

## File format and Database tutorial.

### Outline of document:

1. File Format Representation
2. Search for genes in ENTRAZ:
3. What can we find out about the protein sequence?
4. Is there a structure available?
5. What type of kinetics are we talking about?

### File Format Representation:

These are 3 common ways to collect and represent a gene. These are designed to accommodate the information from more than one sequence.

#### 1. Fasta

```
>name of sequence 1
ATATAT(sequence 1)
>name of sequence 2
GCGCGGC(sequence 2)
```

#### 2. Flat

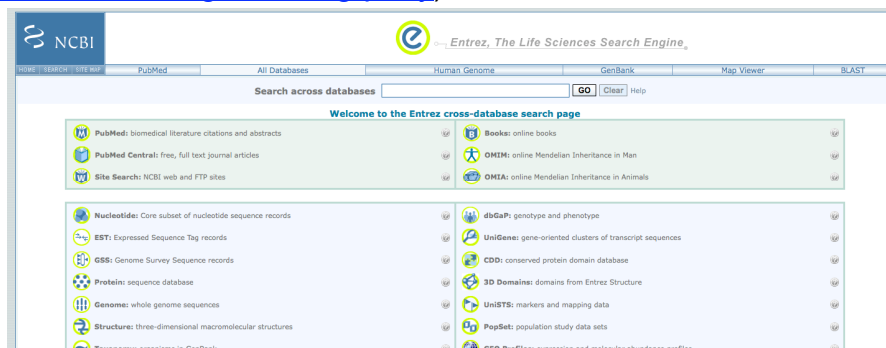
```
LOCUS      seq1
Definition  seq1, 16 base
ORIGIN      1 ATAGC
//
LOCUS      seq2
```

#### 3. EMBL

```
ID seq1
DE seq1, 16 base
SQ 16 base
      ATATAT
//
ID seq 2
```

### Search for genes in ENTRAZ:

(<http://www.ncbi.nlm.nih.gov/sites/gquery>)



## File Format and Database Tutorial

1. Search “Mycobacterium tuberculosis FABG”
2. Select “gene”
3. Click FABG H37Ra/MRA\_2791/NC\_009525.1 hit

The screenshot shows the NCBI Entrez Gene search results for the query "Mycobacterium tuberculosis FABG". The search bar at the top shows the query and the "Gene" filter is selected. Below the search bar, there are tabs for "Limits", "Preview/Index", "History", "Clipboard", and "Details". The "Display" section shows "Summary" selected, with "Show 20" and "Sort by Relevance". The results list shows 141 items, with the first two items displayed. The first item is "fabG" (3-ketoacyl-(acyl-carrier-protein) reductase [Mycobacterium tuberculosis H37Ra]) with other aliases: MRA\_0251, Annotation: NC\_009525.1 (292027..293391, complement), and GenID: 5214935. The second item is "fabG" (3-ketoacyl-(acyl-carrier-protein) reductase [Mycobacterium tuberculosis H37Rv]) with other aliases: Rv2766c, Annotation: NC\_000962.2 (3075588..3076370, complement), and GenID: 887727. A red arrow points to the first item.

4. Browse to bottom of screen, click “YP\_001284131.1 3-ketoacyl-...”

The screenshot shows the NCBI protein information page for the protein YP\_001284131.1. The page is titled "General protein information" and shows the protein name "3-ketoacyl-(acyl-carrier-protein) reductase" and the accession number "YP\_001284131.1". Below the protein name, there is a "COG classification" section with the following information: Description: COG1028 [Q] Dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases), Category: METABOLISM, Group: Secondary metabolites biosynthesis, transport and catabolism, and Catalyzes the first of the two reduction steps in the elongation cycle of fatty acid synthesis. The EC number is 1.1.1.100. Below the protein information, there is a section titled "NCBI Reference Sequences (RefSeq)" with a sub-section "mRNA and Protein(s)". The first entry is "YP\_001284131.1 3-ketoacyl-(acyl-carrier-protein) reductase [Mycobacterium tuberculosis H37Ra]". A red arrow points to this entry.

5. Browse to bottom of flat file, click “CDS”
6. Save Nucleotide sequence as FASTA (copy and paste the sequence into notebook).
7. Take note of the E.C #

## What can we find out about the protein sequence?

([www.expasy.org](http://www.expasy.org))

The screenshot shows the ExPASy Proteomics Server homepage. The page has a header with the ExPASy logo and the text "The ExPASy (Expert Protein Analysis System) proteomics server of the Swiss Institute of Bioinformatics (SIB) is dedicated to the analysis of protein sequences and structures as well as 2-D PAGE (Disclaimer / References / Linking to ExPASy)". Below the header, there are several sections: "Databases" (listing UniProt Knowledgebase, Swiss-2D PAGE, MIAPE, ENZYME, UniPathway, and SWISS-MODEL), "Tools and software packages" (listing Proteomics and sequence analysis tools, Melanie / ImageMaster, and MSlight), "Education and services" (listing The ExPASy FTP server, Swiss-Snap, Popular Science, and e-Proteomics), and "Links" (listing various links to related resources). A red arrow points to the "CDS" link in the "Databases" section.

1. Search for “fabg tuberculosis”
2. Click on the link to the protein (P0A5Y4 FABG\_MYCTU)

## File Format and Database Tutorial

All	Accession	Entry name	Status	Protein names	Gene names
<input type="checkbox"/>	P0A5Y5	FABG_MYCBO	★	3-oxoacyl-[acyl-carrier-protein] reductase (EC 1.1.1.100) (3-ketoacyl-acyl carrier protein reductase)	fabG (fabG1) (Mb1519)
<input type="checkbox"/>	P0A5Y4	FABG_MYCTU	★	3-oxoacyl-[acyl-carrier-protein] reductase (EC 1.1.1.100) (3-ketoacyl-acyl carrier protein reductase)	fabG (fabG1) (Rv1483) (MT1530) (MTCY277.04)
<input type="checkbox"/>	P69166	HSD_MYCBO	★	3-alpha-(or 20-beta)-hydroxysteroid dehydrogenase (EC 1.1.1.53)	fabG3 (Mb2025)
<input type="checkbox"/>	P69167	HSD_MYCTU	★	3-alpha-(or 20-beta)-hydroxysteroid dehydrogenase (EC 1.1.1.53)	fabG3 (Rv2002) (MT2058) (MTCY39.16c)
<input type="checkbox"/>	P66781	Y1350_MYCTU	★	Uncharacterized oxidoreductase Rv1350/MT1393 (EC 1.-.-.-)	fabG2 (Rv1350) (MT1393) (MTCY02B10.14)
<input type="checkbox"/>	P66782	Y1385_MYCBO	★	Uncharacterized oxidoreductase Mb1385 (EC 1.-.-.-)	fabG2 (Mb1385)

- Save the sequence as a FASTA file.
  - Click on FASTA under sequence. (copy and paste the sequence)
- Identify PI, seq length, MW, amino acid composition
  - Change tools from Blast to Protparam and click go.
  - Then click submit

**Sequences**

Sequence	Length	Mass (Da)	Tools
<input type="checkbox"/> P0A5Y4-1 [UniParc]. Last modified March 15, 2005. Version 1. Checksum: 70F6254B0FFCD47	247	25,697	Blast <input type="button" value="go"/>

FASTA

```

MTATATEGAK PPFVSRSLV TGNRGIGLA IAQRLAADGH KVAVTHRGSG APKGLFGVEC
70 80 90 100 110 120
DVTDSDAVDR APTAVEEHQ PVEVLVSNAQ LSADAFLMRM TEEKFEKVIN ANLTGAPRVA
130 140 150 160 170 180
QRASRSMQRN KFGRMIFIGS VSGSWGIGNQ ANYAASKAGV IGMARSIARE LSKANVTANV
190 200 210 220 230 240
VAPGYIDTDM TRALDERIQQ GALQFIPAKR VGTPAEVAGV VSFLASEDAS YISGAVIPVD
GGMGMGHH
  
```

[Hide](#)

- Save data from Protparam.

## Is there a structure available?

([www.pdb.org](http://www.pdb.org))

**RCSB PDB**  
PROTEIN DATA BANK

A MEMBER OF THE PDB MyPDB: Login | Register  
An Information Portal to Biological Macromolecular Structures  
As of Tuesday Jan 27, 2009 there are 55546 Structures | PDB Statistics

CONTACT US | FEEDBACK | HELP | PRINT PDB ID or keyword Author Site Search Advanced Search

**Home Search**

- Home
- Getting Started
- Structural Genomics
- Download Files
- Deposit and Validate
- Dictionaries & File Formats
- Software Tools
- General Education
- Site Tutorials
- BioSync
- General Information
- Acknowledgements
- Frequently Asked Questions

**A Resource for Studying Biological Macromolecules**

The PDB archive contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies. As a member of the [wwPDB](http://www.pdb.org), the RCSB PDB curates and annotates PDB data according to agreed upon standards.

The RCSB PDB also provides a variety of tools and resources. Users can perform simple and advanced searches based on annotations relating to sequence, structure and function. These molecules are visualized, downloaded, and analyzed by users who range from students to specialized scientists.

**Molecule of the Month: Tobacco Mosaic Virus**

Tobacco mosaic virus (TMV) has been at the center of virus research since its discovery over a hundred years ago. TMV was the first virus to be discovered. Late in the 19th century, researchers found that a tiny infectious agent, too small to be a bacterium, was the cause of a disease of tobacco plants. It then took 30 years of work before the nature of this mysterious agent became apparent. In a Nobel-prize-winning study, Wendell Stanley coaxed the virus to form crystals, and discovered that it was composed primarily of protein.

[Read more ...](#) [Previous Features](#)

**PSI Featured Molecule: CBS domain protein TA0289**

Researchers at the PSI MCGS have recently determined the structure of a protein with a new combination of two familiar protein modules, and made the first steps towards uncovering the function of this unusual new family of proteins.

[Read more from PSI SGKB](#) [Previous Features](#)

**News**

- Complete News
- Newsletter
- Discussion Forum
- Job Listings

27-January-2009  
**NJ Science Olympiad Protein Modeling Results**

Many ribonuclease models

- Search "FABG"
- Select 1UZL (MABA FROM MYCOBACTERIUM TUBERCULOSIS)

## File Format and Database Tutorial

The screenshot shows a list of PDB entries. The entry 1uzl is highlighted with a red arrow. The details for 1uzl are as follows:

Entry	Release Date	Exp. Method	Resolution	Classification	Compound	Authors
1uzl	23-Mar-2005	X Ray Diffraction	2.00 Å	Oxidoreductase	Polymer: 1 Molecule: 3-OXOACYL-[ACYL-CARRIER PROTEIN] REDUCTASE Fragment: 54-C TERMINUS, RESIDUES 54-304 Chains: A EC no.: 1.1.1.100	Urch, J.E., Wickramasinghe, S.R., Inglis, K.A., Muller, S., Fairlamb, A.H., Van Aalten, D.M.F.

### 3. Download PDB file and FASTA sequence of 1UZL



The screenshot shows the PDB entry page for 1uzl. The 'Download Files' section is expanded, and the 'FASTA Sequence' link is highlighted with a red arrow. The details for 1uzl are as follows:

Entry	Release Date	Exp. Method	Resolution	Classification	Compound	Authors
1uzl	23-Mar-2005	X Ray Diffraction	1.49 Å	Oxidoreductase	Polymer: 1 Molecule: 3-OXOACYL-[ACYL-CARRIER PROTEIN] REDUCTASE Mutation: YES Chains: A,B EC no.: 1.1.1.100	Cohen-Gonsaud, M., Ducasse, S., Quemard, A., Labesse, G.

### 4. Take note of the E.C #

**What type of kinetics are we talking about?**

(http://www.brenda-enzymes.info/)

EC-Number	Enzyme Name	Organism	Protein	Full text	Advanced Search
<input type="text"/> <input type="button" value="Search"/> Display <input type="text" value="10"/> entries					

New BRENDA release online (7th January 2009)

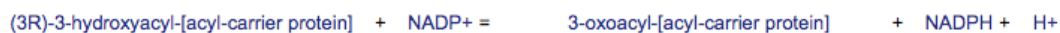
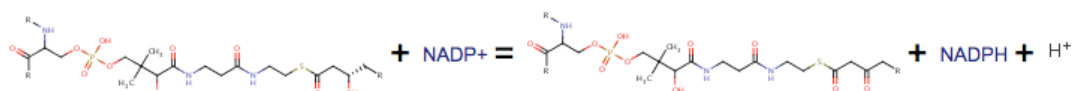
Latest paper on BRENDA (*Nucleic Acids Res.* 37, D588-D592, Jan. 2009) ;  
 BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009.

Nomenclature	Reaction & Specificity	Functional Parameters
Enzyme Names EC Number Common/ Recommended Name Systematic Name Synonyms CAS Registry Number	Pathway Catalysed Reaction Reaction Type Natural Substrates and Products Substrates and Products Substrates Natural Substrate Products Natural Product Inhibitors Cofactors Metals/Ions Activating Compounds Ligands	Km Value Ki Value IC50 Value pI Value Turnover Number Specific Activity pH Optimum pH Range Temperature Optimum Temperature Range
Isolation & Preparation		Organism-related information
Purification Cloned Renatured Crystallization		Organism Source Tissue Localization

1. Search "FABG"
2. Take note of EC #
3. Click first FABG link
4. Go to "synonyms" click on "657270" under literature.

SYNONYMS	ORGANISM	COMMENTARY	LITERATURE
3-ketoacyl acyl carrier protein reductase	-	-	-
3-ketoacyl-ACP(CoA) reductase	<a href="#">Mycobacterium tuberculosis</a>	-	<a href="#">657270</a>
3-ketoacyl-acyl carrier protein reductase	-	-	-
3-ketoacyl-acyl carrier protein reductase	<a href="#">Escherichia coli</a> , <a href="#">Pseudomonas sp.</a>	-	<a href="#">667257</a>
3-oxoacyl-ACP reductase	<a href="#">Plasmodium falciparum</a>	-	<a href="#">667520</a>
3-oxoacyl-[ACP]reductase	-	-	-
ACP reductase	<a href="#">Mycobacterium tuberculosis</a>	-	<a href="#">667708</a>
beta-ketoacyl acyl carrier protein (ACP) reductase	-	-	-
beta-ketoacyl acyl carrier protein reductase	<a href="#">Plasmodium falciparum</a>	-	<a href="#">668574</a>
beta-ketoacyl acyl carrier protein reductase	<a href="#">Pseudomonas aeruginosa</a>	-	<a href="#">656723</a>
beta-ketoacyl reductase	-	-	-
beta-ketoacyl reductase	<a href="#">Homo sapiens</a>	-	<a href="#">669416</a>
beta-ketoacyl reductase	<a href="#">Pseudomonas aeruginosa</a>	-	<a href="#">669470</a>
beta-ketoacyl thioester reductase	-	-	-
beta-ketoacyl-ACP reductase	-	-	-
beta-ketoacyl-ACP reductase	<a href="#">Escherichia coli</a>	-	<a href="#">669013</a>
beta-ketoacyl-ACP reductase	<a href="#">Plasmodium falciparum</a>	-	<a href="#">670774</a>
beta-ketoacyl-ACP reductase	<a href="#">Streptococcus pneumoniae</a>	-	<a href="#">667646</a>

5. Under "Reaction" heading. Click on the first beaker link (show reaction)



6. Identify and record: potential inhibitors of FABG, for mycobacterium optimal pH and temperature.