

NCBI BLAST Services

Protein BLAST

Query: human brain-type creatine kinase, NP_001814

Program: blastp

Database: refseq_protein

Goals:

- Explore creatine kinases in mammals using Reference Sequences.
- Demonstrate the usefulness of organism limits, taxonomy report, and the link to multiple sequence alignment.

Procedure:

- Retrieve **NP_001814** from the Entrez protein service.
- Click "Run BLAST" under the Analyze this sequence portlet.
- Select **refseq_protein** as the database.
- Enter "mammals" in the Organism input box, select from the suggested list
- Click **Algorithm Parameters** and increase the **Max target sequences** to 1000.
- Click **BLAST** to submit the search
- The matches have different types of RefSeq accessions with XP entries representing proteins from gene models.
- Click **Taxonomy report** to see the organism distribution and examine the matches for dog.
- Click the linked accession XP_537561.

Gene models like this one may be incomplete due to missing data in the genome or may represent potential but unsupported splice variants. There are numerous, probably too many, splice variant predicted for the dog genome.

- From the results page, click **Edit and resubmit**
- Check the Exclude **Models (XM/XP)** checkbox
- Click **BLAST** to submit the search.
- The results now contain only NP_ style accessions, experimentally supported gene products. Click **Taxonomy report**, to see the different creatine kinase products found in humans and other mammals.
- Return to the BLAST results, click on the **Distance tree of results** to see a graphic presentation of the relative between the different proteins. The mitochondrial and cytosolic isoforms are two distinct clusters.
- You can extend this search to a Multiple Alignment to obtain a more accurate tree.

Nucleotide BLAST

Query: Macaque CDC20 mRNA, AB168636

Program: nucleotide BLAST page with megablast and blastn

Database: human G+T, refseq_genomic (limit to "marmosets and tamarins")

Purpose:

- Map a sequence onto various genomes
- Compare the speed and sensitivity of various algorithms
- Use the different sorting options in BLAST results
- Use formatting options, CDS feature.

Procedure:

- Retrieve AB168636 from Entrez nucleotide and follow the link to **Run BLAST**.
- Select the **Human genomic + transcripts** database, click **BLAST**.
- Examine the **Graphic summary** and **Descriptions** sections.
-

Notice that there are separate sections for the transcripts and genomic regions. There are two genome assemblies represented: the reference genome, GRCh37; and the alternate assembly, Celera. There are hits to chromosome 9 and chromosome 1 in both assemblies. The retro-transposed pseudogene on Chromosome 9 actually ranks higher than the functional gene because of the single uninterrupted single hit outscores the individual exon hits for the functional gene. Re-sorting the output by **Total score** and/or **Max Ident** bring match to the functional gene to the top of the list.

- Click on the linked score for NT_008470 and examine the alignment to the pseudogene.

Notice the single nearly complete alignment with no introns. The poly-A tail from the mRNA is even present in the genome. This is an example of an apparent retro-transcribed mRNA that has been inserted into the genome.

- Click the linked score for NT_032977 to go to the alignment
- Click **Query Start position** to arrange the matches according to exon order

The first aligned segment starts at position 73 of the mRNA. Megablast misses the first exon hit as well as a match to some related transcripts. Re-running the search with blastn finds this hit. You will need to set the Expect threshold to 1e-6 to avoid additional non-significant matches.

Linking to the Map Viewer

The linked identifiers for the genomic sequences from the Human genomic plus transcript and the Mouse genomic plus transcript searches display the results in the NCBI Map Viewer. This is also the behavior for the organism-specific genome BLAST pages linked at the top of the BLAST homepage.

- Follow the linked identifiers for the hits on chromosome 9 (NT_008470) and chromosome 1 (NT_032977) to display the hits in the map viewer. Make a note of the surrounding genes. You can compare these later to results in the mouse and chimp.

Formatting Options CDS Feature

- Open **Format options** link, check **CDS Features**, click **Reformat**

This adds the translation to the nucleotide alignment if coding regions are annotated on the query or subject (database sequence).

- Examine the alignment to the human transcript NM_001255.

The macaque mRNA sequence has a single base deletion relative to the human transcript. This results in a frame shift making the protein translation diverge at the C- terminus. This is most likely a sequencing error as the other mammalian CDC20 proteins agree with the human sequence. You can use a blastx search with AB168636 to demonstrate this frame shift as well.

Finding matches in Assembled Whole Genome Shotgun genomes

- Click **Edit and resubmit**,
- Change the database to refseq_genomic.
- In the Organism box, type **marmoset**, select **marmosets and tamarins**
- Submit the search

Similar retro-transposed pseudogenes are also present in the white-tufted-ear marmoset. The functional gene appears to be on chromosome 20. Sorting the alignment segments by query start position helps arrange the hits in a more natural order. Since searching/formatting choices are sticky, the translation is automatically displayed.

Align two or more Sequences, Global alignment, and Multiple-alignment

Align 2 sequences

Query 1: Human Albumin, NP_000468

Query 2: Human GC, NP_000574

Program: blastp

Procedure:

- Retrieve NP_000468 from the Entrez protein system.
- Follow the link to **Run BLAST** from the **Analyze this sequence** portlet on the protein record.
- Check the box that reads **Align 2 or more sequences**.
- Enter NP_000574 in the subject sequence box.
- Click BLAST
- Expand and examine the **Dot Matrix View**

Off-diagonal elements show that more than one local alignment is found between these two sequences with a repeated domain structure.

Needleman Wunsch Global Sequence Alignment

Query 1: Human Albumin, NP_000468

Query 2: Human GC, NP_000574

Program: Protein

Procedure:

- Click on the **Global Sequence Alignment Tool** link in the **Specialized BLAST** section of the BLAST homepage.
- Click the **Protein** tab over the Query sequence text area.
- Click the **Align** button

The tool finds a single global alignment between the two sequences.

Align more than two sequences (BLAST) and extend to a multiple-alignment

Query 1: Human Albumin, NP_000468

Query 2: Human AFP, Human AFM, Human GC proteins

NP_001125

NP_001124

NP_000574

Enter these one per line.

Procedure:

- Retrieve NP_000468 from the Entrez protein system.
- Follow the link to **Run BLAST** from the **Analyze this sequence** portlet on the protein record.
- Check the box that reads **Align 2 or more sequences**.
- Enter NP_000574, NP_001125, NP_001124, one accession per line, in the subject sequence box.
- Click BLAST
- From the results click the Multiple Alignment link
- Generate the Phylogenetic Tree from the COBALT results.

Explanatory Notes:

The "Align 2 (or more) sequences" service is now combined with Basic BLAST. Checking the "Align two or more sequences" on the BLAST form will transform the BLAST form to allow direct comparison of two input sequences. This service produces only local alignments since this is BLAST. In cases such as the albumin family used here -- where there is a set of repeated domains, more than one alignment is found. This is easily seen in the dot matrix graphic of the alignments found between albumin and the vitamin D binding protein. The new Needleman-Wunsch alignment tool allows a global comparison of albumin and the vitamin D binding protein and produces the single best alignment that includes all residues.

Entering more than two sequences in the search boxes allows a search against a small custom database. In this case comparing the albumin sequence to the other three members of the family produces pairwise local alignments equivalent to a small database search. As before there are more than one local alignment reported for some sequences. The new COBALT extension to BLAST linked through "Other reports" produces a true global multiple alignment of the four proteins. The Download link at the top of the COBALT output allows export of the alignment for local editing. The Phylogenetic Tree link produces a more accurate distance tree of the albumin protein family than could be obtained from the BLAST alignments. COBALT is available as an extension on all protein BLAST results. A direct interface to COBALT is linked to the "Specialized BLAST" section of the BLAST homepage.

Primer BLAST: Guided Practice

Query: Human FOXP2 mRNA splice variant 2, NM_148898

Organism limit: human

Database: RefSeq RNA

Allow splice variants: off at first then on

- Retrieve NM_148898 from the Entrez nucleotide system.
- Follow the link to **Pick Primers** from the **Analyze this sequence** portlet on the nucleotide record.
- Select **Use new graphic view** at the bottom of the form to see results in the graphical sequence viewer.
- Run the search with the default settings.

This designs primers that only amplify the single splice variant from the background of human mRNA (cDNA). Are all primers clustered in one region?

- Check the first primer pair without a template and use the human reference genome as a background database.
 - o Is this primer pair expected to amplify genomic DNA? Why would it fail?
- Try the search again allowing the primers to amplify splice variants

Explanatory Notes:

The FOXP2 gene has multiple splice variants. It is useful to design primers that will amplify only one to investigate tissue specific expression. Primer BLAST can use information on splice variants to design specific primers with an NCBI mRNA Reference sequence template.

The stringency can be relaxed by selecting "Allow primer to amplify mRNA splice variants". In this case primer pairs can be found that amplify all variants.