

Five Teaching Examples with NCBI BLAST

Handout for NCBI Tech Talk, ASCB 2016

Date/Time: Sunday, December 4, 2016, 5:30-6:30 PM

Location: Theatre 2

Five Teaching Examples with NCBI BLAST

- 1) Identify Unknown Bacteria using the 16S rRNA BLAST Database
 - 2) Identify a PCR Primer Set for Amplifying the Coding Region of an mRNA Transcript
 - 3a,b) Generate Species and Gene Phylogenetic Trees
 - 4) Annotate a Metagenomic Contig
 - 5) Examine Conserved Domains and Solved Structures to Support a Protein Annotation
-

1) Identify Unknown Bacteria using the 16S rRNA BLAST Database

Goal:

Use the 16S rRNA BLAST database to validate the identity of an unknown bacterial sample

Background:

- Useful in microbiology lecture and laboratory courses
- ~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 (and 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.

Procedure:

- From the BLAST home page -- blast.ncbi.nlm.nih.gov – go to Nucleotide BLAST and select the 16S ribosomal RNA sequences database (the next to last entry in the Database pull-down menu).
- The query is one or more sequences of a 16S rRNA region for an “unknown” bacterial sample. Retrieve those sequences from the FTP directory for this course ([ftp://ftp.ncbi.nlm.nih.gov/pub/education/Meeting Presentations/2016/ASCB_BLAST/Bacterial16S.txt](ftp://ftp.ncbi.nlm.nih.gov/pub/education/Meeting_Presentations/2016/ASCB_BLAST/Bacterial16S.txt)). These sequences mimic the results of PCR amplification by a set of microbial universal primers.
- Paste one or more sequences, or upload the file, into the “Enter Query Sequence” box.
- Click the “BLAST” button to run the search.

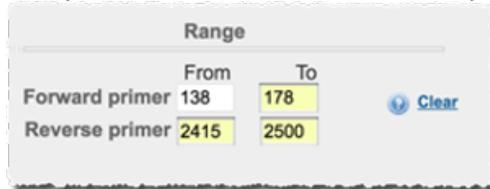
Interpretation:

- It is common for this conserved region to “hit” (be identified as similar to) many bacterial sequences, however by identifying those closest to the input sequence, it is possible to establish a likely identity for the unknown sample.
- A quick view of the results can be seen in the Descriptions section of the BLAST report. For very similar results like these, Max score is often, but not always, the important statistic. Also consider the percent identity and query coverage, and confirm identification by looking in the Alignments section.

used for other lab sessions, such as expression and characterization of the protein, mutagenesis, or promoter analysis.

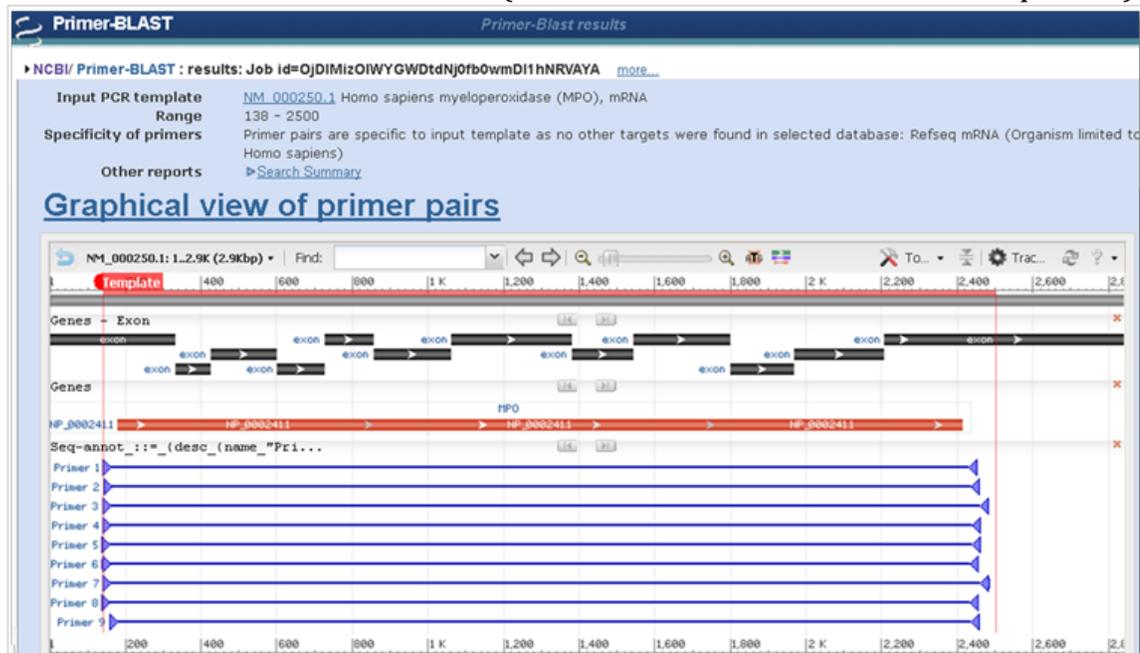
Procedure:

- Retrieve the record for NM_000250, the RefSeq mRNA sequence for Human Myeloperoxidase (MPO): www.ncbi.nlm.nih.gov/nuccore/NM_000250.
- 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; 2237 residues long. Write down these positions because you'll need them in a bit.
- On the right-hand side of the record under "Analyze this sequence" click "Pick Primers" to open the Primer-BLAST page with the accession already provided as the template.
- To amplify the CDS region, the ranges for "Forward" and "Reverse primers" should be set outside of the CDS positions in the record. Set Forward primer from 138 to 178 & Reverse primer from 2415 to 2500. Also, adjust the "PCR product size" so that the entire length amplifies (increase the Max size to 2500).



Range		
	From	To
Forward primer	138	178
Reverse primer	2415	2500

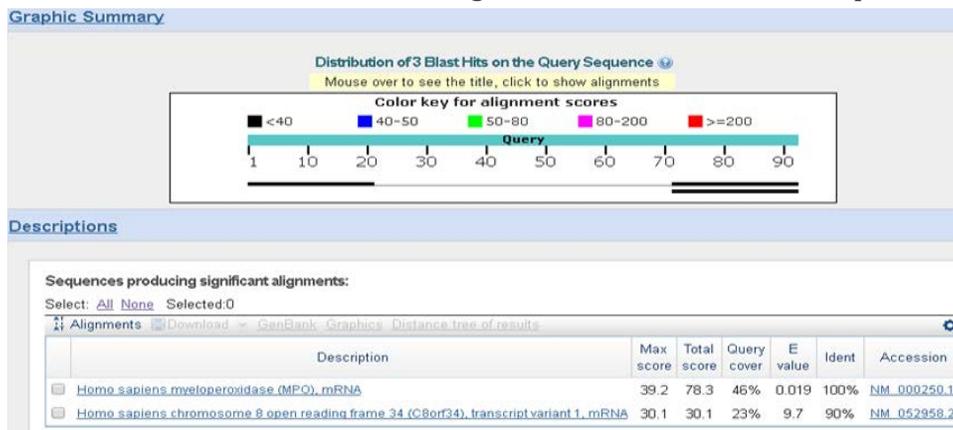
- Click "Get Primers" to run the search. (The default database is human RefSeq mRNA).



Interpretation:

- It is important to find PCR Primer pairs that will amplify only the sequence that is intended. In this case, the selected PCR Primer pairs should amplify only the Human MPO transcript.
- The "Specificity of primers" identified by Primer-BLAST is described at the top of the results page. Also, several primer sets are shown with their key parameters so that the researcher can make his/her own selection.

- Tips for finding primers specific for your template can be found here: www.ncbi.nlm.nih.gov/tools/primer-blast/search_tips.html
- It is most important to have similar T_m values for both primers in a pair, as well as similar 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, so you can relax them if you are not able to find any suitable primers. However, be aware that this may increase the potential for less specific primers.
- To confirm the specificity, you can run a blastn search against the human RefSeq RNA dataset. For the query, copy/paste the two primers and insert a long run of “N”s in between, which allows each primer to be aligned independently in a single search: GGTACAAAGGGGATTGAGCANNNNNNNNNNNNNNNNNNATATACCCCTCACTGCTGCAC. Be sure to use blastn rather than megablast because of the short primer queries.



- In this case one transcript was found by both Forward and Reverse primers – the intended Human MPO mRNA. However the reverse primer did have 90% identity to a region in Human C8orf34 mRNA. This will neither amplify, as both primers do not simultaneously “hit” this transcript, nor will it serve as an efficient “sink” to inhibit the Reverse primer from binding, because of the mismatches. (You get this hit to C8orf34 only if you exclude Models when you submit the search. Bonus points: why do you get another hit when you are searching a smaller dataset?).

3) Generate species and gene phylogenetic trees

Goal:

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

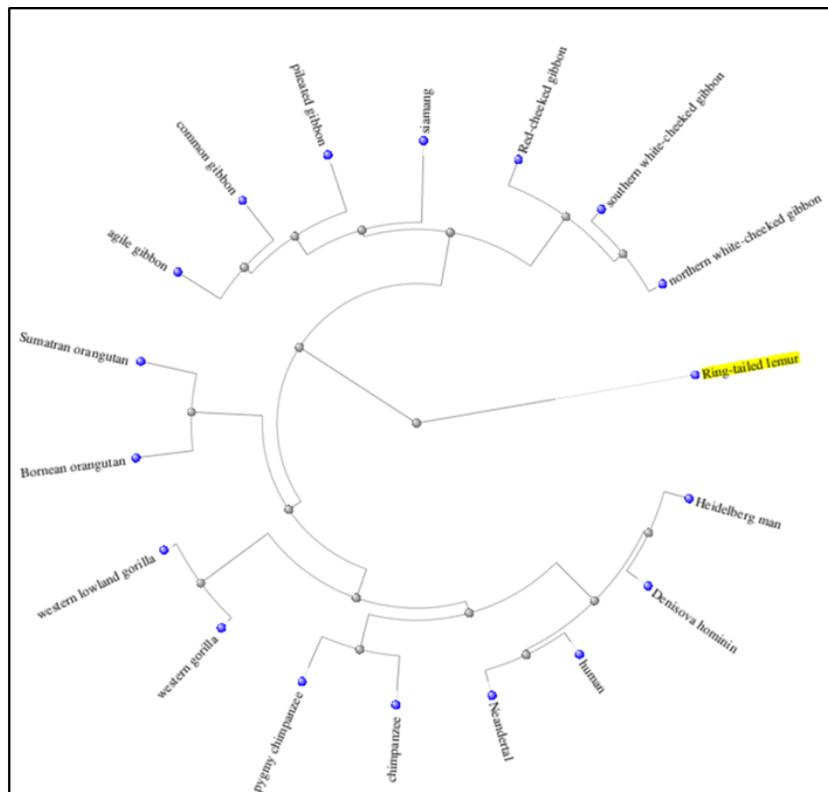
Background:

- Useful in general biology, molecular biology, and vertebrate zoology courses
- ~15 min for each of two examples
- Example 3a generates a phylogeny of apes using complete mitochondrial genome sequences. Example 3b builds a gene (protein) tree for the creatine kinases, a small protein family with four or more members in vertebrate proteomes.

3a) Ape phylogeny

Procedure:

- Retrieve the ring-tailed lemur mitochondrial genome sequence, accession number NC_004025.1, from the Nucleotide database, www.ncbi.nlm.nih.gov/nuccore/NC_004025.1. We can use this sequence as an out group and a query to retrieve and align the ape mitochondrial genomes using blastn.
- Either click “Run BLAST” on that Nucleotide page or open the BLAST home page, go to Nucleotide BLAST and paste the accession into the query box.
- Select the NCBI genomes (chromosomes) database. This database contains completely assembled genetic molecule sequences such as mitochondrial genomes. The information icon “?” provides a link to more information.
- Type “apes” in the Organism box and select the matching taxid (taxid:314295). This restricts searches to only sequences from this taxon in the database.
- Type “mitochondrion[filter]” in the Entrez query box. This will limit to only mitochondrial sequences.
- Finally adjust the BLAST program to “Somewhat similar sequences (blastn)”, expand the “Algorithm parameters” section and set the Word size to 7 and the Expect threshold to 1e-64. These settings find the single longest alignment between the query and each database sequence. Click “BLAST.”
- The results should show 17 nearly full-length matches to the lemur query. These include mitochondrial genome sequences from the gorilla, chimpanzee, 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. The Taxonomy report shows the taxa represented in the output.
- Click the “Distance tree of results” link to generate a tree.



Interpretation:

- The tree supports the two distinct groups of apes: the Great apes (Hominidae) containing humans, chimp, gorilla and orangutan, and the gibbons (Hylobatidae). It also shows the chimp (*Pan troglodytes*) and the bonobo (*Pan paniscus*) as the closest living relatives of humans and the Neanderthal as the closest extinct relative. Keep in mind that this tree is based completely on blastn, pairwise comparisons to the query (lemur) sequence. In this case, this produces a reasonable alignment for generating the tree. However, the most accurate tree 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. The next example uses a true protein multiple alignment through COBALT to generate a protein tree.

3b) Creatine Kinase protein tree

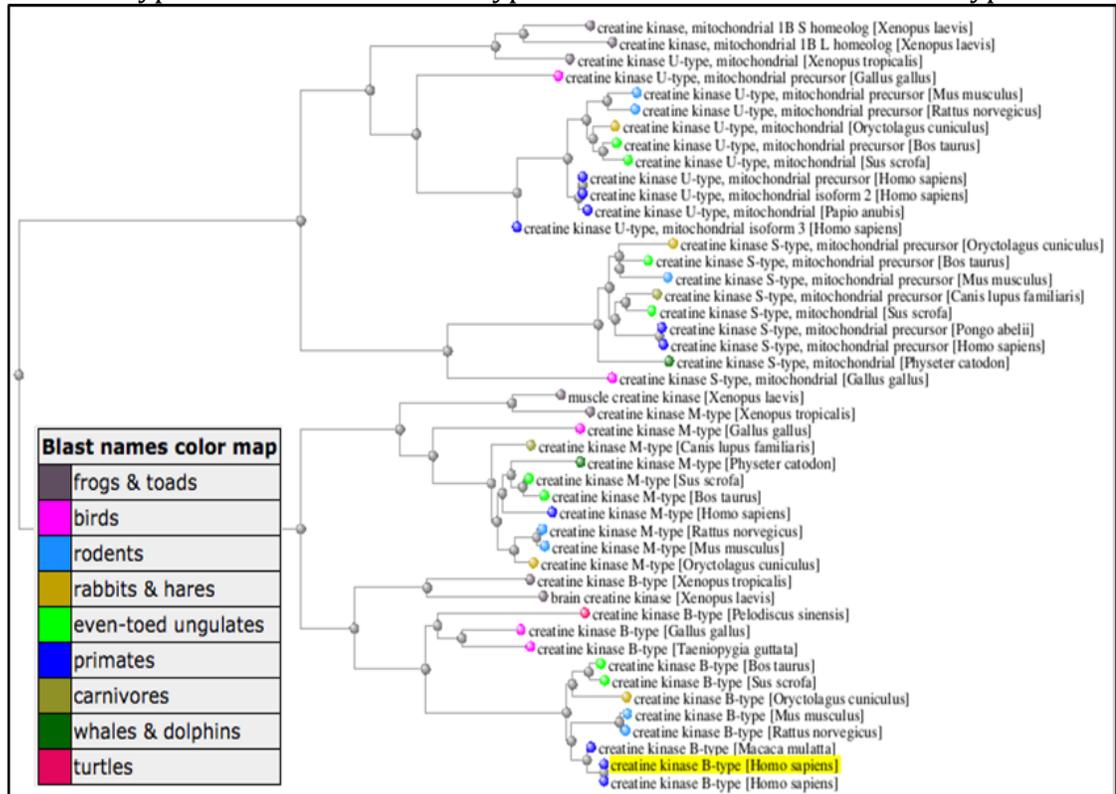
Procedure:

- Retrieve human creatine kinase B-type protein, accession number NP_001814.2, from the Protein database, www.ncbi.nlm.nih.gov/nuccore/NP_001814.2. Use this sequence as a blastp query to retrieve the tetrapod vertebrate creatine kinases, then perform a multiple sequence alignment using COBALT to make a protein tree.
- Either click Run BLAST on that Protein page, or open the BLAST home page, go to Protein BLAST, and paste the accession into the query box.
- Select the Reference proteins database. This database contains NCBI RefSeqs, including those used in or generated by the NCBI genome annotation pipelines. The information icon “?” provides a link to more information.
- Type “tetrapods” in the Organism box and select the matching taxid (taxid:32523). This restricts searches to only sequences from this taxon in the database.
- Check the Exclude box for Model sequences. This will eliminate gene predictions. There are a tremendous number of these models. You can leave this unchecked for a much larger tree.
- Expand the “Algorithm parameters” section and set the Expect threshold to 1e-16. Click the “BLAST” button to run the search.
- The results show matches to 45 proteins from a wide variety of vertebrates including birds, amphibians, turtles and mammals.
- Click the link for “Multiple alignment” to send these 45 proteins to COBALT. COBALT generates a true multiple sequence alignment that includes all residues from all the proteins. Note that the multiple alignment will involve residues that were not present in the BLAST alignments, such as the signal peptide from the mitochondrial targeted proteins.
- From the COBALT results click the “Phylogenetic Tree” link at the top to generate the protein tree. Toggle the “Collapse Mode” to “Show all” to expand all leaf nodes.

Interpretation:

- The tree (see next page) 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 and cytoplasmic) with two kinds genes in each. 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, for example, 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.



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
- ~10 min to complete the exercise
- Other software is often used for large-scale gene prediction and annotation, but blastx nicely illustrates the principles.

Procedure:

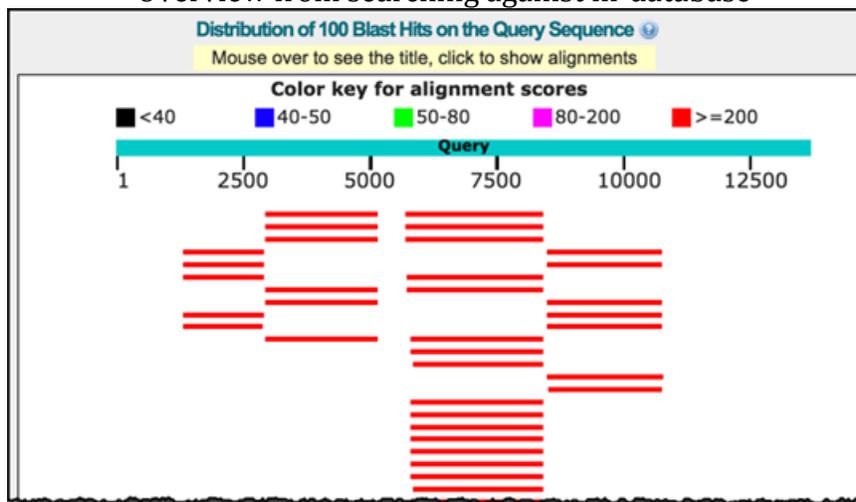
- Retrieve accession number MIZB01000007.1 from the Nucleotide database, www.ncbi.nlm.nih.gov/nuccore/MIZB01000007.1. This is a 13.7 KB contig assembled from 3 Euryarchaea, marine metagenomic sources.
- Either click Run BLAST on that Nucleotide page, or open the BLAST home page and go to blastx, then paste the accession into the query box.
- If you used “Run BLAST,” change the program to blastx (tab at top of page).
- Select the Model Organisms (landmark) database. Why this database? It is small and non-redundant, so the results are more concise, yet it has proteomes representing a wide taxonomic range. The information icon “?” provides a link to more information. For a larger database, such as nr, we suggest setting the Expect threshold to 1e-6 or lower, but that is usually not necessary with this database.

- Use the Organism field to limit the search to archaea (taxid:2157). This limit matches the goals of the BioProject that provided the example query, and creates a cleaner set of results. However, choose such limits carefully depending on your goals. Click “BLAST.”
- You may want to save the RID number found on the results page for later use, although they are saved in Recent Results 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 “BLAST.”

Interpretation:

- The search against the Landmark database 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. The search against nr adds the oligopeptide transporter, OPT family.
- Further analyses, including close examination of conserved domains in the proposed proteins, are advised to confirm the annotation.

Overview from searching against nr database



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.

- ~15 min

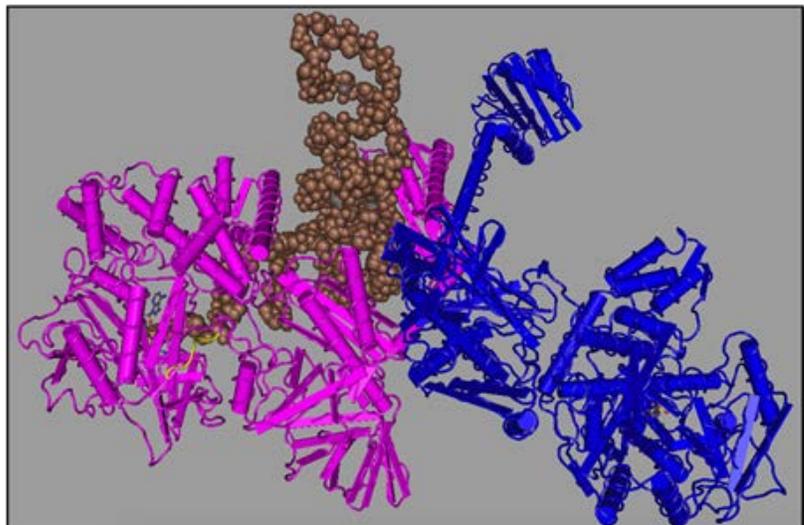
Procedure:

- Run a blastx search with the contig from Example #4, MIZB01000007.1 as the query. Choose the database, Protein Data Bank proteins (pdb), which contains solved structures from NCBI's Structure database. Lower the Expect threshold to 1e-6 to return only the better alignments.
- 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 AlaRS_core domain. Click the checkbox for "Zoom to residue level." By scrolling to the right, you see the residues in 3WQY_A that coincide with the key features of this domain. For example, motif 1 has the residues RIERY.

I		10	20	30
	***
	Feature 2		#####	
gi 6226166	6	TEEVRSKFITYFKAn---	NHTHVPASSLi---	pDNDps
query	62	LDEMREYYLNFFERr---	GHRIERYVva-rw	RTDi-
gi 22096190	60	ISEMREYYLSFFEAR---	GHTRLDRYPVva-rw	RDDi-
gi 22096191	61	VWEAGEEFLRFFERh---	DHEVLDRYPVva-rw	RDDi-
gi 6707751	59	LDEMREKFLRFFEKheiy	PHGRVKRYPVlp-rw	RDDv-
gi 14286171	62	VKEAREKFLSFFEKr---	GHTRIPPKPVla-rw	REDl-
gi 6707749	46	VGEAREAFLSFFEKh---	GHTRVPPRPVva-rw	REDl-
gi 3334348	62	LDEMREYYLNFFERr---	GHGRIERYVPva-rw	RTDi-
gi 22096203	62	LSEMRDAFIKFFEKr---	GHKFLKYPVvp-rw	REDv-
gi 2500963	59	YKEMVKEFINFFKEh---	GHEPIKRAPVtarrw	RDDi-

- To see how these residues match the sequences used to construct this domain, click on the solid bar in the "Specific hits" row, just below the motifs. That opens a page for the AlaRS_core domain.
- In the "Conserved Features/Sites" tab, click on motif 1 and scroll down to the alignments. The 3WQY_A sequence is called the "query" and motif 1 is marked by the # symbols.
- To view the solved structure, go back to the 3WQY_A Protein record, www.ncbi.nlm.nih.gov/protein/3WQY_A. 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 and _B chains, plus a tRNA(ala). 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-drag over RIERY to highlight those residues in yellow on both the sequence and in the viewer.



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 Example #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, www.ncbi.nlm.nih.gov/orffinder. A standalone version of ORFfinder is also available for Linux, [ftp.ncbi.nlm.nih.gov/genomes/TOOLS/ORFfinder/linux-i64/](ftp://ncbi.nlm.nih.gov/genomes/TOOLS/ORFfinder/linux-i64/).

Appendix

Please address questions on the above example cases to:

blast-help@ncbi.nlm.nih.gov

For questions and feedbacks on subjects not related to BLAST, please address them to:

info@ncbi.nlm.nih.gov

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

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

The “Tutorials: BLAST” video playlist from NCBI’s YouTube channel is at:

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

A set of factsheets on popular resources and common tools are available at:

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