

Basic Bioinformatics

Purpose of this protocol: From your literature reading, you might find proteins from other model systems (for example mouse) that have known roles in apoptosis, either upstream or downstream of IAPs. You will need to find out:

- A Search GenBank for the mammalian gene(s) mentioned in the literature, get the GenBank accession numbers of these genes;
- B Search FlyBase to identify Drosophila homologs and to find out more information on these genes – CG number of the gene, coding and 5' and 3'-UTR sequences, and other information on the function of this gene. Also find out if it has alternative isoforms;
- C Search OpenBiosystems to find out if they carry RNAi clones of these genes, and get the clone ID and catalog number;

This protocol has three sections, to answer the above three questions.

Part A: Find genes in GenBank by name.

- 1) Go to NCBI website: <http://www.ncbi.nlm.nih.gov/>. Under “Search”, select “Nucleotide” and type in the gene name and species that you want to search for. Use Booleans, and tags which you can find at the PubMed help page: <http://www.ncbi.nlm.nih.gov/entrez/query/static/help/pmhhelp.html#SearchFieldDescriptionsandTags>
If you know the author who published the paper on the gene you are interested, you can also include the author name in your search along with the gene name and species.

Note: You will usually get more than one result. Read the title of each to screen out the ones that are obviously not what you are looking for. Then click on the accession number and read more details to see if the gene is what you want.

- 2) Once you are at the page showing the gene sequence, there are several things you should pay attention to:

The general output: from top to bottom, the page shows general description of the gene such as which organism, the authors and titles of the literature references listed, the translated protein sequence of this gene, and the gene sequence.

Some specific things you should notice:

Accession Number (write that down, you will need it later);
Pubmed and Medline reference numbers (also author, journal name and title):
click on them to double check this is the gene you are looking for;
Under “Features”, you will find “CDS” and number to the right of it. Say the number is “172...1853”. This means the open reading frame (coding sequence) starts from 172 and ends at 1853 from the nucleic acid sequence given below;

You should save this page (can be .doc or .txt file) and keep one printout of this page for later reference.

- 3) Get the protein sequence of this gene, copy and paste and save the sequence (do a good annotation in your .doc or .txt file, so that you can always go back and find out what exactly you are looking at.

Part B: Do BLAST search to find Drosophila homolog of this gene and get more information on the gene.

- 1) Open a new browser window, and go to the FlyBase website:
<http://www.flybase.org/>. In the middle of this page, you will see the big blue “BLAST” sequence search, click on it.
- 2) Search Flybase using:
Drosophila melanogaster (euchromatin, rel. 4.0)
Feature type: predicted gene (NT)
Program: tblastn
- 3) Paste your protein sequence and hit search, then wait (patiently.....)
- 4) On the next page, the hits are color coded from weak (black) to strong (red), and sorted by strength of homolog from strong to weak. Look at E values – Hits with E-value $<1e-6$ are significant. If you do not see any hit with $E < 1e-6$, you can move on to another gene.
Otherwise, choose the strongest hit which has a FlyBase link (FBgn##), click on it.
- 5) This page gives you a brief report of the gene, including genomic organization, gene product, gene ontology, expression and phenotype, and finally a summary of the gene. Read through the brief summary and make sure it makes sense to you.

The top right side of this page shows the FlyBase ID, write that down. Under “Gene Product”, you can get complete coding region sequence with coding sequence highlighted in red, and 5’ and 3’ UTR sequence in blue. Save the sequence in word or plain text (does not matter) but make sure you highlight each of these regions so you can go back and find them. It is important when you design primers for RT-PCR to detect whether S2 cells express this gene.

On the right hand side, the inset gives you available reports on this particular gene. Click on Abridged report or full report, it brings you to the next page, which will give you CG symbols. Also on this page, you can get protein sequence from Swiss-Prot, as well as domain information within this protein.

Part C: Search OpenBiosystems to find if they carry RNAi clones of these genes, and get the clone ID and catalog number.

- 1) Go to <http://www.openbiosystems.com> , click on Drosophila RNAi collection, and download their current RNAi collection in .xls Excel format.
- 2) Open the file in Excel. With the CG number you searched from FlyBase, and/or with FlyBase ID number, you can search and find the catalog number from the file.