

BLAST and MSA Tutorial

Survey Course in Bio Crystallography and Bioinformatics
Lima, Peru, March 17-29, 2009

BLAST and MSA tutorial.

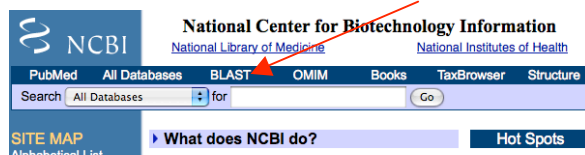
Outline of document:

1. How do we identify orthologs of FABG from *Mycobacterium tuberculosis*?
2. Once we find orthologs how do we compare them?
3. What can we learn from about the evolution of FABG by comparing orthologs?
4. What locations on FABG that are more important then others?

How do we identify orthologs of FABG from *Mycobacterium tuberculosis*? BLAST:

Basic local alignment search tool

0. Go to "NCBI" (www.ncbi.nih.gov)
1. click BLAST-> Protein Blast

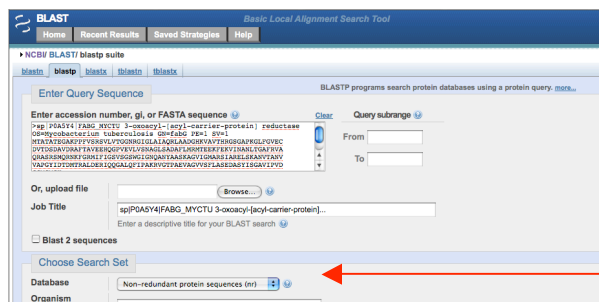


Basic BLAST

Choose a BLAST program to run.

- [nucleotide blast](#) Search a nucleotide database using a nucleotide query
Algorithms: blastn, megablast, discontinuous megablast
- [protein blast](#) Search protein database using a protein query
Algorithms: blastp, psi-blast, phi-blast

2. Paste your sequence of FABG in the box ensure non-redundant database is selected.



3. Blast
4. The return page contains 100 sequences that show homology with our protein. Each hit contains a reference link (ref) and a gene bank (gb) number that identifies each hit. They are ordered by an e-value, a chance value that the query protein and the protein identified from the database match by chance. The smaller the e-value the less likely the 2 proteins were matched by change. A good e-value cutoff of homology is 10^{-5} .
5. What is the first hit? How is this different then the next hit?
6. Find the FABG from "*Mycobacterium ulcerans* Agy99"
 - (don't close this window)

8. Open *ref* link in a new window/ tab

```
>ref|YP_905484.1| 3-oxoacyl-[acyl-carrier protein] reductase, FabG1 [Mycobacterium
ulcerans Agy99]
gb|ABL04013.1| 3-oxoacyl-[acyl-carrier protein] reductase, FabG1 [Mycobacterium
ulcerans Agy99]
Length=255

GENE ID: 4552452 fabG1 | 3-oxoacyl-[acyl-carrier protein] reductase, FabG1
(Mycobacterium ulcerans Agy99) (10 or fewer PubMed links)

Score = 461 bits (1186), Expect = 2e-128, Method: Compositional matrix adjust.
Identities = 226/246 (91%), Positives = 236/246 (95%), Gaps = 0/246 (0%)

Query 2 TATATGAKPPVSRSLVTGGNRGIGLAIAQLAADGHKAVVTHRGSGAPKGLFVCECD 61
      T +A +G KP FVSRSLVTGGNRGIGLAIAQLA DGH+VAVTHRGSGAP+GLFVCECD
Sbjct 10 TESAADGGKPAFVSRSLVTGGNRGIGLAIAQLATDHRVAVVTHRGSGAPGLFVCECD 69

Query 62 VTDSADVDRAPFAVEEHQGPVEVLVSAGLSADAFIMRTEEFKPKVINANTGAFRVAQ 121
      VTD+DAVDRAF VEEHQGPVEVLVSAGLSADAF+RMTEE+PKVI+ANLTGAFRVAQ
Sbjct 70 VTNDADVDRAPFAVEEHQGPVEVLVSAGLSADAFIMRTEEFKPKVIDANLTGAFRVAQ 129

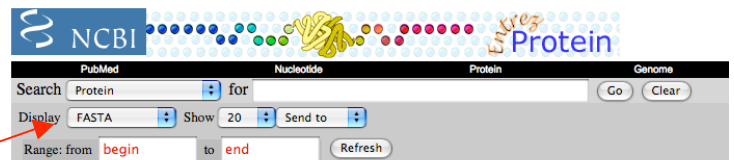
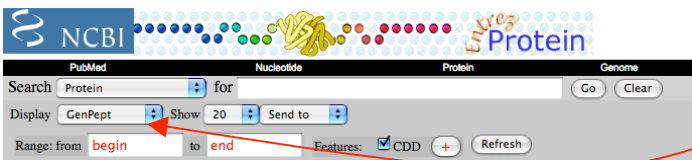
Query 122 RASRSMQRNKKFORMIFIGSVSGWIGNQANYAASKAGVIGMARSLARELSKANVTANVV 181
      RASRSMQR KFOR+IFIGSVSGWIGNQANYAASKAGVIGMARSLARELSR NVTANVV
Sbjct 130 RASRSMQRNKKFORMIFIGSVSGWIGNQANYAASKAGVIGMARSLARELSKANVTANVV 189

Query 182 AFGYIDTDMTRALDERIQGALQFIPAKRVGTAEVAGVVSFLASEDASYISGAVIPVDG 241
      AFGYIDTDMTRALDERIQ+GALQFIPAKRVGT AEVAGVVSFLASEDASYISGAVIPVDG
Sbjct 190 AFGYIDTDMTRALDERIQGALQFIPAKRVGTAEVAGVVSFLASEDASYISGAVIPVDG 249

Query 242 GHMGHGH 247
      GHMGHGH
Sbjct 250 GHMGHGH 255
```

9. Obtain the amino acid sequence of this protein in FASTA format. (save this in your notebook)

- Change “Display GenPept => Display FASTA”



- Copy and paste fasta file

Once we find orthologs how do we compare them?

- Go to “expasy.org”
- Click on “Alignment” link under Tools and software packages, this will link you to the page with all the alignment tools.

Tools and software packages

- Proteomics and sequence analysis tools
 - Identification and characterization (Aldente, FindMod, Popitam, Phenyx, pI/Mw, ProtParam...)
 - DNA -> Protein
 - Similarity searches (BLAST...)
 - Pattern and profile searches (ScanProsite...)
 - Post-translational modification and topology prediction
 - Primary structure analysis
 - Secondary and tertiary structure tools (Swiss-PdbViewer...)
 - Alignment and Phylogenetic analysis
- Melanie / ImageMaster - Software for 2-D PAGE analysis
- MSight - Mass Spectrometry Imager
- Roche Applied Science's Biochemical Pathways

- Click: SIM + LALNVIEW. A binary 2 sequence alignment tool

Sequence alignment

Binary

- SIM + LALNVIEW - Alignment of two protein sequences with SIM, results can be viewed with LALNVIEW
- LALIGN - Finds multiple matching subsegments in two sequences
- Dotlet - A Java applet for sequence comparisons using the dot matrix method

Multiple

- Decrease redundancy - Reduce a set of sequences into a non-redundant set

BLAST and MSA Tutorial

- Put the FABG sequence of Mycobacterium Tuberculosis in as "Sequence 1" and the Mycobacterium ulcerans as "Sequence 2".

SIM - Alignment Tool for protein sequences

SIM (References) is a program which finds a user-defined number of best non-intersecting alignments between two protein sequences or within a sequence. Once the alignment is computed, you can view it using LALNVIEW, a graphical viewer program for pairwise alignments [references].
Note: You can use the ACNUC server to align nucleic acid sequences with a similar tool.

Please enter two sequences. These sequences may either be specified by their SwissProt/EMBL accession numbers (AC), e.g. P08130, or by entry names (ID), e.g. adding your own sequences into the boxes below.

SEQUENCE 1:
☐ Swiss-Prot/EMBL AC or ID:
☐ User-defined sequence Sequence Name:
Paste your sequence below:

SEQUENCE 2:

- Submit these sequences for alignment and save the alignment in your notebook
- Go back to the SIM+LALNVIEW homepage
- Manipulate the default gap opening and extension penalties to find optimum alignment. Set the open penalty to 0 and extension penalty to 0.

Parameters:

Number of alignments to be computed:

Gap open penalty:

Gap extension penalty: (Note about definition of gap penalties.)

Comparison Matrix

- When we adjust the parameters what happens to the alignments?

What can we learn about the evolution of FABG from M. tuberculosis? (Multiple sequence alignment MSA)

- Open a new browser window
 - Go to the expasy home page, www.expasy.org
 - Click alignment
 - Go back to BLAST search results
- Select 20 sequences from different mycobacterium species and other genus. Select sequences that have approximately the same length as our inputted sequence.
 - Select "get sequences"

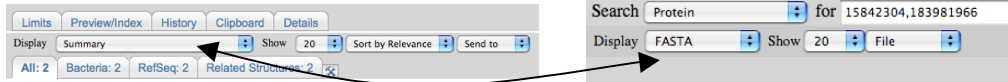
Alignments ☐ Select All

☒ ref|NP_337341.1| G 3-ketoacyl-(acyl-carrier-protein) reductase [Mycobacterium tuberculosis CDC1551]
ref|NP_056434.1| G 3-ketoacyl-(acyl-carrier-protein) reductase [Mycobacterium bovis AF2122/97]
ref|YP_177905.1| G 3-ketoacyl-(acyl-carrier-protein) reductase [Mycobacterium tuberculosis H37Rv]
*21 more sequence titles
Length=260
GENE ID: 925450 fabg | 3-ketoacyl-(acyl-carrier-protein) reductase [Mycobacterium tuberculosis CDC1551] (10 or fewer PubMed links)
Score = 520 bits (1339), Expect = 5e-146, Method: Compositional matrix adjust.
Identities = 260/260 (100%), Positives = 260/260 (100%), Gaps = 0/260 (0%)
Query 1 MTSLDLTGRTAITGASRGICLAIQAQLAAGAHVLTARRQEADEAAQVGDRLGVG 60
Sbjct 1 MTSLDLTGRTAITGASRGICLAIQAQLAAGAHVLTARRQEADEAAQVGDRLGVG 60
Query 61 AHAVEDEARRCVLTLEKPGSDVLLINNAGTNPATGPLEQDHARFAKIFQVNLWAPLM 120
Sbjct 61 AHAVEDEARRCVLTLEKPGSDVLLINNAGTNPATGPLEQDHARFAKIFQVNLWAPLM 120
Query 121 MTSLVVYANMHEGAVVNTASIGMHIQSFANGMYNATKAALIHVTQGLALESPRIYRN 180
Sbjct 121 MTSLVVYANMHEGAVVNTASIGMHIQSFANGMYNATKAALIHVTQGLALESPRIYRN 180
Query 181 AICPQVVTSLAEALWDHEDPLAATIALGRIGEPADIASAVFLVSDAASWITGETHII 240
Sbjct 181 AICPQVVTSLAEALWDHEDPLAATIALGRIGEPADIASAVFLVSDAASWITGETHII 240
Query 241 DGILLGNALGFRAPSTEH 260
Sbjct 241 DGILLGNALGFRAPSTEH 260

☒ ref|YP_001850257.1| G short-chain type dehydrogenase/reductase [Mycobacterium marinum M]
ref|ACC04042.1| G short-chain type dehydrogenase/reductase [Mycobacterium marinum M]
Length=260
GENE ID: 6226212 MMAR_1952 | short-chain type dehydrogenase/reductase [Mycobacterium marinum M] (10 or fewer PubMed links)
Score = 466 bits (1139), Expect = 1e-139, Method: Compositional matrix adjust.
Identities = 226/258 (87%), Positives = 244/258 (94%), Gaps = 0/258 (0%)
Query 1 MTSLDLTGRTAITGASRGICLAIQAQLAAGAHVLTARRQEADEAAQVGDRLGVG 60

BLAST and MSA Tutorial

- This will bring you to NCBI results page
- On the NCBI results page at the top click
 - Change Display =[fasta]
[send to]= file
 - Save the sequence.fasta file to your folder



- Open the sequences file in a text editor.

On the expasy site:

- Open a new window/ tab with each of these sites.

Sequence alignment

Binary

- **SIM + LALNVIEW** - Alignment of two protein sequences with SIM, results can be viewed with **LALNVIEW**
- **LALIGN** - Finds multiple matching subsegments in two sequences
- **Dotlet** - A Java applet for sequence comparisons using the dot matrix method

Multiple

- **Decrease redundancy** - Reduce a set of sequences into a non-redundant set
- **Nomad (Neighborhood Optimization for Multiple Alignment Discovery)** - Ungapped local multiple alignment, optimized
- **CLUSTALW** [At **EBI**, **PBIL**, **My Hits** or at **EMBnet-CH**]
- **KALIGN** - an accurate and fast multiple sequence alignment algorithm [At **Karolinska Institute** or at **EBI**]
- **MAFFT** [At **Kyushu University**, **EBI** or at **MyHits**]
- **Muscle** [At **Berkeley** or at **BioAssist**]
- **T-Coffee** [At **MyHits**, **BioAssist** or at **EBI**]
- **MSA** - at Genestream (IGH)
- **DIALIGN** - Multiple sequence alignment based on segment-to-segment comparison, at University of Bielefeld, Germany
- **Match-Box** - at University of Namur, Belgium - at Washington University
- **Multalin** [At **GenoToul Bioinfo** or at **PBIL**]
- **MUSCA** - Multiple sequence alignment using pattern discovery. at IBM

- Click on: **Clustalw**-> EMBnet-CH
MUSCLE-> Berkeley
T-coffee->EBI
- Copy and Paste the 20 FABG sequences, in fasta format, into the online forms.
- *Goal of alignment*: You should minimize the amount of large gaps and maximize the amount of conserved regions.
- Download and save the alignment from the *ClustalW* site "clustalw (aln)". The alignments from the other sites are used to answer the questions below. Where we will continue to use the clustalw alignment file for further analysis.

Output from clustalW at EMBnet-CH

Here are your search results:

| Multiple alignments | Dendrograms |
|--------------------------------|--------------------------------|
| clustalw (aln) | clustalw (dnd) |
| GCG/MSF | |
| PIR | |
| GDE | |
| phylip | |

Answer these questions:

13. Where are gaps opened and closed?
14. What regions of the sequence alignment are conserved?
15. What regions are variable?

Are there sites in the sequence that have a high evolutionary rate? MSA Sequence analysis

- From expasy.org
- Click ->Alignment analysis -> SVA (Sequence Variability Analyzer)

BLAST and MSA Tutorial

Alignment analysis

- **AMAS** - Analyse Multiply Aligned Sequences
- **Bork's alignment tools** - Various tools to enhance the results of multiple alignments (including consensus building).
- **CINEMA** - Color Interactive Editor for multiple alignments
- **ESPrnt** - Tool to print a multiple alignment
- **MaxAlign** - Post-processing of alignments by removing sequences (taxa) with many gaps
- **PhyloGibbs** - Gibbs motif sampler incorporating phylogeny and tracking statistics
- **SVA** - Sequence Variability Analyser for multiple alignments
- **PVS** - a protein variability server optimized for conserved epitope discovery

- Paste your alignment file sequence into the text box or upload it.
- Click go
 1. Where are the alignment Differences
 2. Where are the alignment Similarities
 3. Save results in lab book.

What does this conservation look like on a structure?

- Go to <http://consurf.tau.ac.il/>

The ConSurf Server

Server for the Identification of Functional Regions in Proteins
Version 3.0 now with *Jmol Viewer* (Sep-06)

- You know from searching the PDB that the FABG from *M. tuberculosis* is 1UZL (1UZM).
-
- Step 1/2: Enter PDB ID: 1UZM with Chain A. ignore any precompiled results that pop up.
- Step 3: Enter the MSA file (alignment file from clustalw)
- Step 4: Enter the ID of the first value in the sequence alignment
- Step 5: Enter email address.
- Step 6: Submit.

The screenshot shows the ConSurf Server web interface. It has several sections for inputting data. Step 1 points to the 'Protein Structure (obligatory)' section where 'Enter the PDB ID:' is set to '1UZL'. Step 2 points to the 'Chain Identifier' dropdown menu which is set to 'A'. Below this is a link 'Enter your own PDB file' with a 'Browse...' button. Step 3 points to the 'User-Provided Multiple Sequence Alignment (optional)' section where 'Enter your own MSA file' has a 'Browse...' button. Step 4 points to the 'Query sequence name in MSA file' input field which contains 'ref|NP_337341.1|'. Step 5 points to the 'User-Provided TREE file (optional)' section where 'Enter your own TREE file (in Newick format)' has a 'Browse...' button. At the bottom, there is a field for 'Please enter your e-mail address:' and a checkbox for 'Send a link to the results by e-mail'. At the very bottom are 'Submit' and 'Clear' buttons.


- Give the submission some time. It will be done when it says "ConSurf job status page –Finished".

BLAST and MSA Tutorial

- Click on “view consurf results with First glance in Jmol”
- Wait for the webpage to load.

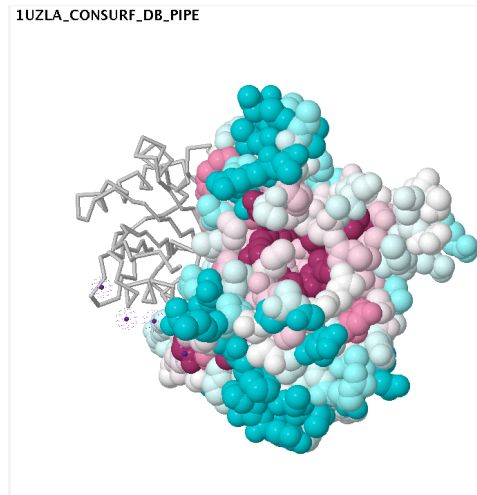
Final Results

[View ConSurf Results with FirstGlance in Jmol](#)
[View ConSurf Results with Protein Explorer \(Windows only\)](#)
(Which is best?)

 Highly recommended for Chimera users:
[View ConSurf results with Chimera](#) ([Download Chimera](#))

Jmol Commands

16. Left mouse button to rotate the molecule.
17. All the options to manipulate the molecule are under a menu through the Right mouse button.



Conservation of amino acids are plotted by color onto the structure.

Where are the highly conserved amino acids located?

Where are the least conserved amino acids located?

Can we build a fingerprint that represents these sequences?

FABG belongs to a larger family of proteins called Short Chain Oxidoreductase enzymes. These enzymes share a structural motif called the Rossmann fold. Which is used to bind the NAD(P) cofactors. They also have many sequence-conserved motifs in common.

Table below describes some of these motifs:

SCOR conserved motifs

| Motifs | Usage |
|----------|-----------------------|
| TGxxxGxG | Nucleotide binding |
| VxNAG | Structural, catalytic |
| YxxxK | Catalytic |
| PGxxxT | Catalytic |
| GG | Structural |

Each amino acid is denoted by a single letter code, the x indicates any amino acid. The usage column defines what the motif is used for.

1. Identify these 5 motifs in your sequence alignment.

2. Approximately where are they?
3. How many amino acids between the TGxxxGxG motif and VxNAG motif? How many between VxNAG and YxxxK? YxxxK and PGxxxT? PGxxxT and GG?
TGxxxGxG [] VxNAG
VxNAG[] YxxxK
YxxxK[] PGxxxT
PGxxxT[] GG
4. Compare these motifs with those generated from SVA analysis.
5. Since we know the motifs are present in our sequence and we know how many positions between each motif we can generate a fingerprint for this family that can be used to search for other members.