DATA ACQUISITION FROM BIO-DATABASES AND BLAST

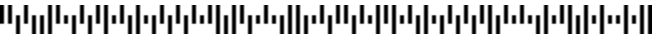
Natapol Pornputtapong

18 January 2018

PHARMACY INFORMATICS

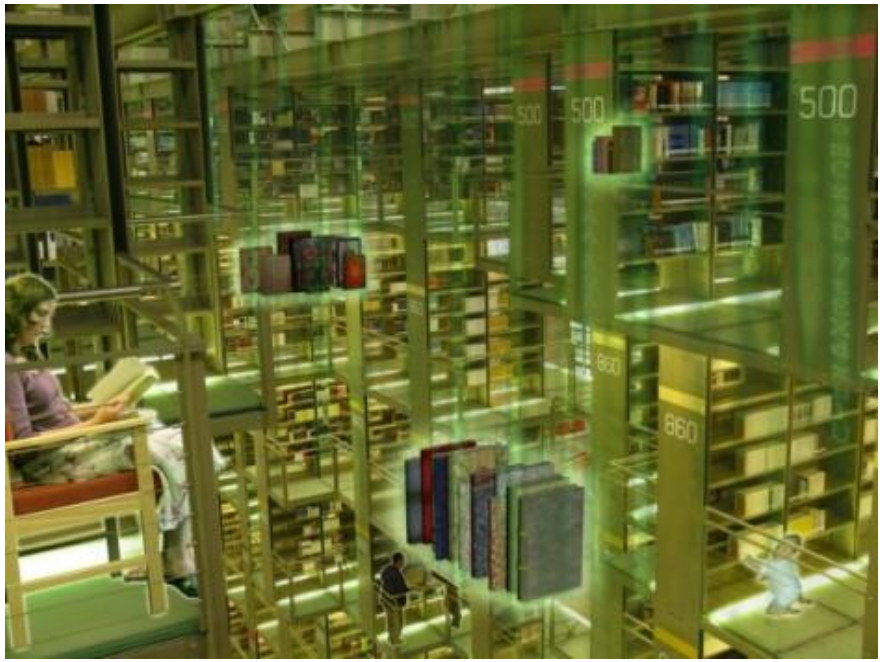
CC BY NC SA

DATABASE



- Collections of data
- To share – multi-user interface
- To prevent data loss
- To make sure to get the right things

LIBRARY -> DIGITAL LIBRARY



DATABASE: A LIBRARY OF DATA

Database

- Files, Tables, Records
- Data structure
- Database management system
- Programming interface
- User interface

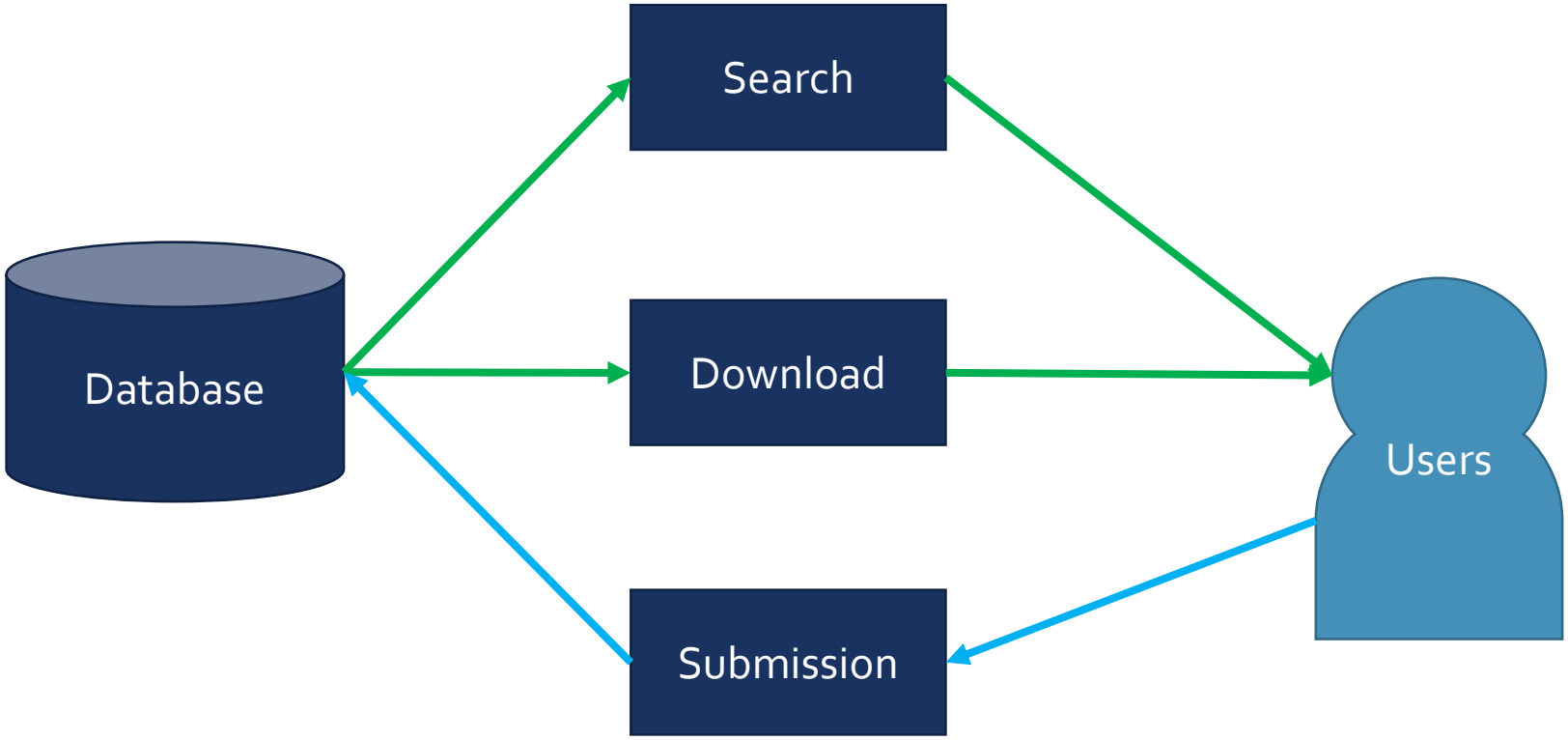
Library

- Books
- building, shelves
- Librarian
- Protocols, SOPs
- Services

ADVANTAGE OF DATABASE

- Data integrity
- Smaller space
- Data availability
- Speed

DATABASE FOR USERS



HOW TO CHOOSE DATABASE?

- **1695** bio-databases in ***NAR* online Molecular Biology Database Collection** in **15** categories

DATA CONTENT

- Literature

- DNA sequence

GenBank

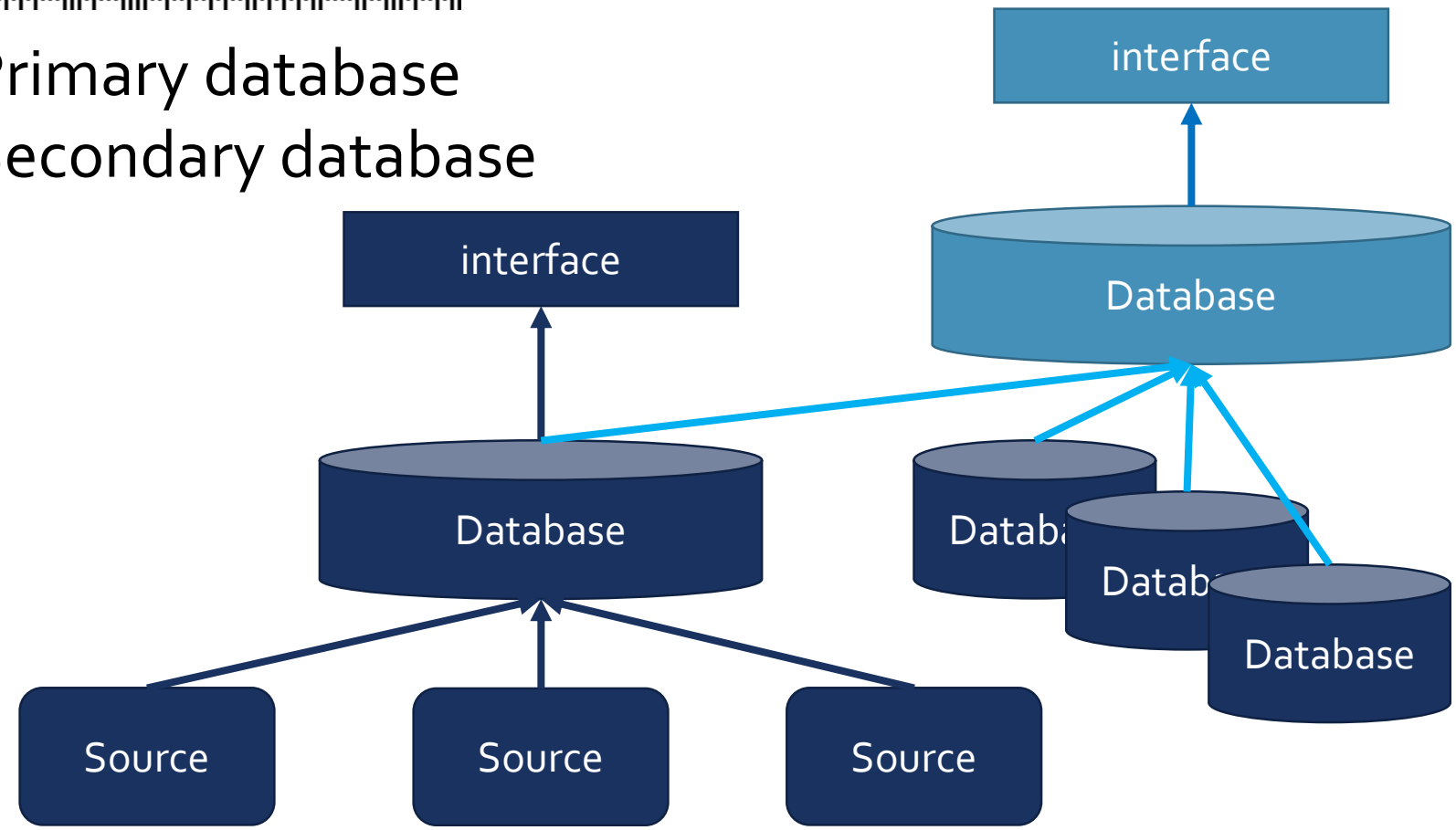
- Protein sequence

RefSeq

TrEMBL

CONCEPTS OF DATABASE

- Primary database
- Secondary database



PRIMARY & SECONDARY DB

	Primary database	Secondary database
Synonyms	Archival database	Curated database; knowledgebase
Source of data	Direct submission of experimentally-derived data from researchers	Results of analysis, literature research and interpretation, often of data in primary databases
Examples	•ENA, GenBank and DDBJ (nucleotide sequence) •ArrayExpress Archive and GEO (functional genomics data) •Protein Data Bank (PDB; coordinates of three-dimensional macromolecular structures)	•InterPro (protein families, motifs and domains) •UniProt Knowledgebase (sequence and functional information on proteins) •Ensembl (variation, function, regulation and more layered onto whole genome sequences)

DATA COLLECTION CRITERIA

NCBI GenBank

NCBI RefSeq

GenBank ® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences

RefSeq: NCBI Reference Sequence Database

A comprehensive, integrated, non-redundant, well-annotated set of reference sequences including genomic, transcript, and protein.

UniProtKB

UniProt Knowledgebase

Swiss-Prot (556,388)

Manually annotated and reviewed.

TrEMBL (102,248,261)

Automatically annotated and not reviewed.

creative commons

Attribution (by)

All CC licenses require that others who use your work in any way must give you credit the way you request, but not in a way that suggests you endorse them or their use. If they want to use your work without giving you credit or for endorsement purposes, they must get your permission first.

ShareAlike (sa)

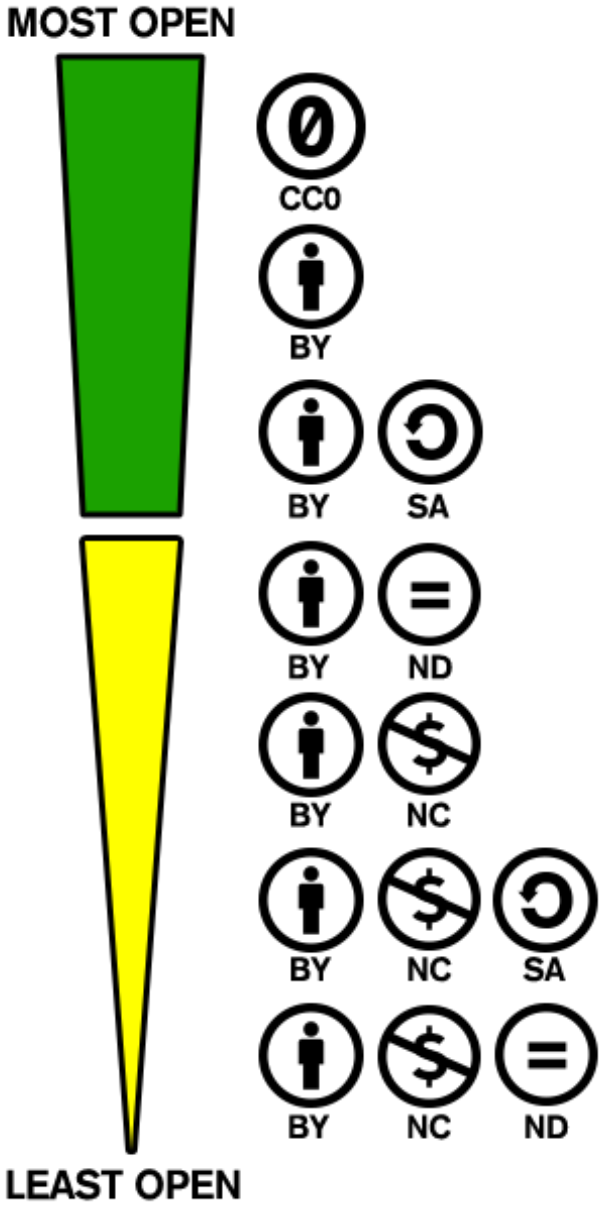
You let others copy, distribute, display, perform, and modify your work, as long as they distribute any modified work on the same terms. If they want to distribute modified works under other terms, they must get your permission first.

NonCommercial (nc)

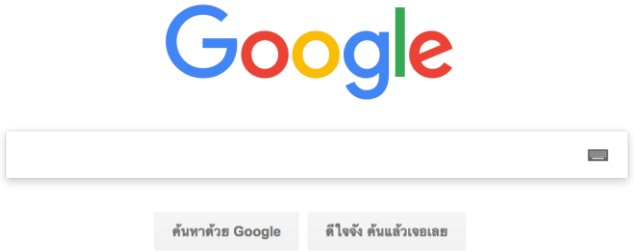
You let others copy, distribute, display, perform, and (unless you have chosen NoDerivatives) modify and use your work for any purpose other than commercially unless they get your permission first.

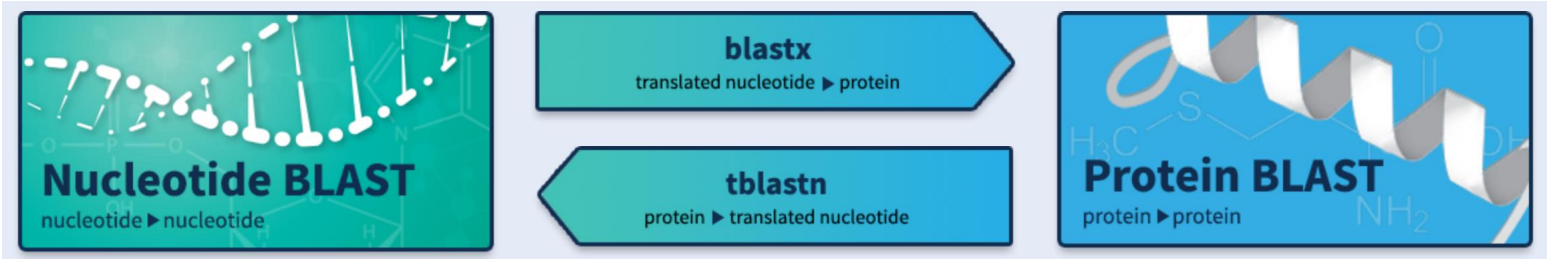
NoDerivatives (nd)

You let others copy, distribute, display and perform only original copies of your work. If they want to modify your work, they must get your permission first.



ACCESSIBILITY: TOOLS & INTERFACES

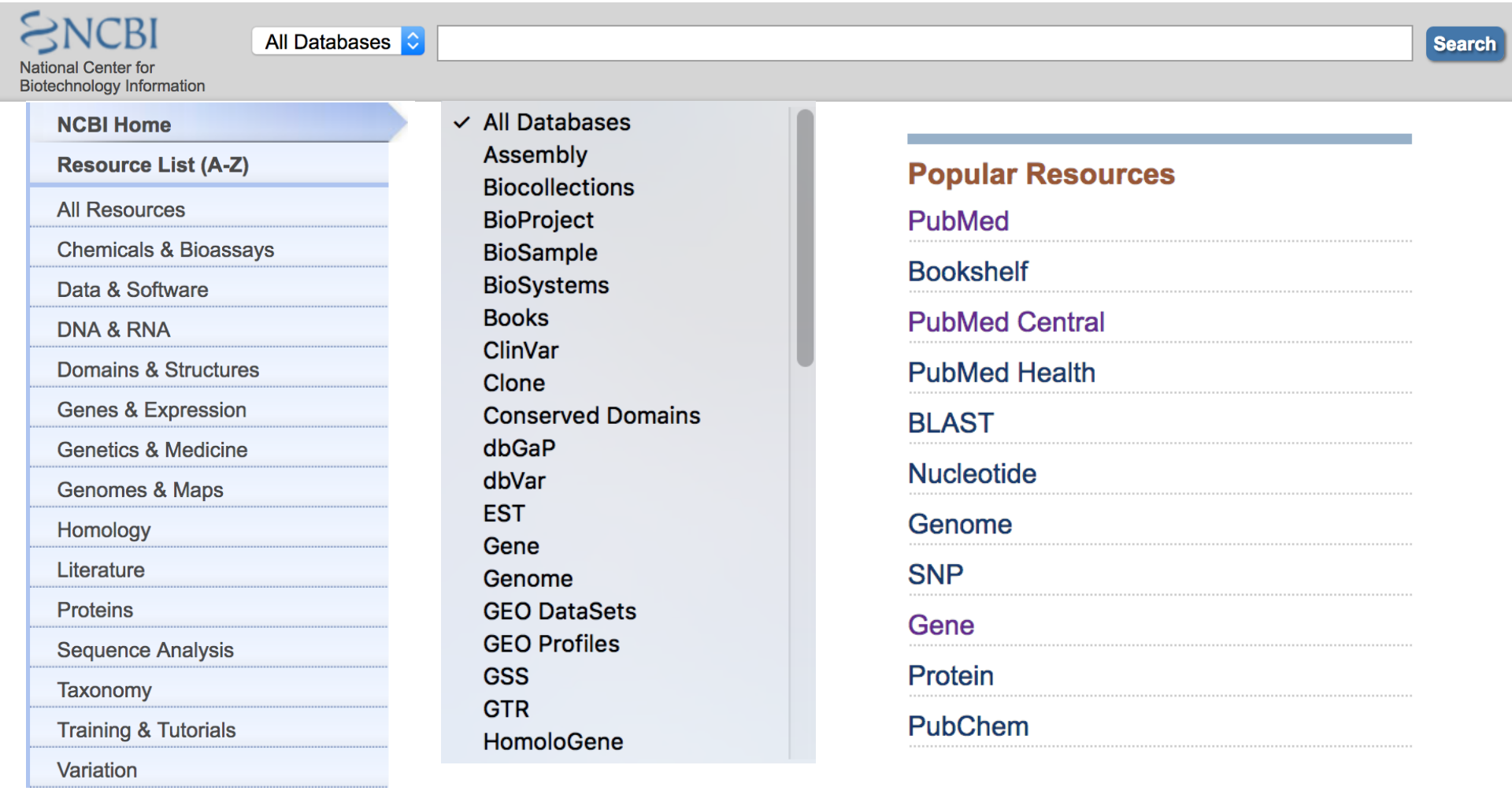




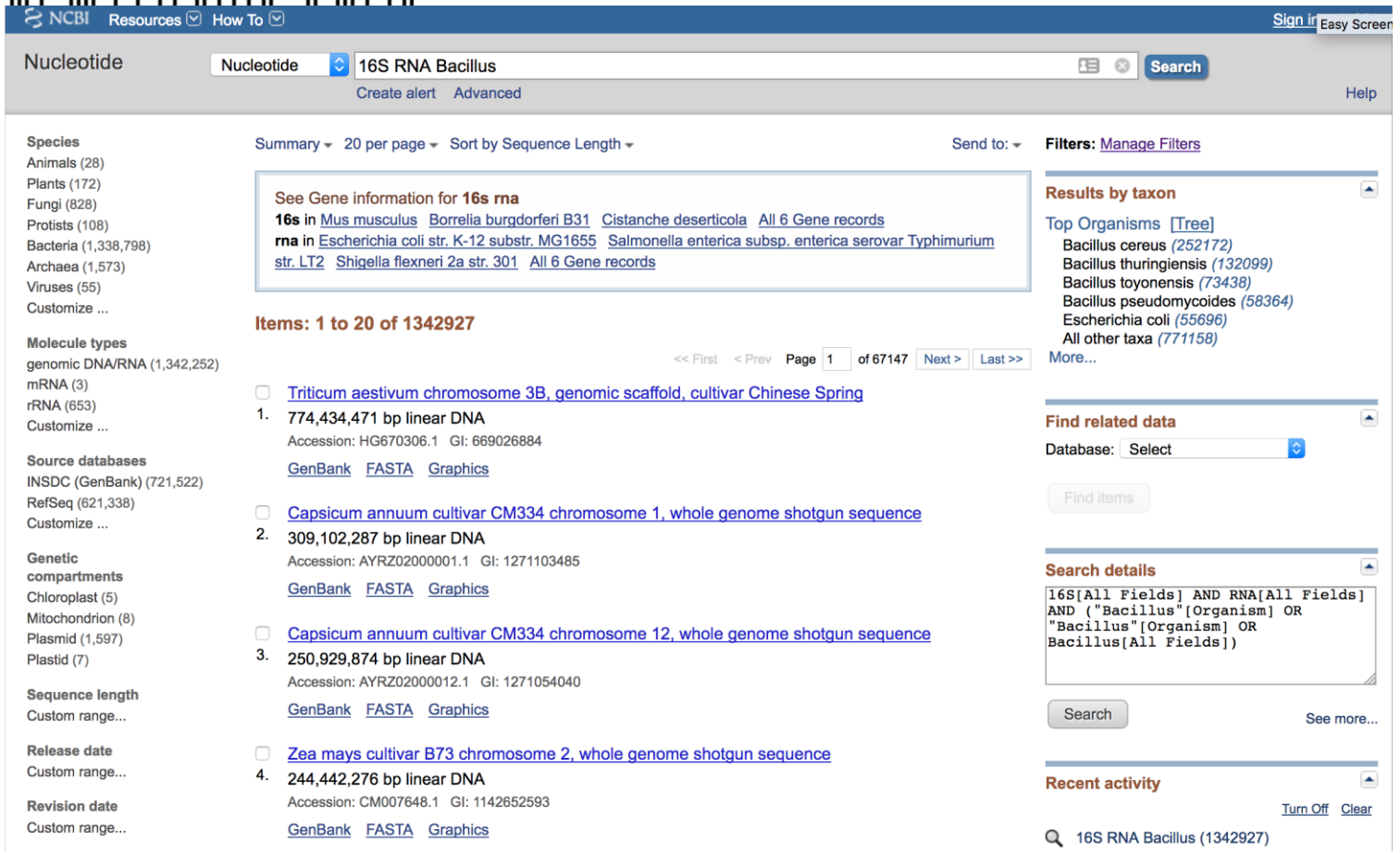
NCBI Entrez

RESTful interface to the ENA

NCBI SEARCH TOOL



SIMPLE SEARCH



BOOLEAN OPERATOR

Nucleotide

Nucleotide "16S RNA" AND Bacillus Search

Create alert Advanced Help

Species
Bacteria (731)
Archaea (16)
Viruses (2)
Customize ...

Molecule types
genomic DNA/RNA (747)
Customize ...

Source databases
INSDC (GenBank) (717)

Summary 20 per page Sort by Sequence Length Send to: **Filters:** Manage Filters

See Gene information for **16s rna**
16s in Mus musculus Borrelia burgdorferi B31 Cistanche deserticola All 6 Gene records
rna in Escherichia coli str. K-12 substr. MG1655 Salmonella enterica subsp. enterica serovar Typhimurium str. LT2 Shigella flexneri 2a str. 301 All 6 Gene records

Items: 1 to 20 of 751

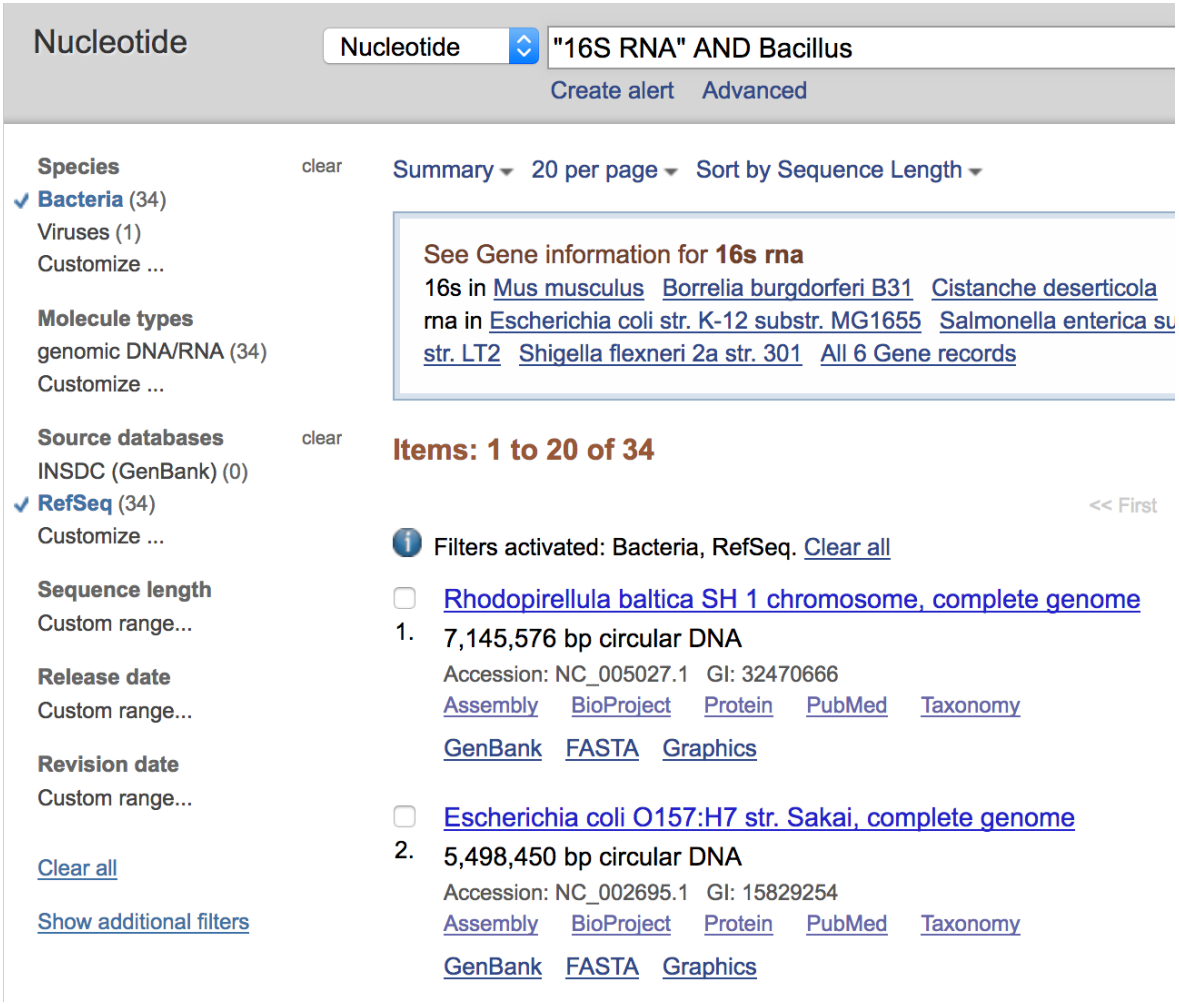
<< First < Prev Page 1 of 38 Next > Last >>

Results by taxon

Top Organisms [Tree]
Bacillus cereus *(61)*
Bacillus thuringiensis *(39)*
Acinetobacter baumannii *(26)*
Erwinia amylovora *(14)*
Bacillus subtilis *(12)*
All other taxa *(599)*
More...

FILTER

- Limit with filter
- Advanced search builder



RESULTS

GenBank ▾ Send to: ▾

Bacillus subtilis strain IAM 12118 16S ribosomal RNA, complete sequence

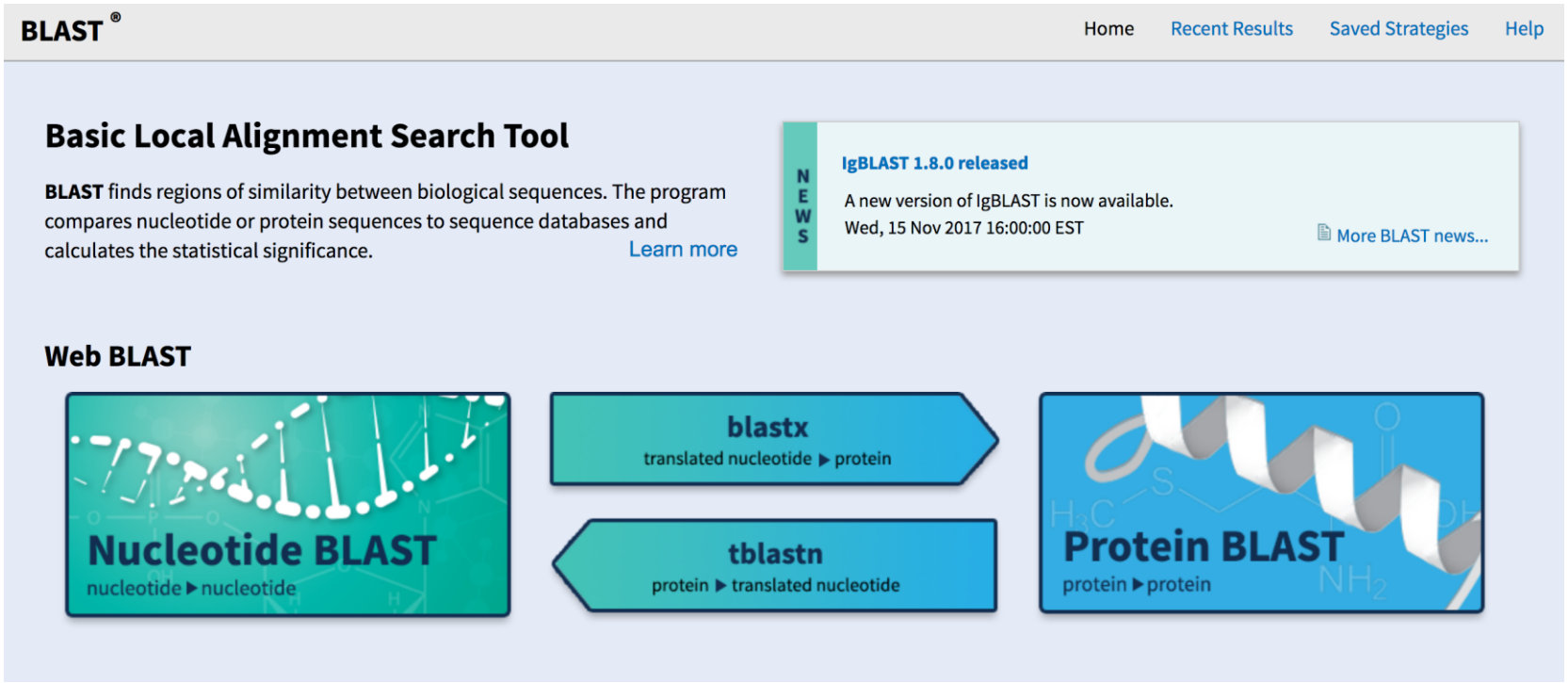
NCBI Reference Sequence: NR_112116.2

FASTA Graphics

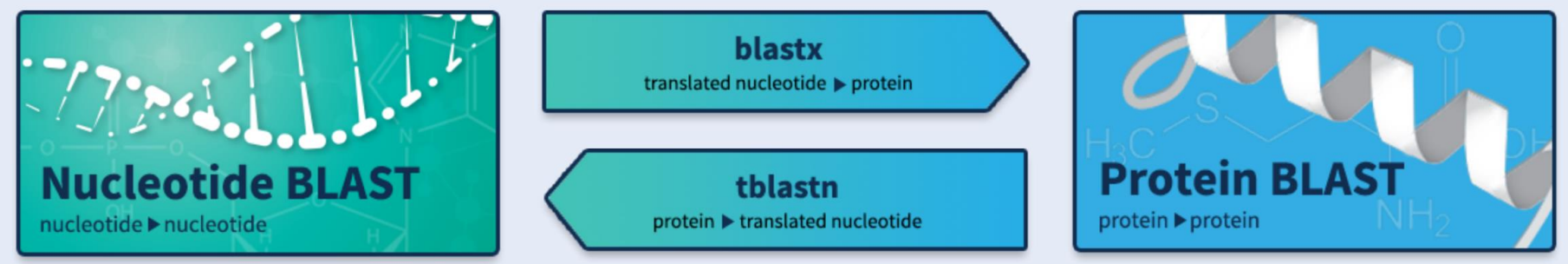
Go to:

```
LOCUS       NR_112116              1550 bp    rRNA    linear   BCT 25-OCT-2017
DEFINITION  Bacillus subtilis strain IAM 12118 16S ribosomal RNA, complete
            sequence.
ACCESSION   NR_112116
VERSION     NR_112116.2
DBLINK      Project: 33175
            BioProject: PRJNA33175
KEYWORDS    RefSeq.
SOURCE      Bacillus subtilis
  ORGANISM  Bacillus subtilis
            Bacteria; Firmicutes; Bacilli; Bacillales; Bacillaceae; Bacillus.
REFERENCE   1
  AUTHORS   Goto,K.
  TITLE     16S rRNA gene sequence of spore-forming bacteria
  JOURNAL   Published Only in Database (2003)

COMMENT     REVIEWED REFSEQ: This record has been curated by NCBI staff. The
            reference sequence was derived from AB042061.
            On Oct 25, 2017 this sequence version replaced gi:631250919.
FEATURES             Location/Qualifiers
     source          1..1550
                     /organism="Bacillus subtilis"
                     /mol_type="rRNA"
                     /strain="IAM 12118"
                     /culture_collection="IAM:12118"
                     /type_material="type strain of Bacillus subtilis"
                     /db_xref="taxon:1423"
     rRNA            1..1550
                     /product="16S ribosomal RNA"
ORIGIN
        1 ttatcggaga gtttgatcct ggctcaggac gaacgctggc ggcgtgccta atacatgcaa
       61 gtcgagcgga cagatgggag cttgctccct gatgttagcg gcggacgggt gagtaacacg
      121 tgggtaacct gcctgtaaga ctgggataac tccgggaaac cggggctaat accggatggt
      181 tgtttgaacc gcatggttca aacataaaag gtggcttcgg ctaccactta cagatggacc

     1201 tcaaatcatc atgcccctta tgacctgggc tacacacgtg ctacaatgga cagaacaaag
     1261 ggcagcgaaa ccgcgaggtt aagccaatcc cacaaatctg ttctcagttc ggatcgcagt
     1321 ctgcaactcg actgcgtgaa gctggaatcg ctagtaatcg cggatcagca tgccgcggtg
     1381 aatacgttcc cgggccttgt acacaccgcc cgtcacacca cgagagtttg taacacccga
     1441 agtcggtgag gtaacctttt aggagccagc cgccgaaggt gggacagatg attggggtga
     1501 agtcgtaaca aggtagccgt atcggaaggt gcggctggat cacctccttt
//
```

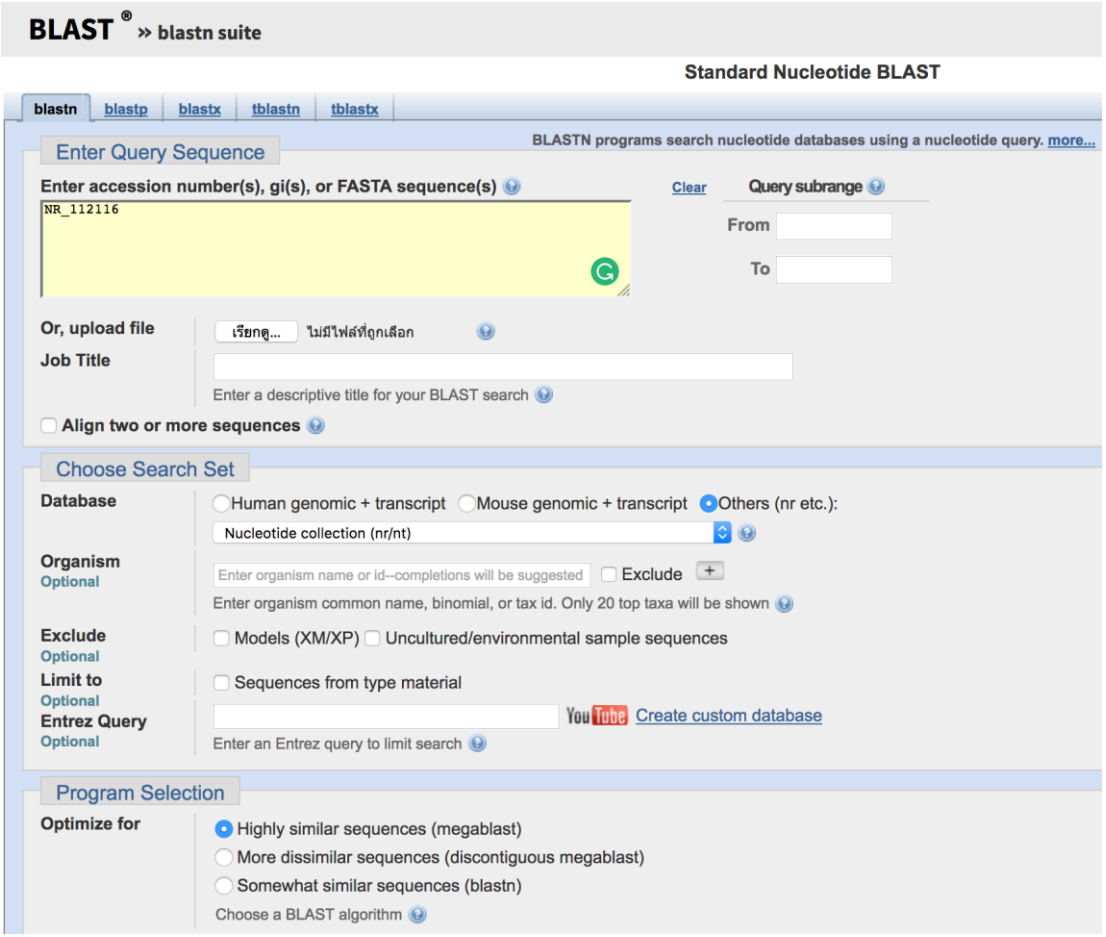
BLAST: BASIC LOCAL ALIGNMENT SEARCH TOOL

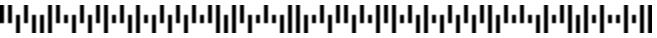


MAJOR BLAST PROGRAMS



BLAST SEARCH





Edit and Resubmit | Save Search Strategies | ▸Formatting options | ▸Download | How to read this page | Blast report description

Job title: NR_112116:Bacillus subtilis strain IAM 12118...

RID 5PCMFZ2M014 (Expires on 01-15 17:35 pm)

Query ID NR_112116.2
Description Bacillus subtilis strain IAM 12118 16S ribosomal RNA, complete sequence
Molecule type nucleic acid
Query Length 1550

Database Name nr
Description Nucleotide collection (nt)
Program BLASTN 2.7.1+ ▸Citation

Other reports: ▸Search Summary [Taxonomy reports] [Distance tree of results] [MSA viewer]

Graphic Summary

Distribution of the top 200 Blast Hits on 100 subject sequences

Mouse over to see the title, click to show alignments

Color key for alignment scores

<40 | 40-50 | 50-80 | 80-200 | >=200

Query

1 300 600 900 1200 1500

Descriptions

Sequences producing significant alignments:

Select: All None Selected:0

Alignments | Download | GenBank | Graphics | Distance tree of results

Description	Max score	Total score	Query cover	E value	Ident	Accession
Bacillus subtilis strain IAM 12118 16S ribosomal RNA, complete sequence	2863	2863	100%	0.0	100%	NR_112116.2
Bacillus subtilis gene for 16S ribosomal RNA, partial sequence	2863	2863	100%	0.0	100%	AB042061.1
Bacillus subtilis subsp. subtilis strain NCD-2 chromosome, complete genome	2857	28464	100%	0.0	99%	CP023755.1
Bacillus subtilis strain TLO3 chromosome, complete genome	2857	28362	100%	0.0	99%	CP023257.1
Bacillus subtilis strain GS 188 genome	2857	28319	100%	0.0	99%	CP022391.1

Alignments

Download | GenBank | Graphics ▼ Next ▲ Previous Descriptions

Bacillus subtilis strain IAM 12118 16S ribosomal RNA, complete sequence

Sequence ID: NR_112116.2 Length: 1550 Number of Matches: 1

Related Information

Range 1: 1 to 1550 GenBank Graphics ▼ Next Match ▲ Previous Match

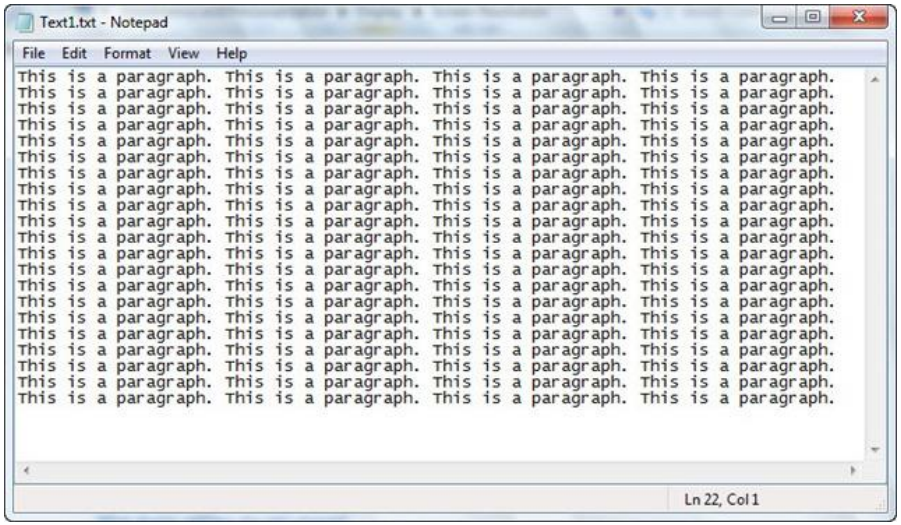
Score	Expect	Identities	Gaps	Strand
2863 bits(1550)	0.0	1550/1550(100%)	0/1550(0%)	Plus/Plus

```
Query  1    TTATCGGAGAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCCTAATACATGCAA  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1    TTATCGGAGAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCCTAATACATGCAA  60

Query  61   GTCGAGCGGACAGATGGGAGCTTGCTCCCTGATGTTAGCGGCGGACGGGTGAGTAACACG  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61   GTCGAGCGGACAGATGGGAGCTTGCTCCCTGATGTTAGCGGCGGACGGGTGAGTAACACG  120
```

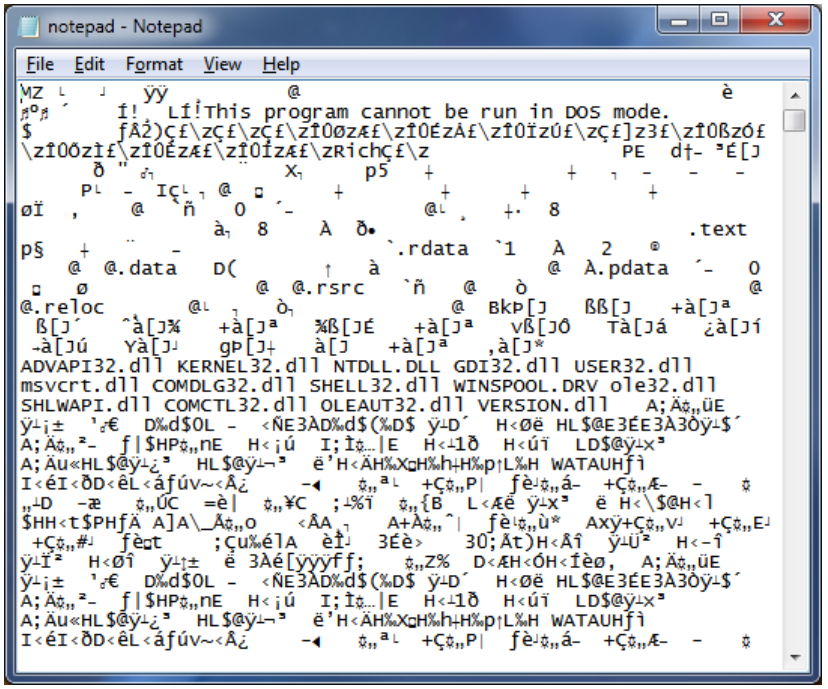
OTHER BLAST PROGRAMS

WORLD OF FILES

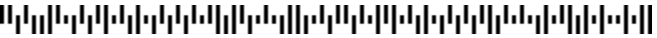
Text files



Binary files

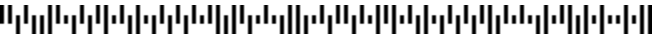


TEXT FILES: WORLD OF FORMATS



- MS Words: .doc, .docx, .rtf, .txt
- Sequence: FastA (.fasta), Genbank (.gbk)
- Protein structure: PDB (.pdb)

FASTA FORMAT



```
>P01013 GENE X PROTEIN (OVALBUMIN-RELATED)
QIKDLLVSSSTDLDTTLVLVNAIYFKGMWKTAFNAEDTREMP
FHVTKQESKPVQMMCMNNSFNVATLPAEKMKILELPFASGDL
SMLVLLPDEVSDLERIEKTINFEKLTEWTNPNTMEKRRVKVY
LPQMKIEEKYNLTSVLMALGMTDLFIPSANLTGISSAESLKI
SQAVHGAFMELSEDGIEMAGSTGVIEDIKHSPESEQFRADHP
FLFLIKHNPTNTIVYFGRYWSP


>…
```

GENBANK FORMAT

GenBank ▾ Send to: ▾

Bacillus subtilis strain IAM 12118 16S ribosomal RNA, complete sequence

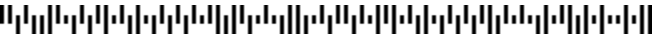
NCBI Reference Sequence: NR_112116.2

FASTA Graphics

Go to:

```
LOCUS       NR_112116              1550 bp    rRNA    linear   BCT 25-OCT-2017
DEFINITION  Bacillus subtilis strain IAM 12118 16S ribosomal RNA, complete
            sequence.
ACCESSION   NR_112116
VERSION     NR_112116.2
DBLINK      Project: 33175
            BioProject: PRJNA33175
KEYWORDS    RefSeq.
SOURCE      Bacillus subtilis
  ORGANISM  Bacillus subtilis
            Bacteria; Firmicutes; Bacilli; Bacillales; Bacillaceae; Bacillus.
REFERENCE   1
  AUTHORS   Goto,K.
  TITLE     16S rRNA gene sequence of spore-forming bacteria
  JOURNAL   Published Only in Database (2003)

COMMENT     REVIEWED REFSEQ: This record has been curated by NCBI staff. The
            reference sequence was derived from AB042061.
            On Oct 25, 2017 this sequence version replaced gi:631250919.
FEATURES             Location/Qualifiers
     source          1..1550
                     /organism="Bacillus subtilis"
                     /mol_type="rRNA"
                     /strain="IAM 12118"
                     /culture_collection="IAM:12118"
                     /type_material="type strain of Bacillus subtilis"
                     /db_xref="taxon:1423"
     rRNA            1..1550
                     /product="16S ribosomal RNA"
ORIGIN
        1 ttatcggaga gtttgatcct ggctcaggac gaacgctggc ggcgtgccta atacatgcaa
       61 gtcgagcgga cagatgggag cttgctccct gatgttagcg gcggacgggt gagtaacacg
      121 tgggtaacct gcctgtaaga ctgggataac tccgggaaac cggggctaat accggatggt
      181 tgtttgaacc gcatggttca aacataaaag gtggcttcgg ctaccactta cagatggacc

     1201 tcaaatcatc atgcccctta tgacctgggc tacacacgtg ctacaatgga cagaacaaag
     1261 ggcagcgaaa ccgcgaggtt aagccaatcc cacaaatctg ttctcagttc ggatcgcagt
     1321 ctgcaactcg actgcgtgaa gctggaatcg ctagtaatcg cggatcagca tgccgcggtg
     1381 aatacgttcc cgggccttgt acacaccgcc cgtcacacca cgagagtttg taacacccga
     1441 agtcggtgag gtaacctttt aggagccagc cgccgaaggt gggacagatg attggggtga
     1501 agtcgtaaca aggtagccgt atcggaaggt gcggctggat cacctccttt
//
```

NEXUS FORMAT



```
#NEXUS
BEGIN DATA;
     DIMENSIONS NTAX=8 NCHAR=1202;
     FORMAT MISSING=? DATATYPE=PROTEIN GAP=-;
     OPTIONS GAPMODE=MISSING;
MATRIX
[                                          10         20 ...]
[                                 ---------|---------|-...]
Homo_sapiens_4379045              TERLVLPPPDPLDLPLRAVEL...
Pan_troglodytes_114606536         TERLVLPPPDPLDLPLRAVEL...
Ailuropoda_melanoleuca_301788522  TERLVLPPPDPLDLPLRPVEL...
Mus_musculus_87252727             TERLVLPPLDPLNLPLRALEV...
Danio_rerio_113678409             MDKIDLPPVGPDDLPLSLLEM...
Xenopus_tropicalis_301627725      MNTLDLSNRDPLDLPLSVLEL...
Monodelphis_domestica_126309591   TERLVLPPRGPLDLPLCALEL...
Canis_familiaris_73972333         TERLALPPPDPLDLPLRPVEL...;
END;
```

NEXT

QUESTION?