



## Teaching NCBI Resources

## Through Case Studies



# Five Examples for NCBI BLAST

## Introduction

BLAST programs from NCBI are powerful sequence alignment tools widely used in the analyses of biological sequences. In this booklet, we will work through five representative approaches that apply different functions provided by NCBI BLAST services ([blast.ncbi.nlm.nih.gov](http://blast.ncbi.nlm.nih.gov)) to address specific biological questions. Those examples will help you familiarize yourself with the web interface, better understand the capability of different BLAST programs, and learn the different result presentation formats plus their applications. With diverse sequence collections, you should be able to create your own working examples for use in your own teaching, or use different combinations of demonstrated capabilities to address your research needs.

## Table of Content

<b>Introduction</b>	... p.2
<b>Table of Content</b>	... p.2
<b>Case 1: Identify Unknown Bacteria using the 16S rRNA BLAST Database</b>	... p.3
<b>Case 2: Identify a PCR Primer Set for Amplifying the Coding Region of an mRNA Transcript</b>	... p.4
<b>Case 3: Generate Species and Gene Phylogenetic Trees</b>	... p.6
i) Ape Phylogeny	
ii) Creatine Kinase Protein Tree	
<b>Case 4: Annotate a Metagenomic Contig</b>	... p.10
<b>Case 5: Examine Conserved Domains and Solved Structures to Support a Protein Annotation</b>	... p.11
<b>Appendix</b>	... p.11

## Case 1: Identify Unknown Bacteria Using the 16S rRNA BLAST Database

### Goal

To validate the identity of an unknown bacterial sample using the 16S rRNA BLAST database

### Background

- Useful in microbiology lectures and laboratory courses
- About 5 minutes to complete the exercise
- A common exercise in microbiology is to identify a bacterial sample based on biochemical and growth properties. Another approach is to use targeted PCR amplification and analysis of genomic variations in the 16S rRNA gene to validate and even identify microbial samples. The 16S rRNA gene has a very conserved sequence overall that maintains its structure and role as a scaffold for the small subunit of the ribosome. This enables the use of “universal” primers for PCR amplification, but the gene contains a few regions of variability that allow it to be exploited for identification of microbial species.

### Steps

- From the BLAST home page ([blast.ncbi.nlm.nih.gov](http://blast.ncbi.nlm.nih.gov)), click **Nucleotide BLAST** to open the search page (A). You may want to use Reset Page (B) to return to default settings.
- Choose one or more sequences of a 16S rRNA region for an “unknown” bacterial sample. Retrieve those sequences from the FTP directory for this course ([bit.ly/2suC3lx](http://bit.ly/2suC3lx)). These sequences mimic the results of PCR amplification by a set of microbial universal primers.
- Paste one or more sequences into the **Enter Query Sequence** box (C), or upload the file through **Choose file** button (D).
- Select the **16S ribosomal RNA sequences** database using the database pull-down menu (E).
- Click the **BLAST** button to run the search.

Search results are shown on the next page.

The screenshot shows the NCBI BLAST Nucleotide BLAST interface. Key elements include:

- BLAST** logo and navigation links (Home, Recent Results, Saved Strategies, Help).
- Standard Nucleotide BLAST** header.
- Enter Query Sequence** section with a text input field (B) and a **Choose File** button (D).
- Choose Search Set** section with a **Database** dropdown menu (E) set to **16S ribosomal RNA sequences (Bacteria and Archaea)**.
- Program Selection** section with **Optimize for** options: **Highly similar sequences (megablast)**, **More dissimilar sequences (discontiguous megablast)**, and **Somewhat similar sequences (blastn)**.
- BLAST** button at the bottom left.

Annotations A through E are placed on the page to indicate the steps described in the text. A red arrow points from the dropdown menu in the 'Choose Search Set' section to the '16S ribosomal RNA sequences (Bacteria and Archaea)' option.

## Case 1 (cont.)

### Interpretation

- It is common for this conserved region to match many bacterial sequences. However, by identifying those closest to the query sequence, it is possible to establish a likely identity for the unknown sample.
- The **Descriptions** section of the BLAST report provides a quick view of the results (A). For very similar results like these, Max score is often, but not always, the important statistic. Consider the percent identity and query coverage, and confirm identification by looking in the **Alignments** section (B).
- You can use the **Formatting options** menu near the top of the results page to more easily compare the alignments (such as) by changing the **Alignment View** to one of the *...with dots for identities* formats (no illustrated).
- Finally, the **Distance tree of results** link (C) provides a quick, BLAST-based phylogenetic tree of the alignments. This is another way to identify the sequence that is most similar to the unknown sample.

**BLAST** » blastn suite » RID-894HX93S014

Home Recent Results Saved Strategies Help

BLAST Results

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

Job title: 7 sequences (Shigella dysenteriae)

Results for: 1:|c|Query\_141321 Shigella dysenteriae(1487bp)

RID: 8965H3M1014 (Expires on 02-16 03:14 am)

Query ID: |c|Query\_51349  
Description: unknown Shigella  
Molecule type: nucleic acid  
Query Length: 1487

Database Name: rRNA\_tpestrains/prokaryotic\_16S\_ribosomal\_RNA  
Description: 16S ribosomal RNA (Bacteria and Archaea)  
Program: BLASTN 2.8.0+ Citation

Other reports: Search Summary Taxonomy reports Distance tree of results

Graphic Summary

Descriptions

Sequences producing significant alignments:

Select: All None Selected: 0

Description	Max score	Total score	Query cover	E value	Ident	Accession
Shigella dysenteriae strain ATCC 13313 16S ribosomal RNA gene, partial sequence	2747	2747	100%	0.0	100%	NR_026332.1
Shigella flexneri strain ATCC 29903 16S ribosomal RNA gene, partial sequence	2669	2669	100%	0.0	99%	NR_026331.1
Escherichia fergusonii strain ATCC 35469 16S ribosomal RNA, complete sequence	2658	2658	100%	0.0	99%	NR_074902.1
Escherichia marmotae strain HT073016 16S ribosomal RNA, partial sequence	2652	2652	100%	0.0	99%	NR_136472.1
Citrobacter amalonaticus strain CECT 863 16S ribosomal RNA gene, partial sequence	2542	2542	100%	0.0	98%	NR_104823.1
Kosakonia sacchari strain SP1 16S ribosomal RNA, partial sequence	2529	2529	100%	0.0	97%	NR_118333.1
Enterobacter massiliensis strain JC163 16S ribosomal RNA gene, partial sequence	2525	2525	100%	0.0	97%	NR_125600.1
Salmonella enterica subsp. enterica strain Ty2 16S ribosomal RNA, partial sequence	2519	2519	100%	0.0	97%	NR_074799.1
Salmonella enterica subsp. arizonae strain ATCC 13314 16S ribosomal RNA gene, partial sequence	2519	2519	100%	0.0	97%	NR_041696.1
Salmonella enterica subsp. enterica strain LT2 16S ribosomal RNA, partial sequence	2514	2514	100%	0.0	97%	NR_074910.1

Download GenBank Graphics

Shigella dysenteriae strain ATCC 13313 16S ribosomal RNA gene, partial sequence

Sequence ID: NR\_026332.1 Length: 1487 Number of Matches: 1

Range 1: 1 to 1487 GenBank Graphics

Score	Expect	Identities	Gaps	Strand
2747 bits(1487)	0.0	1487/1487(100%)	0/1487(0%)	Plus/Plus

Query 1 TGGCTCAGATTGAACGCTGGCGGCAAGCCTAACACATGCAAGTCGAACGGTAACAGAAAAG 60  
Sbjct 1 TGGCTCAGATTGAACGCTGGCGGCAAGCCTAACACATGCAAGTCGAACGGTAACAGAAAAG 60

Query 61 CAGCTTGTGTTTGTCTGACGAGTGGCGGACGGGTGAGTAATGCTGGGAAACTGCCTGAT 120  
Sbjct 61 CAGCTTGTGTTTGTCTGACGAGTGGCGGACGGGTGAGTAATGCTGGGAAACTGCCTGAT 120

Query 121 GGAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATAACGTCGCAAGACCAAGAG 180  
Sbjct 121 GGAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATAACGTCGCAAGACCAAGAG 180

Find: all

Tools Upload

primates

Select

Collapse

Reroot tree

Show subtree

Show Alignment

leaf-count: 21

enterobacteria | 9 leaves

enterobacteria | 7 leaves

enterobacteria | 52 leaves

enterobacteria | 14 leaves

enterobacteria | 5 leaves

enterobacteria | 10 leaves

Escherichia marmotae strain HT073016 16S ribosomal RNA, partial sequence

unknown Shigella

Shigella dysenteriae strain ATCC 13313 16S ribosomal RNA gene, partial sequence

Trabusiella guamensis strain NBRC 103172 16S ribosomal RNA gene, partial sequence

Hover over a node or a collapsed branch triangle to see a popup menu (insert) for tree manipulation.

The display has non-query containing branches collapsed.

Success Nodes 201(0 selected) View port at (0,0) of 926x827 8.005



## Case 2: Identify a Pair of PCR Primers for Amplifying the Coding Region of an mRNA

### Goal

Use Primer-BLAST to find a PCR primer pair that can be used in the laboratory to amplify a coding region of an mRNA

### Background

- Useful in molecular biology and biochemistry courses
- About 10 minutes to complete the exercise

A common laboratory exercise in molecular biology and biochemistry courses is to design PCR primers for a target sequence. This target sequence is often a protein coding sequence that can be subsequently ligated into an expression plasmid and then used for other lab sessions, such as expression and characterization of the protein, mutagenesis, or promoter analysis.

### Steps

- Retrieve [NM\\_000250](#), the RefSeq mRNA sequence for Human Myeloperoxidase, MPO (A) from the Nucleotide database
- Use the web browser's *Find in page* function (ctrl+F) to find the **CDS** feature. The coding sequence (CDS) for this gene starts at position 178 and ends at position 2415. It is 2238 nucleotides long (B). Write down these positions because you'll need them in subsequent steps.
- On the right side of the record under **Analyze this sequence**, click **Pick Primers** (C) to open the Primer-BLAST page with the accession already entered as the template.
- To amplify the CDS region, set the ranges for **Forward** and **Reverse primers** outside of the CDS positions in the record. For forward primer range set the range from 138 to 178. For the reverse primer, set the range from 2415 to 2500 (D). Also, adjust the **PCR product size** by increasing the **Max** size to 2500 (E), so that the entire CDS amplifies.

- Click **Get Primers** to run the search (not illustrated), which uses the default database RefSeq mRNA database limited to Human.

**Homo sapiens myeloperoxidase (MPO), mRNA** (A)

NCBI Reference Sequence: NM\_000250.1

FASTA Graphics

Go to: (x)

LOCUS NM\_000250 3215 bp mRNA linear PRI 22-JAN-2018

DEFINITION Homo sapiens myeloperoxidase (MPO), mRNA.

ACCESSION NM\_000250

VERSION NM\_000250.1

FEATURES

source	Location/Qualifiers
1..3215	/organism="Homo sapiens" /mol_type="mRNA" /db_xref="taxon:9606" /chromosome="17" /map="17q22"
gene	1..3215 /gene="MPO" /note="myeloperoxidase" /db_xref="GeneID:4353" /db_xref="HGNC:7218" /db_xref="MIM:606989"
exon	1..331 /gene="MPO" /inference="alignment:Splign:2.0.8"
misc_feature	127..129 /gene="MPO" /note="upstream in-frame stop codon"
<b>CDS</b>	<b>178..2415</b> /gene="MPO" /EC_number="1.11.2.2" /codon_start=1 /product="myeloperoxidase precursor" /protein_id="NP_000241.1"

(B)

**CDS 178..2415 /gene="MPO"**

Change region shown

Customize view

Analyze this sequence

Run BLAST

Pick Primers (C)

Highlight Sequence Features

Find in this Sequence

Show in Genome Data Viewer

Range

	From	To
Forward primer	138	178
Reverse primer	2415	2500

(D)

Reset page Save search parameters Retrieve recent results Publication Finding specific primers

PCR Template

Enter accession, gi, or FASTA sequence (A refseq record is preferred) Clear

NM\_000250.1

Or, upload FASTA file Choose File No file chosen

Range

	From	To
Forward primer	138	178
Reverse primer	2415	2500

Primer Parameters

Use my own forward primer (5'->3' on plus strand) Clear

Use my own reverse primer (5'->3' on minus strand) Clear

PCR product size

Min	Max
70	2500 (E)

# of primers to return

10

Primer melting temperatures (T<sub>m</sub>)

Min	Opt	Max	Max T <sub>m</sub> difference
57.0	60.0	63.0	3

## Case 2 (cont.)

### Interpretation

- It is important to find PCR Primer pairs that will amplify only the sequence intended. In this case, the selected PCR Primer pairs should amplify only the Human MPO transcript.
- The top of the results page (A) summarizes the Primer BLAST search. The **Detailed primer reports** section (B) lists a set of primer pairs and their key characteristics for you to select from.
- You can find suggestions on how to get primers specific to your template from the tips page, at [www.ncbi.nlm.nih.gov/tools/primer-blast/search\\_tips.html](http://www.ncbi.nlm.nih.gov/tools/primer-blast/search_tips.html)
- It is most important for both primers in a pair to have similar T<sub>m</sub> values and GC percentages when possible. In addition, self-complementarity should be low to prevent primers binding to themselves and each other, rather than the template.
- By default, Primer BLAST uses stringent parameters. You can relax them if you are not able to find any suitable primers. Be aware that this may increase the potential of misprimed amplification due to annealing to secondary annealing sites on other templates.

**Primer-BLAST** Primer-Blast results

NCBI/Primer-BLAST : results: Job id=-PlnXxwjEYs2tRSwGdAwgmPLIbB02DqtTw [more...](#)

**Input PCR template** NM\_000250.1 Homo sapiens myeloperoxidase (MPO), mRNA  
**Range** 138 - 2500  
**Specificity of primers** Primer pairs are specific to input template as no other targets were found in selected database: Refseq mRNA (Organism limited to Homo sapiens)

**Other reports** [Search Summary](#)

### Graphical view of primer pairs

Genes - Exon

NP\_000241.1

Primer pairs for job -PlnXxwjEYs2tRSwGdAwgmPLIbB02DqtTw

Primer 1  
Primer 2  
Primer 3  
Primer 4  
Primer 5

### Detailed primer reports

**Primer pair 1**

Sequence (5'→3')	Template strand	Length	Start	Stop	T <sub>m</sub>	GC%	Self complementarity	Self 3' complementarity
Forward primer	Plus	21	141	161	59.99	52.38	4.00	2.00
Reverse primer	Minus	21	2451	2431	59.58	52.38	5.00	3.00

Product length 2311

Products on intended target  
 >NM\_000250.1 Homo sapiens myeloperoxidase (MPO), mRNA

product length = 2311  
 Forward primer 1 GGTACAAAGGGGATTGAGCA 21  
 Template 141 ..... 161  
 Reverse primer 1 ATATACCCCTCACTGCTGCAC 21  
 Template 2451 ..... 2431

**Primer pair 2**

Sequence (5'→3')	Template strand	Length	Start	Stop	T <sub>m</sub>	GC%	Self complementarity	Self 3' complementarity
Forward primer	Plus	21	143	163	59.99	52.38	4.00	2.00
Reverse primer	Minus	21	2457	2437	59.58	52.38	5.00	3.00

Product length 2315

Products on intended target  
 >NM\_000250.1 Homo sapiens myeloperoxidase (MPO), mRNA

product length = 2315  
 Forward primer 1 TACAAAGGGGATTGAGCAGC 21  
 Template 143 ..... 163  
 Reverse primer 1 GCCCAGATATACCCCTCACTG 21  
 Template 2457 ..... 2437

### Graphic Summary

Distribution of the top 2 Blast Hits on 1 subject sequences

Mouse over to see the title, click to show alignments

**Color key for alignment scores**

Score Range	Color
<40	Black
40-50	Blue
50-80	Green
80-200	Magenta
>=200	Red

Query

1 10 20 30 40 50

### 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
Homo sapiens myeloperoxidase (MPO), mRNA	39.9	79.8	72%	0.020	100%	NM_000250.1

To confirm the specificity, you can run a Nucleotide BLAST search, with program set to *blastn*, against the human RefSeq RNA dataset. To force BLAST to align each primer independently in a single search, while preserving their spatial relationship, make sure you provide the primer set in this example format, **GGTACAAAGGGGATTGAG-CANNNNNNNNNNNNNNNNNATATACCCCTCACTGCTGCAC**. In the result display, the primers matching to different parts of the same target sequence have a thin line connecting them (C).

## Case 3: Generate Species and Gene Phylogenetic Trees

### Goal

Use blastn and blastp to find homologous molecules and generate distance trees.

### Background

- Useful in general biology, molecular biology, and vertebrate zoology courses
- About 15 min for each of two examples
- Example **i)** generates a phylogeny of apes using complete mitochondrial genome sequences. Example **ii)** builds a gene (protein) tree for the creatine kinases, a small protein family with four or more members in vertebrate proteomes.

### i) Ape Phylogeny

#### Steps:

- Retrieve the ring-tailed lemur mitochondrial genome sequence, accession number [NC\\_004025.1](#), from the Nucleotide database. You can use this sequence as a query to retrieve and align the ape mitochondrial genomes using blastn.
- Click **Run BLAST** (A) on that Nucleotide page to load the blast search form.
- Select the **RefSeq Genomic sequences (refseq\_genomic)** as the database (B). This database contains all genomic sequences from NCBI's RefSeq project. The information icon ? links to a detailed description of the database.
- Paste in the following list of accessions for apes mitochondrial genomes to the **Entrez Query** box (C):  
`NC_001643 OR NC_001644 OR NC_001645 OR NC_001646 OR NC_002082 OR NC_002083 OR NC_011120 OR NC_011137 OR NC_012920 OR NC_013993 OR NC_014042 OR NC_014045 OR NC_014047 OR NC_014051 OR NC_018753 OR NC_021957 OR NC_023100 OR NC_033882 OR NC_033883 OR NC_033884 OR NC_033885`
- Adjust the BLAST program to **More dissimilar sequences** (D), expand the **Algorithm parameters** section and set the **Expect** threshold to  $1e-64$  (E). A page with the above setting is at <http://bit.ly/2qBBJo4>
- Click **BLAST** button to submit the search.
- The results showed nearly full-length matches to the lemur query. These include mitochondrial genome sequences from gorillas, chimpanzees, orangutans, gibbons, and four distinct taxa in the genus *Homo* – modern humans, plus three extinct groups: the *Neanderthal* and *Denisovan* hominids as well as *Homo heidelbergensis* (not shown). The Taxonomy report (F) shows the taxa represented in the output.
- Click the **"Distance tree of results"** link to generate a tree.

## Case 3 (cont.)

### Interpretation:

- The tree supports the two distinct groups of apes: the Great apes (*Hominidae*, **A**) containing humans, chimpanzees, gorillas and orangutans, and gibbons (*Hylobatidae*, **B**). It also shows the chimpanzee (*Pan troglodytes*) and the bonobo (*Pan paniscus*) as the closest living relatives of humans and the Neanderthal as the closest extinct relative (**C**).
- Note that this tree is based completely on blastn's local and pairwise comparisons to the query (lemur) sequence. It produces a reasonable alignment for generating the tree due to overall conservation in the mitochondrion genomes for this group of organisms. The most accurate tree, however, requires a true multiple sequence alignment (using a tool such as **MUSCLE**) for nucleotide sequences. NCBI does not have a separate nucleotide multiple alignment tool. Example ii) below uses a true protein multiple alignment through COBALT to generate a protein tree.



### ii) Creatine Kinase protein tree

#### Steps

- Retrieve human creatine kinase B-type protein, [NP\\_001814.2](#), from the Protein database. Use it as a blastp query to retrieve the tetrapod vertebrate creatine kinases, then perform a multiple sequence alignment for the matched sequences using COBALT to make a protein tree.
- Click **Run BLAST** on that Protein page to load the blastp search form (as described in Case 3i).
- Change the database to **Reference proteins**, which contains NCBI RefSeqs used in or generated by the NCBI genome annotation pipelines. Click the information icon **?** to see more information.
- Type **tetrapods** in the Organism box and select **taxid:32523** from the list. This limits the search to sequences from this group of organisms.
- Eliminate predicted entries by checking the **Exclude** box for **Models** to get a smaller tree.
- Expand the **Algorithm parameters** section and set the Expect threshold to **1e-55** to further limit hits returned by BLAST. Click the **BLAST** button to run the search.



## Case 3 (cont.)

### ii) Creatine Kinase Protein Tree

#### Steps

- In the results page, click **Multiple alignment (A)** to send matched proteins to COBALT for true multiple sequence alignment. Note that the multiple alignment will involve residues that were not present in the local BLAST alignments, such as the signal peptide from the mitochondrial targeted proteins.
- From the COBALT results, click the **Phylogenetic Tree (B)** at the top to generate the protein tree.

**BLAST Results**

Your search is limited to records that include: tetrapods (taxid:32523) ; and exclude: models (XM/XP) ▶ Full Entrez Query

Edit and Resubmit Save Search Strategies ▶ Formatting options ▶ Download YouTube How to read this page Blast report

**Job title: (2) - NP\_001814:creatine kinase B-type [Homo sapiens]**

RID 8MWR11F5014 (Expires on 02-20 13:46 pm)

Query ID	NP_001814.2	Database Name	refseq_protein
Description	creatine kinase B-type [Homo sapiens]	Description	NCBI Protein Reference Sequences
Molecule type	amino acid	Program	BLASTP 2.8.0+ ▶ Citation
Query Length	381		

Other reports: ▶ Search Summary [Taxonomy reports] [Distance tree of results] [Multiple alignment] [MSA viewer]

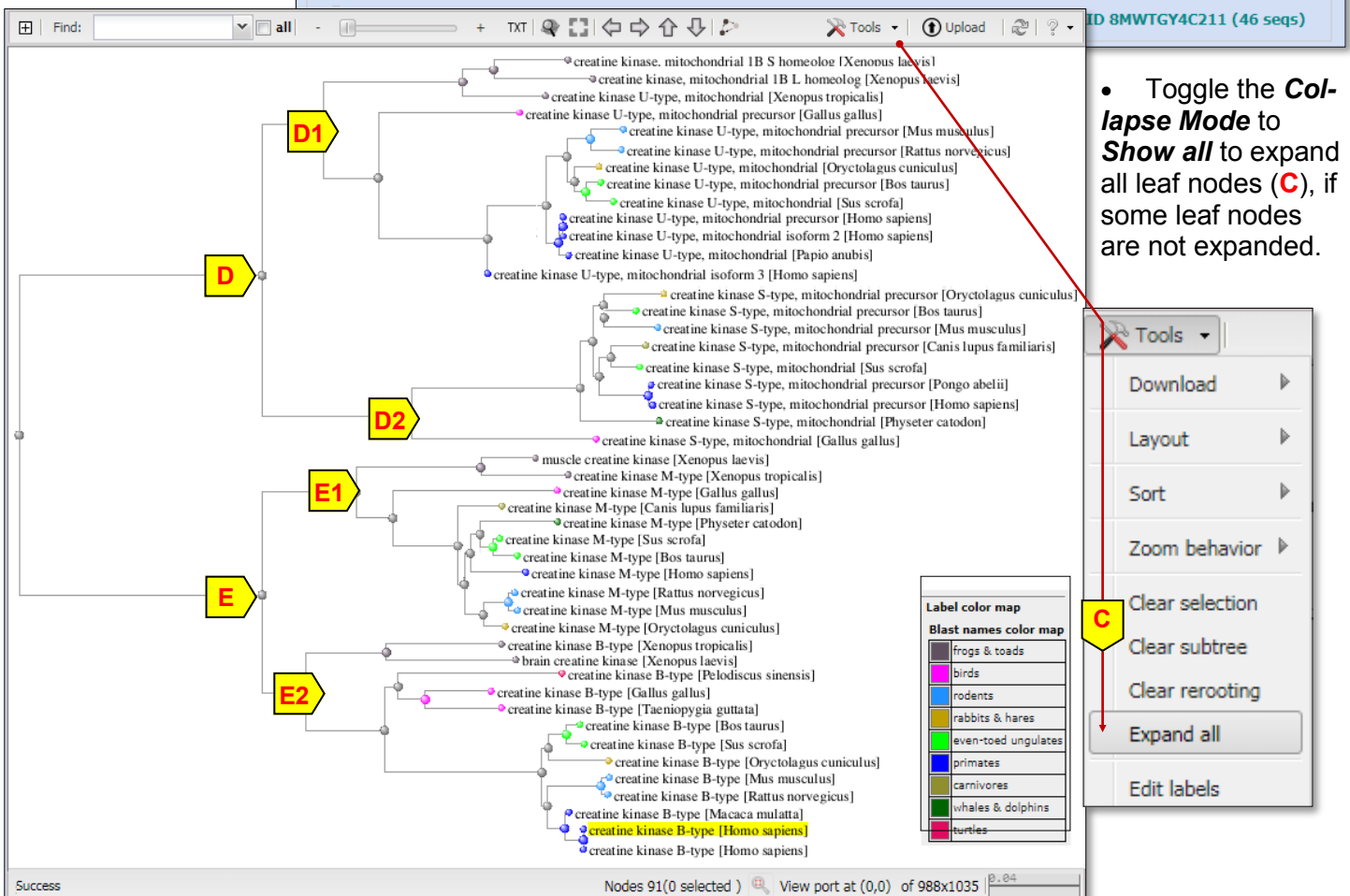
**COBALT** Constraint-based Multiple Alignment Tool

Phylogenetic Tree Edit and Resubmit Back to Blast Results ▶ Download

Multiple alignment

Home Recent Results Help

8MWTGY4C211 (46 seqs)



- Toggle the **Collapse Mode to Show all** to expand all leaf nodes (C), if some leaf nodes are not expanded.

#### Interpretation

- The resulting tree (right) is complicated by the presence of multiple isoforms from the same gene in a particular species. However, there are clearly two distinct groups of proteins, mitochondrial (D) and cytoplasmic (E), with two types of genes in each, U-type (D1) and S-type (D2) for mitochondrial group, and M-type (E1) and B-Type (E2) for Cytoplasmic group. This is a good example of a gene (protein) tree as compared to a species tree. Notice that mouse and human have proteins in all four groups, and that the mouse M-type is more similar to the human M-type than it is to the mouse U-type. Within a particular protein type though, tetrapod relationships are about as expected. For instance, the mouse M-type is closer to the rat M-type than either is to the human M-type.

## Case 4: Annotate a Metagenomic Contig

### Goal:

Use blastx to find potential proteins/genes on a genomic contig.

### Background

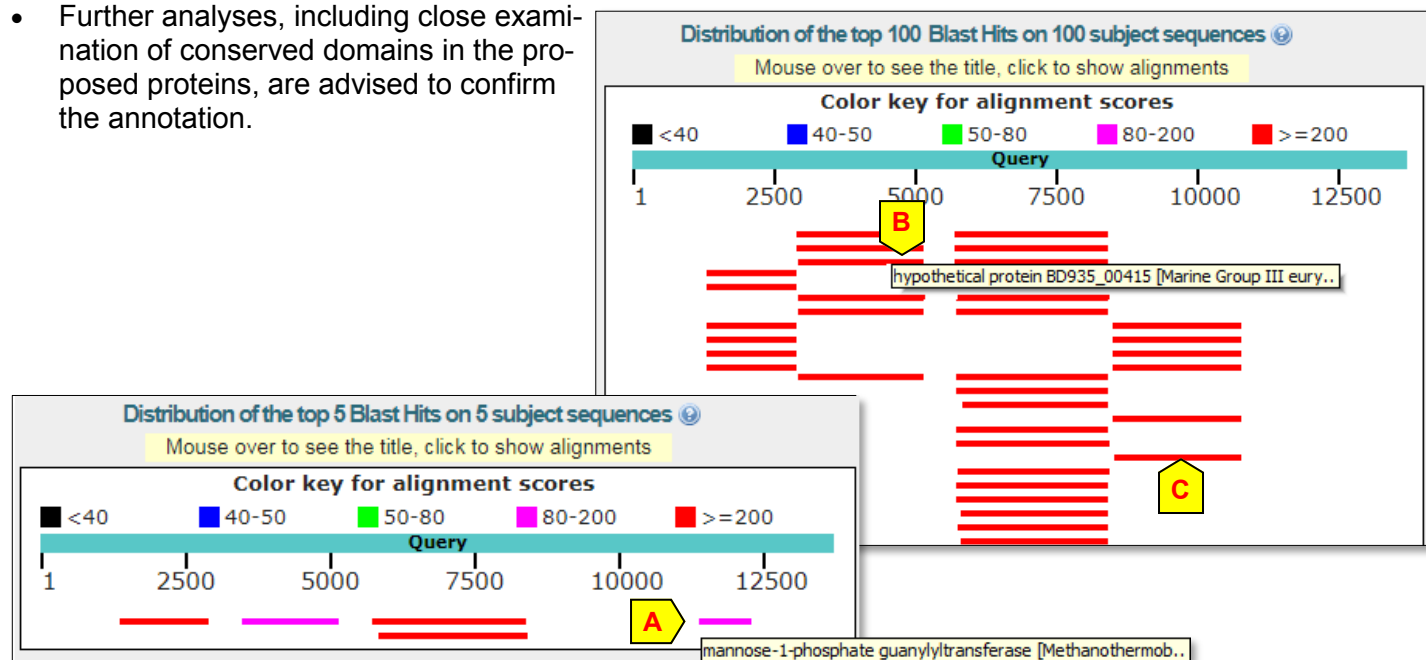
- Useful in molecular biology and microbiology courses
- About 10 min to complete the exercise
- Other software is often used for large-scale gene prediction and annotation, but blastx nicely illustrates the principles.

### Steps

- Retrieve accession number [MIZB01000007.1](#) from the Nucleotide database. This is a 13.7 kb genomic contig assembled from an Euryarchaeal marine metagenomic reads.
- Click **Run BLAST** on that Nucleotide page, click the **blastx** tab at the top to change the program.
- Select the **Model Organisms (landmark)** database. The reason for selecting this database is that it is small and non-redundant, with proteomes from a wide taxonomic group, thus your search will be quick and the result will be more concise. The information icon ? provides more information. **Tips:** For a larger database, such as *nr*, we suggest setting the **Expect** threshold to 1e-6 (10 to the minus 6th power) or lower, but that is usually not necessary with the *landmark* database.
- Use the **Organism** field to limit the search to *archaea (taxid:2157)* to match the organism source of the query, and creates a cleaner set of results. However, choose such limits carefully to match the goals you want to achieve. Click the **BLAST** button to submit the search.
- You may want to save the Request ID (RID) number found on the results page for later use, although they are saved in **Recent Results** page for the current browser session. All RIDs expire after about 36 hours.
- We want to compare results with a search against the protein *nr* database. Click the **Edit and resubmit** link near the top of the page, then change the database to *nr*. Set the **Expect** threshold to 1e-6, check that the Organism limit remains, then click the **BLAST** button.

### Interpretation

- The search against the Landmark database (**A**) suggests four possible genes on the contig, including: Hef nuclease, DNA topoisomerases, alanine-tRNA ligase, and mannose-1-phosphate guanylyltransferase.
- The value of searching multiple databases: the search against Landmark identifies the Hef nuclease, which was only labeled “hypothetical protein” in the search against *nr* (**B**). The search against *nr* adds the oligopeptide transporter, OPT family (**C**).
- Further analyses, including close examination of conserved domains in the proposed proteins, are advised to confirm the annotation.



## Case 5: Examine Conserved Domains and Solved Structures to Support a Protein Annotation

### Goals:

- Use blastx to find a structure record related by sequence to your annotated protein
- Confirm that your protein contains important sequence motifs for the conserved domain
- View these motifs on the solved structure using the iCn3D or Cn3D viewers.

### Background:

- Useful in many biology courses.
- About 15 min

### Steps

- Run a blastx search with the contig from Case 4, [MIZB01000007.1](#) as the query.
- Choose **Protein Data Bank proteins (pdb)** as the database, which contains solved structures from NCBI's Structure database. Lower the **Expect** threshold to 1e-6 to return only the better alignments.
- On the results page, we'll focus on the best hit, an alanyl-tRNA synthetase (another name for alanine-tRNA ligase). In the Descriptions section of the results page, click the Accession link for [3WQY\\_A](#), the crystal structure of Archaeoglobus fulgidus alanyl-trna synthetase in complex with wild-type tRNA(ala).
- In the Protein record for 3WQY\_A, click **Identify Conserved Domains** under **Analyze this Sequence**.
- Notice the match to AlaRS\_core domain, when viewed in **Full Results** mode (**A**).
- To see how residues in the motifs match the sequences used to construct this domain, click on the AlaRS\_core bar (**B**) in the "Specific hits" row to open the domain.
- In the **Conserved Features/Sites** tab, click on **motif 1** and **Scroll to Sequence Alignment Display** (**C**), the 3WQY\_A sequence is called the **query** and motif 1 is marked by the # symbols.

Conserved domains on [gi|1101737691|gb|OIR21235.1|] A

alanine--tRNA ligase [Marine Group III euryarchaeote CG-Epi1]

### Protein Classification

**alanine--tRNA ligase** (domain architecture ID 11486935)  
alanine--tRNA ligase catalyzes the attachment of alanine to the 3' OH group of ribose of tRNA(Ala)

### Graphical summary

Zoom to residue level [show extra options >](#)

Query seq. 1 125 250 375 500

active site motif 1 motif 2 motif 3

Specific hits alaS

AlaRS\_core B

### Sequence Alignment

include consensus sequence ?

Reformat Format: **Hypertext** Row Display: **up to 10** Color Bits: **2.0 bit** Type Selection: **the most similar members**

Feature 2	gi 6226166	6	TEEVRSKFITYFKAn	---	NHTHVPASSLi	--	pDP
query	58	LDEAREAF	LRFFFEK	n	---	KHTRVDRASV	varwRR
gi 22096190	60	ISEMREY	YLSFFFE	Ar	---	GHTRLDRYPV	varwRR
gi 22096191	61	VWEAGEE	FFRFFER	h	---	DHEVLDRYPV	varwRR
gi 6707751	59	LDEMREK	FLRFFFE	Kheiy	PHGRVKRYPV	prwRR	
gi 14286171	62	VKEAREK	FLSFFEK	r	---	GHTRIPPKPV	larwRED
gi 6707749	46	VGEAREA	FLSFFEK	h	---	GHTRVPPRPV	varwRED
gi 3334348	62	LDEMREY	YLNFFER	r	---	GHGRIERYPV	varwRTD
gi 22096203	62	LSEMRDA	FIKFFEK	r	---	GHKFLKPYPV	vprwRED
gi 2507425	5	TAEIRQA	FLDFHFS	k	---	GHQVVASSSL	v--pHNDpt
						LLFTNAGMNQ	FKDVF-LGLDKR
						YsRATT	SQRCVRA
						aggk	hndl
							78

Conserved Features/Sites ? [PubMed References ?](#)

active site motif 1 motif 2 motif 3

Feature 2: motif 1

Evidence:

- Comment: characteristic motif of class II is part of dimer interface

[Scroll to Sequence Alignment Display](#)

C

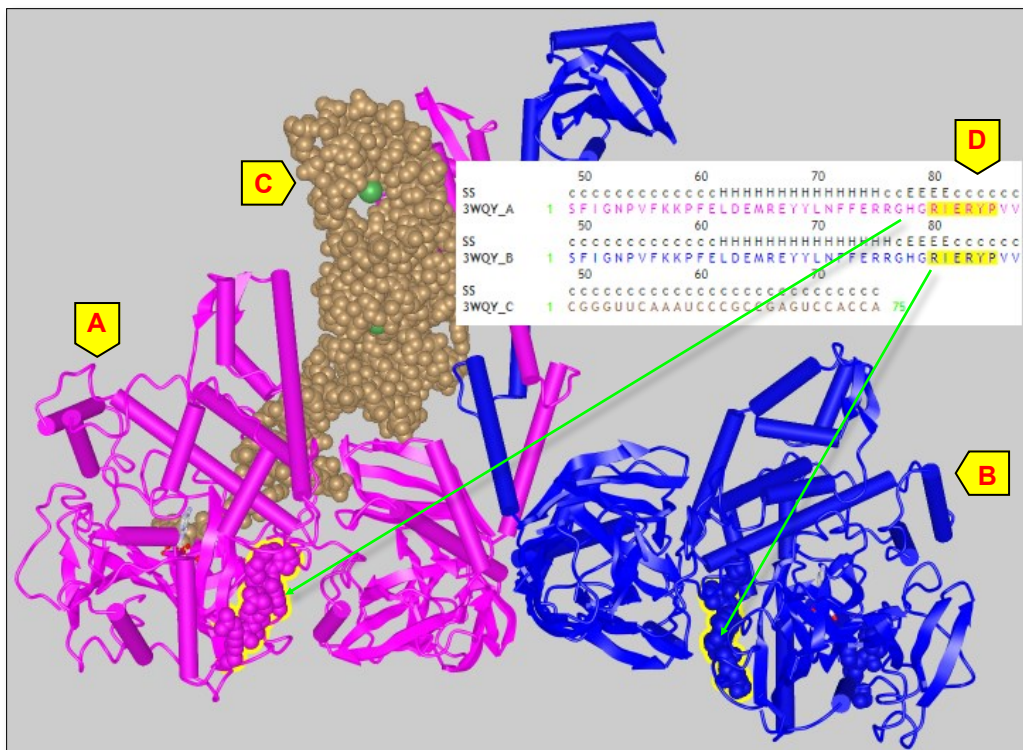
## Case 5 (cont.)

### Steps (cont)

- To view the solved structure, go back to the [3WQY\\_A](#) protein record. Click on the thumbnail graphic of the structure on the right side of the page. Four structure records are shown. Click on the “**View in iCn3D**” link for PDB ID: 3WQY (the second record).
- The structure contains both the \_A (A) and \_B (B) chains, plus a tRNA(ala) (C). Locate the motif 1 residues in the sequence by using the browser’s “find in page” function to search for RIERY. You find this motif at residues 80-84. Click and drag over RIERY to highlight those residues in yellow on both the sequence and in the viewer (D).

### Interpretation

- You can examine the other features in the 3WQY sequence to confirm their similarity to the members of the conserved domain, and do the same analysis with the other domains. This type of analysis increases confidence in the proposed annotation.
- Starting with a blastx search against the pdb database is one of several approaches that lead to conserved domains and structure records for your annotated proteins. You can also use the records for the proteins found in a blastx search against nr or the landmark database, such as WP\_010877290, an alanine-tRNA ligase identified in Case 4.
- If an ORF finding tool is in your workflow, that also will identify potential coding regions. Our web-based ORFfinder tool accepts nucleotide sequences up to 50 KB, and allows you to directly submit ORFs to the blastp service, <https://www.ncbi.nlm.nih.gov/orffinder>. A standalone version of ORFfinder is also available for Linux, <https://ftp.ncbi.nlm.nih.gov/genomes/TOOLS/ORFfinder/linux-i64/>.



## Appendix

Please address questions on the above example cases to the NCBI blast-help group:

[blast-help@ncbi.nlm.nih.gov](mailto:blast-help@ncbi.nlm.nih.gov)

For questions and feedback on subjects not related to BLAST, email:

[info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov)

You can check the NCBI Learn page for links to help documents, information on webinars, and workshops:

<https://www.ncbi.nlm.nih.gov/learn/>

The “Tutorials: BLAST” video playlist from NCBI’s YouTube channel can be found at:

<https://www.youtube.com/playlist?list=PLH-TjWpFfWrtjzMCivUe-YbrlleFQIKMq>

Factsheets on popular resources and common tools are available at:

<https://ftp.ncbi.nlm.nih.gov/pub/factsheets/README.html>