**Using NCBI BLAST**

**Identifying sequences**

Michael Crichton’s fantasy about cloning dinosaurs, *Jurassic Park*, contains a putative dinosaur DNA sequence. Use basic nucleotide BLAST against the nucleotide database, nr, to identify the real source of the following sequence from the novel. You can retrieve the sequence from the NCBI ftp site:


**Select, copy and paste the sequence into the BLAST form window and run the search against the nr(nt) database. Use the default Megablast algorithm.**

What is the sequence that Michael Crichton used?

This search is an example of the most common use of nucleotide-nucleotide BLAST: sequence identification, establishing whether an exact match for a sequence is already present in the database.

Mark Boguski, who was at the NCBI at the time, noticed this obvious contaminant and supplied Crichton with a better sequence for the sequel, *The Lost World.*

You can also retrieve this sequence from the NCBI ftp site:


**Select, copy and paste the sequence into the BLAST form window and run the search.**

Identify the most likely source of this sequence using nucleotide-nucleotide BLAST.

Mark imbedded his name in the sequence he provided. To see Mark’s name, use the translating BLAST (blastx) page with the sequence. (Look for MARK WAS HERE NIH).

The most important use of the translating BLAST services is to look for similar proteins (identify potential homologs) in other species.

**Short Nucleotide Sequences and Algorithm Parameters**

A frequent use of nucleotide-nucleotide BLAST is to check the specificity oligonucleotides for hybridization or PCR. The goal most people have when doing this is to make sure that the primer will give a unique product from the target genome or cDNA population. Because BLAST is local and searches both strands, one can simply concatenate a pair of +/- strand primers and use them
in a single search. You can try the traditional method below with this set of primers then use the new PrimerBLAST tool, linked under “Specialized BLAST” to check the primers.

Combine the following pair of candidate PCR primers in a nucleotide-nucleotide search against the nr(nt) database. Be sure to choose blastn (Somewhat similar sequences) as the BLAST program under “Program selection.”

F12 GTCAAGTGCACTCGTCAG
R8   TTGAGAGATGGATTGCTC

To prevent false matches that overlap the forward and reverse primer sequences, type ten or more “n’s” between the sequences when using them as a query.

GTCAAGTGCACTCGTCAGnnnnnnnnnTTGAGAGATGGATTGCTC

Retrieve the results and identify the gene amplified by these primers.

What is the predicted size of the product that would be amplified by PCR from cDNA (RT-PCR)? How could you distinguish the products amplified from genomic DNA versus cDNA?

You can also try these primers against the human genomic plus transcript database to get a clearer view of the product predicted from genomic DNA in the Map Viewer.

Now try these modified primers in standard nucleotide-nucleotide BLAST. There is one mismatch in each near the middle.

F12_mod GTCAAGTGCGACTCGTCAG
R8_mod TTGAGAGATGTATTGCTC

GTCAAGTGCGACTCGTCAGnnnnnnnnnTTGAGAGATGTATTGCTC

Notice that the previous hits are completely missing. This is because the default word size setting requires an exact match of 11 before extensions can occur. A mismatch in the middle of a 21-mer will prevent any initial word hits. There is an automatic adjustment for short sequences that will allow these hits with mismatches to be found. However the sequence with the linking “n’s” is too long to trigger the adjustment.

Run the search again with the forward and reverse primers as separate sequences. Copy and paste the following FASTA formatted primers in the search box.

> F12_mod
GTCAAGTGCGACTCGTCAG
>R8_mod
TTGAGAGATGTATTGCTC
Your results should now display a message that your search parameters were adjusted to search for a short input sequence, and you should see results for both primers. Notice that although there are now hits, the original hits are still missing. This is because the expect value of the mismatch hits is above 10.

You can manually adjust search parameters to short sequence setting through the “Algorithm parameters” section of the nucleotide BLAST form. After adjusting these, the search with the concatenated mismatched primer will work.

Go back to the BLAST form. Click on the reset page link at the top to restore the default settings. Then select blastn under “Program Selection” and expand the “Algorithm parameters” section of the form. Make the following changes.

- Uncheck the box next to “Automatically adjust parameters for short input sequences.”
- Increase the expect threshold to 100.
- Set the Word size to 7
- Set the Match Mismatch Scores to 1, -3
- Uncheck any Filter options

Now run the search again with the concatenated mismatch primers.

GTCAAGTGGCgACTCCGTCAgnnnnnnnnTTGAGAGATGtATTGTTGCTC
Do you find the original hits now?

**Protein-protein BLAST and Short Peptides: ELVIS lives**

As the database grows, so does the number of chance occurrences of amino acid motifs that spell out words or people’s names in single-letter amino acid codes. One such name motif is ELVIS. In this example we will count the number of occurrences of ELVIS in the default protein database. The automatic adjustment of search parameters will allow us to find matches with this short peptide.

**Type ELVIS in the search box on the blastp form.**

**Expand the Algorithm parameters section and adjust the number of Max target sequences to 1000 or more to include all Elvises.**

**Run the search.**

What is the expect value for an exact match to ELVIS? The number of Elvises increases in a linear fashion with the size of the database in accordance with the random behavior of protein sequences.

**Click on the “Edit and Resubmit” link at the top of the BLAST form. Examine the Algorithm parameters section to see how the settings were adjusted to search with this short peptide.**
PSI-BLAST and Conserved Domains

The histidine kinase-like ATPase domain (HATPase_c) is present in a wide variety of proteins with quite different functions. These include bacterial sensor histidine kinases, DNA mismatch repair proteins, topoisomerases, DNA gyrase and 90 KDa heat shock protein homologs. We can use PSI-BLAST to demonstrate the similarity among these proteins that is not apparent with ordinary BLAST.

Use the human DNA mismatch repair protein MLH1 (NP_000240) in an ordinary blastp search and examine the conserved domain results to verify the presence of the HATPase_c domain.

From the results of the above search, click the “Edit and Resubmit” link and make the following changes to prepare to run a PSI-BLAST search with just the region of MLH1 that corresponds to the HATPase_c domain

- Set the query subrange in the boxes on the right hand side of the form. Use 32 as the “From” coordinate and 122 as the “To” coordinate.
- Change the database to “swissprot.”
- Change the “Program Selection to PSI-BLAST.”
- Expand the “Algorithm parameters” section and set the “Max target sequences” to 5000.

Now click the BLAST button to run the first iteration of PSI-BLAST and examine the results.

The results are just the blastp results that are formatted for PSI-BLAST. Notice that the descriptions section of the results is divided into two sections. The upper section contains the sequence with alignments that will be used to generate the position specific score matrix in the next iteration of PSI-BLAST. These sequence alignments have e-values less than 0.005. This cut-off is empirically determined to give good results in PSI-BLAST searches. All of the proteins above this threshold in the first iteration are DNA mismatch repair proteins PMS, MutL and HexB homologs. Just below the PSI-BLAST threshold with e-values ranging from 0.008 to 6.0 are several bacterial signaling histidine kinases. Some of these have marginally significant e-values in ordinary BLAST but many are not distinguishable from chance matches.

Now click the “Run PSI-BLAST iteration 2” button to run the second iteration of PSI-BLAST and examine the results.

There are now new proteins less than the 0.005 threshold. Notice that these are now marked with a “New” graphic while the proteins found in the previous iteration are marked with a green ball. Many of the new proteins are topoisomerases or DNA gyrase. There are also many more gyrase and topoisomerases just above the 0.005 threshold.

Retrieve a few of the new proteins in Entrez by clicking on the linked identifier and verify that that they contain the HATPase_c domain.
Click the “Run PSI-BLAST iteration 3” button to run the third iteration of PSI-BLAST and examine the results.

Again there are new proteins, not only more gyrases and topoisomerases, but also signaling histidine kinases and HSP90 chaperonins.

Continue to run PSI-BLAST iterations until you have collected some plant phytochrome and ethylene receptor proteins below the 0.005 threshold.

These are plant signaling proteins. As these results show, plant ethylene receptors and phytochromes are related by sequence similarity to the two component sensor kinase system of bacteria.

Demonstrate the similarity between the HATPase_c domain of the E. coli sensor protein PhoR (PHOR_ECOLI, P08400) and plant ethylene receptors by performing a first iteration PSI search against swissprot. Use a query subrange on PhoR of 318 to 421.

Now, continue to run PSI-BLAST iterations until the plant phytochromes appear.

The number of iterations should be fewer than when using the MLH1 protein as a query.

Retrieve the protein record for an ethylene receptor (ETR1_LYCES, ETR1_ARATH) and a phytochrome (PHYA_ARATH, PHY_PICA) by following the link to Entrez. Compare their domain structures by following the links to the pre-computed Conserved Domains results.

What three domains do they have in common?

Retrieve a protein record for one of the bacterial sensor proteins (PHOR_ECOLI) and examine its domain structure.

Notice that the plant proteins and the bacterial protein all contain the histidine kinase domain (HATPase_c) and the HisKA (phosphoacceptor) domain. In the classic two component bacterial system, the HisKA domain is phosphorylated on a conserved histidine residue by the HATPase_c domain in response to an external signal. This phosphate is then transferred from the HisKA domain of the sensor protein to a conserved receiver domain on a separate response regulator protein. In the case of PhoR the response regulator is PhoB.

Retrieve the E. coli PhoB protein (PHOB_ECOLI, P0AFJ5) and examine its domain structure as before.

Notice the presence of the receiver domain (REC) and a DNA binding effector domain (trans_reg_C) in PhoB. In the plant ethylene receptors examined previously there is a receiver domain is on the receptor itself, but the effector domain present in PhoB is lacking. The plant ethylene receptors apparently mediate their effects through the MAP-kinase pathway. Unlike the ethylene receptors, the phytochromes function as serine/threonine kinases but also appear to share an ancestry with bacterial histidine kinases.
Translating BLAST searches, mining polymorphisms

The prion protein is found in high concentrations in the brains of humans and other mammals. In certain degenerative neurological diseases, prion proteins aggregate into polymers. Several of these prion diseases seem to be transmissible. Perhaps the most remarkable aspect of these is that the infectious agent appears to be an aberrant form of the prion protein itself. Bovine spongiform encephalopathy (BSE) is one of the transmissible prion diseases that has received much recent notority. There are a number of polymorphisms that have been identified in the prion proteins for several mammals, notably human, mouse, and sheep. Some of these are associated with inherited prion diseases and some with susceptibility to transmissible forms.

Retrieve the SWISS-PROT record for the human prion protein (PRIO_HUMAN, P04156) and look at the FEATURE table to see the various polymorphisms.

Notice the methionine / valine polymorphism at position 129. The amino acid at this position affects the particular disease phenotype when another disease causing mutation is present. People who are heterozygous at this position appear to be more resistant to kuru, one of the transmissible encephalopathies. There is population genetic evidence that their may have been balancing selection for heterozygotes at this position during human evolution. The EST data for human represents a large number of individuals and can be used as a resource for identifying nucleotide polymorphisms. In this case, we can investigate the prevalence of the two alleles at position 129 of the prion protein in the EST data for human. We will use one of the formatting options to make the different alleles easier to identify.

Set up and run this search by following these steps:

- From the BLAST homepage, link to the tblastn form “Search translated nucleotide database using a protein query.”
- Type the prion protein accession number, P04156, in the search text area.
- Use the “Query subrange” boxes to use only residues 100 to 160.
- Choose the “Expressed sequence tags (est)” database.
- Type “human” in the Organism limit box and choose human (taxid:9606) from the resulting list to limit to human sequences.
- Open the Algorithm parameters section and set the Max target sequences to 1000.
- Turn off the “Low complexity” filter option.
- Click the BLAST button to run the search.
- Immediately click the “Formatting options” link at the top of the intermediate page.
- Set the alignment view to “Query-anchored with dots for identities.” This is a stacked pairwise alignment format that makes it easy to see changes relative to the query sequence in all the database hits at once.
Click the “View report” button to display the results.

Look at the alignments to see how the query-anchored format helps to investigate changes in sequences. Find position 129 in the query. Which amino acid is most prevalent at position 129?

WGS and Trace Archive Data in Entrez and BLAST

Verify that nearly all of the rabbit DNA records in the NCBI database are whole genome shotgun. You can retrieve all nucleotide rabbit sequences by using the Limits tab and setting the field restriction in the pull-down list to organism. You can further limit to genomic DNA through the “molecule” pull-down list.

How many records are there?

Follow the link to the “CoreNucleotide” results before continuing. Now restrict to whole genome shotgun records by adding the following query term to your search.

wgs[Properties]

The overall search performed now is


The first record is the master record for the project that gathers all of the contigs. You can get only this record by adding wgs_master[Properties] to the search.

Retrieve the first contig record in your list and verify that it is unannotated—no genes or other features.

Using BLAST, Spidey and Splign to annotate wgs

You can find the genomic sequences corresponding to a rabbit (Orycolagus cuniculus) mRNA sequence by using BLAST to search the wgs database. A sequence that demonstrates this is the rabbit apolipoprotein A-1 mRNA (NM_001101687).

- From the BLAST homepage select the blastn page
- Type NM_001101687 in the “Search box” and select wgs as the database.
- Use the Organism limit feature to limit to rabbit (taxid:9986)
- Expand the Algorithm parameters section and set the e-value threshold to 1e-12.
- Run the search and re-format your results using the “CDS feature” option and “Pairwise with identities” Alignment view option.
Your search should hit one wgs contig (AAGW01335306). How many exons did you identify in each?

**Use the sort by “Query start position” to put the exons in the genomic order on AAGW01335306.**

This is a rather primitive gene model because it does not constrain the alignment breaks to splice junctions.

**Use the same mRNA and genomic sequences as above, make gene models using the spliced alignment tools Spidey and Splign and compare them to the BLAST results.**

The spliced alignment tools place two of the exon-intron boundaries at slightly different points than BLAST alignments.

**Trace Archive**

Some sequences are only available through the NCBI trace archive.


These data can be retrieved by species code or trace number. The most important way to search these is through the Trace archive megablast pages. Both standard (contiguous) megablast and cross-species megablast are available. These are linked through the BLAST tab on the main trace archive page or through the BLAST homepage in the Specialized BLAST section.
We can use the cross-species page to find an HSP70 gene homolog in the sea lamprey (Petromyzon marinus) traces

- Go to the BLAST homepage and choose the trace archive search from the Specialized BLAST section.

- Enter the accession number for the human HSP70 1A mRNA Reference Sequence (NM_005345) in the search box on the BLAST form.

- Choose Petromyzon marinus-WGS as the database and set blastn as the program.

- Click the BLAST button to run the search.

Because HSP70 is well conserved it is easy to find homologs in the sea lamprey at the nucleotide level using the human sequence. Many less well-conserved genes may only be identified at the protein level. Unfortunately the large size of the trace databases makes translating searches impractical.

New BLAST Displays

TreeView

The treview display in BLAST will not always produce reasonable phylogenetic species trees or gene trees because the alignments are not multiple sequence alignments and don’t necessarily
include all residues. Nevertheless searches with complete mitochondrial genomes often reproduce accepted phylogenetic groupings.

**From the BLAST homepage, choose the blastn page.** Select the RefSeq genomic database from the database pull-down list and put the accession for the wolf mitochondrial genome (NC_008092) in the “Search” box as a query.

The Refseq genomic database contains chromosome (NC_) RefSeqs including plastid genomes, mitochondrial genomes and chromosomes for prokaryotic genomes.

**Use the following Entrez limit to restrict to the mammalian order carnivora (dogs, cats, seals, hyenas, weasels etc.).**

carnivores[organism] NOT gene in genomic[properties]

This last term, “NOT gene in genomic[properties]”, eliminates hits to mitochondrial insertion sequences present in the dog genome.

**Run the search. Click on the “Distance Tree of Results” link under the BLAST graphic to display the tree. Compare the groupings to the classification of the carnivores in the NCBI Taxonomy database.**

The family groupings correspond to those in the tree. However, many families of carnivores are not represented because the mitochondrial genomic sequences are not available yet.

**New View of Results and Genome and Transcript Databases**

The new human genome and transcript database provides direct access to the human genome through the main BLAST page. The new view options provide a more organized and sortable presentation of the results.

**Use nucleotide-nucleotide BLAST (blastn) to search the human genome plus transcript database with the human alcohol lactate dehydrogenase B (LDHB) transcript (NM_002300).**

**Use the new sorting options and summary statistics to identify the functional multi-exon gene by sorting using the “Total Score” column.**

On which chromosome is the functional gene? How many exons does it have?

**Use the “Sort alignments” feature to sort by “Query start position” to get the exons in genomic order.**

On which chromosome is the longest retrocopy pseudogene?
Follow the linked identifiers to the human Map Viewer to display the hits for both the functional gene and the retrocopy pseudo gene.