**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

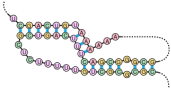
• Nucleic acids sequence formats.

**Practical session #2**

**Nucleotides Proteins**

**Primary structures vs. Primary databases**

**Primary structure Secondary structure Tertiary structure**

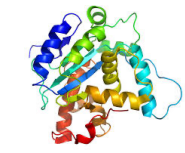
>XN\_003444 

AACCACTTAGA

CAGATAGACAG

ATAGAGACAAA

AGCTTT

>XP\_003444 

MATGWYALAAN

CNAAAFKLAGEC

NVTQALAANFMI

LIANLAAY

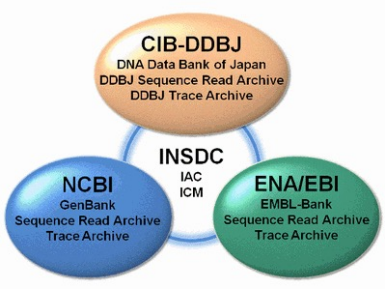
**Primary (archival) nucleotide sequence databases**

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) o Example: https://www.ncbi.nlm.nih.gov/nuccore/AM743169.1

• Archival in nature

o Historical

o Reflective of submitter point of view (subjective) o Redundant

• GenBank Data

o Direct submissions (traditional records)

o Batch submissions (EST, GSS, STS)

o ftp accounts (genome data)

• Three collaborating databases (all data from INSDC) o GenBank

o DNA Database of Japan (DDBJ)

o European Molecular Biology Laboratory (EMBL) Database



**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** Nucleotide (NCBI)

GenBank EMBL DDBJ genbank[Filter] EMBL[Filter] ddbj[Filter]

**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)**

o Free text search 

o Advanced searches (search tags and regular expressions)

o List of identifiers (Batch *Entrez*)

**Search sequence databases using a sequence query**

o BLAST (Basic Local Alignment Search Tool)

**Web service** (**Programmatic data access**)

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

**File Transfer Protocol (FTP)**

o 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 **Others**

PAT - patent sequences

gbdiv\_pri[PROP]

**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)

gbdiv\_htg[PROP]

HTG – High Throughput Genomic (incomplete sequences of genomic clones) STS – Sequence Tagged Site (PCR-based mapping)

HTC - High-throughput cDNA

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’ 3’ tcgaagaaaccgtttgccaa 5’**

**Orientation to read**

**5’aaccgtttgccaaagaagct 3’**

**Direct**

**Complement**

**Reverse complement**

**It's all relative**

What information is requested from authors when submitting a sequence?

• The sequence

Search Field Descriptions

• Automatic assignment of a unique **accession number** [accession] • Source 

• The organism [organism]

Primary key

• 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.,……. 

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/

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

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

Fields (identifiers) in columns

s

w

o

r

n

i

)s

e

**Accession (ID)**

**Sequence Mol type**

**Organism Authors Genes Feat**

How all this information is stored?

c

n

e

How is it shown to users?

u

q

e

s

(

s

e

i

r

t

n

E



**GenBank Flat Files (.gbff)**

**LOCUS HUMCYPB 1868 bp mRNA linear PRI 27-APR-1993** accession

**DEFINITION Human cytochrome P-450 1 mRNA, complete cds, clone Hp1-2.**

****Identification codes

**ACCESSION M17398 J03472 SOURCE Homo sapiens (human) ORGANISM Homo sapiens**

organism

division

**Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.**

**REFERENCE 1 (bases 1 to 1868)**

**AUTHORS Okino,S.T., Quattrochi,L.C., Pendurthi,U.R., McBride,O.W. and Tukey,R.H.**

**TITLE Characterization of multiple human cytochrome P-450 1 cDNAs. The chromosomal localization of the gene and evidence for alternate RNA**

****Source and References

**splicing**

s

**JOURNAL J. Biol. Chem. 262 (33), 16072-16079 (1987)** r

e

**FEATURES Location/Qualifiers**

i

f

**source 1..1868**

i

t

**/organism="Homo sapiens"**

type of molecule

n

journal

**/mol\_type="mRNA"**

e

**/db\_xref="taxon:9606"**

d

I

**CDS 2..1183**

**/note="cytochrome P-450 1" /codon\_start=1**

**/protein\_id="AAA35740.1"**

**/db\_xref="GI:181328"**

CDS coordinates protein name

Feature Table

**/translation="MEPFVVLVLCLSFMLLFSLWRQSCRRRKLPPGPTPLPIIGNMLQ** 

**IDVKDICKSFTNFSKVYGPVFTVYFGMNPIVVFHGYEAVKEALIDNGEEFSGRGNSPI**

**SQRITKGLGIISSNGKRWKEIRRFSLTNLRNFGMGKRSIEDRVQEEAHCLVEELRKTK**

protein sequence

**ASPCDPTFILGCAPCNVICSVVFQKRFDYKDQNFLTLMKRFNENFRILNSPWIQVCNN**

**FPLLIDCFPGTHNKVLKNVALTRSYIREKVKEHQASLDVNNPRDFMDCFLIKMEQEKD**

**NQKSEFNIENLVGTVADLFVAGTETTSTTLRYGLLLLLKHPEVTAKVQEEIDHVIGRH**

**RSPCMQDRSHMPYTDAVVHEIQRYSDLVPTGVPHAVTTDTKFRNYLIPKSFDNKIMLA**

**A"**

**ORIGIN 337 bp**

**1 aatggaacct tttgtggtcc tggtgctgtg tctctctttt atgcttctct tttcactctg 61 gagacagagc tgtaggagaa ggaagctccc tcctggcccc actcctcttc ctattattgg 121 aaatatgcta cagatagatg ttaaggacat ctgcaaatct ttcaccaatt tctcaaaagt 181 ctatggtcct gtgttcaccg tgtattttgg catgaatccc atagtggtgt ttcatggata 241 tgaggcagtg aaggaagccc tgattgataa tggagaggag ttttctggaa gaggcaattc 301 cccaatatct caaagaatta ctaaaggact tggaatcatt tccagcaatg gaaagagatg 361 gaaggagatc cggcgtttct ccctcacaaa cttgcggaat tttgggatgg ggaagaggag 421 cattgaggac cgtgttcaag aggaagctca ctgccttgtg gaggagttga gaaaaaccaa 481 ggcttcaccc tgtgatccca ctttcatcct gggctgtgct ccctgcaatg tgatctgctc 541 cgttgttttc cagaaacgat ttgattataa agatcagaat tttctcaccc tgat...... //**

****

****DNA sequence

**FASTA file**

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

**>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.

**FASTA file (ENA)**

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

**>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



**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)**

o Free text search 

o Advanced searches (search tags and regular expressions)

o List of identifiers (Batch *Entrez*)

**Search sequence databases using a sequence query**

o BLAST (Basic Local Alignment Search Tool)

**Web service** (**Programmatic data access**)

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

**File Transfer Protocol (FTP)**

o 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**

****

**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*) )

BLAST Basic programs

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:**

****https://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo\_BLASTGuide.pdf

**NCBI Datasets** 

https://www.ncbi.nlm.nih.gov/datasets/docs/v2/getting\_started/

