prion protein"
/codon_start=1
/protein_id="AAA19664.1"
/db_xref="GI:190470"
/translation="MLVLFVATWSDLGLCKKRPKPGGWNTGGSRYPGQGSPGGNRYPP
QGGGGWGQPHGGGWGQPHGGGWGQPHGGGWGQPHGGGWGQGGGTHSQWNKPSKPKTNM
KHMAGAAAGAVVGGLGGYMLGSAMSRPIIHFGSDYEDRYYRENMHRYPNQVYYRPMDE
YSNQNNFVHDCVNITIKQHTVTTTTKGENFTETDVKMMERVVEQMCITQYERESQAYY
QRGSSMVLFSSPPVILLISFLIFLIVG"
ORIGIN 171 bp upstream of SmaI site; chromosome 20.
1 cgagcagcca aggttcgcca taatgactgc tctcggtcgt gaggagagga gaagctcgcg
61 gcgccgcggc tgctggatgc tggttctctt tgtggccaca tggagtgacc tgggcctctg
121 caagaagcgc ccgaagcctg gaggatggaa cactgggggc agccgatacc cggggcaggg
181 cagccctgga ggcaaccgct acccacctca gggcggtggt ggctgggggc agcctcatgg
241 tggtggctgg gggcagcctc atggtggtgg ctgggggcag ccccatggtg gtggctgggg
301 acagcctcat ggtggtggct ggggtcaagg aggtggcacc cacagtcagt ggaacaagcc
2161 tgaagtgtct aatgcattaa cttttgtaag gtactgaata cttaatatgt gggaaaccct
2221 tttgcgtggt ccttaggctt acaatgtgca ctgaatcgtt tcatgtaaga atccaaagtg
2281 gacaccatta acaggtcttt gaaatatgca tgtactttat attttctata tttgtaactt
2341 tgcatgttct tgttttgtta tataaaaaaa ttgtaaatgt ttaatatctg actgaaatta
2401 aacgagccaa gatgagcacc
//


Header
Feature Table
Sequence
EMBL
NUCLEOTIDE SEQUENCE DATABASE
ID HSPRPOA standard; mRNA; HUM; 2420 BP.
XX
AC M13667;
XX
SV M13667.1
XX
DT 19-SEP-1987 (Rel. 13, Created)
DT 04-MAR-2000 (Rel. 63, Last updated, Version 6)
XX
DE Human prion protein 27-30 mRNA, complete cds.
XX
KW amyloid; prion protein; sialoglycoprotein.
XX
OS Homo sapiens (human)
OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC Eutheria; Euarchontoglires; Primates; Catarrhini; Hominidae; Homo.
XX
RN [1]
RP 1-2420
RX PUBMED; 3014653.
RA Liao Y.-C.J., Lebo R.V., Clawson G.A., Smuckler E.A.;
RT "Human prion protein cDNA: molecular cloning, chromosomal mapping, and
RT biological implications.";
RL Science 233(4761):364-367(1986).
XX
CC A single prion protein gene is found on chromosome 20 per haploid
CC genome.
XX
FH Key Location/Qualifiers
FH
FT source 1..2420
FT /db_xref="taxon:9606"
FT /mol_type="mRNA"
FT /organism="Homo sapiens"
FT mRNA <1..2420
FT /note="PrP mRNA"
FT CDS 77..814
FT /codon_start=1
FT /db_xref="PDB:1Z0720"
FT /db_xref="GOA:P04156"
FT /db_xref="HGNC:9449"
FT /db_xref="PDB:1E1G"
FT /db_xref="PDB:1E1J"
FT /db_xref="PDB:1E1P"
FT /db_xref="PDB:1E1S"
FT /db_xref="PDB:1E1U"
FT /db_xref="PDB:1E1W"
FT /db_xref="PDB:1FMC"
FT /db_xref="PDB:1FO7"
FT /db_xref="PDB:1H0L"
FT /db_xref="PDB:1HJM"
FT /db_xref="PDB:1HJN"
FT /db_xref="PDB:1I4M"
FT /db_xref="PDB:1OEB"
FT /db_xref="PDB:1OE1"
FT /db_xref="PDB:1QLX"
FT /db_xref="PDB:1QLZ"
FT /db_xref="PDB:1QM0"
FT /db_xref="PDB:1QM1"
FT /db_xref="PDB:1QM2"
FT /db_xref="PDB:1QM3"
FT /db_xref="UniProtKB/Swiss-Prot:P04156"
FT /note="prion protein"
FT /gene="PRNP"
FT /protein_id="AAA19664.1"
FT /translation="MEVLFVATWSDLGLCKKRPKPGGWNTGGSRYPGQGSPGGSRYPPQ
FT GGGGWGQPHGGGWGQPHGGGWGQPHGGGWGQPHGGGWGQGGGTHSQWNKPSKPKTNMKH
FT MAGAAAAGAVVGGLGGYMLGSAMSRPIIHFGSDYEDRYYRENMHRYPNQVYYRPMDEYSN
FT QNNFVHDCVNITIKQHTVTTTTKGENFTETDVKMMERVVEQMCITQYERESQAYYQRGS
FT SMVLFSSPPVILLISFLIFLIVG"
XX
SQ Sequence 2420 BP; 669 A; 500 C; 583 G; 668 T; 0 other;
cgagcagcca aggttcgcca taatgactgc tctcggtcgt gaggagagga gaagctcgcg 60
gcgccgcggc tgctggatgc tggttctctt tgtggccacc tggagtgacc tgggcctctg 120
caagaagcgc ccgaagcctg gaggatggaa cactgggggc agccgatacc cggggcaggg 180
cagccctgga ggcaaccgct accccacctca gggcggtggt ggctgggggc agcctcatgg 240
tgaagtgtct aatgcattaa cttttgtaag gtactgaata cttaatatgt gggaaacctt 2220
tttgcgtggt ccttaggctt acaatgtgca ctgaatcgtt tcatgtaaga atccaaagtg 2280
gacaccatta acaggtcttt gaaatatgca tgtactttat atttctctata tttgtaactt 2340
tgcatgttct tgttttgtta tataaaaaaa ttgtaaatgt ttaatatctg actgaaatta 2400
aacgagccaa gatgagcacc 2420
//

GenBank Format Example: Header Section

```
LOCUS       NM_197856               1580 bp    mRNA    linear   PLN 09-NOV-2004
DEFINITION  Oryza sativa (japonica cultivar-group) putative
            carnitine/acylcarnitine translocase (OSJNBa0003O19.7), mRNA.
ACCESSION   NM_197856
VERSION     NM_197856.1  GI:37537071
KEYWORDS    .
SOURCE      Oryza sativa (japonica cultivar-group)
  ORGANISM  Oryza sativa (japonica cultivar-group)
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; Liliopsida; Poales; Poaceae;
            Ehrhartoideae; Oryzeae; Oryza.
REFERENCE   1  (bases 1 to 1580)
  AUTHORS   .
  CONSRTM   The Rice Chromosome 10 Sequencing Consortium
  TITLE     In-depth view of structure, activity, and evolution of rice
            chromosome 10
  JOURNAL   Science 300 (5625), 1566-1569 (2003)
   PUBMED   12791992
REFERENCE   2  (bases 1 to 1580)
  AUTHORS   Buell,C.R., Wing,R.A., McCombie,W.R., Messing,J. and Yuan,Q.
  TITLE     Direct Submission
  JOURNAL   Submitted (05-MAY-2003) The Institute for Genomic Research, 9712
            Medical Center Dr, Rockville, MD 20850, USA
COMMENT     PROVISIONAL REFSEQ: This record has not yet been subject to final
            NCBI review. This record is derived from an annotated genomic
            sequence (NT_080068).
            COMPLETENESS: not full length.
```

GenBank Format Example: Feature Table Section



The EMBL feature table is the same with an identifier of FT on each line.

GenBank/EMBL/DDBJ Feature Table: Feature Definitions

http://www.ncbi.nlm.nih.gov/collab/FT/

Definition for mRNA feature:

Feature Key	**mRNA**
Definition	messenger RNA; includes 5'untranslated region (5'UTR), coding sequences (CDS, exon) and 3'untranslated region (3'UTR);
Optional qualifiers	/allele="text" /citation=[number] /db_xref="<database>:<identifier>" /evidence=<evidence_value> /function="text" /gene="text" /label=feature_label /locus_tag="text" (single token) /map="text" /note="text" /old_locus_tag="text" (single token) /operon="text" /product="text" /pseudo /standard_name="text" /usedin=accnum:feature_label

Some features have mandatory qualifiers.

Genbank/EMBL/DDBJ Feature Table: Feature Location

http://www.ncbi.nlm.nih.gov/collab/FT/

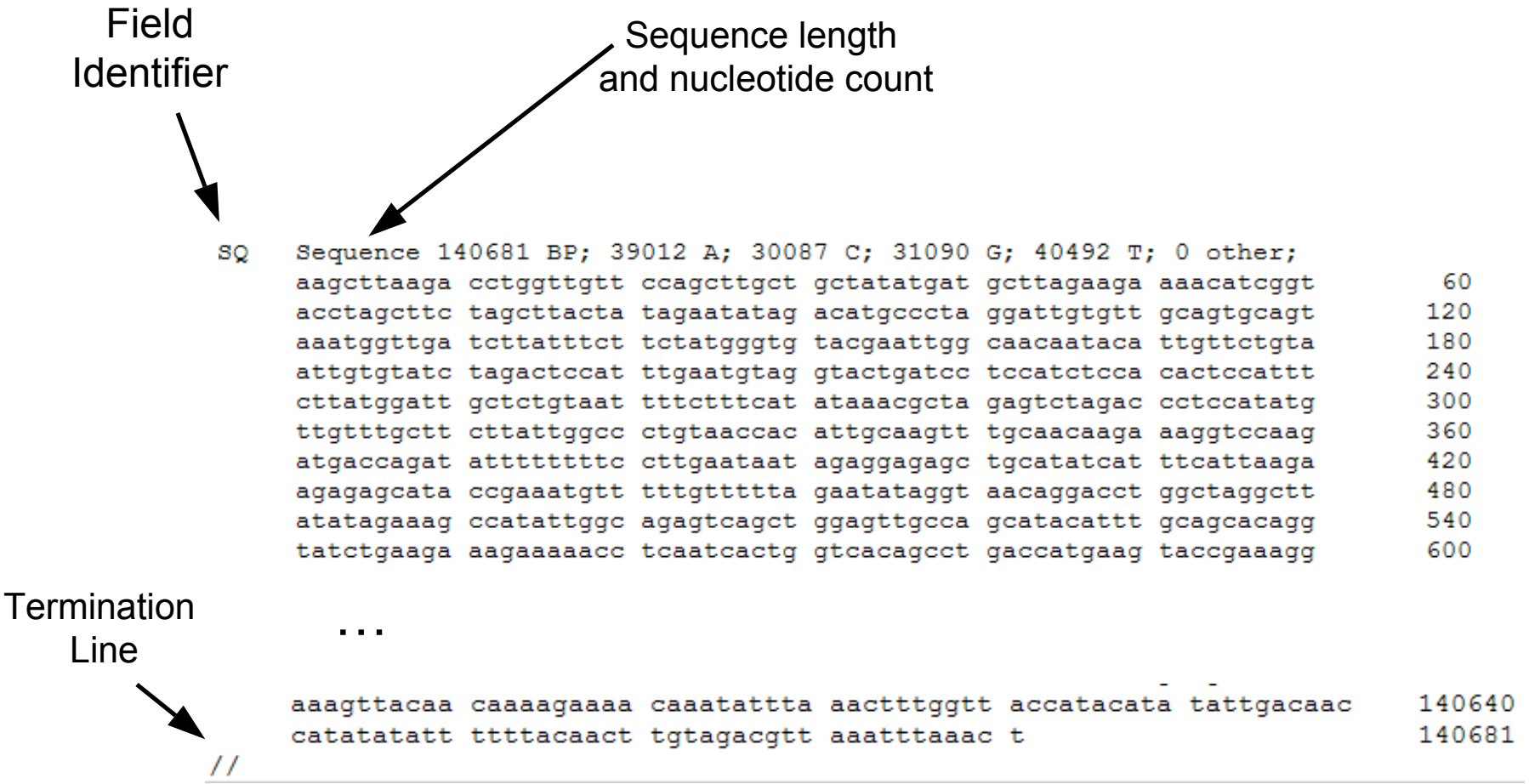
- Feature located at a single base in the seqeunce:
  - misc_feature 176564
- Feature located between two bases:
  - misc_feature 54365^54366
- Feature located in a continuous range:
  - exon 1294..5763
- 'Fuzzy' location of a feature:
  - promoter (1500.1505)..1700
- Feature on complementary strand:
  - mRNA complement(54..3765)
- Feature composed of several segments of sequence:
  - CDS join(10453..12948,13754..15932)
- Combinations of the above are possible.

GenBank Format Example: Sequence Section

EMBL Format Example: Sequence Info

```
ID   AC078894   standard; genomic DNA; PLN; 140681 BP.
XX
AC   AC078894;
XX
SV   AC078894.11
XX
DT   09-AUG-2000 (Rel. 64, Created)
DT   14-APR-2005 (Rel. 83, Last updated, Version 13)
XX
DE   Oryza sativa chromosome 10 BAC OSJNBa0096G08 genomic sequence, complete
DE   sequence.
XX
KW   HTG.
XX
OS   Oryza sativa (japonica cultivar-group)
OC   Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
OC   Spermatophyta; Magnoliophyta; Liliopsida; Poales; Poaceae; Ehrhartoideae;
OC   Oryzeae; Oryza.
XX
RN   [1]
RP   1-140681
RA   Buell C.R., Yuan Q., Ouyang S., Liu J., Gansberger K., Kim M.M.,
RA   Overton II L.L., Bera J.J., Tsitrin T., Krol M.I., Jarrahi B.B., Jin S.S.,
RA   Koo H., Zismann V., Hsiao J., Blunt S., Vanaken S.S., Utterback T.T.,
RA   Feldblyum T.V., Yang Q.Q., Haas B.J., Suh B.B., Peterson J.J.,
RA   Quackenbush J., White O., Salzberg S.L., Fraser C.M.;
RT   "Oryza sativa chromosome 10 BAC OSJNBa0096G08 genomic sequence";
RL   Unpublished.
XX
RN   [2]
RP   1-140681
RA   Buell R.;
RT   ;
RL   Submitted (08-AUG-2000) to the EMBL/GenBank/DDBJ databases.
RL   The Institute for Genomic Research, 9712 Medical Center Dr, Rockville, MD
RL   20850, USA
XX
RN   [3]
RP   1-140681
RA   Buell R.;
RT   ;
RL   Submitted (15-NOV-2001) to the EMBL/GenBank/DDBJ databases.
RL   The Institute for Genomic Research, 9712 Medical Center Dr, Rockville, MD
RL   20850, USA
RN   [4]
RP   1-140681
RA   Buell R.;
RT   ;
RL   Submitted (25-APR-2002) to the EMBL/GenBank/DDBJ databases.
RL   The Institute for Genomic Research, 9712 Medical Center Dr, Rockville, MD
RL   20850, USA, rbuell@tigr.org
XX
RN   [5]
RP   1-140681
RA   Buell R.;
RT   ;
RL   Submitted (18-DEC-2002) to the EMBL/GenBank/DDBJ databases.
RL   The Institute for Genomic Research, 9712 Medical Center Dr, Rockville, MD
RL   20850, USA
XX
RN   [6]
RP   1-140681
RA   Buell R.;
RT   ;
RL   Submitted (20-DEC-2002) to the EMBL/GenBank/DDBJ databases.
RL   The Institute for Genomic Research, 9712 Medical Center Dr, Rockville, MD
RL   20850, USA, rbuell@tigr.org
XX
CC   On Dec 20, 2002 this sequence version replaced gi:27228824.
CC   Address all correspondence to:rice@tigr.org
CC   BAC clone OSJNBa0096G08 is from Oryza sativa chromosome 10
CC   The orientation of the sequence is from SP6 to T7 end of the BAC
CC   clone.
CC   Genes were identified by a combination of several methods: Gene
CC   prediction programs including Fgenesh (http://www.softberry.com/),
CC   genscan and Genscan+ (Chris Burge,
CC   http://CCR-081.mit.edu/GENSCAN.html), GeneMarkHMM (Mark Borodovsky,
CC   http://genemark.biology.gatech.edu/GeneMark/), and GeneSplicer
CC   (Mihaela Pertea and Steven Salzberg, contact mpertea@tigr.org),
CC   searches of the complete sequence against a peptide database and
CC   the plant EST database at TIGR (http://www.tigr.org/tdb/tgi.shtml).
CC   Annotated genes are named to indicate the level of evidence for
CC   their annotation. Genes with similarity to other proteins are named
CC   after the database hits. Genes without significant peptide
CC   similarity but with EST similarity are named as  unknown  proteins.
CC   Genes without protein or EST similarity, that are predicted by more
CC   than two gene prediction programs over most of their length are
CC   annotated as  hypothetical  proteins. Genes encoding tRNAs are
CC   predicted by tRNAscan-SE (Sean Eddy,
CC   http://genome.wustl.edu/eddy/tRNAscan-SE/). Simple repeats are
CC   identified by repeatmasker (Arian Smit,
CC   http://ftp.genome.washington.edu/RM/RepeatMasker.html).
```

EMBL Format Example: Sequence



SWISSPROT/TrEMBL Format

- Very similar to EMBL format. Feature table is extended to capture structural features and biochemical information about the protein.

```
FT   SIGNAL        1     18       By similarity.
FT   CHAIN        19    247       Chloroplast ATP synthase a chain.
FT   TRANSMEM     39     58       Potential.
FT   TRANSMEM     97    115       Potential.
FT   TRANSMEM    134    153       Potential.
FT   TRANSMEM    221    240       Potential.
SQ   SEQUENCE   247 AA;  27291 MW;  540649B34778E585 CRC64;
     MNIIPCSIKT LKGLYDISGV EVGQHFYWQI GGFQIHAQVL ITSWVVITIL LGSVIIAVRN
     PQTIPTDGQN FFEYVLEFIR DLSKTQIGEE YGPWVPFIGT MFLFIFVSNW SGALLPWKII
     QLPHGELAAP TNDINTTVAL ALLTSAAYFY AGLSNKGLSY FEKYIKPTPI LLPINILEDF
     TKPLSLSFRL FGNILADELV VVVLVSLVPL VVPIPVMFLG LFTSGIQALI FATLAAAYIG
     ESMEGHH
//
```

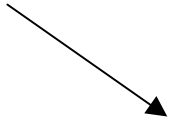
Formats for Sequence Analysis

- The database flat file formats are unwieldy for sequence analysis.
  - Sometimes you need just the sequence for analysis
  - Other times you need to work with the annotations in the database or generated by sequence analysis programs
  - Rarely do you need all of the metadata
- Many formats have been created over the years for this purpose
- FASTA format is the most common sequence format.

FASTA Format

A sample FASTA sequence record from a sequence DB:

Definition Line



```
>gi|46849661|gb|AC137924.3| Oryza sativa (japonica cultivar-group) chromosome 11 clone OSJNBa0095K08, complete sequence
AAGCTTTGACGATACATGCATTATAAATGTGCAGTGTGACCTCTACCACTTCATCCACCATGAGTGCTAT
CATGTCAAAGGACGATTCTTCGACGCAGAGAGCATTCTGGCTACAAGCGAAGCTTACAAGAATCTTCAGG
AGTGGAACAACCGAACAACGCGATGCCATAATATAATTAGATAGTGTACTCGAAGTGACAATGTAAAAAA
TATTTTAACATTTTGATGACACTATTCACTTATGTAATATATGTATATTTGTCAGTATCAAATGTTTGTT
AGTTACTAATATTATTTAAGCATCAATTAATTAATTAATTGGAACATTATTTATCACAAATTATTATTGT
AACATTTTTTCAAATTTTTTACCTATTTTCGCAGGCGGCGTCATCTCATTTTTTCAGGCGGACTAAGATA
TATTTTCCCAGGCAGCGTGTCTAGCTCCAGTCCGCTAGGGAAAATGATCTTCCCAGGCGGACCTCCTACC
```

Width of sequence rows usually 60,70,72 or 80 cols.

Note:The MultiFASTA Format is composed of FASTA records concatenated together.

FASTA Format: Definition Line

The minimum standard for a FASTA definition line is a ‘>’ immediately followed by a sequence identifier. White space followed by a comment may optionally be added.

Example:
>TA347833

The sequence databases follow a convention for composition of a sequence identifier for a FASTA formatted record.

FASTA Format: Sequence Identifiers

- GenBank/EMBL/DDBJ
  - gi|gi_number|gb|accession.version|locus
  - gi|gi_number|embl|accession.version|locus
  - gi|gi_number|dbj|accession.version|locus
- NCBI Reference Sequence
  - ref|accession|locus
- PIR
  - pir|entry
- SWISSPROT
  - sp|accession|locus
- PDB
  - pdb|entry|chain

- This list is not comprehensive, there are others out there.

Multiple Sequence Alignment(MSA) Formats

- MSA formats are needed to analyze and store the results of multiple sequence alignment from a programs such as ClustalW or MUSCLE.

- The MSA formats need to preserve information about the alignment such as gaps and substitutions.

Aligned FASTA Format

Gaps in the alignments are represented by dashes (-).

```
>PCXB_PSEPU P00437 Protocatechuate 3,4-dioxygenase beta chain (EC 1.13.11.3) (3,4-PCD).
PAQDNSRFVIR----------------DRNW--HPKALTPD-------------YKTSIA
RSPRQALVSIP----QSISETTGPNFSHLGFGAH---------------------DHDLL
LNFNNGGLPIGERIIVAGRVVDQYGKPVPNTLVEMWQANAGGRYRHKNDRYLAPLDPNFG
GVGRCLTDSDGYYSFRTIKPGPYPWRNGPNDWRPAHIHFGISGPSIATKLITQLYFEGDP
L----IPMCPIVKSIANPEAVQ-QLIAKLDMNNANPMD-------------------CLA
YRFD----IVLRGQRKTHFENC-------------------
>Q9ZFA1 Q9ZFA1 Protocatechuate 3,4-dioxygenase beta subunit (EC 1.13.11.3).
MTLTQHDIDLEIAAEHATYEKRVADGAPVEH--HPRRDYAP-------------YRSSTL
RHPKQPPVTIDVSKDPELVELASPAFGERDITEI---------------------DNDLT
RQ--HNGEPIGERITVSGRLLDRDGRPIRGQLVEIWQANSAGRYAHQREQHDAPLDPNFT
GVGRTLTDDEGGYHFTTVQPGPYPWRNHVNAWRPAHIHFSMFGSAFTQRLVTQMYFPSDP
L----FPYDPIIQS-VTDDAARQRLVATYDHSLSVPEF-------------------SMG
YHWD----IVLDGPHATWIEEGR------------------
>PCXB_BURCE P15110 Protocatechuate 3,4-dioxygenase beta chain (EC 1.13.11.3) (3,4-PCD).
---MDSPTILT----------------PRDWPSHPAYVHPD-------------YRSSVK
RGPTRPMIPLK----ERLRDQYAPVYGAEDLGPL---------------------DHDLT
KNAVKNGEPLGERIVVTGRVLDEGGKPVRNTLVEVWQANAAGRYVHKVDQHDAPLDPNFL
GAGRCMTDAEGRYRFLTIKPGAYPWGNHPNAWRPNHIHFSLFGDYFGSRLVTQMYFPGDP
L----LAYDPIFQ--GTPEAARDRLISRFSLDTTEEGH-------------------ALG
YEFD----IVLRGRDATPMER--------------------
```

ClustalW Format

A common MSA format is the alignments from the ClustalW program.
Most phylogenetic programs can take ClustalW alignments as input.

```
CLUSTAL W (1.74) multiple sequence alignment


ATP7B_MOUSE          MDPRKNLASVGTMPEQERQVTAKE-ASRKILSKLALPGRPWEQSMKQSFAFDNVGYEGGL 59
ATP7B_RAT            ------------MPEQERKVTAKE-ASRKILSKLALPTRPWGQSMKQSFAFDNVGYEGGL 47
ATP7B_HUMAN          ------------MPEQERQITAREGASRKILSKLSLPTRAWEPAMKKSFAFDNVGYEGGL 48
ATP7B_OVIS_ARIES     -----------MKPEEERPIIDREKASRRILSKLFQP------AMKQSFAFDNNGYEDDL 43
                                  **:** :  :* ***:*****  *      :**:****** ***..*

ATP7B_MOUSE          DSTSSSPAATD-VVNILGMTCHSCVKSIEDRISSLKGIVNIKVSLEQGKHTVRYVPSVMN 118
ATP7B_RAT            DSTCFILQLTTGVVSILGMTCHSCVKSIEDRISSLKGIVSIKVSLEQGSATVKYVPSVLN 107
ATP7B_HUMAN          DGLGPSSQVATSTVRILGMTCQSCVKSIEDRISNLKGIISMKVSLEQDSATVKYVPSVVC 108
ATP7B_OVIS_ARIES     DGVCPS-QTAAGTISIVGMTCQSCVKSIEGRVSSLKGIVSIKVSLEQSSAEVRYVPSVVS 102
                     *.       :  .: *:****:*******.*:*.****:.:******..  *:*****:

ATP7B_MOUSE          LQQICLQIEDMGFEASAAEGKAASWPSRSSPAQEAVVKLRVEGMTCQSCVSSIEGKIRKL 178
ATP7B_RAT            LQQICLQIEDMGFEASAAEGKAASWPSRSSPAQEAVVKLRVEGMTCQSCVSSIEGKIRKL 167
ATP7B_HUMAN          LQQVCHQIGDMGFEASIAEGKAASWPSRSLPAQEAVVKLRVEGMTCQSCVSSIEGKVRKL 168
ATP7B_OVIS_ARIES     LMQICHQIEDMGFQASVAEGKATSWASRVSPTSEAVVKLRVEGMTCQSCVSSIEGKIGKL 162
                     * *:* ** ****:** *****:**.**  *:.**********************: **
```

Downloading from NCBI

http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi

- “Entrez” is NCBI’s downloading service.
- Convenient to download one or 1000 sequences.
- Accepts complex search queries.
- Accepts lists of accession numbers.
- Allows downloading of numerous formats.

Searching at Entrez

http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helpentrez.chapter.EntrezHelp
http://www.ncbi.nlm.nih.gov/entrez/query/static/help/helpdoc.html

- Boolean terms: AND, OR, NOT.
- Limit search of specific terms to specific fields: [orgn], [accession], [author], [gene], [keyword], [journal], [slen]
- Via the web interface, limits can be imposed. (Find sequences published since January 1997.)

GFF2 Format for Annotation

http://www.sanger.ac.uk/Software/formats/GFF/

GFF = General Feature Format
Tab delimited, easy to work with.
Many annotation viewers accept this format in various ‘dialects’.

Columns:

1. Reference Sequence: base seq to which the coordinated are anchored
2. Source: source of the annotation
3. Type: Type of feature
4. Start
5. End (Start is always less than End)
6. Score: Used for holding numerical scores (similarity, etc)
7. Strand: “+’,“-”, or ‘.” if unstranded
8. Phase: Signifies codon phase for CDS features
9. Group: Group feature belongs to. Also attributes such as name and alias

```
Chr1    TIGR_annot_DB   gene    9523    12619   .       +       .       Gene LOC_Os01g01030; Note "Multicopper oxidase, putative"; Alias "11667.t00003"
Chr1    TIGR_annot_DB   mRNA    9523    12619   .       +       .       mRNA LOC_Os01g01030.1; Gene LOC_Os01g01030; Note "11667.m00004".
Chr1    TIGR_annot_DB   five_prime_UTR  9523    9575    .       +       .       mRNA LOC_Os01g01030.1.
Chr1    TIGR_annot_DB   CDS     9576    10615   .       +       0       mRNA LOC_Os01g01030.1.
Chr1    TIGR_annot_DB   CDS     10708   11073   .       +       2       mRNA LOC_Os01g01030.1.
Chr1    TIGR_annot_DB   CDS     11161   11239   .       +       2       mRNA LOC_Os01g01030.1.
Chr1    TIGR_annot_DB   CDS     11771   11973   .       +       0       mRNA LOC_Os01g01030.1.
Chr1    TIGR_annot_DB   CDS     12068   12161   .       +       2       mRNA LOC_Os01g01030.1.
Chr1    TIGR_annot_DB   three_prime_UTR 12162   12619   .       +       .       mRNA LOC_Os01g01030.1.
```

GFF3 Format

http://song.sourceforge.net/gff3.shtml

Extension of GFF by the Sequence Ontology (SO) and GMOD Projects

A much needed extension to GFF/GFF2:
Allows hierarchies more than one level deep
Separated Group membership and feature name/ID
Attributes take the form of “Key=Value” pairs
Feature can belong to more than one group

```
11667   PASA2   gene    13123   16979   .       +       .       ID=11667.t00004;Name="[pasa:asmbl_8,status:4]".
11667   PASA2   mRNA    13123   13778   .       +       .       ID=11667.m00005,update,status:[pasa:asmbl_8,status:4];Parent=11667.t00004.
11667   PASA2   CDS     13401   13778   .       +       0       Parent=11667.m00005,update,status:[pasa:asmbl_8,status:4].
11667   PASA2   mRNA    14185   14276   .       +       .       ID=11667.m00005,update,status:[pasa:asmbl_8,status:4];Parent=11667.t00004.
11667   PASA2   CDS     14185   14276   .       +       0       Parent=11667.m00005,update,status:[pasa:asmbl_8,status:4].
11667   PASA2   mRNA    14360   15060   .       +       .       ID=11667.m00005,update,status:[pasa:asmbl_8,status:4];Parent=11667.t00004.
11667   PASA2   CDS     14360   15060   .       +       2       Parent=11667.m00005,update,status:[pasa:asmbl_8,status:4].
11667   PASA2   mRNA    15303   15373   .       +       .       ID=11667.m00005,update,status:[pasa:asmbl_8,status:4];Parent=11667.t00004.
11667   PASA2   CDS     15303   15373   .       +       1       Parent=11667.m00005,update,status:[pasa:asmbl_8,status:4].
11667   PASA2   mRNA    15770   15859   .       +       .       ID=11667.m00005,update,status:[pasa:asmbl_8,status:4];Parent=11667.t00004.
11667   PASA2   CDS     15770   15859   .       +       0       Parent=11667.m00005,update,status:[pasa:asmbl_8,status:4].
11667   PASA2   mRNA    15944   16123   .       +       .       ID=11667.m00005,update,status:[pasa:asmbl_8,status:4];Parent=11667.t00004.
11667   PASA2   CDS     15944   16123   .       +       0       Parent=11667.m00005,update,status:[pasa:asmbl_8,status:4].
11667   PASA2   mRNA    16333   16431   .       +       .       ID=11667.m00005,update,status:[pasa:asmbl_8,status:4];Parent=11667.t00004.
11667   PASA2   CDS     16333   16395   .       +       0       Parent=11667.m00005,update,status:[pasa:asmbl_8,status:4].
11667   PASA2   mRNA    16536   16979   .       +       .       ID=11667.m00005,update,status:[pasa:asmbl_8,status:4];Parent=11667.t00004.
11667   PASA2   gene    19643   23696   .       +       .       ID=11667.t00005;Name="[pasa:asmbl_14,status:8]".
```

A Few Final Words About Text Editors, Excel, Windows and Unix

- True text editors must be used when working with sequence records. Word processors such as MS word introduce formatting and control characters.

- Excel files must be saved in Tab Delimited format to be truly portable.

- Unix and MS Windows (and DOS) use different characters to indicate a new line in a text file. If you open a sequence file in notepad and see the text in a long string with boxes where the line breaks should be, the file uses Unix line ending.

- Using a text editor such as Emacs or JEdit will allow you to open the file properly in Windows, otherwise utilities exist for convert between Unix and DOS line endings (dos2unix, unix2dos).