

## NCBI Exercises

The Zebrafish gene product u-boot is a homolog of the human PR-domain protein 1 PRDM1 and the mouse Blimp1. The following exercises use the u-boot mRNA and protein sequences to explore NCBI resources.

1. Use the Entrez system to find the u-boot mRNA sequence. Perform a global search for u-boot and retrieve the results for the Nucleotide database. Then follow the link to the CoreNucleotide subset. You should retrieve the GenBank record and the RefSeq mRNA record. It also retrieves the annotated whole genome shotgun supercontig from the zebrafish genome assembly. From the links menu of either mRNA record follow the related sequences link. This provides a list of similar nucleotide sequences. Notice that the only related sequences are other zebrafish records. The nucleotide similarity cutoff is set very stringently, and it is unusual to find homologs from any other species in this list. There are two additional records that you did find not in the original search. One of these, AL928573, is the finished sequence of a BAC clone from the zebrafish genome project. Retrieve this record and examine the FEATURES table. Notice that the u-boot gene is not annotated on this record. You could not have found it by performing any kind of text search for u-boot; you found it only because of the pre-computed sequence similarity feature. AL928573 contains all of the exons of the u-boot gene. You can demonstrate this for yourself using the mRNA to genomic alignment tool Spidey.

<http://www.ncbi.nlm.nih.gov/spidey>

You can also try the newly public Splign tool. This is the mRNA to genomic alignment tool used in NCBI genome annotations for zebrafish and other genomes.

<http://www.ncbi.nlm.nih.gov/sutils/splign/>

Notice that we now have annotated this gene on the zebrafish map viewer. Follow the map viewer link from RefSeq mRNA record to see the graphical exon-intron structure and other features in this region of the zebrafish genome. The graphical alignments of the mRNA sequences including ESTs shown in the Map Viewer are created using Splign.

2. Link to the protein record from the GenBank u-boot record (AY497217). Follow the related sequences link. Notice the very large number of related records. The stringency setting for the protein neighbors is set so that it is easy to find homologs in other species. In this case, the very large number of neighbors is caused by the zinc finger domains that are present in a wide variety of proteins that bind to DNA. Many of the proteins in the list share similarity only with the zinc finger region of the u-boot protein. The

list of related sequences is in order of decreasing BLAST score; thus the records near the top of the list are those that usually share similarity along their entire lengths. You can easily find the human homolog, PR-domain zinc finger protein 1, and the mouse homolog, B-lymphocyte-induced maturation protein 1, near the top of the list. Notice that the list is redundant, and because of differences in nomenclature it can be confusing to try to pick out those records that represent the same protein. An additional problem is that older records do not have the organism information displayed in the summary.

The BLink link, in many cases, is a better option for identifying homologs in other species. Follow the BLink link from the u-boot protein record (AAR87139). BLink provides an output that is equivalent to a protein-protein BLAST search against the nr protein database. The output is limited to the top 200 BLAST hits. Keep in mind that there are over 19 thousand related proteins, in part because of the zinc fingers as mentioned earlier. Because of the top 200 limit, important homologs may be missing from this output. The two conserved domains present in this protein, the SET domain and the C2H2 zinc finger, are previewed at the top of the graphical overview. The graphical overview allows you to pick out the proteins that share similarity over the entire protein rather than in just the zinc finger regions. To get more detailed information about the related proteins, you can click on the hyperlinked BLAST score to see the BLAST 2 Sequences alignment. The list of related proteins can be made simpler by clicking the “Best Hits” button at the top of the page. This will display the single best protein match for each species in the list. This is a very fast way to identify the closest homolog in another species. The human PRMD1 and mouse Blimp1 can be easily found. To see what groups of organisms are involved, click on the Taxonomy Report or the Common Tree buttons. From there you can link to the Taxonomy browser for detailed information and links to all molecular records for that organism. For example: What kind of organism is *Tetraodon nigroviridis*?

3. BLink also allows you to see search results against different databases. You can use the pre-computed results against the proteins from PDB to quickly find a structure for the zinc finger. Choose PDB from the “Keep only” drop-down menu from the BLink results from the preceding exercise and press “Display.” This shows alignments of u-boot with PDB proteins. The proteins from PDB will have links to the structure database. Follow the link to the 1UBDC protein record. This protein sequence was extracted from the structure record. The presence of these sequences from structures in the sequence database combined with the related sequences feature provides a rapid mechanism for finding potential structural models for proteins without experimentally determined structures. Follow the “Structure” link from the 1UBDC protein record. Follow the linked identifier to the “Structure summary”.

An equivalent and shorter method for finding a related structure is provided in the Links menu of protein records as the “Related Structure” link. Go back to the protein record for u-boot (AAR87139) and from the links menu click on “Related Structure”. The resulting output shows graphical alignments with the structure identifiers linked on the left hand side. From here you can follow the 1UBD\_C link directly to the Structure database to see the structure summary.

This structure contains one protein chain, (chain C), a zinc finger protein, bound to a segment of double stranded DNA (chains A and B). Click the view 3D structure button to display the structure in the NCBI structure viewer Cn3D. If the viewer is installed, it launches automatically. If it is not installed, you can follow the link to “*Get Cn3D 4.1*”. This leads to instructions for downloading and installing the viewer.

The viewer clearly shows the way the zinc fingers lay in the major groove of the double stranded DNA. You can use the Style menu to look at different aspects of the structure. Use the “Edit global style” dialog to change the way the DNA and protein are rendered. Identify and highlight the amino acid side chains involved in coordinating the zinc ions.

4. The *Drosophila* protein enhancer-of-zeste is one of the original proteins identified with the SET domain. The zebrafish u-boot also contains a special kind of SET domain, the PR-domain also present in many mammalian proteins. However, ordinary BLAST finds no significant similarity between most PR-domain proteins and enhancer-of-zeste. Perform a protein-protein BLAST search against the swissprot protein database using enhancer-of-zeste (accession P42124) as a query. Focus only on the SET domain by setting the coordinates 626 to 742 in the “Set subsequence” boxes. To focus on finding animal homologs, limit the database by typing the following in the “Limit by Entrez query” box:

animals[Organism]

Retrieve the BLAST results and verify that there are no mammalian proteins with PR-domain in their titles. If present, you can recognize them by their swissprot identifiers. These are of form PRDM2\_RAT, PRDM2\_HUMAN, etc.

We can now demonstrate using PSI-BLAST that the PR-domain mammalian proteins are significant matches to enhancer-of-zeste. Go back to the BLAST formatting window and check the box that is labeled “Format for PSI-BLAST.” Click format to reformat your results. On the results, there is now line separating significant hits that will be used in construction of the Position Specific Score Matrix (PSSM) from the non-

significant hits. Press the “Run PSI-BLAST iteration 2” button. Be sure you can still see the formatting window when you do this; PSI-BLAST will now refresh the formatting window providing a new Request ID for the position-specific results. Continue running PSI-BLAST iterations in this manner until PR-domain containing proteins appear as significant hits. You have essentially created a PSSM that is equivalent to one used for the CD search to find this relationship. Keep in mind that the CD results that you got with the initial BLAST search already told you of the presence of this domain.