Public Databases in Health and Life Sciences

“the potential to translate big data into big discovery”

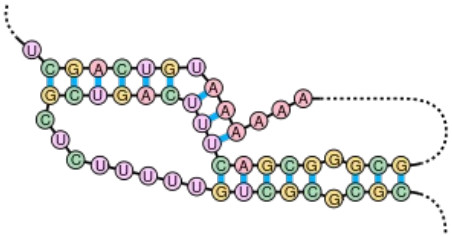
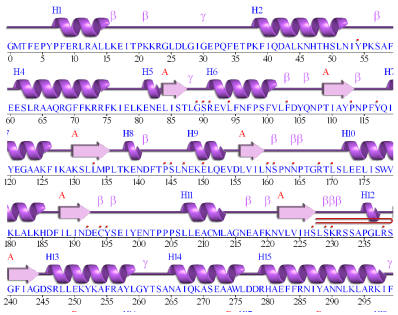
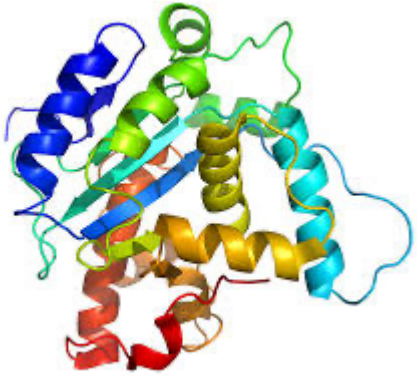
Public Databases in Health and Life Sciences

Topic 2. Biological raw data in public sequence databanks.

- Repositories for primary nucleotide sequences.
- Nucleic acid submission and accession: The GenBank and the ENA.
- Tools for online databases search and retrieval (part II).
- Models for sequence-related information
- Nucleic acids sequence formats.

Practical session #2

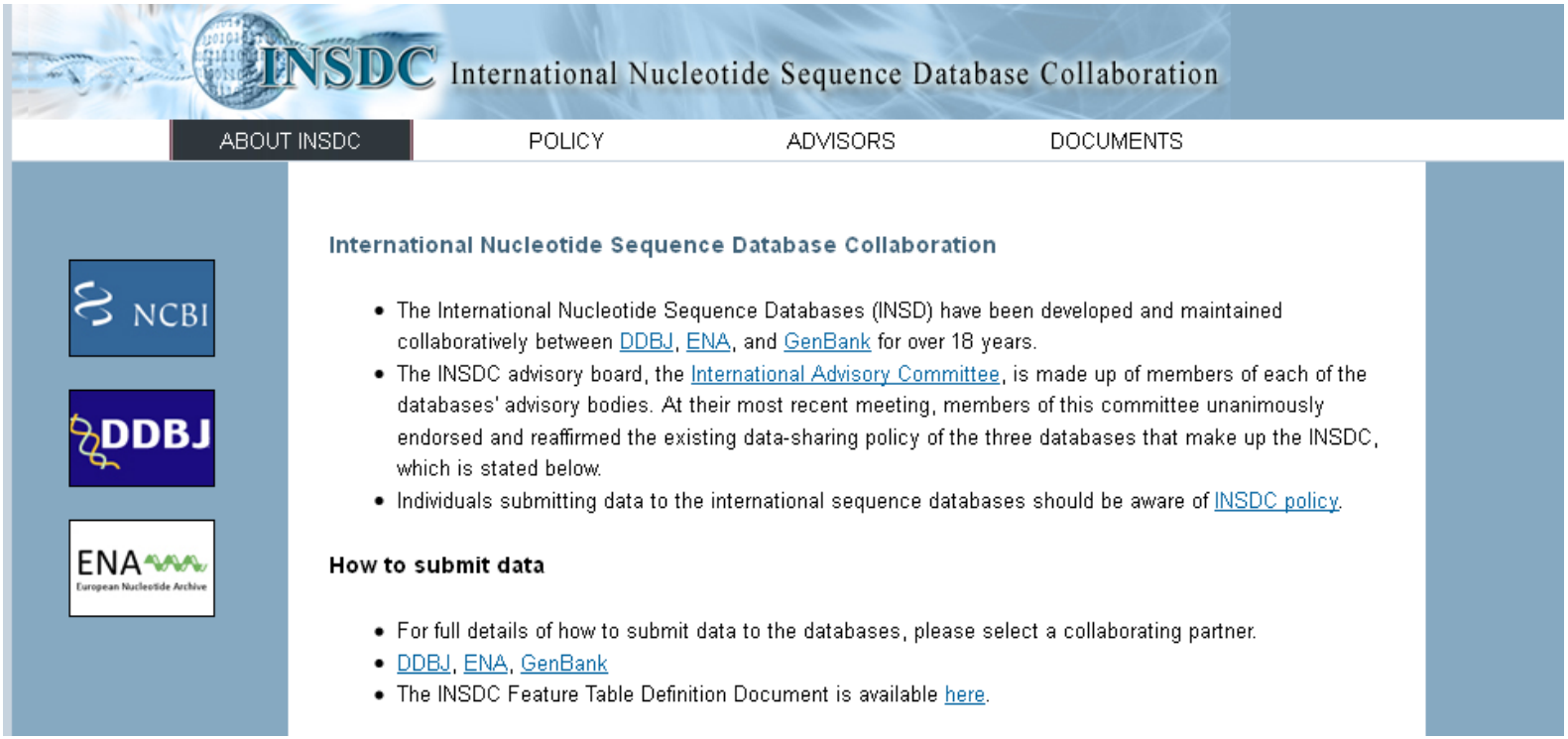
Primary structures vs. Primary databases

	Primary structure	Secondary structure	Tertiary structure
Nucleotides	>XN_003444 AACCACTTAGA CAGATAGACAG ATAGAGACAAA AGCTTT		
Proteins	>XP_003444 MATGWYALAAN CNAAAFKLAGEC NVTQALAANFMI LIANLAAY		

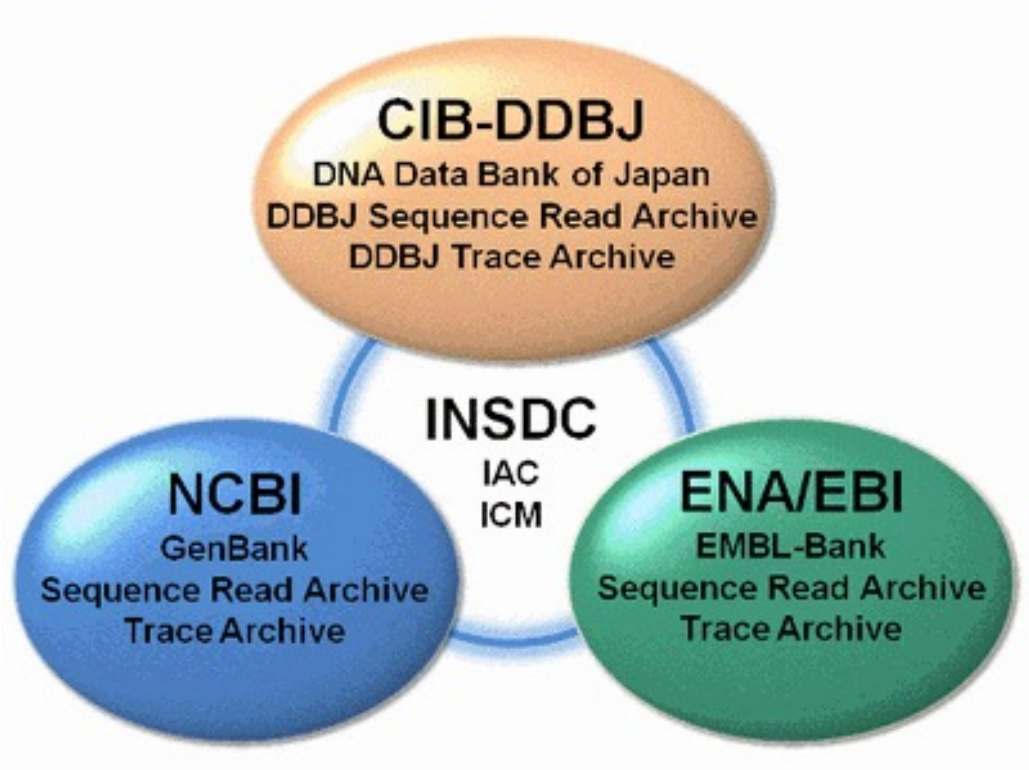
Primary (archival) nucleotide sequence databases

The International Nucleotide Sequence Database (INSDC) (http://www.insdc.org/) consists of the following databases.

DDBJ (DNA Data Bank of Japan)
EMBL/ENA (European Bioinformatics Institute)
GenBank (National Center for Biotechnology Information)



Primary (archival) nucleotide sequence databases



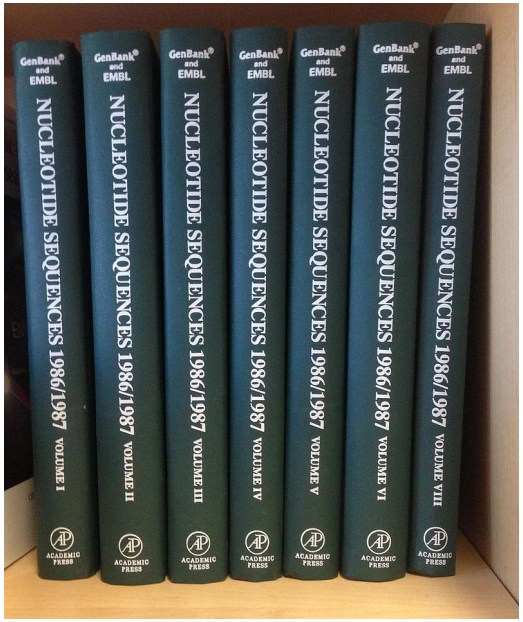
- The three databases are synchronized on a daily basis
- The accession numbers are consistent.
- There are no legal restriction in the usage of these databases. However, there are some patented sequences in the database

Sequence submission to nucleotide databases

- Direct submissions from the authors:
  - Free submissions.
  - Authors can annotate the sequences.
  - Only minor staff supervision and quality assurance checks.
- Submissions through the Internet:
  - Web forms / Web services.
  - Email.
- Sequences shared/exchanged between the 3 centers on a daily basis:
  - The sequence content of the banks is identical.

What is GenBank ?

- GenBank is the NIH (NCBI) genetic sequence database
- Nucleotide only sequence database (Beware: GenPept)
  - Example: https://www.ncbi.nlm.nih.gov/nuccore/AM743169.1
- Archival in nature
  - Historical
  - Reflective of submitter point of view (subjective)
  - Redundant
- GenBank Data
  - Direct submissions (traditional records)
  - Batch submissions (EST, GSS, STS)
  - ftp accounts (genome data)
- Three collaborating databases (all data from INSDC)
  - GenBank
  - DNA Database of Japan (DDBJ)
  - European Molecular Biology Laboratory (EMBL) Database

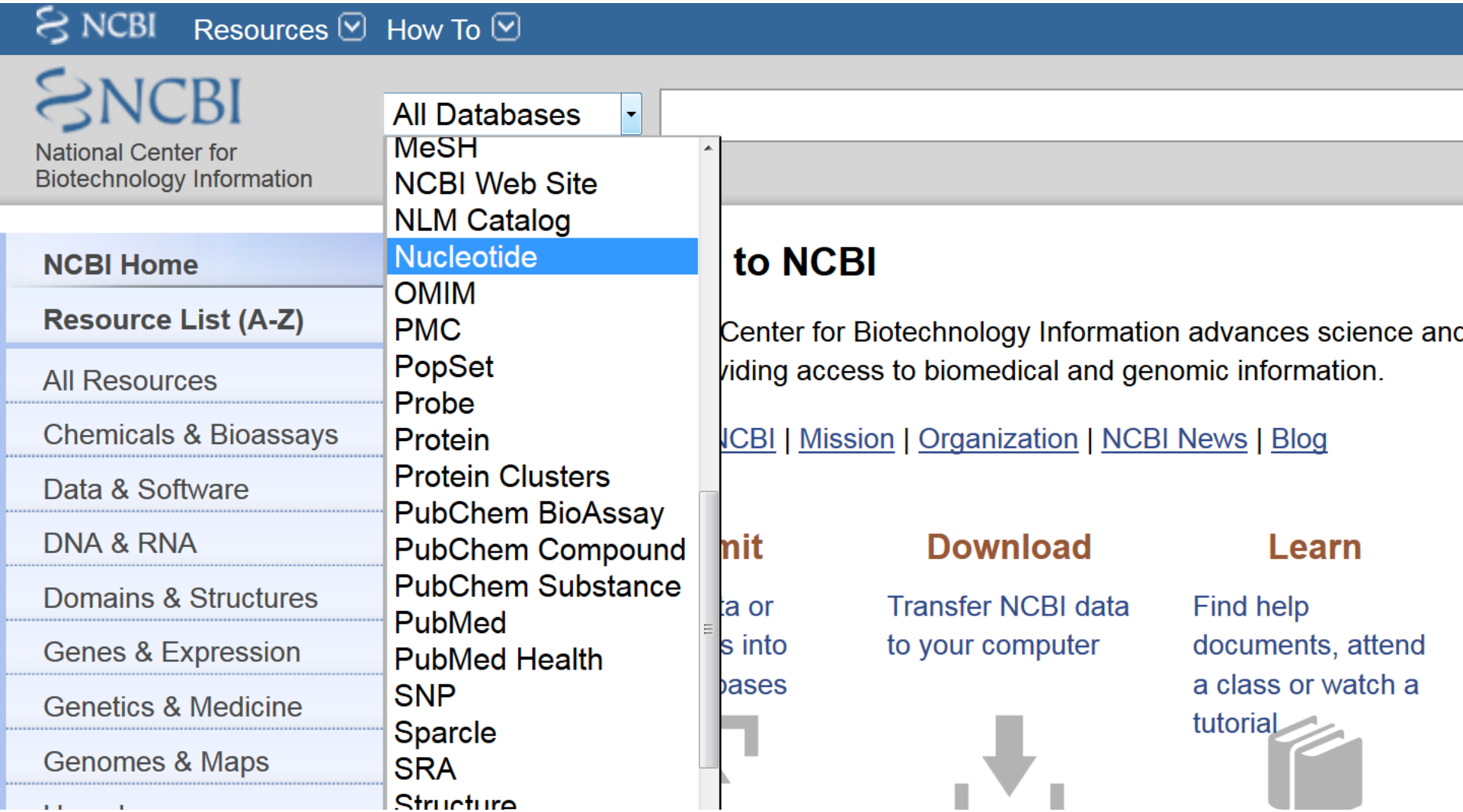

GenBank® and EMBL
NUCLEOTIDE SEQUENCES 1986/1987 VOLUME I
NUCLEOTIDE SEQUENCES 1986/1987 VOLUME II
NUCLEOTIDE SEQUENCES 1986/1987 VOLUME III
NUCLEOTIDE SEQUENCES 1986/1987 VOLUME IV
NUCLEOTIDE SEQUENCES 1986/1987 VOLUME V
NUCLEOTIDE SEQUENCES 1986/1987 VOLUME VI
NUCLEOTIDE SEQUENCES 1986/1987 VOLUME VII
ACADEMIC PRESS


GenBank

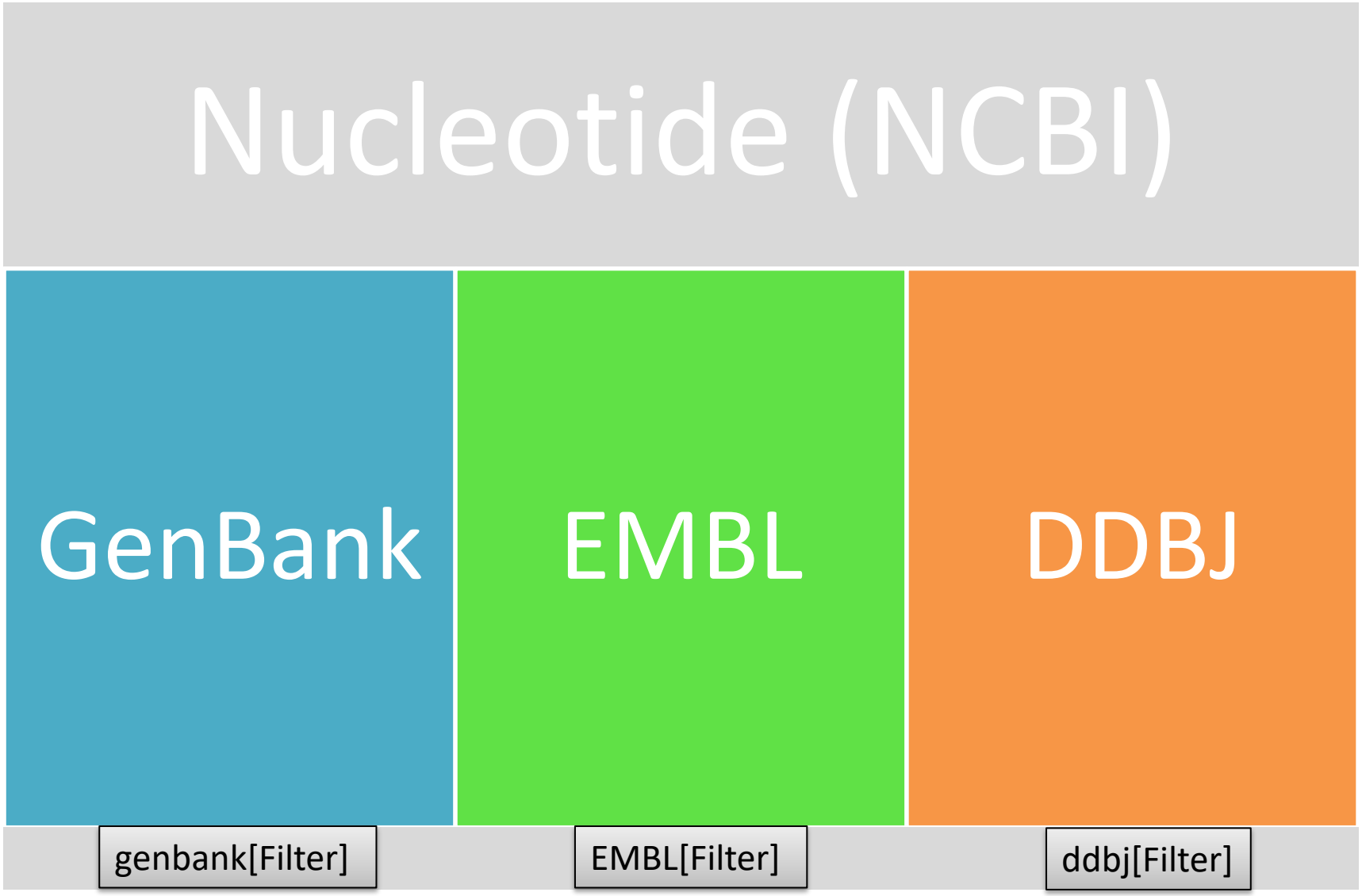

RELEASE
100.0
GenBank

The GenBank at the National Center for Biotechnology Information (NCBI)

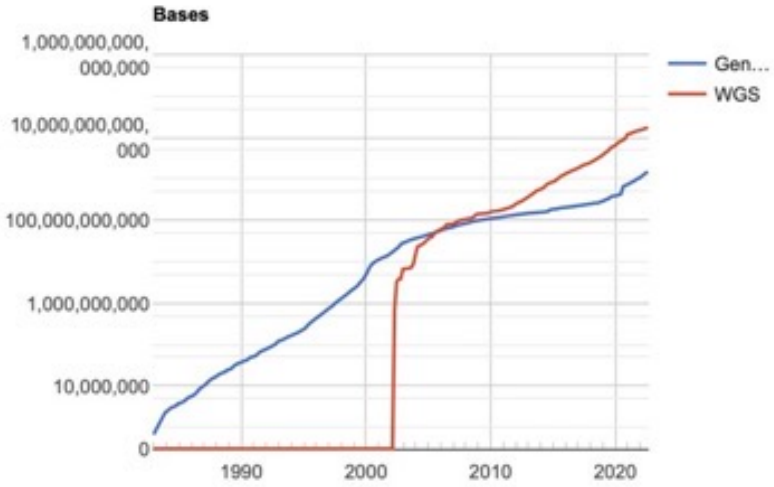
http://www.ncbi.nlm.nih.gov/nuccore

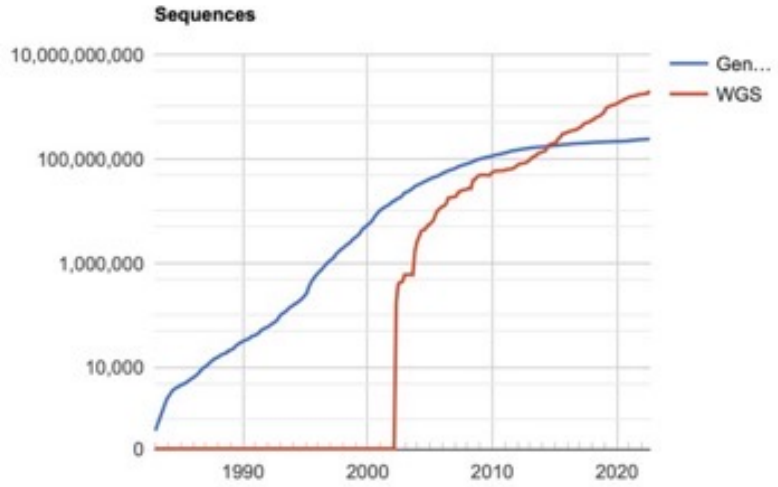


The source databases for NCBI nucleotide and protein sequences



Growth of GenBank (1982-2022)



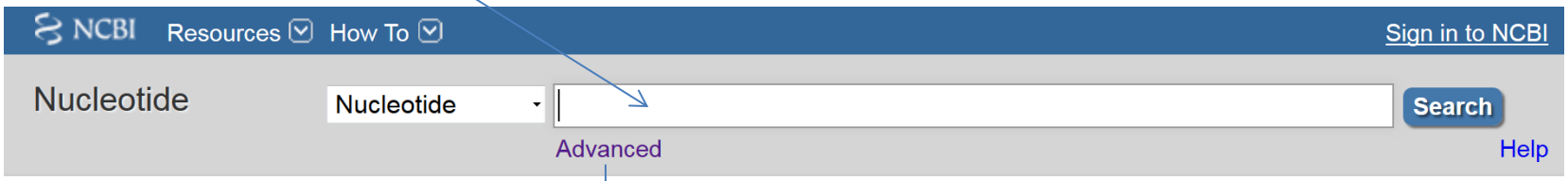


See the current GenBank release notes for up-to-date information.
http://ftp.ncbi.nih.gov/genbank/gbrel.txt
And GenBank statistics page:
https://www.ncbi.nlm.nih.gov/genbank/statistics/

There are several ways to search and retrieve data from GenBank.

Human Web interface (web based, small scale)

- Free text search



- Advanced searches (search tags and regular expressions)

- List of identifiers (Batch *Entrez*)

Search sequence databases using a sequence query

- BLAST (Basic Local Alignment Search Tool)

Web service (Programmatic data access)

- *Entrez* Utilities Web Service (NCBI): The E-utilities

File Transfer Protocol (FTP)

- Flat files (bulk data download)

ftp://ftp.ncbi.nlm.nih.gov/genbank
http://ftp.ncbi.nlm.nih.gov/genbank

GenBank organization: taxonomical divisions

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
VRL - viral sequences
PHG - bacteriophage sequences
SYN - synthetic sequences
UNA - unannotated sequences
ENV - environmental sampling sequences

gbdiv_pri[PROP]

Others

PAT - patent sequences

GenBank organization: high-throughput or functional divisions and other sub-databases

EST – Expressed Sequence Tag (1st pass single read cDNA)
GSS – Genome Survey Sequences (1st pass single read gDNA)
HTG – High Throughput Genomic (incomplete sequences of genomic clones)
STS – Sequence Tagged Site (PCR-based mapping)
HTC - High-throughput cDNA

gbdiv_htg[PROP]

TSA - Transcriptome shotgun data (transcriptome shotgun assembly projects)
WGS - Whole genome shotgun data (whole genome shotgun projects)

"wgs"[Properties]

ftp://ftp.ncbi.nih.gov/genbank/wgs/

Others

CON - Contains no sequence data, but rather instructions for the assembly of reads.

GenBank vs Nucleotide

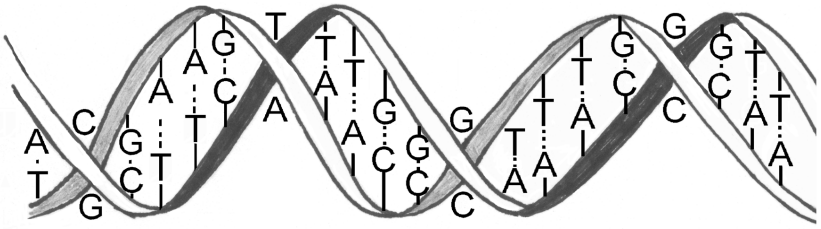
Nucleotide: a collection of sequences from several sources, including GenBank, RefSeq, TPA and PDB. Genome, gene and transcript sequence data provide the foundation for biomedical research and discovery.

GenBank: An archival database of primary nucleotide sequences that were directly sequenced by the submitter.
RefSeq: A curated, non-redundant database that includes genomic DNA, transcript (RNA), and protein products, for major organisms. The sequence data are derived from GenBank primary data, and the annotation is computational, from published literature, or from domain experts. All RefSeq ids have a prefix
TPA: A database designed to capture experimental or inferential results that support submitter-provided annotation for sequence data that the submitter did not directly determine but derived from GenBank primary data.

PDB: Repeat with me: ”PDB is not a protein database” (2UVG, 4E1U)

The DNA sequence as a line of text

agcttctttggcaaacggtt



Orientation to read

5′ agcttctttggcaaacggtt 3′ **Direct**

3′ tcgaagaaaccgtttgccaa 5′ **Complement**

Orientation to read

5′aaccgtttgccaaagaagct 3′ **Reverse complement**

It's all relative

What information is requested from authors when submitting a sequence?

Search Field Descriptions

- The sequence
  - Automatic assignment of a unique **accession number** [accession]
- Source
  - The organism [organism]
  - Type of molecule [properties]
- References
  - Authors [author]
  - Publication [journal]
- Coding Sequence (CDS) features
  - Gene [gene name]
  - Coordinates (automatic translation - protein sequence)
  - Genetic elements [feature key]
  - Protein [protein name]
  - Comments [text word]
- etc., etc.,…….

Primary key

Identifiers

Advanced Searches

Search Field Descriptions and Tags at NCBI

[Accession]	**[ACCN]**	The accession number assigned by NCBI.
[All Fields]	**[ALL]**	All terms from all search fields in the database.
[Author]	**[AU]**	All authors from all references in the records.
[EC/RN Number]	**[ECNO]**	Enzyme Commission (EC) number for an enzyme activity.
[Feature Key]	**[FKEY]**	Biological features listed in the Feature Table of the sequence records.
[Gene Name]	**[GENE]**	Gene names annotated on database records.
[Issue]	**[ISS]**	The issue number of the journals cited on sequence records
[Journal]	**[JOUR]**	The name of the journals cited on sequence records.
[Keyword]	**[KYWD]**	Keywords applied by submitter
[Modification Date]	**[MDAT]**	The date of most recent modification of a sequence record.
[Organism]	**[ORGN]**	The scientific and common names for the complete taxonomy of organisms
[Properties]	**[PROP]**	Molecular type, source database, and other properties of the sequence
[Protein Name]	**[PROT]**	The names of protein products as annotated on sequence records.
[Publication Date]	**[PDAT]**	The date that records were made public in Entrez.
[Sequence Length]	**[SLEN]**	The total length of the sequence
[Text Word]	**[WORD]**	Text on a sequence record that is not indexed in other fields.
[Title]	**[TI]**	Words and phrases found in the title of the sequence record.
ETC……..	**ETC……**	ETC…….

https://www.ncbi.nlm.nih.gov/books/NBK49540/

Nucleotide Advanced Search Builder

History deleted.

https://www.ncbi.nlm.nih.gov/books/NBK25500/

(Cytochrome-b[Gene Name]) AND human[Primary Organism]

Edit

Builder

Gene Nai | Cytochrome-b | Show index list

AND | Primary (| human | Show index list

AND | Show index list

AND | Show index list

Searc

- Accession
- All Fields
- ✓ Assembly
- Author
- BioProject
- BioSample
- Breed
- Component Accession
- Cultivar
- Division
- EC/RN Number
- Feature key
- Filter
- Gene Name
- Isolate
- Issue
- Journal
- Keyword
- Modification Date
- Organism
- Page Number
- Primary Accession
- Primary Organism
- Properties
- Protein Name
- Publication Date
- SeqID String
- Sequence Length
- Strain
- Substance Name
- Text Word
- Title
- Volume

History

There is

You are h … otide Database

GETTIN … | RCES … | POPULAR | FEATURED

NCBI Edu … | & Bioassays … | PubMed | Genetic Testing Registry

NCBI Hel … | ftware … | Bookshelf | GenBank

NCBI Har … | NA … | PubMed Central | Reference Sequences

Training & … | & Structures … | BLAST | Gene Expression Omnibus

Submit D … | Expression … | Nucleotide | Genome Data Viewer

& Medicine … | Genome | Human Genome

& Maps … | SNP | Mouse Genome

Gene | Influenza Virus

Protein | Primer-BLAST

PubChem | Sequence Read Archive

Sequence Analysis …

Entrez Programming Utilities Help [Internet]. < Prev Next >

Show details

Contents

Search this book

The E-utilities In-Depth: Parameters, Syntax and More

Eric Sayers, PhD.

Author Information

Created: May 29, 2009; Last Update: October 24, 2018.

Estimated reading time: 18 minutes

Introduction

Go to:

This chapter serves as a reference for all supported parameters for the E-utilities, along with accepted values and usage guidelines. This information is provided for each E-utility in sections below, and parameters and/or values specific to particular databases are discussed within each section. Most E-utilities have a set of parameters that are required for any call, in addition to several additional optional parameters that extend the tool's functionality. These two sets of parameters are discussed separately in each section.

General Usage Guidelines

Go to:

Please see Chapter 2 for a detailed discussion of E-utility usage policy. The following two parameters should be included in all E-utility requests.

tool

Name of application making the E-utility call. Value must be a string with no internal spaces.

email

E-mail address of the E-utility user. Value must be a string with no internal spaces, and should be a valid e-mail address.

api_key – enforced in December 2018

In December 2018, NCBI will begin enforcing the practice of using an API key for sites that post more than 3 requests per second. Please see Chapter 2 for a full discussion of this new policy.

E-utilities DTDs

Go to:

With the exception of EFetch, the E-utilities each generate a single XML output format that conforms to a DTD specific for that utility. Links to these DTDs are provided in the XML headers of the E-utility returns.

ESummary version 2.0 produces unique XML DocSums for each Entrez database, and as such each Entrez database has a unique DTD for version 2.0 DocSums. Links to these DTDs are provided in the version 2.0 XML.

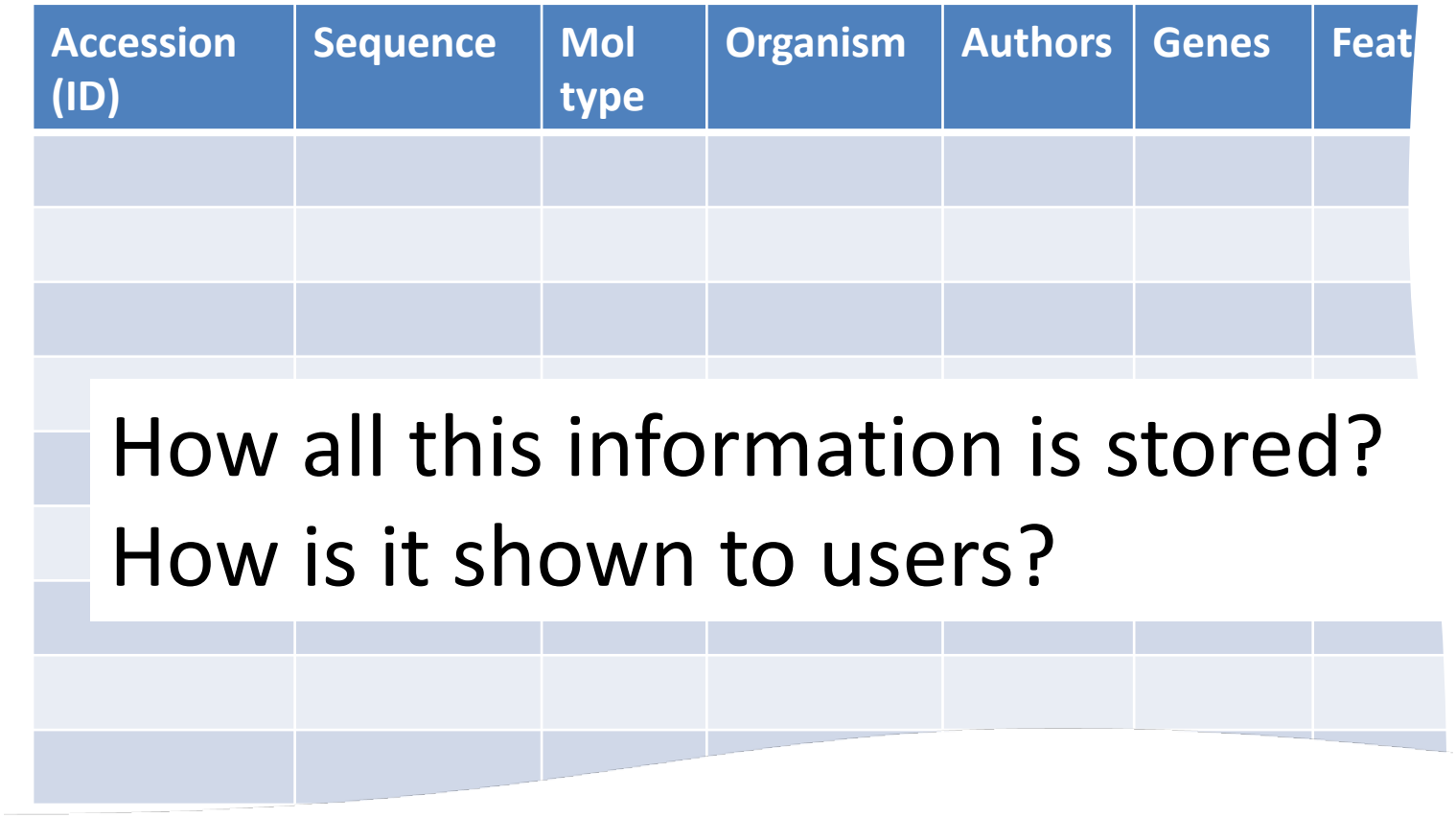
EFetch produces output in a variety of formats, some of which are XML. Most of these XML formats also conform to DTDs or schema specific to the relevant Entrez database. Please follow the appropriate link below for the PubMed DTD:

- PubMed DTD June 2018 – current PubMed DTD
- PubMed DTD January 2019 – forthcoming DTD

GenBank as any other database is a body of information stored in two dimensions (rows and columns) = Tables

Fields (identifiers) in columns

Entries (sequences) in rows



ASN.1 Send to:

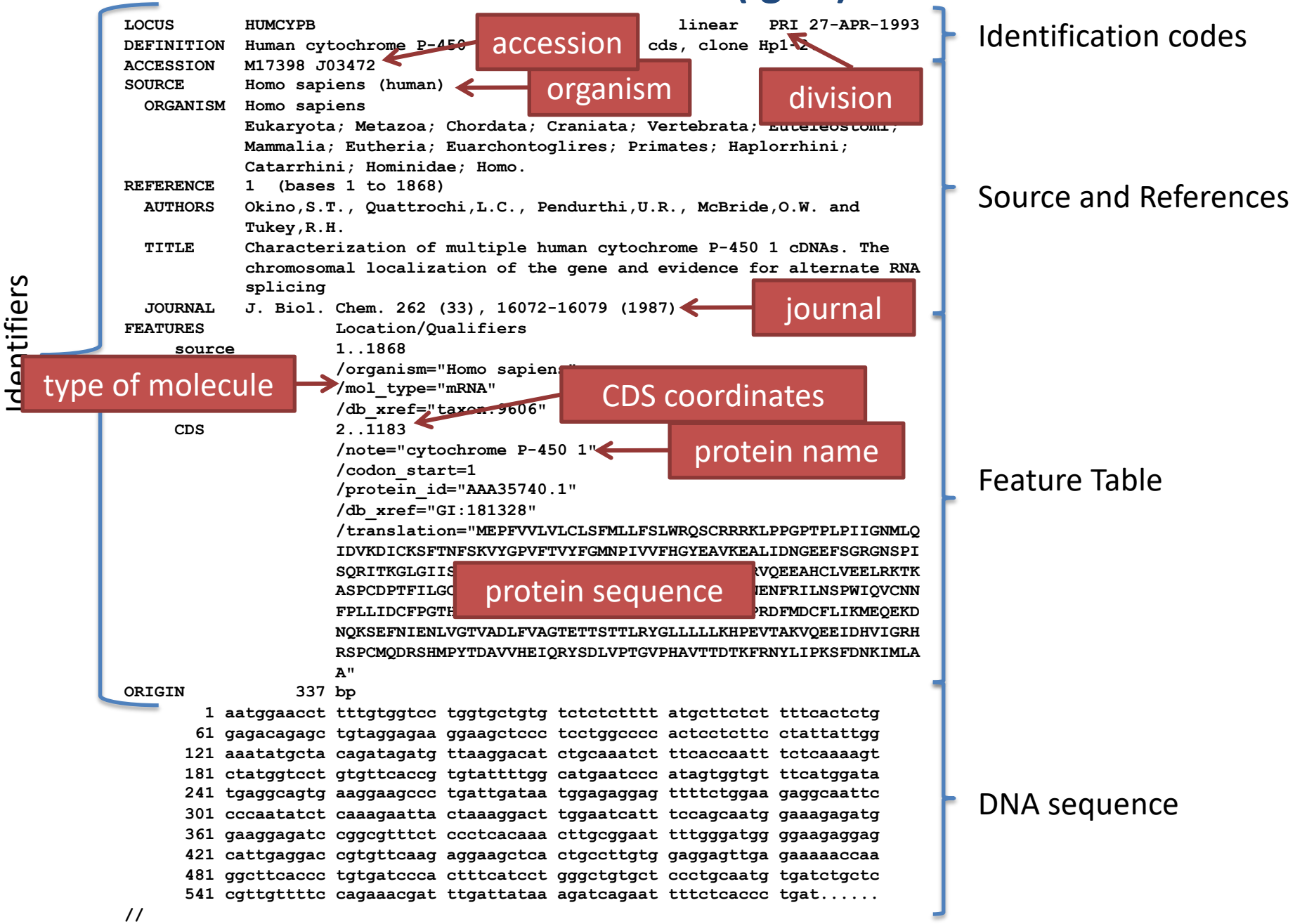
Human mitochondrial cytochrome b gene, partial cds

GenBank: M28016.1

GenBank FASTA Graphics

```
Seq-entry ::= set {
  level 1,
  class nuc-prot,
  descr {
    source {
      genome mitochondrion,
      org {
        taxname "Homo sapiens",
        common "human",
        db {
          {
            db "taxon",
            tag id 9606
          }
        },
        orgname {
          name binomial {
            genus "Homo",
            species "sapiens"
          },
          lineage "Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
 Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
 Catarrhini; Hominidae; Homo",
          gcode 1,
          mgcode 2,
          div "PRI"
        }
      },
      subtype {
        {
          subtype cell-line,
          name "Maja"
        },
        {
          subtype cell-type,
          name "lymphoblastoid"
        },
        {
          subtype tissue-lib,
          name "Provided by Drs John Trowsdale and Janet Lee"
        }
      }
    },
    pub {
      pub {
        gen {
          serial-number 1
        },
        muid 86064879,
        article {
          title {
            name "Serendipitous cloning of a mitochondrial cDNA and its
 polymorphism."
          },
          authors {
            names std {
              {
                name name {
                  last "Spurr",
                  initials "N.K."
                }
              },
              {
                name name {
                  last "Bodmer",
                  initials "W.F."
                }
              }
            }
          },
          from journal {
            title {
              iso-jta "Mol. Biol. Med.",
              ml-jta "Mol Biol Med",
              issn "0735-1313",
              name "Molecular biology & medicine."
            },
            imp {
              date std {
                year 1984,
                month 8
              },
              volume "2",
              issue "4",
              pages "239-249",
              language "eng"
            }
          },
          ids {
            pubmed 6100559,
            medline 86064879
          }
        },
        pmid 6100559
      },
      comment "simple staff_entry"
    },
    create-date std {
      year 1993,
      month 8,
      day 3
    },
    update-date std {
      year 1992,
      month 3,
      day 19
    }
  },
  seq-set {
    seq {
      id {
        genbank {
          name "HUMMTCYTB",
          accession "M28016",
          version 1
        },
        gi 337203
      },
      descr {
        title "Human mitochondrial cytochrome b gene, partial cds.",
        molinfo {
          biomol mRNA
        },
        genbank {
          source "Mitochondrion Homo sapiens (organelle Mitochondrion Homo
 sapiens) (library: Provided by Drs John Trowsdale and Janet Lee) cDNA to mRNA.",
          keywords {
            "cytochrome b"
          },
          entry-date std {
            year 1992,
            month 3,
            day 19
          }
        }
      },
      inst {
        repr raw,
        mol rna,
        length 337,
        strand ss,
        seq-data ncbi2na '095838CFD73D97110F758D6D5C101CA29B5F95CF1CD4D74D7243
0D54D75333501040930CFD951C250D1FCF875C96485D74F705E0DA28414B09C57FC534FA10B24D
6C731F4410D70D70C50737570F80100C740EA5C0'H
      },
      annot {
        {
          data ftable {
            {
              data gene {
                locus "cytochrome b"
              },
              location int {
                from 0,
                to 336,
                id gi 337203
              }
            }
          }
        }
      },
      seq {
        id {
          genbank {
            accession "AAA31851",
            version 1
          },
          gi 552606
        },
        descr {
          title "cytochrome b, partial (mitochondrion) [Homo sapiens]",
          molinfo {
            biomol peptide,
            tech concept-trans,
            completeness no-ends
          }
        },
        inst {
          repr raw,
          mol aa,
          length 112,
          seq-data ncbieaa "KPEWYFLFAYTILRSVPNKLGGVLALLLSILILAMIPILHMSKQQSMMFRPL
SQSLYWLLAADLLILTWIGGQPVSYPFTIIGQVASVLYFTTILILMPTISLIENKMLKWA",
          hist {
            replaces {
              date std {
                year 1994,
                month 10,
                day 3
              },
              ids {
                gi 337204
              }
            }
          }
        },
        annot {
          {
            data ftable {
              {
                data prot {
                  name {
                    "cytochrome b"
                  }
                },
                partial TRUE,
                location int {
                  from 0,
                  to 111,
                  id gi 552606,
                  fuzz-from lim lt,
                  fuzz-to lim gt
                }
              }
            }
          }
        }
      }
    },
    annot {
      {
        data ftable {
          {
            data cdregion {
              frame one,
              code {
                id 2
              }
            },
            partial TRUE,
            product whole gi 552606,
            location int {
              from 0,
              to 336,
              id gi 337203,
              fuzz-from lim lt,
              fuzz-to lim gt
            }
          }
        }
      }
    }
  }
}
```

GenBank Flat Files (.gbff)



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

FASTA file

```
>M17398.1 Human cytochrome P-450 1 mRNA, complete cds, clone Hp1-2
AATGGAACCTTTTGTGGTCCTGGTGCTGTGTCTCTCTTTTATGCTTCTCTTTTCACTCTGGAGACAGAGC
TGTAGGAGAAGGAAGCTCCCTCCTGGCCCCACTCCTCTTCCTATTATTGGAAATATGCTACAGATAGATG
TTAAGGACATCTGCAAATCTTTCACCAATTTCTCAAAAGTCTATGGTCCTGTGTTCACCGTGTATTTTGG
CATGAATCCCATAGTGGTGTTTCATGGATATGAGGCAGTGAAGGAAGCCCTGATTGATAATGGAGAGGAG
TTTTCTGGAAGAGGCAATTCCCCAATATCTCAAAGAATTACTAAAGGACTTGGAATCATTTCCAGCAATG
GAAAGAGATGGAAGGAGATCCGGCGTTTCTCCCTCACAAACTTGCGGAATTTTGGGATGGGGAAGAGGAG
CATTGAGGACCGTGTTCAAGAGGAAGCTCACTGCCTTGTGGAGGAGTTGAGAAAAACCAAGGCTTCACCC
TGTGATCCCACTTTCATCCTGGGCTGTGCTCCCTGCAATGTGATCTGCTCCGTTGTTTTCCAGAAACGAT
TTGATTATAAAGATCAGAATTTTCTCACCCTGATGAAAAGATTCAATGAAAACTTCAGGATTCTGAACTC
CCCATGGATCCAGGTCTGCAATAATTTCCCTCTACTCATTGATTGTTTCCCAGGAACTCACAACAAAGTG
CTTAAAAATGTTGCTCTTACACGAAGTTACATTAGGGAGAAAGTAAAAGAACACCAAGCATCACTGGATG
TTAACAATCCTCGGGACTTTATGGATTGCTTCCTGATCAAAATGGAGCAGGAAAAGGACAACCAAAAGTC
AGAATTCAATATTGAAAACTTGGTTGGCACTGTAGCTGATCTATTTGTTGCTGGAACAGAGACAACAAGC
ACCACTCTGAGATATGGACTCCTGCTCCTGCTGAAGCACCCAGAGGTCACAGCTAAAGTCCAGGAAGAGA
TTGATCATGTAATTGGCAGACACAGGAGCCCCTGCATGCAGGATAGGAGCCACATGCCTTACACTGATGC
TGTAGTGCACGAGATCCAGAGATACAGTGACCTTGTCCCCACCGGTGTGCCCCATGCAGTGACCACTGAT
ACTAAGTTCAGAAACTACCTCATCCCCAAGAGCTTTGATAACAAGATAATGCTGGCTGCATAAAACTAGG
GCACAACCATAATGGCATTACTGACTTCCGTGCTACATGATGACAAAGAATTTCCTAATCCAAATATCTT
TGACCCTGGCCACTTTCTAGATAAGAATGGCAACTTTAAGAAAAGTGACTACTTCATGCCTTTCTCAGCA
GGAAAACGAATTTGTGCAGGAGAAGGACTTGCCCGCATGGAGCTATTTTTATTTCTAACCACAATTTTAC
AGAACTTTAACCTGAAATCTGTTGATGATTTAAAGAACCTCAATACTACTGCAGTTACCAAAGGGATTGT
TTCTCTGCCACCCTCATACCAGATCTGCTTCATCCCTGTCTGAAGAATGCTAGCCCATCTGGCTGCTGAT
CTGCTATCACCTGCAACTCTTTTTTTATCAAGGACATTCCCACTATTATGTCTTCTCTGACCTCTCATCA
AATCTTCCCATTCACTCAATATCCCATAAGCATCCAAACTCCATTAAGGAGAGTTGTTCAGGTCACTGCA
CAAATATATCTGCAATTATTCATACTCTGTAACACTTGTATTAATTGCTGCATATGCTAATACTTTTCTA
ATGCTGACTTTTTAATATGTTATCACTGTAAAACACAGAAAAGTGATTAATGAATGATAATTTAGTCCAT
TTCTTTTGTGAATGTGCTAAATAAAAAGTGTTATTAATTGCTGGTTCA
```

The FASTA format

• FASTA is the name of a popular sequence alignment-and-database-scanning program created by W.R. Pearson and D.J. Lipman in 1988 [PNAS, 85:2444-2448].

• The sequences used by FASTA have to obey the following format:

> (mandatory) definition line (My_identifiers)
ARCGTCRGCKINTANDRGCKINTAND
CKINTANDARCGTCRGCKINTANDRG
CKINTAND

• The line starting with > (the definition line) should contains a short definition or identifiers. The lines that follow it contain the DNA or protein sequence (in one-letter code). Usually, this line has a specific format decided by each database

• Sequences are expected to be represented in the standard IUB/IUPAC amino acid and nucleic acid codes. No numbers allowed! (blank characters are ignored)

• Because FASTA is easy to parse, It is now the default input format for much sequence analysis software.

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

FASTA file (ENA)

```
>ENA|M17398|M17398.1 Human cytochrome P-450 1 mRNA, complete cds, clone Hp1-2.
AATGGAACCTTTTGTGGTCCTGGTGCTGTGTCTCTCTTTTATGCTTCTCTTTTCACTCTG
GAGACAGAGCTGTAGGAGAAGGAAGCTCCCTCCTGGCCCCACTCCTCTTCCTATTATTGG
AAATATGCTACAGATAGATGTTAAGGACATCTGCAAATCTTTCACCAATTTCTCAAAAGT
CTATGGTCCTGTGTTCACCGTGTATTTTGGCATGAATCCCATAGTGGTGTTTCATGGATA
TGAGGCAGTGAAGGAAGCCCTGATTGATAATGGAGAGGAGTTTTCTGGAAGAGGCAATTC
CCCAATATCTCAAAGAATTACTAAAGGACTTGGAATCATTTCCAGCAATGGAAAGAGATG
GAAGGAGATCCGGCGTTTCTCCCTCACAAACTTGCGGAATTTTGGGATGGGGAAGAGGAG
CATTGAGGACCGTGTTCAAGAGGAAGCTCACTGCCTTGTGGAGGAGTTGAGAAAAACCAA
GGCTTCACCCTGTGATCCCACTTTCATCCTGGGCTGTGCTCCCTGCAATGTGATCTGCTC
CGTTGTTTTCCAGAAACGATTTGATTATAAAGATCAGAATTTTCTCACCCTGATGAAAAG
ATTCAATGAAAACTTCAGGATTCTGAACTCCCCATGGATCCAGGTCTGCAATAATTTCCC
TCTACTCATTGATTGTTTCCCAGGAACTCACAACAAAGTGCTTAAAAATGTTGCTCTTAC
ACGAAGTTACATTAGGGAGAAAGTAAAAGAACACCAAGCATCACTGGATGTTAACAATCC
TCGGGACTTTATGGATTGCTTCCTGATCAAAATGGAGCAGGAAAAGGACAACCAAAAGTC
AGAATTCAATATTGAAAACTTGGTTGGCACTGTAGCTGATCTATTTGTTGCTGGAACAGA
GACAACAAGCACCACTCTGAGATATGGACTCCTGCTCCTGCTGAAGCACCCAGAGGTCAC
AGCTAAAGTCCAGGAAGAGATTGATCATGTAATTGGCAGACACAGGAGCCCCTGCATGCA
GGATAGGAGCCACATGCCTTACACTGATGCTGTAGTGCACGAGATCCAGAGATACAGTGA
CCTTGTCCCCACCGGTGTGCCCCATGCAGTGACCACTGATACTAAGTTCAGAAACTACCT
CATCCCCAAGAGCTTTGATAACAAGATAATGCTGGCTGCATAAAACTAGGGCACAACCAT
AATGGCATTACTGACTTCCGTGCTACATGATGACAAAGAATTTCCTAATCCAAATATCTT
TGACCCTGGCCACTTTCTAGATAAGAATGGCAACTTTAAGAAAAGTGACTACTTCATGCC
TTTCTCAGCAGGAAAACGAATTTGTGCAGGAGAAGGACTTGCCCGCATGGAGCTATTTTT
ATTTCTAACCACAATTTTACAGAACTTTAACCTGAAATCTGTTGATGATTTAAAGAACCT
CAATACTACTGCAGTTACCAAAGGGATTGTTTCTCTGCCACCCTCATACCAGATCTGCTT
CATCCCTGTCTGAAGAATGCTAGCCCATCTGGCTGCTGATCTGCTATCACCTGCAACTCT
TTTTTTATCAAGGACATTCCCACTATTATGTCTTCTCTGACCTCTCATCAAATCTTCCCA
TTCACTCAATATCCCATAAGCATCCAAACTCCATTAAGGAGAGTTGTTCAGGTCACTGCA
CAAATATATCTGCAATTATTCATACTCTGTAACACTTGTATTAATTGCTGCATATGCTAA
TACTTTTCTAATGCTGACTTTTTAATATGTTATCACTGTAAAACACAGAAAAGTGATTAA
TGAATGATAATTTAGTCCATTTCTTTTGTGAATGTGCTAAATAAAAAGTGTTATTAATTG
CTGGTTCA
```

Used for both nucleotides and proteins

```
> My nucleotide sequence
AATGGAACCTTTTGTGGTCCTGGTGCTGTGTCTCTCTTTTATGCTTCTCTTTTCACTCTGGAGACAGA
GCTGTAGGAGAAGGAAGCTCCCTCCTGGCCCCACTCCTCTTCCTATTATTGGAAATATGCTACAGATA
GATGTTAAGGACATCTGCAAATCTTTCACCAATTTCTCAAAAGTCTATGGTCCTGTGTTCACCGTGTA
TTTTGGCATGAATCCCATAGTGGTGTTTCATGGATATGAGGCAGTGAAGGAAGCCCTGATTGATAATG
GAGAGGAGTTTTCTGGAAGAGGCAATTCCCCAATATCTCAAAGAATTACTAAAGGACTTGGAATCATT
TCCAGCAATGGAAAGAGATGGAAGGAGATCCGGCGTTTCTCCCTCACAAACTTGCGGAATTTTGGGAT
GGGGAAGAGGAGCATTGAGGACCGTGTTCAAGAGGAAGCTCACTGCCTTGTGGAGGAGTTGAGAAAAA
CCAAGGCTTCACCCTGTGATCCCACTTTCATCCTGGGCTGTGCTCCCTGCAATGTGATCTGCTCCGTT
GTTTTCCAGAAACGATTTGATTATAAAGATCAGAATTTTCTCACCCTGATGAAAAGATTCAATGAAAA
CTTCAGGATTCTGAACTCCCCATGGATCCAGGTCTGCAATAATTTCCCTCTACTCATTGATTGTTTCC
CAGGAACTCACAACAAAGTGCTTAAAAATGTTGCTCTTACACGAAGTTACATTAGGGAGAAAGTAAAA
GAACACCAAGCATCACTGGATGTTAACAATCCTCGGGACTTTATGGATTGCTTCCTGATCAAAATGGA
GCAGGAAAAGGACAACCAAAAGTC

> My protein sequence
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTN
LVEWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIK
DFLGLLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALF
LSIVILGLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIIL
AFLPIAGXIENY
```

Sequence formats

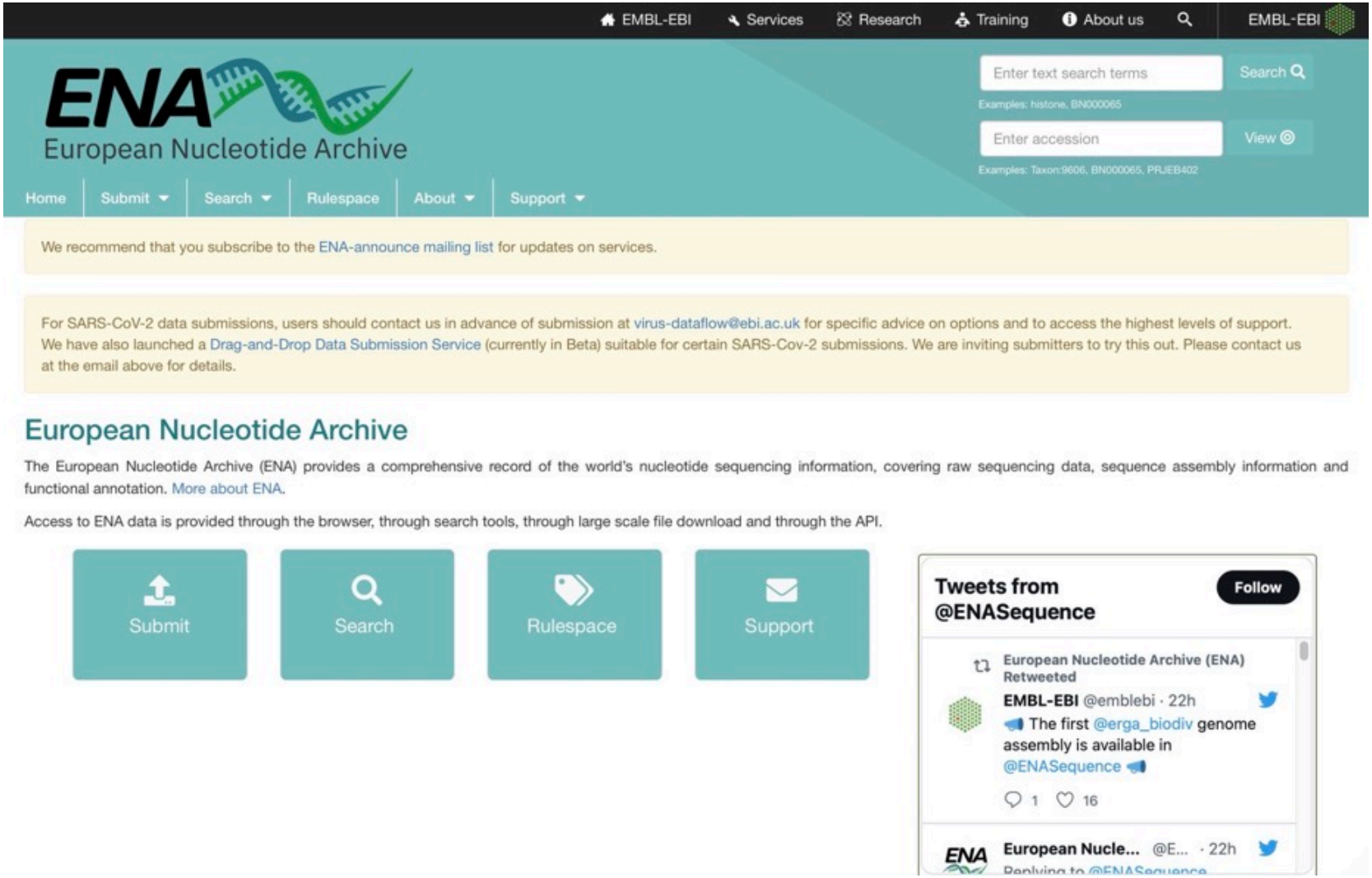
ASN.1
DNAStrider
EMBL
Fitch
GCG
GenBank/GB
IG/Stanford
MSF
NBRF
Olsen
PAUP/NEXUS
Pearson/Fasta
Phylip
PIR/CODATA
Plain/Raw
Pretty
Zuker

NOTE:

- FASTA is a popular sequence format

The EMBL-EBI nucleotide repository
ENA

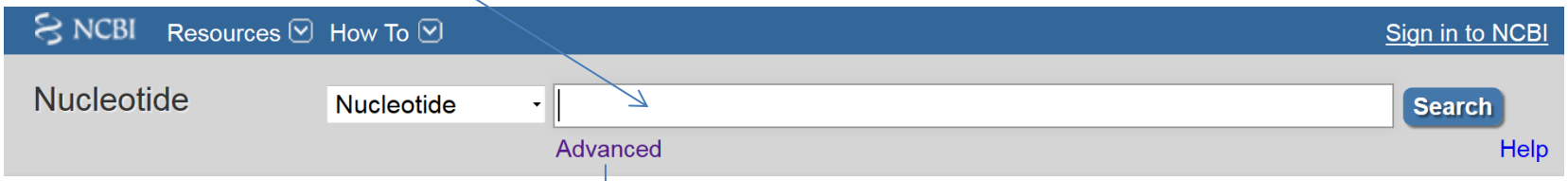
http://www.ebi.ac.uk/ena/



There are several ways to search and retrieve data from GenBank.

Human Web interface (web based, small scale)

- Free text search



- Advanced searches (search tags and regular expressions)

- List of identifiers (Batch *Entrez*)

Search sequence databases using a sequence query

- BLAST (Basic Local Alignment Search Tool)

Web service (Programmatic data access)

- *Entrez* Utilities Web Service (NCBI): The E-utilities

File Transfer Protocol (FTP)

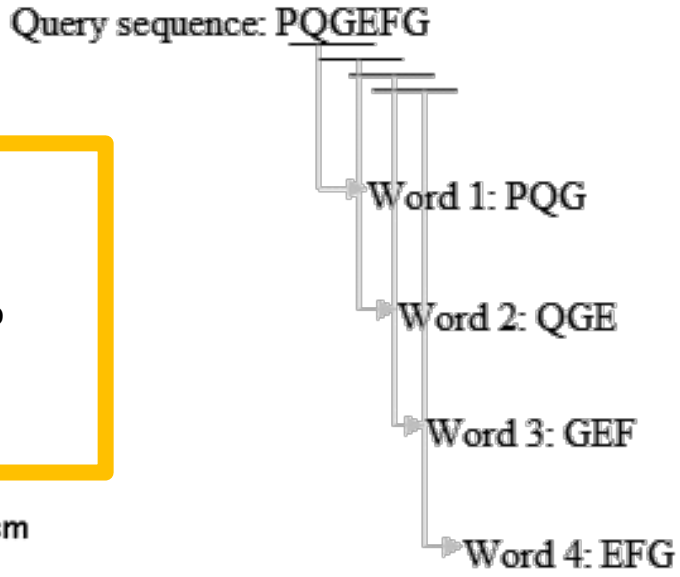
- Flat files (bulk data download)

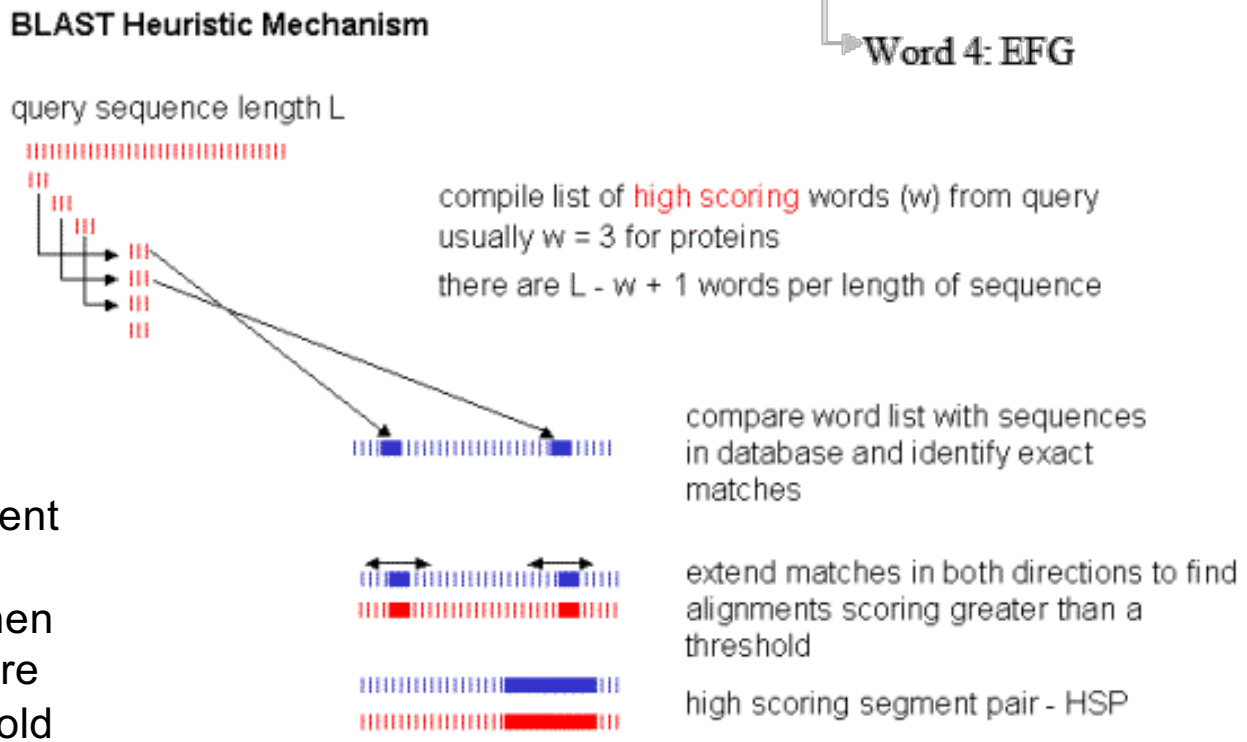
ftp://ftp.ncbi.nlm.nih.gov/genbank
http://ftp.ncbi.nlm.nih.gov/genbank

How does BLAST work?

Generate a list of k-words

Not all of them, “low-complexity” regions (pe:PPCDPPPPPKDKKKKDDGPP) are not used to generate k-words





- Search for these ”words"

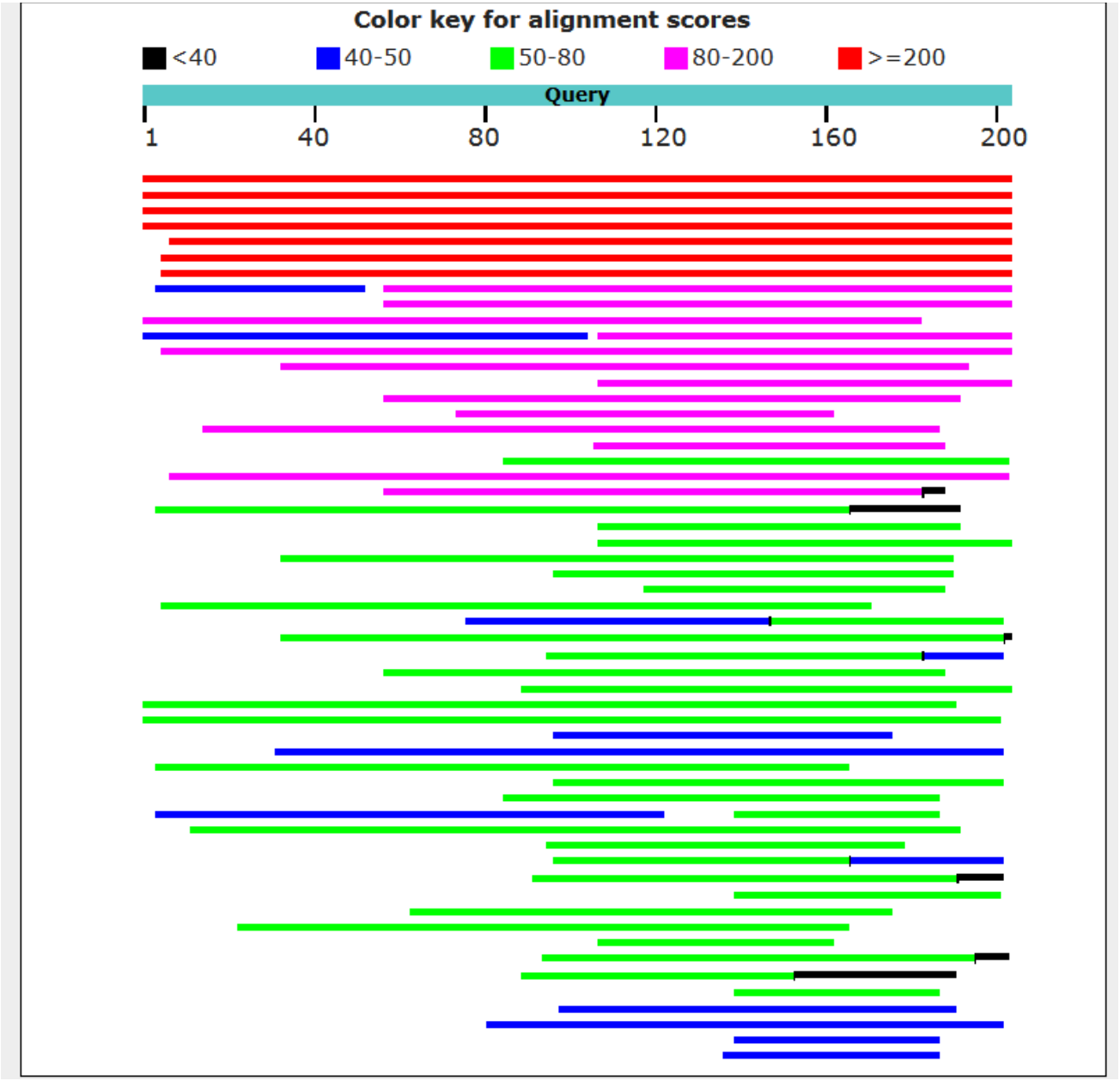
- Extend the alignment

- Stop extension when the acumulated score falls below a threshold (*dropoff value*)

BLOSUM 62

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

The BLAST output: The NCBI graphical format



The BLAST output

- Produces local alignments: only a portion of each sequence could be aligned.
- Calculates similarity for biological sequences.
  - Score, % identity and % similarity
- Uses statistical theory to determine if a match might have occurred by chance. (Expected value (*E*))

PREDICTED: putative uncharacterized protein FLJ44672 [Gorilla gorilla gorilla]

Sequence ID: XP_004045538.1 **Length:** 213 **Number of Matches:** 1

Range 1: 1 to 213 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
281 bits(720)	7e-94	Compositional matrix adjust.	180/213(85%)	183/213(85%)	10/213(4%)

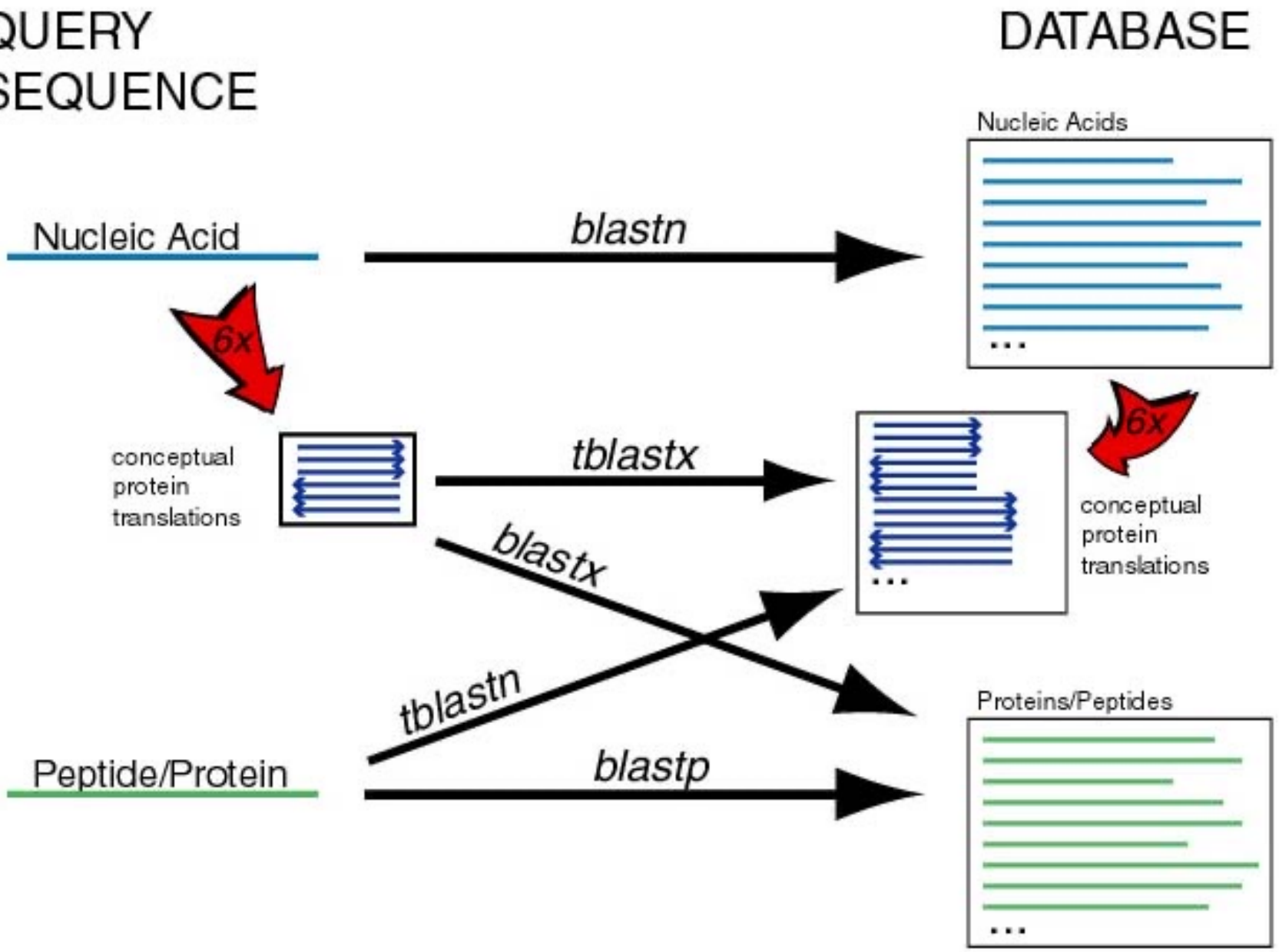
```
Query  1    MPLLASPGPAPACWRPVEAQPLPQQWALHAHLLPRRGLLGPGSRLGAAPAGPAPASR-PS  59
            MPLLASPGPAPACWRP+EAQPLPQQWALHAHL PRRGLLGPGSR+GAAP GPAPASR PS
Sbjct  1    MPLLASPGPAPACWRPLEAQPLPQQWALHAHLSPRRGLLGPGSRVGAAPTGPAPASRRPS  60

Query  60   GGQAHASGRPLPAWRLLLCMGSRSCTSSGRPLQTQL---------FLPAASVGPDCRQVG  110
            GGQAHASGRPLPAWRLLLCMGSR CTSS RPLQ Q           FLPAAS GPDCRQVG
Sbjct  61   GGQAHASGRPLPAWRLLLCMGSRICTSSSRPLQAQQAQQALQAQLFLPAASAGPDCRQVG  120

Query  111  LSRDSSCLPAASVGPDCHQVGLSMDSSCLPVTSVGPSRPQVGLPRPSSGLSAASPSAKVP  170
            LSRDSSCLPAAS GPDC QVGLS DSSCLP  SVGPSRPQVGLPRPSSGLSAAS SAKVP
Sbjct  121  LSRDSSCLPAASAGPDCRQVGLSRDSSCLPSASVGPSRPQVGLPRPSSGLSAASSSAKVP  180

Query  171  RVSLSRPRSSCLPVASFGPAQFMPPGGLPRPHF  203
            RV LSR  SSCLPVASF PAQ +PPGGL RP F
Sbjct  181  RVRLSRSSSSCLPVASFSPAQLIPPGGLSRPRF  213
```

BLAST Basic programs

Web BLAST https://blast.ncbi.nlm.nih.gov/Blast.cgi

nucleotide blast	Search a nucleotide database using a nucleotide query
protein blast	Search protein database using a protein query
blastx	Search protein database using a translated nucleotide query
tblastn	Search translated nucleotide database using a protein query
tblastx	Search translated nucleotide database using a translated nucleotide query

The BLAST family of programs



BLAST databases:

Nucl. Database	Content
nt (default)	All GenBank + EMBL + DDBJ + PDB sequences, excluding sequences from PAT, EST, STS, GSS, WGS, TSA and phase 0, 1 or 2 HTGS sequences. Non-redundant, records with identical sequences collapsed into a single entry.
rRNA/ITS data-bases	A collection of four databases: a 16S Microbial rRNA sequences from NCBI's Targeted Loci Projects, an 18S and a 26S RNA rRNA dataabses for fungi, plus an ITS database for fungi.
refseq_rna	Curated (NM_, NR_) plus predicted (XM_, XR_) sequences from NCBI Reference Sequence Project.
refseq_ representative_ genomes	NCBI RefSeq Reference and Representative genomes across broad taxonomy groups including eukaryotes, bacteria, archaea, viruses and viroids. These genomes are among the best quality genomes available with minimum redundancy - one genome per species for eukaryotes and diverse isolates for the same species for others.
Refseq genome	This database contains NCBI RefSeq genomes across all taxonomy groups. It contains only the top-level sequences, i.e. the longest sequences representing any given part of the genomes, to reduce redundancy.
wgs	Assemblies of Whole Genome Shotgun sequences.
est	Database of GenBank + EMBL + DDBJ sequences from EST division
Human G+T	The genomic sequences plus curated and predicted RNAs from the current build of the human genome.
Mouse G+T	The genomic sequences plus curated and predicted RNAs from the current build of the mouse genome.
est	Database of GenBank + EMBL + DDBJ sequences from EST division
TSA	Transcriptome Shotgun Assemblies, assembled from RNA-seq SRA data
SRA	Nextgen sequences from NCBI's Sequence Read Archive (SRA), limit to specific subset required.
HTGS	Unfinished High Throughput Genomic Sequences; Sequences: phases 0, 1 and 2.
pat	Nucleotides from the Patent division of GenBank.
pdb	Nucleotide sequences from the 3-dimensional structure records from Protein Data Bank.
refseq_genomic	All genomic sequences from NCBI Reference Sequence Project, highly redundant.
Prot. Database	**Content**
nr (default)	Non-redundant GenBank CDS translations + RefSeq + PDB + SwissProt + PIR + PRF, excluding those in PAT, TSA, and env_nr.
refseq_protein	Protein sequences from NCBI Reference Sequence project.
Landmark	The landmark database includes proteomes from representative genomes spanning a wide taxonomic range
swissprot	Last major release of the UniProtKB/SWISS-PROT protein sequence database (no incremental updates).
pat	Proteins from the Patent division of GenBank.
pdb	Protein sequences from the 3-dimensional structure records from the Protein Data Bank.
env_nr	Protein sequences translated from the CDS annotation of metagenomic nucleotide sequences.
tsa_nr	Protein sequences translated from CDSs annotated on transcriptome shotgun assemblies.

https://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo_BLASTGuide.pdf

NCBI Datasets

https://www.ncbi.nlm.nih.gov/datasets/docs/v2/getting_started/

NCBI Datasets is a new resource that lets you easily gather data from across NCBI databases. You have the choice of getting the data through three interfaces:

- NCBI Datasets website
- Command-line tools
- API : Accessible with UNIX tools (such as wget and curl).