Teaching NCBI Resources

Through Case Studies

Five Examples for NCBI BLAST

Contact: info@ncbi.nlm.nih.gov

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) 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

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

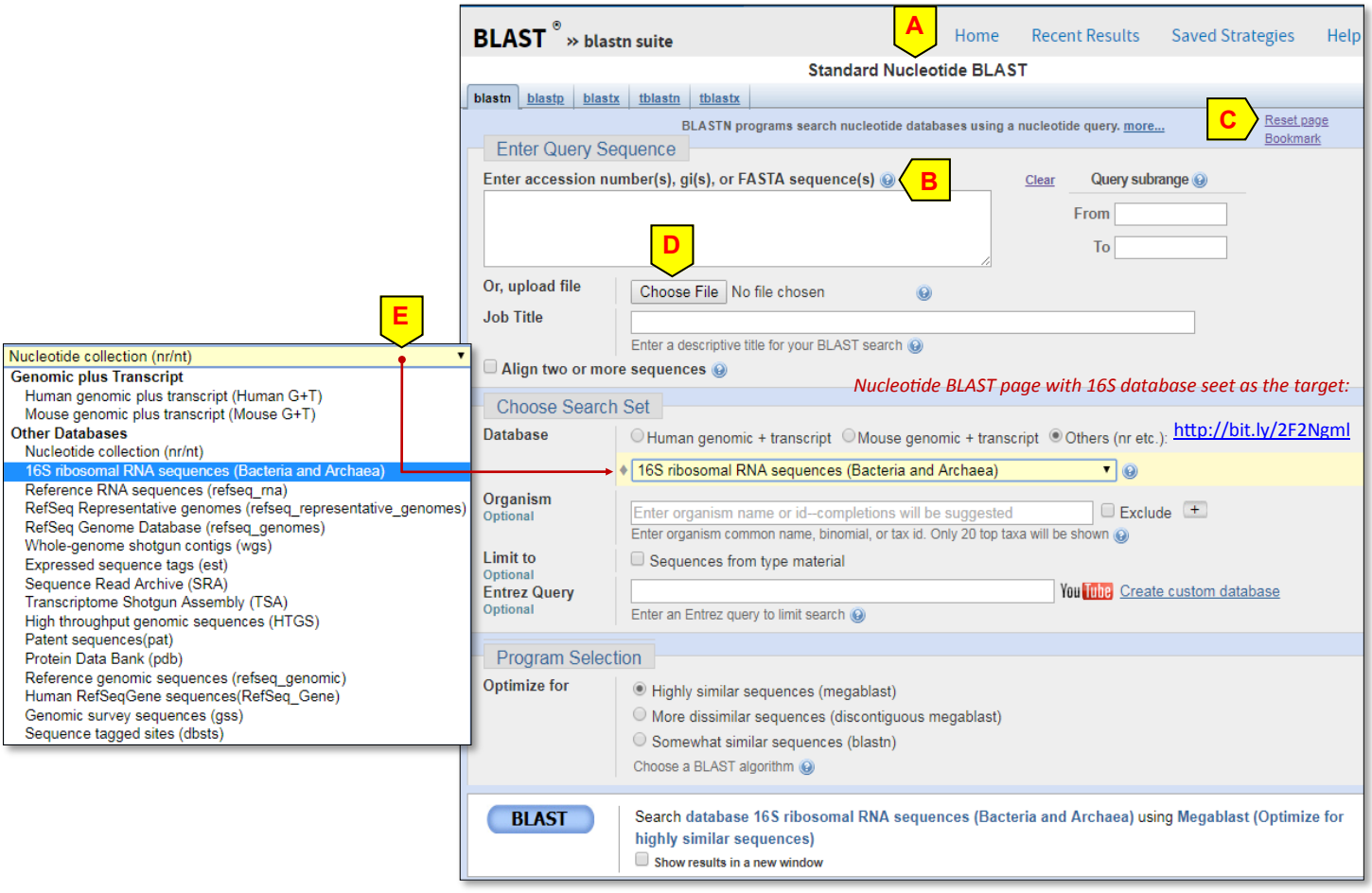
Background

- 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), 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/2suC3Ix). 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.

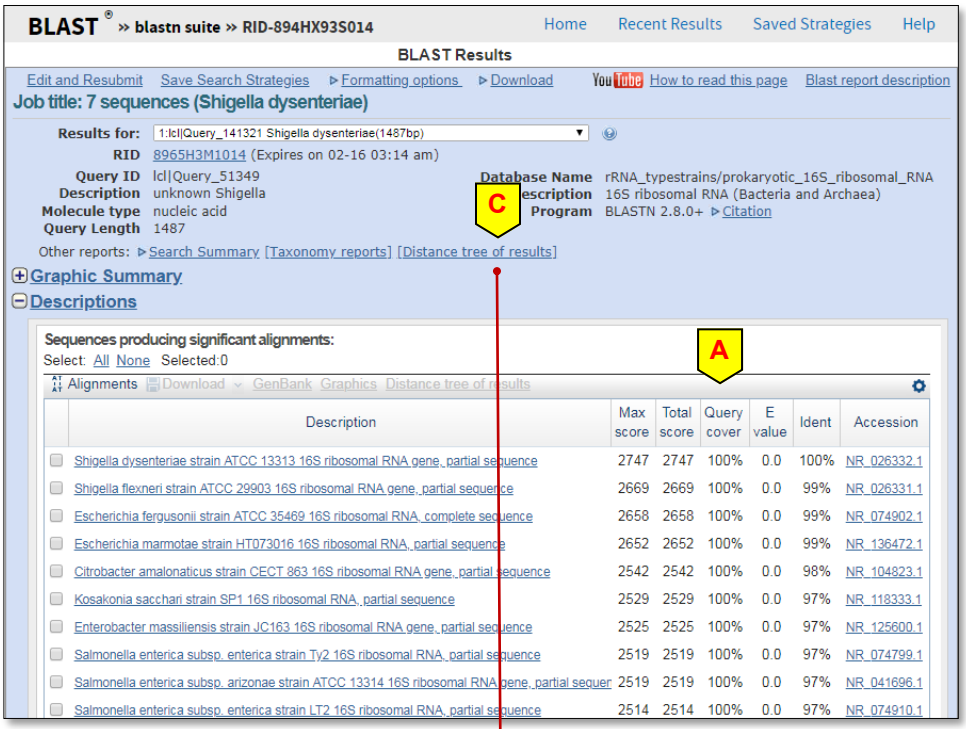
Search results are shown on the next page.

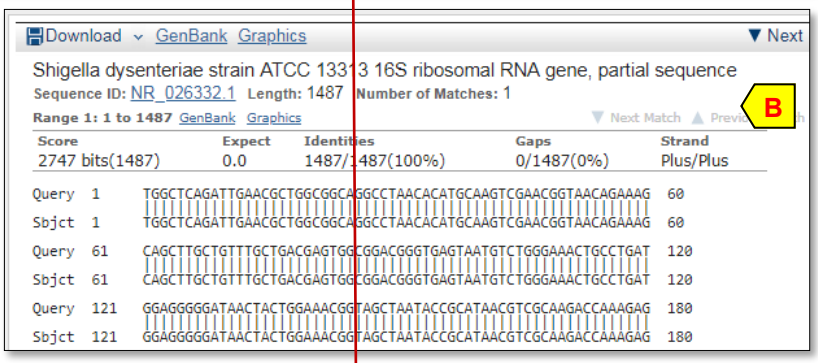


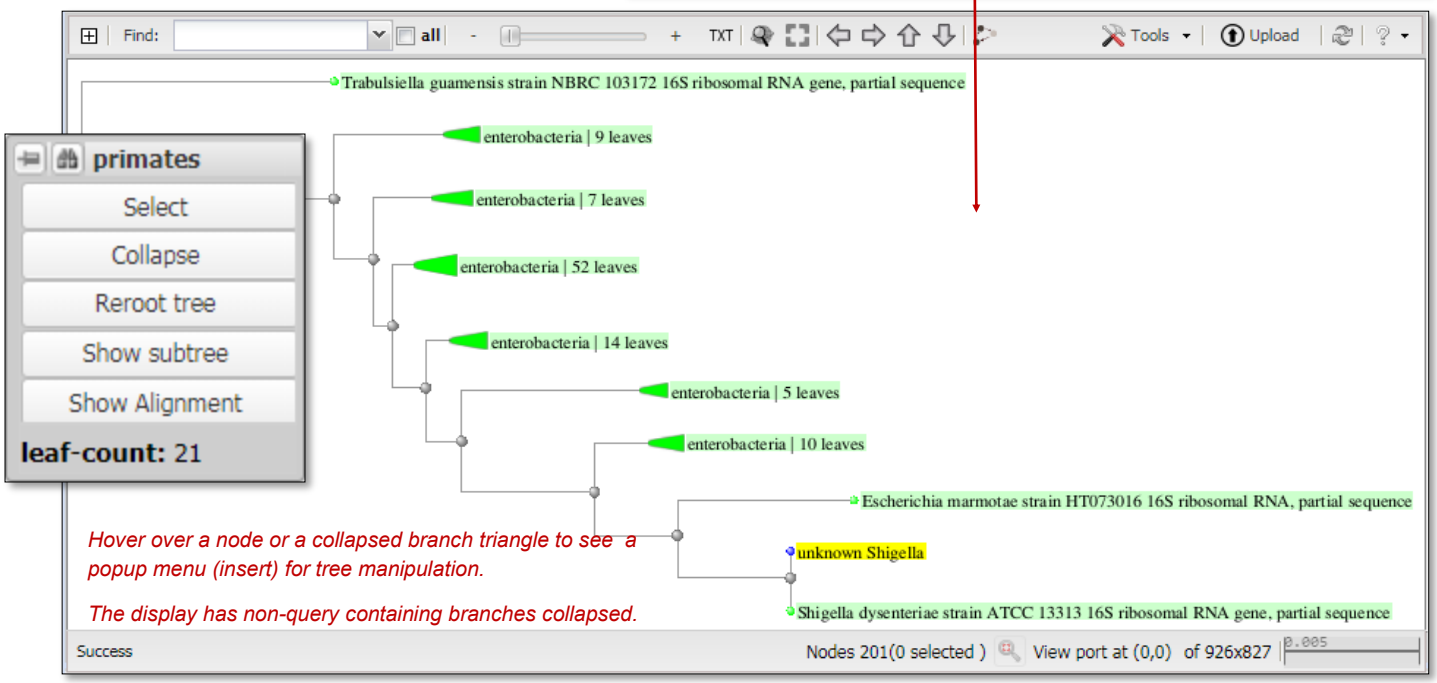
Case 1 (cont.)

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.







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

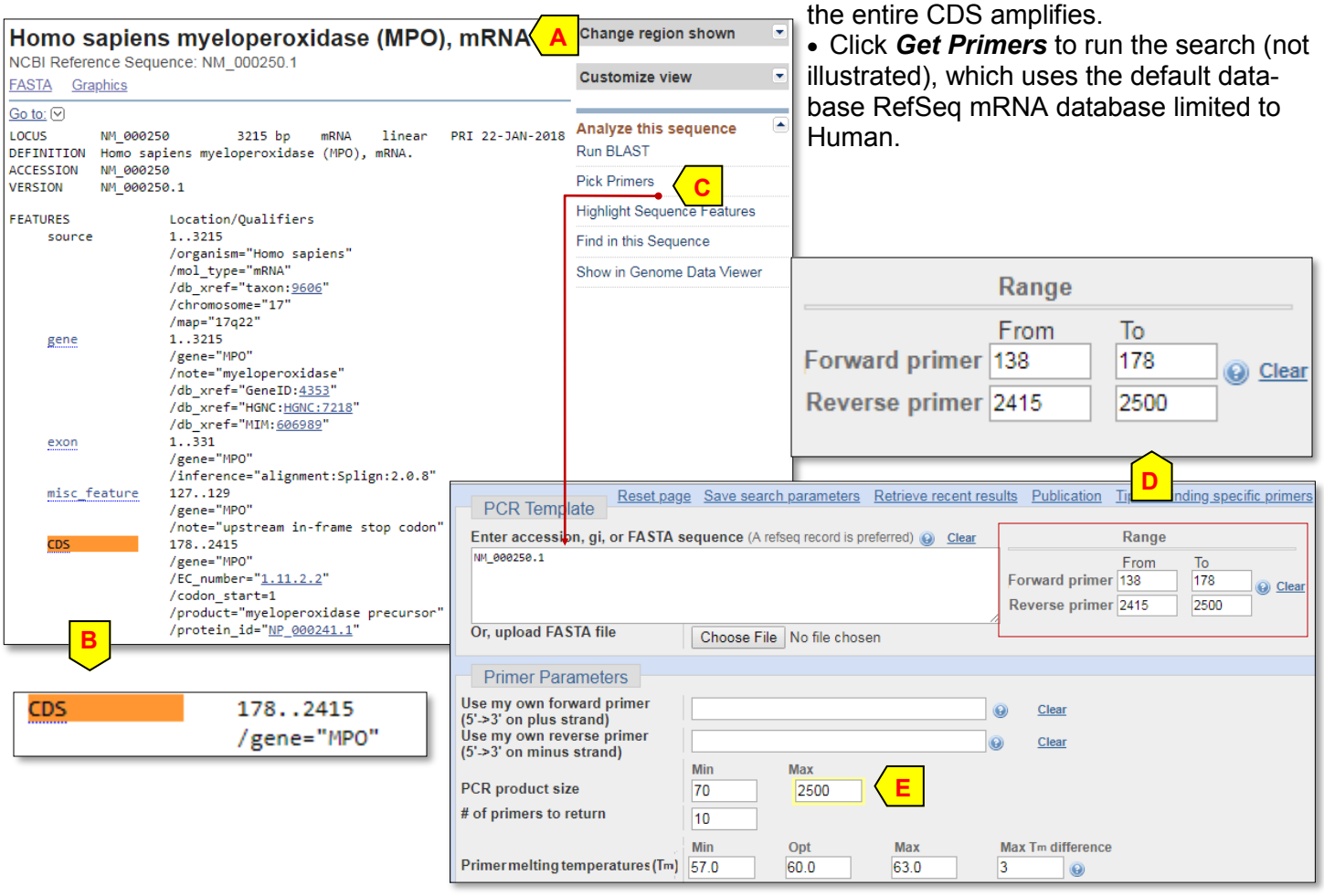
Background

- 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

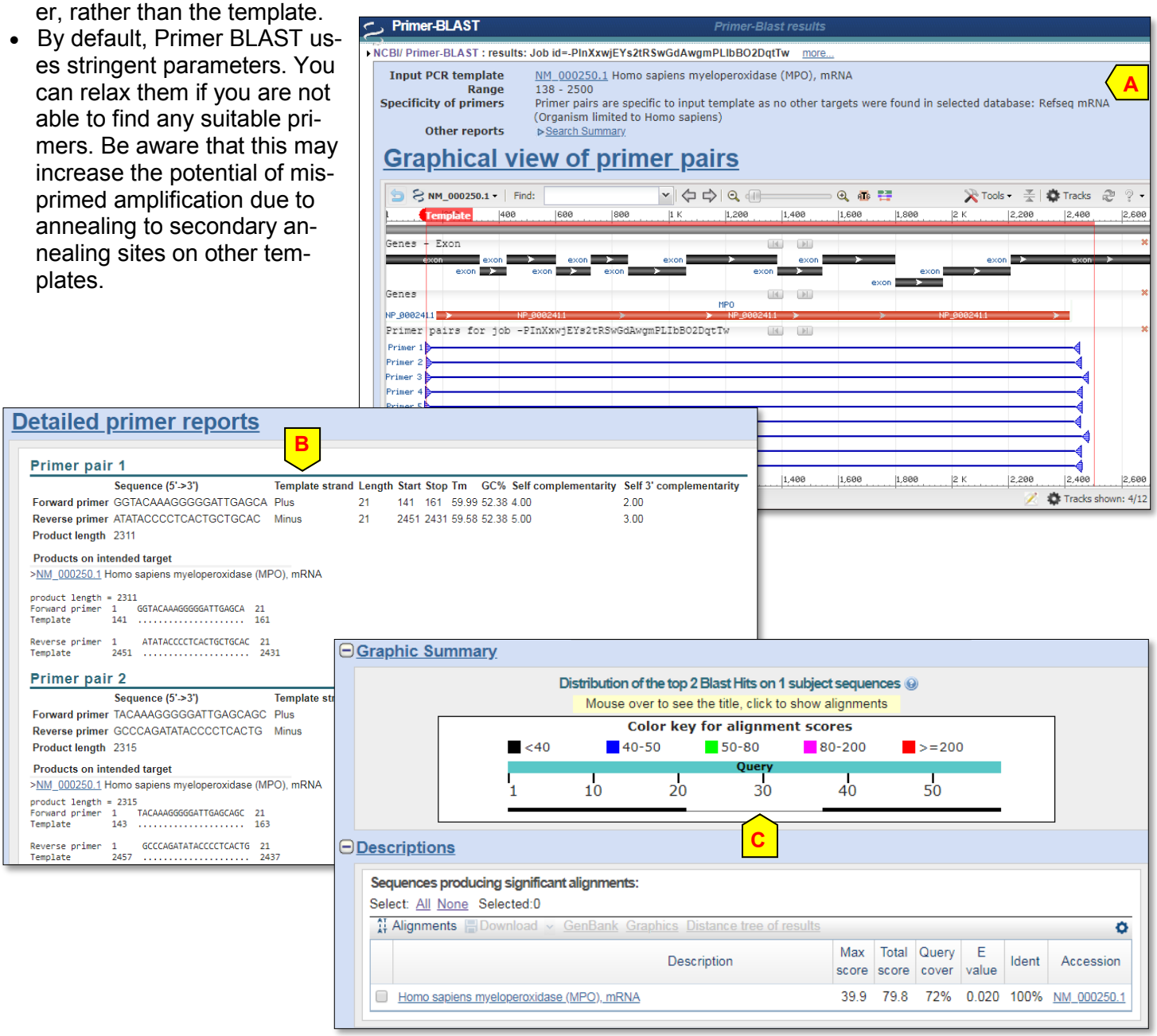
- Retrieve NM_000250, the RefSeq mRNA sequence for Human Myeloperoxidase, MPO (**A**) from the Nucleotide database
- Use the web browser's ***Find in page*** function (ctrl+F) to find the ***CDS*** feature. The coding sequence (CDS) for this gene starts at position 178 and ends at position 2415. It is 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.



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
- It is most important for both primers in a pair to have similar Tm 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.



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, GGTACAAAGGGGGATTGAGCANNNNNNNNNNNNNNNNATATACCCCTCACTGCTGCAC. 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

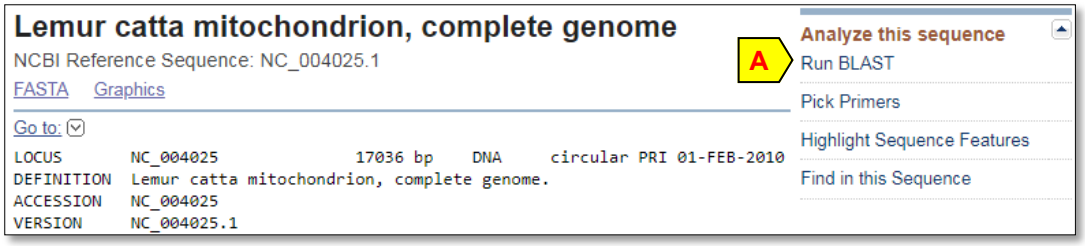
Goal

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

Background

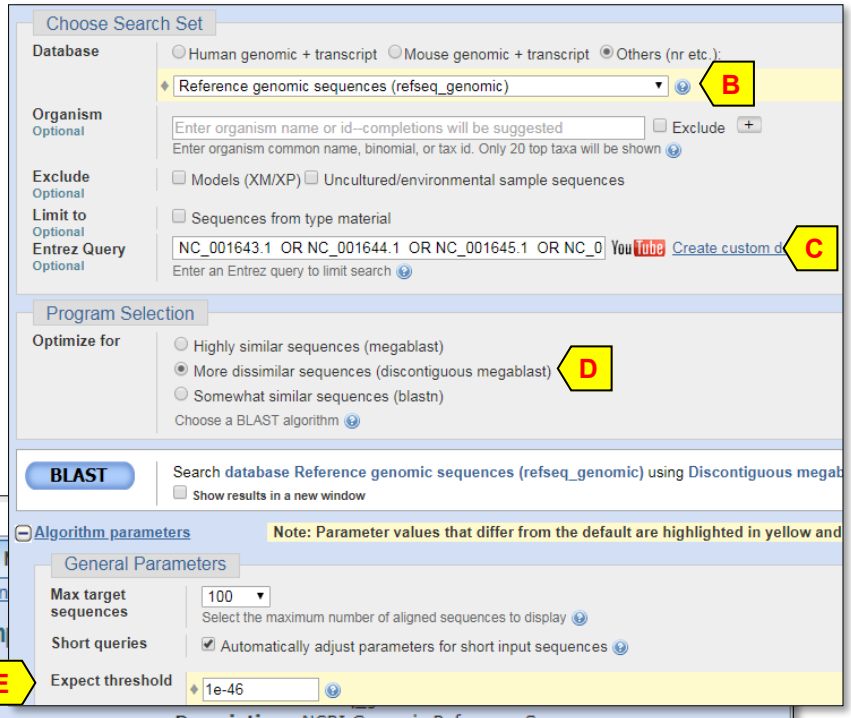
- Useful in general biology, molecular biology, and vertebrate zoology courses
- About 15 min for each of two examples
- Example **i)** generates a phylogeny of apes using complete mitochondrial genome sequences. Example **ii)** builds a gene (protein) tree for the creatine kinases, a small protein family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

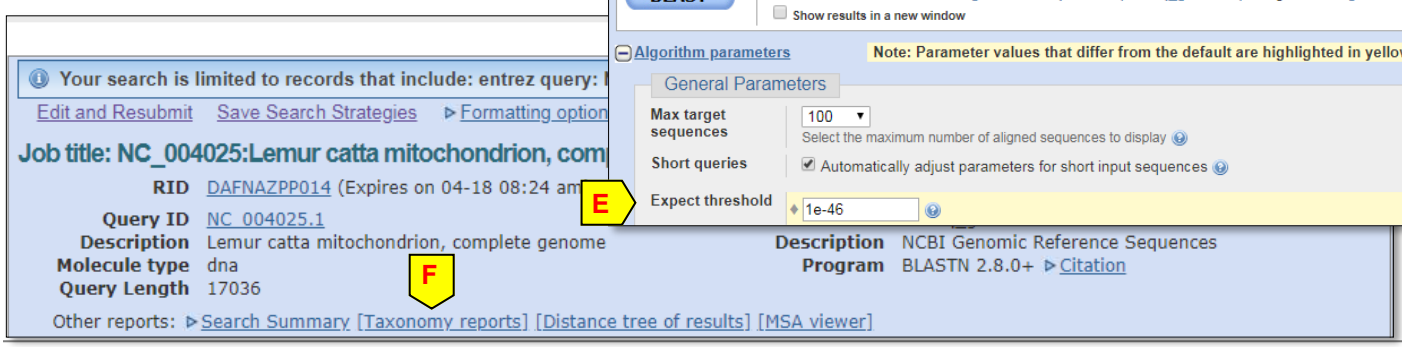
i) Ape Phylogeny



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.

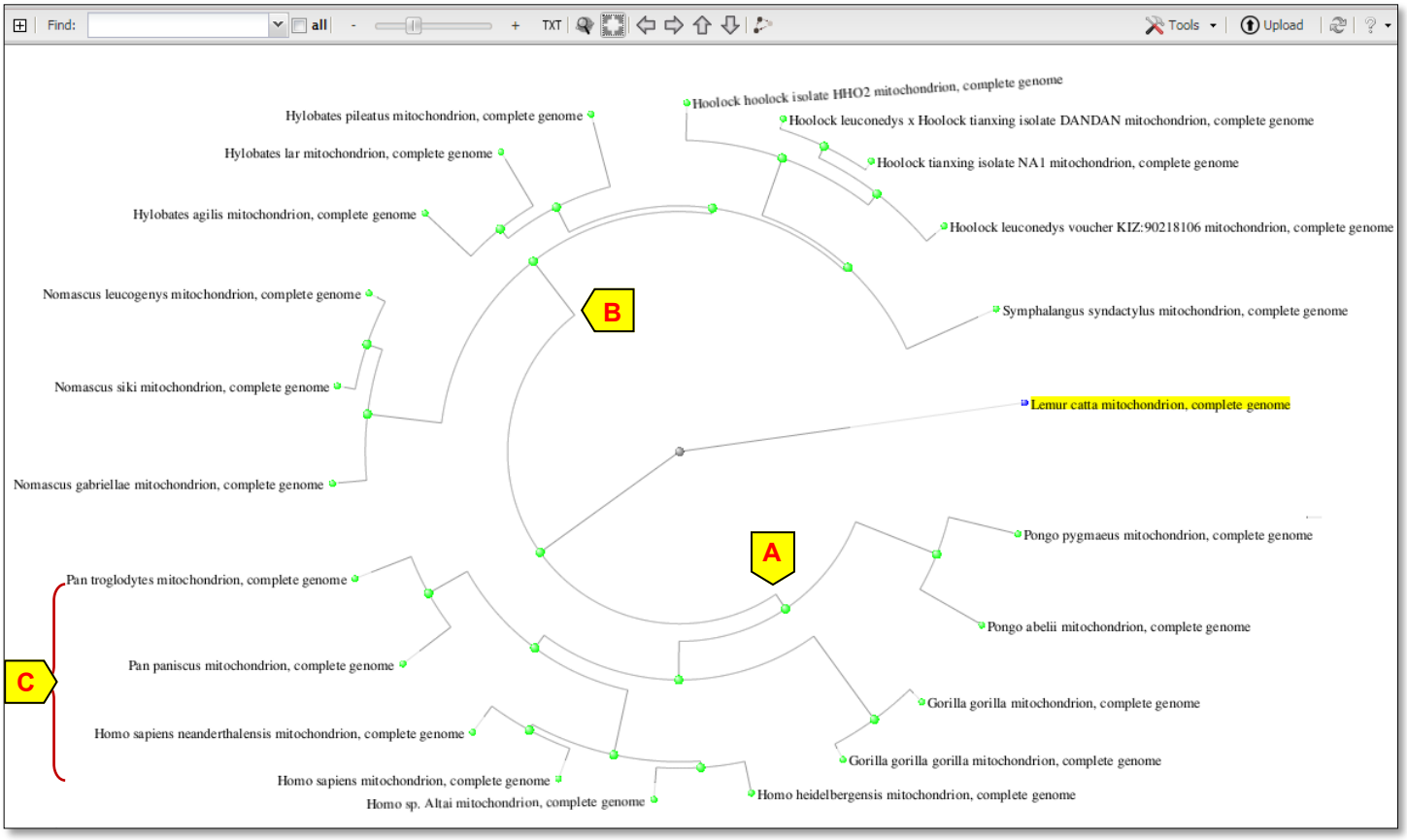




Case 3 (cont.)

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 pairwised 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

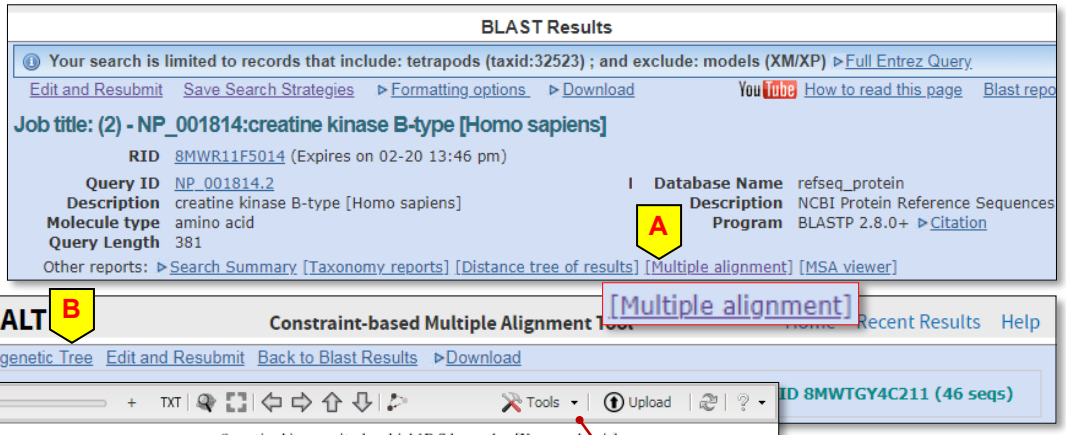
- Retrieve human creatine kinase B-type protein, NP_001814.2, from the Protein database. Use it as a blastp query to retrieve the tetrapod vertebrate creatine kinases, then perform a multiple sequence alignment for the matched sequences using COBALT to make a protein tree.
- Click ***Run BLAST*** on 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.

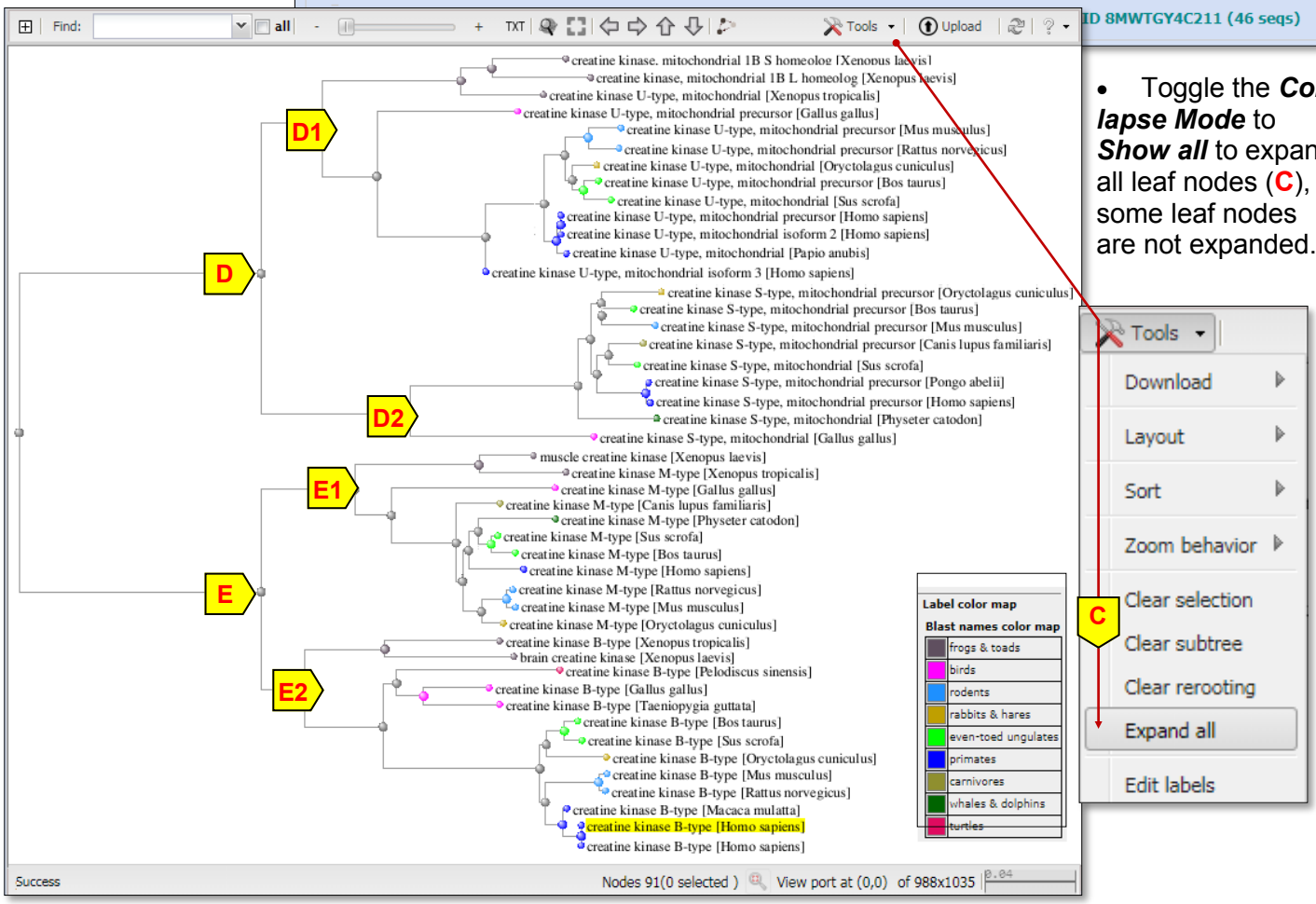
Case 3 (cont.)

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.





- 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.

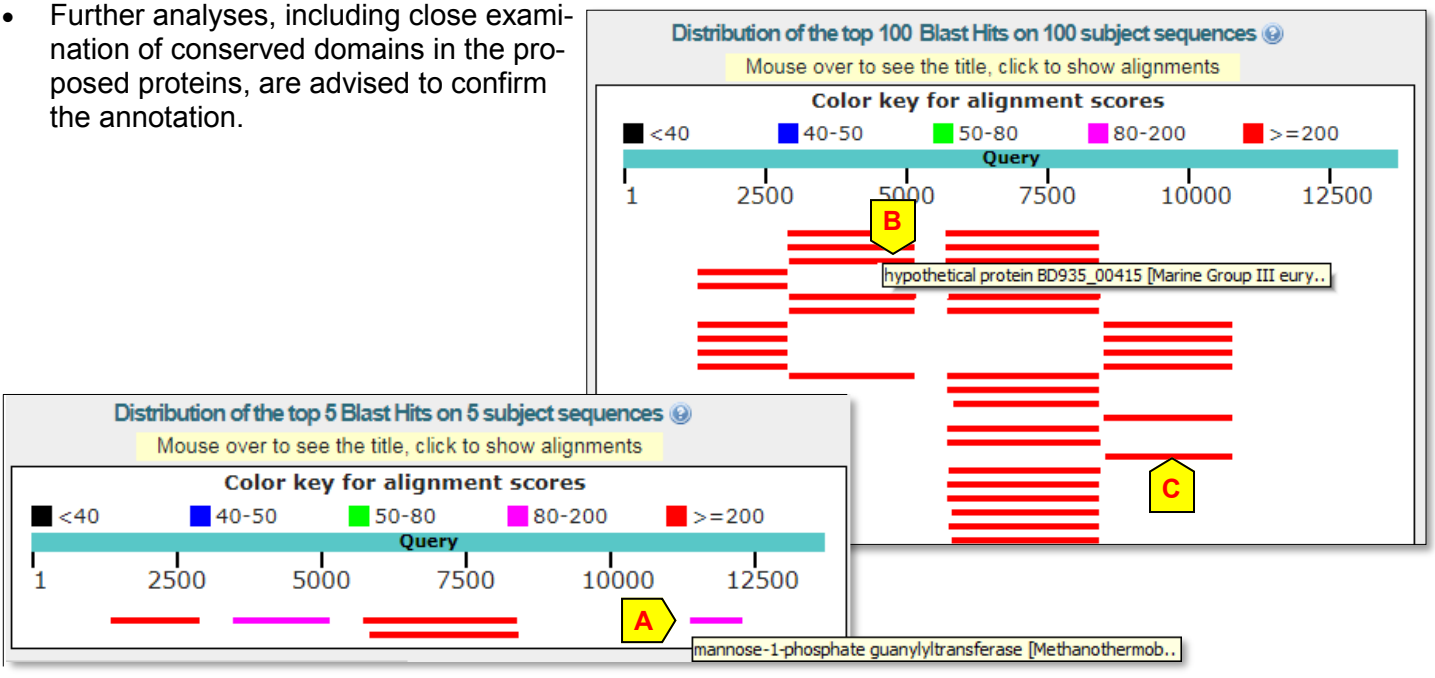
Case 4: Annotate a Metagenomic Contig

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:

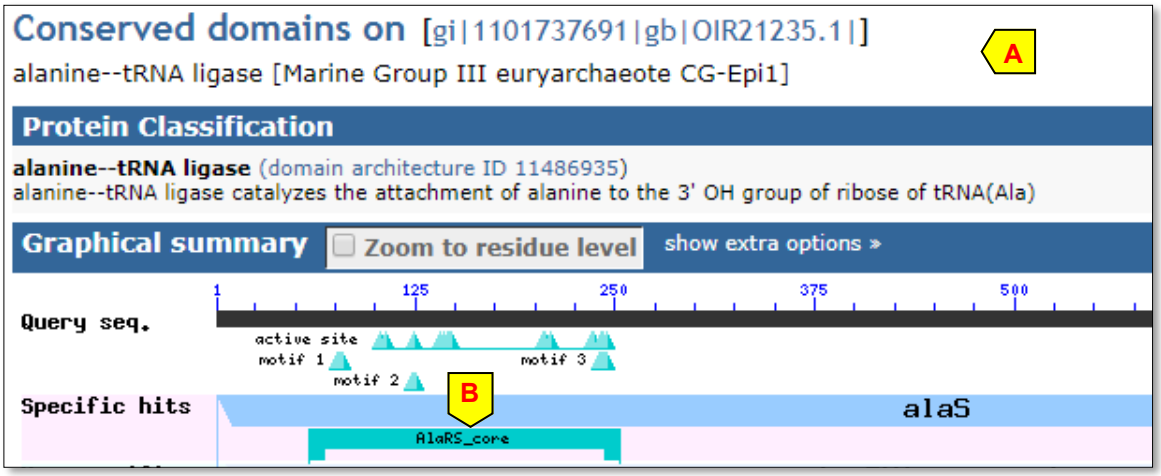
- Use blastx to find a structure record related by sequence to your annotated protein
- Confirm that your protein contains important sequence motifs for the conserved domain
- View these motifs on the solved structure using the iCn3D or Cn3D viewers.

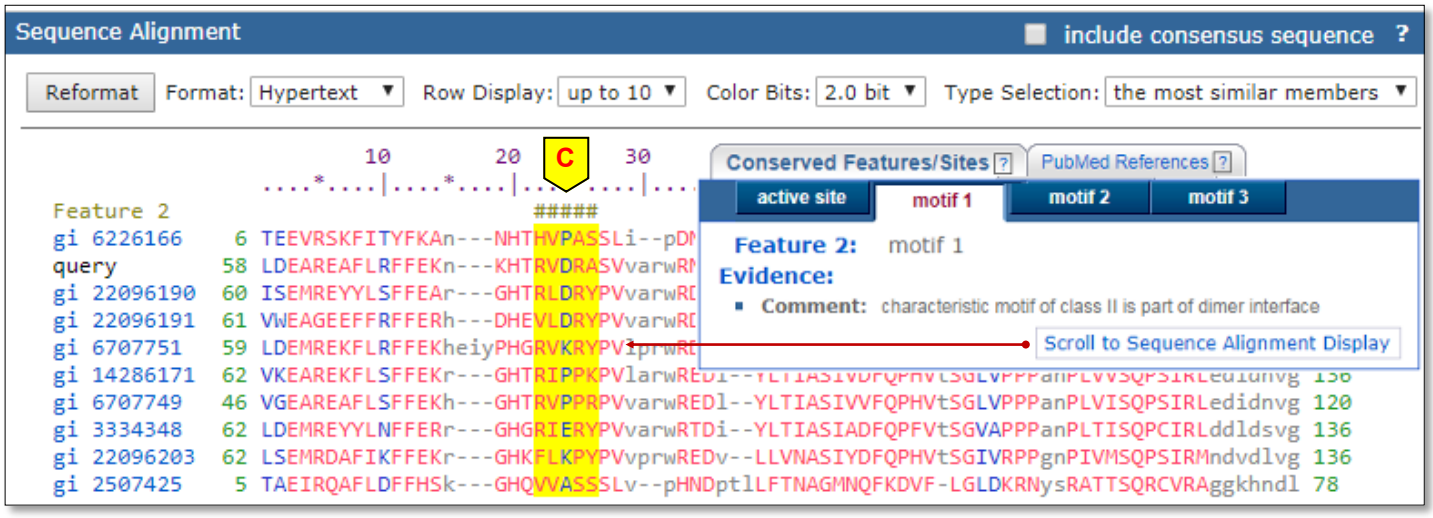
Background:

- 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.



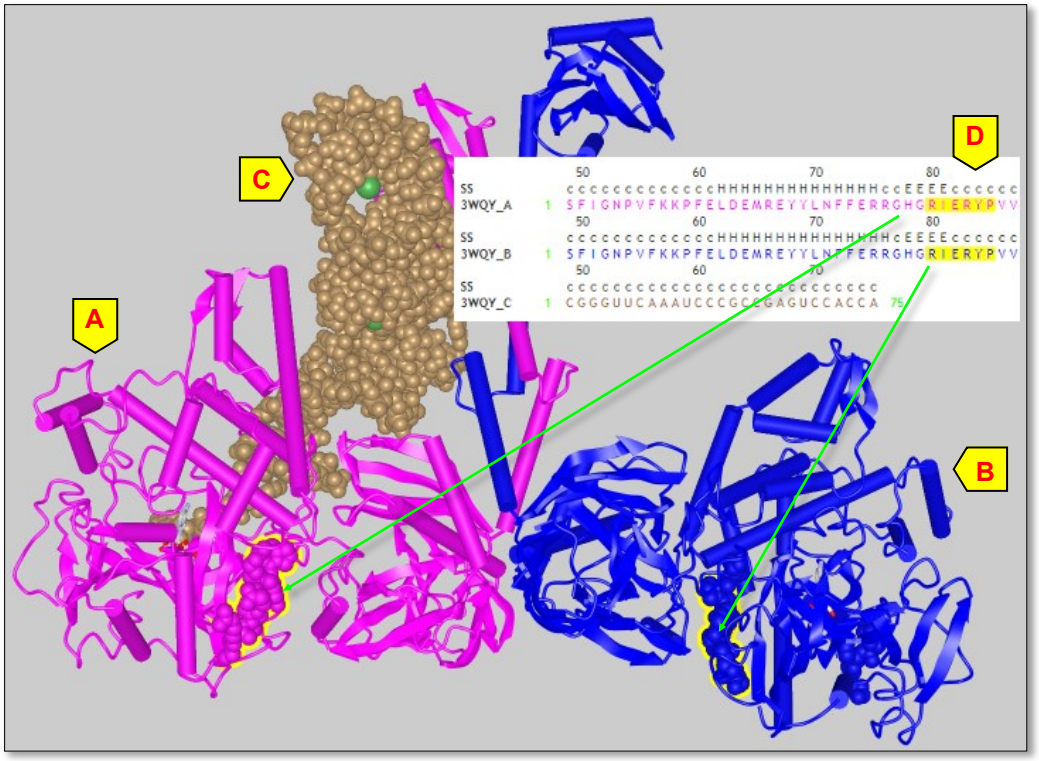


Case 5 (cont.)

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 verison 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
For questions and feedback on subjects not related to BLAST, email:
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-YbrIleFQlKMq
Factsheets on popular resources and common tools are available at:
https://ftp.ncbi.nlm.nih.gov/pub/factsheets/README.html