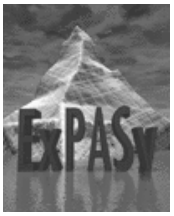
Database Systems

in Bioinformatics
Exercises



Several Exercises

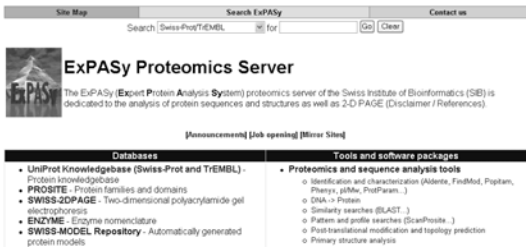
- Retrieve Protein Sequences by **SWISS-PROT**
- Compare Protein Sequences by **BLAST**
- Use **Entrez** for Searching
- Search in the **Tree of Life**

ExPASy and SWISS-PROT

- Created and managed by one of the pioneers of protein informatics: **Prof. Amos Bairoch**
- **ExPASy server**: world-leading resource of protein information
- **SWISS-PROT**: name of the database
- **Numerous** analysis **tools**

Example 1

1. Point your favorite Internet browser to the website **www.expasy.org**
2. Inform yourself about the abbreviation **ExPASy**



20.08.07

Example 1

3. Inform yourself about the **Site Map** and change again to the homepage.

Example 1

4. Click on UniProt Knowledgebase (Swiss-Prot and TrEMBL)
5. Type here **`dUTPase coli`** and click on **Go**

Example 1

You get a large amount of information about the dUTPase protein of E. Coli

Search in UniProt Knowledgebase (Swiss-Prot and TrEMBL) for: dUTPase coli

UniProtKB/Swiss-Prot Release 54.0 of 24-Jul-2007
UniProtKB/TrEMBL Release 37.0 of 24-Jul-2007

- Number of sequences found in UniProt Knowledgebase (Swiss-Prot[S] and TrEMBL[E]): **3**
- Note that the selected sequences can be saved to a file to be later retrieved; to do so, go to the bottom of this page.
- For more directed searches, you can use the Sequence Retrieval System SRS.

Search in UniProtKB/Swiss-Prot: There are matches to 3 out of 276256 entries

DUT_ECO57 (**P64007**)
Deoxyuridine 5'-triphosphate nucleotidohydrolase (EC 3.6.1.23) (dUTPase) (dUTP pyrophosphatase). (GENE: Name=dut; OrderedLocusNames=Z5064, ECs4515) - Escherichia coli O157:H7

Example 1

7. Click e.g. on **DUT_ECOLI (P06968)**

Note: most headings are clickable, even if they don't appear as links. They link to the user manual or other documents.

Entry information	
Entry name	**DUT_ECOLI**
Primary accession number	**P06968**
Secondary accession number	Q2M7V4
Integrated into Swiss-Prot on	April 1, 1988
Sequence was last modified on	April 1, 1988 (Sequence version 1)
Annotations were last modified on	July 24, 2007 (Entry version 78)
Name and origin of the protein	
Protein name	**Deoxyuridine 5'-triphosphate nucleotidohydrolase**
Synonyms	**EC 3.6.1.23** **dUTPase** **dUTP pyrophosphatase**
Gene name	**Name: dut** Synonyms: dnaS, sof OrderedLocusNames: b3640, JW3615
From	Escherichia coli [TaxID: 562] [HAMAP proteome]
Taxonomy	Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Escherichia.
Protein existence	1: Evidence at protein level;

A SWISS-PROT entry consists (roughly described) of four parts: top, top-middle, bottom-middle and last section.

Example 1

Top section = general information:

- Entry name
- Unique identifier (this code is worth writing down because it is used to cross-reference related entries in other databases)
- Further entry information

Entry information	
Entry name	**DUT_ECOLI**
Primary accession number	**P06968**
Secondary accession number	Q2M7V4
Integrated into Swiss-Prot on	April 1, 1988
Sequence was last modified on	April 1, 1988 (Sequence version 1)
Annotations were last modified on	July 24, 2007 (Entry version 78)

Example 1

Top-middle section = biochemical description

- Standard name
- International Enzyme Committee number
- Synonyms
- List of bibliographic references

Name and origin of the protein	
Protein name	**Deoxyuridine 5'-triphosphate nucleotidohydrolase**
Synonyms	**EC 3.6.1.23** **dUTPase** **dUTP pyrophosphatase**
Gene name	**Name: dut** Synonyms: dnaS, sof OrderedLocusNames: b3640, JW3615
From	Escherichia coli [TaxID: 562] [HAMAP proteome]
Taxonomy	Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Escherichia.
Protein existence	1: Evidence at protein level;
References	

Example 1

Bottom-middle section

- Series of links to various functional classification schemes

Cross-references		
Sequence databases		
EMBL	X01714; CAA25859.1; -; Genomic_DNA.	[EMBL / GenBank / DDBJ] [CoDingSequence]
	V01578; CAA24897.1; -; Genomic_DNA.	[EMBL / GenBank / DDBJ] [CoDingSequence]
	L10328; AAA61993.1; -; Genomic_DNA.	[EMBL / GenBank / DDBJ] [CoDingSequence]
	U00096; AAC76664.1; -; Genomic_DNA.	[EMBL / GenBank / DDBJ] [CoDingSequence]
	AP009048; BAE77652.1; -; Genomic_DNA.	[EMBL / GenBank / DDBJ] [CoDingSequence]
PIR	A30388; WPECDU.	

Example 1

Last section = sequence section

- Keywords
- Length
- Molecular weight

Sequence information

Length: **151 AA** [This is the length of the unprocessed precursor] Molecular weight: **16155 Da** [This is the MW of the unprocessed precursor] CRC64: **98FA3DE0BC70FFB2** [This is a checksum on the sequence]

```
        10         20         30         40         50         60
MKKIDVKILD PRVGKEFPLP TYATSGSAGL DLRACLNDAV ELAPGDTTLV PTGLAIHIAD
        70         80         90        100        110        120
PSLAAMMLPR SGLGHKHGIV LGNLVGLIDS DYQGQLMISV WNRGQDSFTI QPGERIAQMI
       130        140        150
FVPVVQAEFN LVEDFDATDR GEGGFGHSGR Q
```

IP06968 in FASTA format

Example 1

8. To get **FASTA** format, click the FASTA format button (right on the bottom)

```
>PO6968|DUT_ECOLI Deoxyuridine 5'-triphosphate nucleotidohydrolase - Escherichia coli.
MKKIDVKILDPRVGKEFPLPTYATSGSAGLDLRACLNDAVELAPGDTTLVPTGLAIHIAD
PSLAAMMLPRSGLGHKHGIVLGNLVGLIDSDYQGQLMISVWNRGQDSFTIQPGERIAQMI
FVPVVQAEFNLVEDFDATDRGEGGFGHSGRQ
```

9. For further studies (in our next example) **copy** the **amino-acid sequence** and **paste** it into the simple text file *dutpasecoli.txt*.

Exercise 1

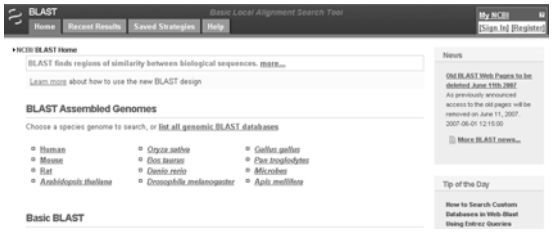
Find the **Human Insulin** Protein

BLAST

- = **B**asic **L**ocal **A**lignment **S**earch **T**ool
- **Great sequence comparison tool**
- Information can be used for:
  prediction of protein function, 3-D-structure, identification of homologues in other organisms
- Server: National Center for Biotechnologie Information = **NCBI**

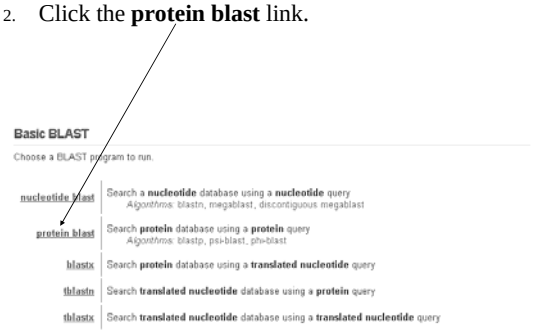
Example 2

1. Point your browser to the website **www.ncbi.nih.gov/BLAST/**



Example 2

2. Click the **protein blast** link.



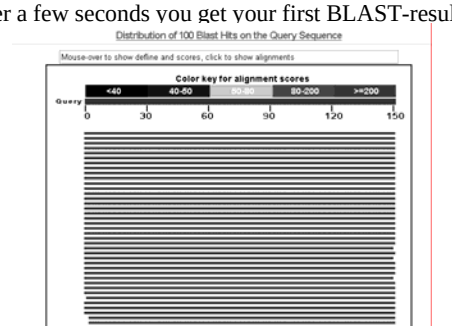
Example 2

The following **query screen** appears

3. Copy and paste the protein sequence from *dutpasecoli.txt* into the search window, and click the BLAST button on the bottom of the page, and wait!

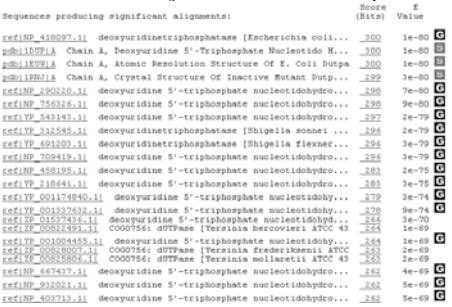
Example 2

After a few seconds you get your first BLAST-result



Example 2

4. Scroll down until you reach a list of sequences.

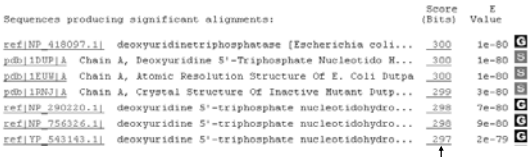


Sequences producing significant alignments:		Score (Bits)	E Value
ref|NP_418097.1|	deoxyuridinetriphosphatase [Escherichia coli...	300	1e-80
pdb|1DUP|A	Chain A, Deoxyuridine 5'-Triphosphate Nucleotido H...	300	1e-80
pdb|1EUW|A	Chain A, Atomic Resolution Structure Of E. Coli Dutpa	300	1e-80
pdb|1PNJ|A	Chain A, Crystal Structure Of Inactive Mutant Dutp...	299	3e-80
ref|NP_290220.1|	deoxyuridine 5'-triphosphate nucleotidohydro...	298	7e-80
ref|NP_756326.1|	deoxyuridine 5'-triphosphate nucleotidohydro...	298	9e-80
ref|YP_543143.1|	deoxyuridine 5'-triphosphate nucleotidohydro...	297	2e-79
ref|YP_312545.1|	deoxyuridinetriphosphatase [Shigella sonnei ...	296	2e-79
ref|YP_691203.1|	deoxyuridinetriphosphatase [Shigella flexner...	296	3e-79
ref|NP_709419.1|	deoxyuridine 5'-triphosphate nucleotidohydro...	296	3e-79
ref|NP_458195.1|	deoxyuridine 5'-triphosphate nucleotidohydro...	283	2e-75
ref|YP_218641.1|	deoxyuridine 5'-triphosphate nucleotidohydro...	283	3e-75
ref|YP_001174840.1|	deoxyuridine 5'-triphosphate nucleotidohy...	279	3e-74
ref|YP_001337632.1|	deoxyuridine 5'-triphosphate nucleotidohy...	278	9e-74
ref|ZP_01537436.1|	deoxyuridine 5'-triphosphate nucleotidohyd...	246	3e-70
ref|ZP_00822491.1|	COG0756: dUTPase [Yersinia bercovieri ATCC 43	244	1e-69
ref|YP_001004455.1|	deoxyuridine 5'-triphosphate nucleotidohy...	244	1e-69
ref|ZP_00828009.1|	COG0756: dUTPase [Yersinia frederiksenii ATCC	243	2e-69
ref|ZP_00825806.1|	COG0756: dUTPase [Yersinia mollaretii ATCC 43	243	2e-69
ref|NP_667437.1|	deoxyuridine 5'-triphosphate nucleotidohydro...	242	4e-69
ref|NP_932021.1|	deoxyuridine 5'-triphosphate nucleotidohydro...	242	5e-69
ref|NP_403713.1|	deoxyuridine 5'-triphosphate nucleotidohydro...	242	5e-69

Example 2

- List contains the sequences with significant similarity
- Ranked by decreasing score values
- Using a protein sequence taken from a database (like in this example) the best matching protein is the one that you started with
- Score value depends on the length of the most similar segments between 2 sequences

Example 2



Sequences producing significant alignments:		Score (Bits)	E Value
ref|NP_418097.1|	deoxyuridinetriphosphatase [Escherichia coli...	300	1e-80
pdb|1DUP|A	Chain A, Deoxyuridine 5'-Triphosphate Nucleotido H...	300	1e-80
pdb|1EUW|A	Chain A, Atomic Resolution Structure Of E. Coli Dutpa	300	1e-80
pdb|1PNJ|A	Chain A, Crystal Structure Of Inactive Mutant Dutp...	299	3e-80
ref|NP_290220.1|	deoxyuridine 5'-triphosphate nucleotidohydro...	298	7e-80
ref|NP_756326.1|	deoxyuridine 5'-triphosphate nucleotidohydro...	298	9e-80
ref|YP_543143.1|	deoxyuridine 5'-triphosphate nucleotidohydro...	297	2e-79

5. Click one of the underlined scores.
   This brings you further down in the page, to see the alignment between your query sequence and the matching sequence of the protein corresponding to this score.

Example 2

Query sequence Matching residues

Database sequence

Exercise 2

a) What protein is given by the following sequence (you find it in the file *sequence.txt*).

```
GDVEKGKKIFIMKCAQCHTVEKGGKHK
TGPNLHGLFGRKTGGAPGYBYTAANKN
KGIIWGEDTLMEYLENPKKYIPGTKMI
FVGIKKKEERADLIAYLKKATNE
```

b) Change some letters in the sequence and try again.

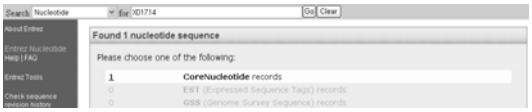
Entrez

- The Life Sciences **Search Engine**
- integrated, text-based search and retrieval system
- used at **NCBI** for the major databases
- including PubMed, Nucleotide and Protein Sequences, Protein Structures, Complete Genomes, Taxonomy, and others

Example 3

1. Point your favorite Internet browser to the website **http://www.ncbi.nlm.nih.gov/entrez/** →

2. Choose *Nucleotide* in the *Search* pull-down menu, type the GenBank ID *X01714* in the *for* window and click *Go* →



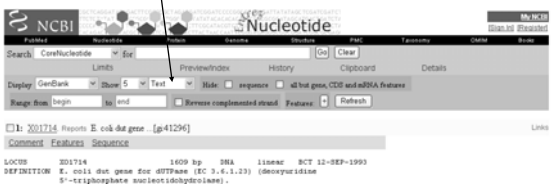
28.08.07

Example 3

3. Click the hyperlink *1* → a result page appears with a short definition for the nucleotide sequence entry
4. Click the *X01714* hyperlink →

Example 3

5. Choose *Text* in the *Show* pull-down menu to generate a **flat file** (poor text format)



6. Save your entry as a text document (with the name *EscherColi.txt*) on your hard drive by choosing *File* → *Save as* by your browser's main menu

Example 3

First Section of the flat file: **Header**

```
LOCUS       X01714 1609 bp    DNA     linear   BCT 12-SEP-1993
DEFINITION  E. coli dut gene for dUTPase (EC 3.6.1.23) (deoxyuridine
            5'-triphosphate nucleotidohydrolase).
ACCESSION   X01714
VERSION     X01714.1  GI:41296
KEYWORDS    dUTPase; unidentified reading frame.
SOURCE      Escherichia coli
  ORGANISM  Escherichia coli
            Bacteria; Proteobacteria; Gammaproteobacteria;
            Enterobacteriales;
            Enterobacteriaceae; Escherichia.
REFERENCE   1  (bases 1 to 1609)
  AUTHORS   Lundberg,L.G., Thoresson,H.O., Karlstrom,O.H. and Nyman,P.O.
  TITLE     Nucleotide sequence of the structural gene for dUTPase of
            Escherichia coli K-12
  JOURNAL   EMBO J. 2 (6), 967-971 (1983)
  PUBMED    6139280
COMMENT     Data kindly reviewed (25-NOV-1985) by L. Lundberg.
```


Example 3

- **LOCUS**: locus name, size of the Nucleotide sequence in base pairs, nature of the molecule (here DNA), its topology (linear or circular)
- **DEFINITION**: short definition, here *E. coli dut gene*
- **ACCESSION**: unique identifier
- **VERSION**: synonymous or past numbers
- **KEYWORDS**: terms that characterize the entry

Example 3

- **SOURCE**: common name of the relevant organism to which the sequence belongs
- **ORGANISM**: more complete identification of the organism
- **REFERENCE**: authors, titles, journals, ...
- **COMMENT**: info that doesn't fit in the previous sections

31 / 38

Example 3

Middle Section: **Feature table**

```
FEATURES             Location/Qualifiers
     source          1..1609
                     /organism="Escherichia coli"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:562"
     promoter        286..291 /note="-35 region"
     promoter        310..316 /note="-10 region"
     misc_feature    322..324 /note="put. transcription start region"
     RBS             330..333
                     /note="put. rRNA binding site"
     CDS             343..798
                     /note="unnamed protein product; dUTP-ase (aa 1-151)"
                     /codon_start=1
                     /transl_table=11
                     /protein_id="CAA25859.1"
                     /db_xref="GI:41297"
                     /db_xref="GOA:P06968"
                     /db_xref="InterPro:IPR008180"
     ...
```

32 / 38

Example 3

- **source**: indicates the origin of specific regions of the sequence
- **promoter**: shows the precise coordinates of a promoter element, in X01714 a -35 box is indicated from position 286 to 291
- **Misc feature**: (miscellaneous feature) indicates the putative location of the transcription start (mRNA synthesis)
- **RBS**: (Ribosome Binding Site) indicates the location of the last upstream element
- **CDS**: (CoDing Segment) describes the gene open reading frame

33 / 38

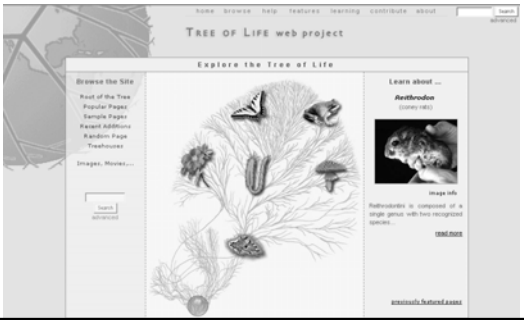
Exercise 3

Find the complete genome for the mitochondrion of the blue whale

34 / 38

Example 4: Tree of Life

1. Point your browser to the website **http://tolweb.org/** →



35 / 38

Example 4: Tree of Life

2. Click the **Root of the Tree** link →

36 / 38

Example 4: Tree of Life

Find the **path from the origin to human** by clicking

- Eukaryotes
- Animals
- Bilateria
- Deuterostomia
- Chordata
- Craniata
- Vertebrata
- Gnathostomata
- Sarcopterygii
- Terrestrial Vertebrates
- Amniota
- Synapsida
- Therapsida
- Mammalia
- Eutheria
- Primates
- Catarrhini
- Hominidae
- Homo
- Homo sapiens

Literature

- Claverie, J-M, Notredame, C.: *Bioinformatics for Dummies*. Wiley Publishing, Inc. 2003

example	pages
1	46 – 49
2	63 – 67
3	78 – 82