Biological Data Formats

Biological Data: Characteristics

- Biological data is ‘data’ or ‘measurements’ collected from biological sources, which is stored or exchanged in a digital form.
- The amount and range of variability in data is high.
- Representations of the same data by different biologists can be different.

Types of Data

- Broad and diverse :
  - Sequence
  - annotated sequence feature
  - gene expression data
  - alignment data
  - protein cluster data
  - 3 –dimensional structure information
  - Images
  - Graphs
  - semi-structured/unstructured text
- New bio-analytical procedures and progresses add new data types and so add more instability to the data types.

Data Organization

- Challenges to store and maintain the huge complex biological data:
  - Amounts of data increased almost exponentially in the last decade.
  - New types of data are coming into existence with evolving biological concepts.
  - Lack of standardization in nomenclature in biological data.
- Data Storage : Flat files and relational databases
  - Most of the biological data (estimated ~70%) are stored in text form.
  - Rest resides in different databases, ranging from indexed files to specialized relational databases.

Distribution of DATA

- Biological knowledge seems to be distributed among specialized databases/data sources.
- Each database has its own complex data structures reflecting the scientific concept they model.
- Many data sources have overlapping data elements with conflicting definition.
- Data sources are non-standard and often not well documented.
- Integration and conversion of data from heterogeneous data sources are very important for effective use of the biological information.
- Important to interpret the various data formats, downloading data from various data sources and conversion of the data to integrate information.

Disparate DATA sources

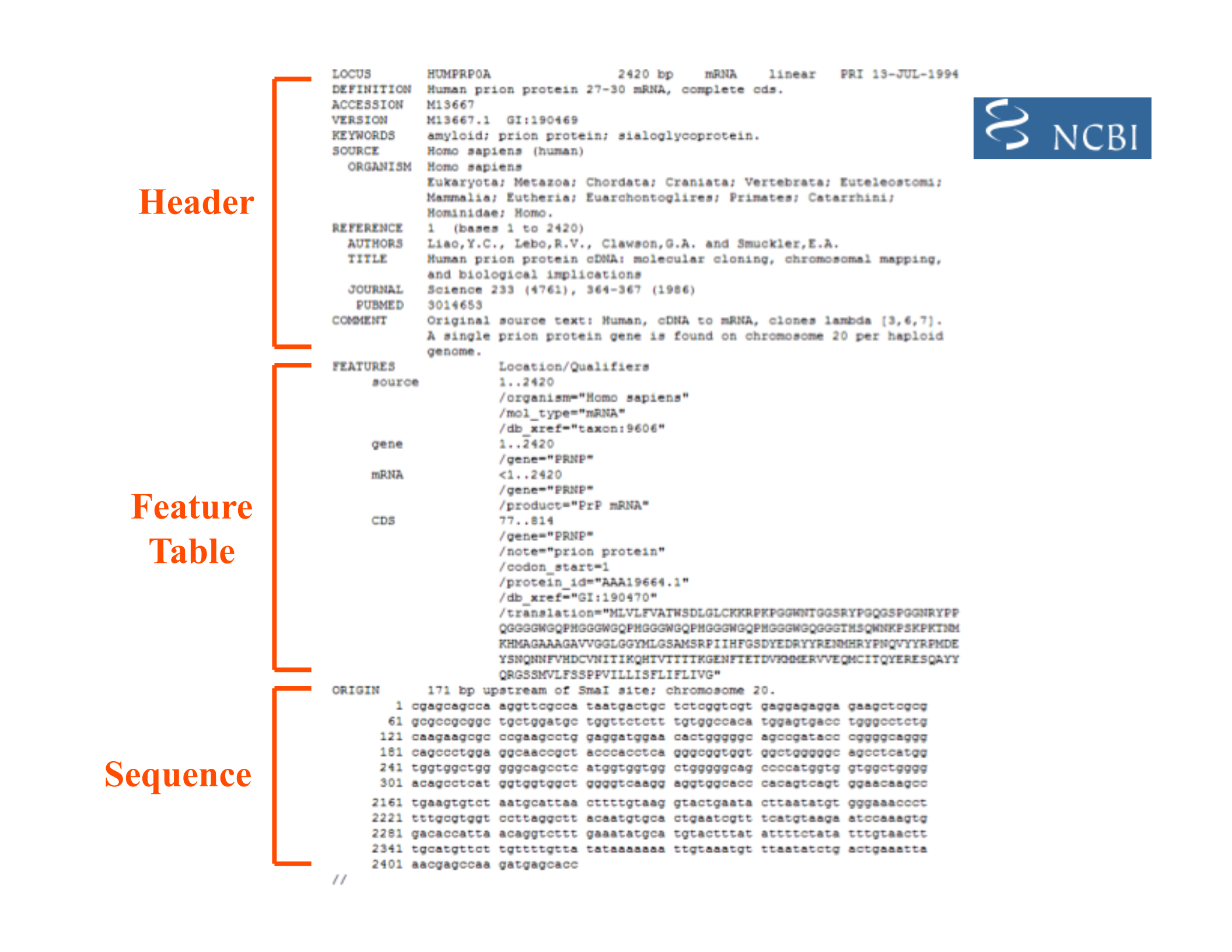
- Partial list of hundreds of databases exist today.
- Each database has its own complex data structures reflecting the scientific concept they model.
  - Genbank/EMBL
  - Swissprot – protein-curated, minimum redundancy
  - KEGG – cellular pathways
  - PDB – protein structures
  - PIR – protein database
  - BIND – protein-protein interaction

Heterogeneous Data Formats

- Sequence
  - Flat File
  - FASTA (multi-FASTA)
  - XML
- Annotation
  - GFF
  - XML
  - Flat File
- Multi-Sequence Alignment
  - ClustalW

Flat File Format

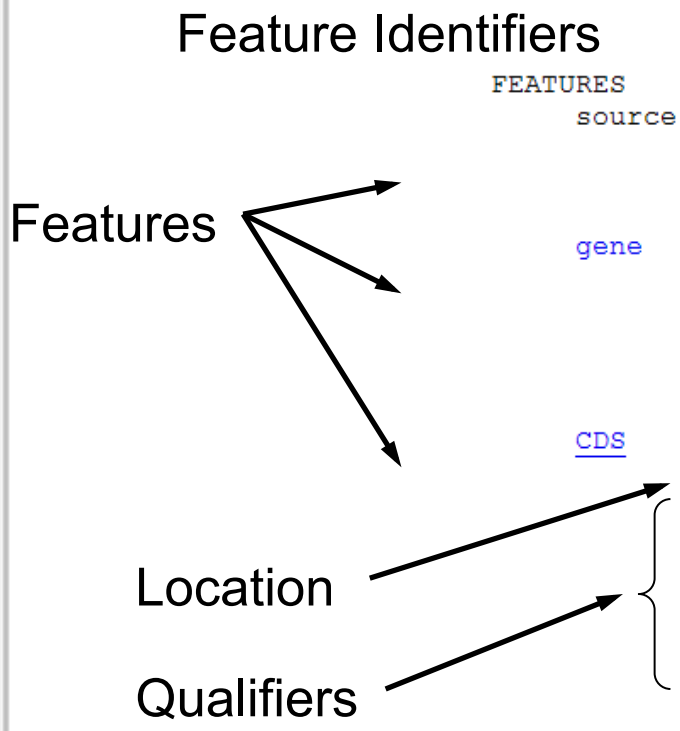
- GenBank, EMBL and DDBJ formed a collaboration in 1986.
- Sequence data moved to a defined flat file format with a shared feature table format and annotation standards.
- The flat file format from the sequence databases are still used today to access and display sequence and annotation.


Header
Feature Table
Sequence
NCBI
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
//

GenBank : Header

```
LOCUS       TVU35243                1804 bp    mRNA    linear   INV 30-JAN-2006
DEFINITION  Trichomonas vaginalis AP65-3 adhesin mRNA, complete cds.
ACCESSION   U35243
VERSION     U35243.1  GI:1209523
KEYWORDS    .
SOURCE      Trichomonas vaginalis
  ORGANISM  Trichomonas vaginalis
            Eukaryota; Parabasalidea; Trichomonada; Trichomonadida;
            Trichomonadidae; Trichomonadinae; Trichomonas.
REFERENCE   1  (bases 1 to 1804)
  AUTHORS   O'Brien,J.L., Lauriano,C.M. and Alderete,J.F.
  TITLE     Molecular characterization of a third malic enzyme-like AP65
            adhesin gene of Trichomonas vaginalis
  JOURNAL   Microb. Pathog. 20 (6), 335-349 (1996)
   PUBMED   8831829
REFERENCE   2  (bases 1 to 1804)
  AUTHORS   Alderete,J.F.
  TITLE     Direct Submission
  JOURNAL   Submitted (01-SEP-1995) John F. Alderete, University of Texas
            Health Science Center at San Antonio, Microbiology, 7703 Floyd Curl
            Drive, San Antonio, TX 78284-7758, USA
REFERENCE   3  (bases 1 to 1804)
  AUTHORS   Mundodi,V., Kucknoor,A.S., Klumpp,D.J., Chang,T.H. and
            Alderete,J.F.
  TITLE     Silencing the ap65 gene reduces adherence to vaginal epithelial
            cells by Trichomonas vaginalis
  JOURNAL   Mol. Microbiol. 53 (4), 1099-1108 (2004)
   PUBMED   15306014
```

GenBank: Features




```
FEATURES             Location/Qualifiers
     source          1..1580
                     /organism="Oryza sativa (japonica cultivar-group)"
                     /mol_type="mRNA"
                     /db_xref="taxon:39947"
     gene            1..1580
                     /locus_tag="OSJNBa0003O19.7"
                     /note="similar to carnitine acylcarnitine translocase
                     GB:CAB55356 GI:5851675 (Homo sapiens); EST C26405,
                     AU093530, AU093531, AU108742 from this gene"
                     /db_xref="GeneID:3054652"
     CDS             79..969
                     /locus_tag="OSJNBa0003O19.7"
                     /codon_start=1
                     /product="putative carnitine/acylcarnitine translocase"
                     /protein_id="NP_922838.1"
                     /db_xref="GI:37537072"
                     /db_xref="GeneID:3054652"
                     /translation="MGDVVKDLVAGTVGGAANLIVGHPFDTIKVKLQSQPTPAPGQFP
                     KYAGAVDAVKQTIATEGPRGLYKGMGAPLATVAAFNALLFTVRGQMEALLRSEPGQPL
                     TVNQQVVAGAGAGVAVSFLACPTELIKCRLQAQSALAEAAAASGVALPKGPIDVAKHV
                     VREAGMKGLFKGLVPTMGREVPGNAVMFGVYEGTKQYLAGGQDTSNLGRGSLILSGGL
                     AGAVFWLSVYPTDVVKSVIQVDDYKKPRYSGSVDAFKKILAADGVKGLYKGFGPAMAR
                     SVPANAATFLAYEITRSALG"
```

GenBank: Sequence

Sequence Field Identifier

```
ORIGIN
        1 ttttagatta aagatgctcg catcttcagt cgctgctcca gtccgcaaca tctgcagggc
       61 taagctccca gctctcaaga caggaatgac cctccttcag gatggtgatc tttccaaggg
      121 ctctgctttc acaaaggaag aacgtgatcg ccttaacctt cgcggtctcc tcccatacaa
      181 ggtcttcaca aaggatgaac aagctgctcg tatccgccgc cagttcgagt tgatgccaac
      241 accactcctc aagtacatct tcctcgctaa cgagcgtgag aaaaactcac agtccttctg
      301 gagattcctc ttcacacacc caccaacaga gacaatgcca gttctctaca caccaacagt
      361 tggtgaagcc tgccagaagt gggctacaca ccgccagtca taccgtggca tctacatcac
      421 accagaagac tctggcaaga tcaaggacat cctccgcaac tacccacgcc aggacatccg
      481 ctgcatcgtc gttacagatg gtggccgtat cctcggtctc ggtgatctcg gtgcttccgg
      541 ccttggtatc ccagtcggca agcttatgct ttacacactc atcggtcagg tccatccaga
      601 tcagacactc ccagtccagt tagatatggg tacagaccgc aaggaaatcc tcgccgaccc
      661 actctaccac ggctggcgcc atccaagaat acgtggccca gaacacacaa agttcgttgc
      721 cgagttcgtt gatgctgtca aggaagtctt tggcgagaca tgccttgtcc agttcgaaga
      781 tttcgaaatg gaaactgctt tcaagcttct tgatcacttc cgctggcgct gcaactgctt
      841 caacgatgat atcgaaggca cagctgccgt cgctgctgct acactcgctt ccgctacaca
      901 catggaaggc gttccagatc tcaagaacca gaagatcatc ttcatcggcg ctggctctgc
      961 tgctacaggc attgctaacc tcatcgttga tatggctgtt tcccgcggtg gcatctcacg
     1021 caaggatgct gagagaaaca tcatcatgtt cgatcacaag ggtatggtcc atgctgaccg
     1081 taaggatctc tacgacttca acaagccata catgcacgac atggaagtct acggctccgt
     1141 ccttgagggt gtcaagaagt tcaaggctac atgcgtcatc ggcgtttctg gtgttccagg
     1201 actcatcaca aaggaaatcg tccaggctac atgcgctaac tgcgagcgcc cagtcatcat
     1261 gccactttcc aacccaacag tcaaggctga agctaagcca cacgatgtct accagtggtc
     1321 caatggcaag gccctctgcg ctacaggctc tccattccca gttgagacag tcaacggaaa
     1381 gaagacaatc acagctcagg ctaacaactc ctggatcttc ccagctgtcg gctacgccct
     1441 cgttacaaca cgcgctcgcc actgcccagg caaggtcttc gaagttgctg ctgaatccct
     1501 tgcttccctt gttaagaagg aagaccacga tatgggcaac cttctcccac cactcgacaa
     1561 gatccgtgag tactcattcg gcatcgccct cgatgttgct aagtacctca tcaagaacga
     1621 gctcgccaca gctctcccac caaagggcac agagctcaag gactggctca aggctcagct
     1681 cttcgatcca caggctgaat acgagcaact ctactaagca gtttttaaaa ctctttcaat
     1741 tgtctttgaa aaaaaaaaaa aaaaaaaaaa aaaaaaaaaa aaaaaaaaaa aaaaaaaaaa
     1801 aaaa
//
```

Termination Line

SWISSPROT/TrEMBL Format

- Swiss-Prot/ TrEMBL - protein sequence database.
- 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
//
```


13

FASTA Format : Significance

- 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 output generated by sequence analysis programs.
- Many formats have been created over the years for this purpose.
- FASTA format is the most common sequence format.

FASTA Format – Example

A single FASTA sequence record from a sequence database:

- Definition Line or Header begins with '>'

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

- Width of sequence rows usually 60 letters.

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

15

FASTA Format: Definition Line

- Definition line has the important identifier for the sequence.
- 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

16

XML: Characteristics

- XML = eXtensible Markup Language
- Tag based like HTML
- Human readable
- Have inherent hierarchical data structures
- Easy to use for data exchange
- Many bioinformatics software tools are XML-compliant
- The data to be contained is described using a Document type Definition (DTD) or an XML schema.

XML: Example

```
<?xml version="1.0" encoding="UTF-8"?>
<?format DECIMAL="."?>
<!DOCTYPE Bsml PUBLIC "-//EBI//Labbook, Inc. BSML DTD//EN"
                "http://www.ebi.ac.uk/xembl/dtd/BSML2_2.DTD">
<bsml>
<definitions>
<sequences>
<sequence id="AB12345" title="AB12" molecule="dna"
length="500" topology="linear"
strand="ds"
representation="raw">
<seq-data>acgtacgtacgtacgtacgtcgcgaacgccg
taact...</seq-data>
</sequence>
</sequences>
</definitions>
</bsml>
```

Root Tag

Start Tag

Attributes

Character Data

End Tag

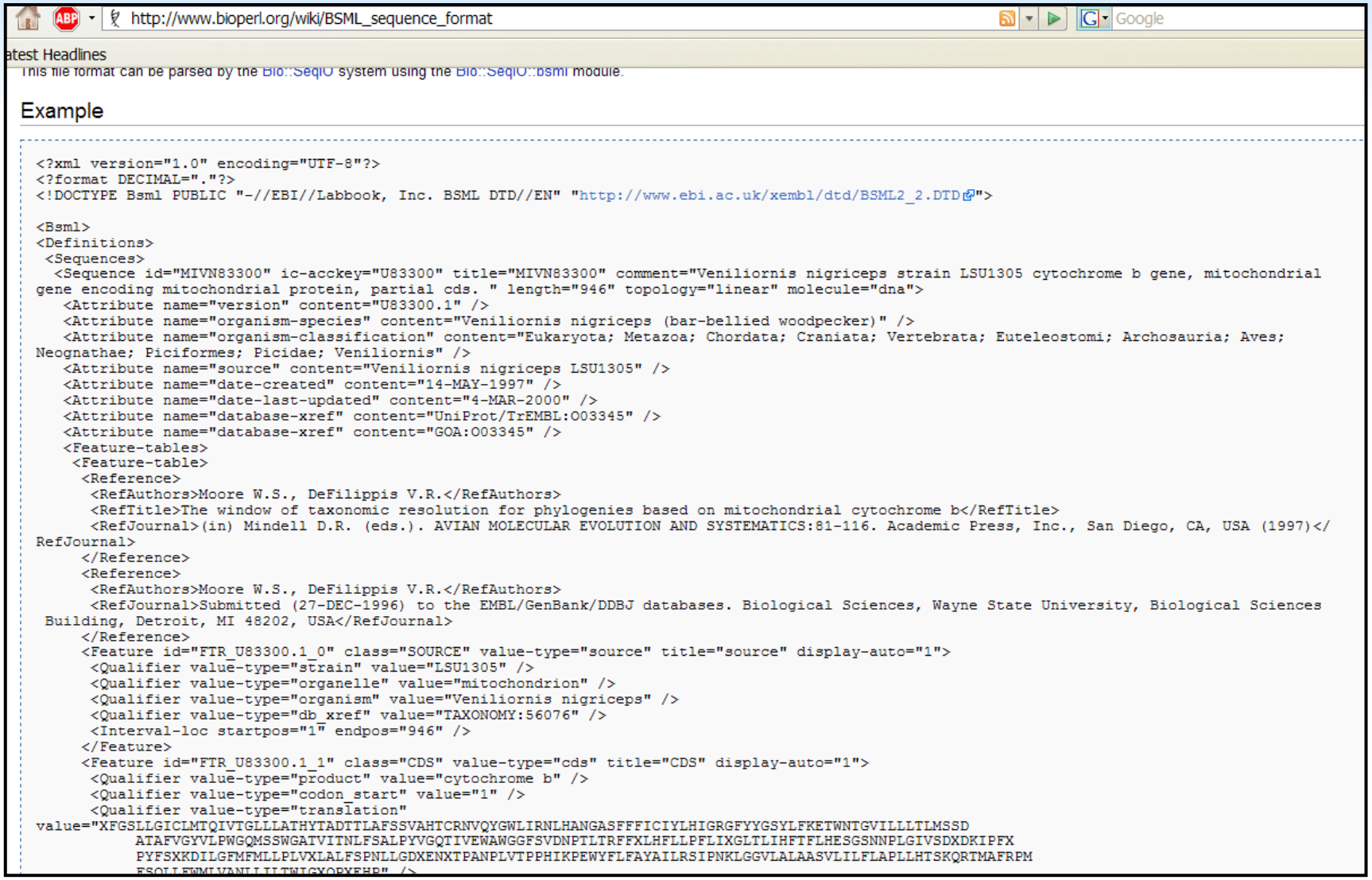
End Root Tag

Bioinformatics Sequence Markup Language (BSML)

- XML Formats in Bioinformatics
  - Specially designed for a wide variety of information attached to biological sequences
- Allows modularization and improves portability.
- Flexible mean of describing any element of a sequence, sequence alignment, etc.
- BSML Objects can be downloaded from
  - http://bsml.sourceforge.net/ .

19

BSML


http://www.bioperl.org/wiki/BSML_sequence_format
Google
atest Headlines
This file format can be parsed by the Bio::SeqIO system using the Bio::SeqIO::bsml module.
Example
<?xml version="1.0" encoding="UTF-8"?>
<?format DECIMAL="."?>
<!DOCTYPE Bsml PUBLIC "-//EBI//Labbook, Inc. BSML DTD//EN" "http://www.ebi.ac.uk/xembl/dtd/BSML2_2.DTD">
<Bsml>
<Definitions>
<Sequences>
<Sequence id="MIVN83300" ic-acckey="U83300" title="MIVN83300" comment="Veniliornis nigriceps strain LSU1305 cytochrome b gene, mitochondrial gene encoding mitochondrial protein, partial cds. " length="946" topology="linear" molecule="dna">
<Attribute name="version" content="U83300.1" />
<Attribute name="organism-species" content="Veniliornis nigriceps (bar-bellied woodpecker)" />
<Attribute name="organism-classification" content="Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Archosauria; Aves; Neognathae; Piciformes; Picidae; Veniliornis" />
<Attribute name="source" content="Veniliornis nigriceps LSU1305" />
<Attribute name="date-created" content="14-MAY-1997" />
<Attribute name="date-last-updated" content="4-MAR-2000" />
<Attribute name="database-xref" content="UniProt/TrEMBL:O03345" />
<Attribute name="database-xref" content="GOA:O03345" />
<Feature-tables>
<Feature-table>
<Reference>
<RefAuthors>Moore W.S., DeFilippis V.R.</RefAuthors>
<RefTitle>The window of taxonomic resolution for phylogenies based on mitochondrial cytochrome b</RefTitle>
<RefJournal>(in) Mindell D.R. (eds.). AVIAN MOLECULAR EVOLUTION AND SYSTEMATICS:81-116. Academic Press, Inc., San Diego, CA, USA (1997)</RefJournal>
</Reference>
<Reference>
<RefAuthors>Moore W.S., DeFilippis V.R.</RefAuthors>
<RefJournal>Submitted (27-DEC-1996) to the EMBL/GenBank/DDBJ databases. Biological Sciences, Wayne State University, Biological Sciences Building, Detroit, MI 48202, USA</RefJournal>
</Reference>
<Feature id="FTR_U83300.1_0" class="SOURCE" value-type="source" title="source" display-auto="1">
<Qualifier value-type="strain" value="LSU1305" />
<Qualifier value-type="organelle" value="mitochondrion" />
<Qualifier value-type="organism" value="Veniliornis nigriceps" />
<Qualifier value-type="db_xref" value="TAXONOMY:56076" />
<Interval-loc startpos="1" endpos="946" />
</Feature>
<Feature id="FTR_U83300.1_1" class="CDS" value-type="cds" title="CDS" display-auto="1">
<Qualifier value-type="product" value="cytochrome b" />
<Qualifier value-type="codon_start" value="1" />
<Qualifier value-type="translation"
value="XFGSLLGICLMTQIVTGLLLATHYTADTTLAFSSVAHTCRNVQYGWLIRNLHANGASFFFICIYLHIGRGFYYGSYLFKETWNTGVILLLTLMSSD
ATAFVGYVLPWGQMSSWGATVITNLFSALPYVGQTIVEWAWGGFSVDNPTLTRFFXLHFLLPFLIXGLTLIHFTFLHESGSNNPLGIVSDXDKIPFX
PYFSXKDILGFMFMLLPLVXLALFSPNLLGDXENXTPANPLVTPPHIKPEWYFLFAYAILRSIPNKLGGVLALAASVLILFLAPLLHTSKQRTMAFRPM
ESQLLFWMLVANLLILTWIGXQPVEHP" />
J. Craig Venter
INSTITUTE
20

GFF: an Exchange Format for Feature Description

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

GFF = General Feature Format

Tab delimited, easy for data parsing and processing.

Many annotation viewers accept this format in various ‘dialects’.

Fields:

1. Reference Sequence: base seq to which the coordinates 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. Frame: Signifies codon phase for coding sequence (CDS) features
9. Other attributes or/and comments

```
SEQ1    EMBL    atg      103    105    .      +      0
SEQ1    EMBL    exon     103    172    .      +      0
SEQ1    EMBL    splice5  172    173    .      +      .
SEQ1    netgene splice5  172    173    0.94   +      .
SEQ1    genie   sp5-20   163    182    2.3    +      .
SEQ1    genie   sp5-10   168    177    2.1    +      .
SEQ2    grail   ATG      17     19     2.1    -      0
```

GFF3 Format

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

Extension of GFF by the Sequence Ontology (SO) and Generic Model Organism Database (GMOD) Projects

-- Allows hierarchies more than one level deep

-- Separated group membership and feature name/ID

-- Attributes take the form of “Key=Value” pairs

```
##gff-version   3
##sequence-region   ctg123 1 1497228
ctg123 . gene            1000  9000  .  +  .  ID=gene00001;Name=EDEN
ctg123 . TF_binding_site 1000  1012  .  +  .  Parent=gene00001
ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00001;Parent=gene00001
ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00002;Parent=gene00001
ctg123 . mRNA            1300  9000  .  +  .  ID=mRNA00003;Parent=gene00001
ctg123 . exon            1300  1500  .  +  .  Parent=mRNA00003
ctg123 . exon            1050  1500  .  +  .  Parent=mRNA00001,mRNA00002
ctg123 . exon            3000  3902  .  +  .  Parent=mRNA00001,mRNA00003
ctg123 . exon            5000  5500  .  +  .  Parent=mRNA00001,mRNA00002,mRNA00003
ctg123 . exon            7000  9000  .  +  .  Parent=mRNA00001,mRNA00002,mRNA00003
ctg123 . CDS             1201  1500  .  +  0  ID=cds00001;Parent=mRNA00001
ctg123 . CDS             3000  3902  .  +  0  ID=cds00001;Parent=mRNA00001
ctg123 . CDS             5000  5500  .  +  0  ID=cds00001;Parent=mRNA00001
ctg123 . CDS             7000  7600  .  +  0  ID=cds00001;Parent=mRNA00001
ctg123 . CDS             1201  1500  .  +  0  ID=cds00002;Parent=mRNA00002
ctg123 . CDS             5000  5500  .  +  0  ID=cds00002;Parent=mRNA00002
ctg123 . CDS             7000  7600  .  +  0  ID=cds00002;Parent=mRNA00002
ctg123 . CDS             3301  3902  .  +  0  ID=cds00003;Parent=mRNA00003
ctg123 . CDS             5000  5500  .  +  2  ID=cds00003;Parent=mRNA00003
ctg123 . CDS             7000  7600  .  +  2  ID=cds00003;Parent=mRNA00003
ctg123 . CDS             3391  3902  .  +  0  ID=cds00004;Parent=mRNA00003
ctg123 . CDS             5000  5500  .  +  2  ID=cds00004;Parent=mRNA00003
ctg123 . CDS             7000  7600  .  +  2  ID=cds00004;Parent=mRNA00003
```

ClustalW Alignment Format

A common multi-sequence alignment format is the alignments written by 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
                     * *:* ** ****:** *****:**.**  *:.**********************: **
```

What About Other Formats?

- Other Phylogenetics Analysis programs:
  - PAUP, Phylip, GCG/Pileup
- Pairwise alignment output – BLAST, sim4, BLAT, GMAP
- Genefinding softwares – variety of custom formats
- Tabulated/tab delimited summary formats

Format Conversion

- There are several options:
  - stand-alone tools
  - web based tools
  - the programming option

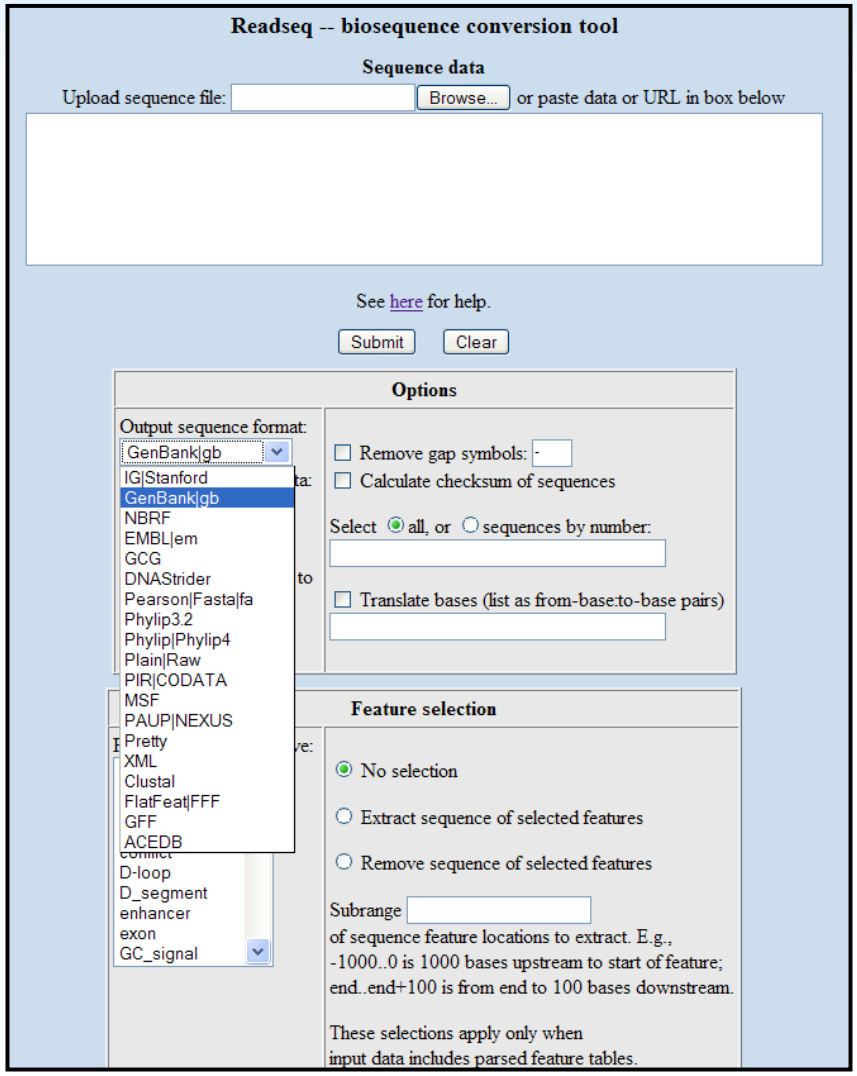
- Example: If you have to submit a sequence to genbank, you can convert the sequence to genbank format using the genbank conversion software.

25

Format Conversion Using "ReadSeq"

http://iubio.bio.indiana.edu/soft/molbio/readseq/java/

- Java based tool.
- Converts between many of the formats discussed in this lecture.
- Offered as command line interface or offered as a web based tool.



Summary

- It is important that you understand the complex and heterogeneous nature of the data.
- Familiarize yourself with commonly used data formats.
- Be aware of the Genbank, BSML and different flavors of GFF format.
- Keep in mind that there are many different custom data formats specifically for gene finding software.

27