

“Searching For Fossil Genes”

Follow-up Activity:

Using ORF-finder to Identify GULO Sequences as Genes or Pseudogenes

David Form

Nashoba Regional High School

Bolton, MA

In the bioinformatics research activity, “Searching for Fossil Genes”, you used nucleotide BLAST to determine if various mammals have the intact GULO gene in their genome. This information was used to predict if these mammals can synthesize their own vitamin C. Protein BLAST was used to determine if these genes could be translated to produce a protein product. A more direct way to determine if a genetic sequence can be translated into a protein product is to analyze the sequence with NCBI’s Open Reading Frame Finder, or ORF finder.

An open reading frame, or ORF, is a genetic sequence that begins with a start codon, ends with a stop codon. An ORF is called a coding sequence (CDS) if it is actually translated into a protein product by the cell. ORF-finder identifies ORF’s in a genetic sequence in six reading frames, three for each DNA strand. Protein BLAST is used to determine if the translated ORF matches known protein sequences, in this case it will be the GULO protein. Failure of any ORF in your genetic sequence to match the GULO protein will indicate that mutations in the sequence have disrupted the CDS; the gene is now a pseudogene!

Procedure:

1. Go to the NCBI homepage @www.ncbi.nlm.nih.gov.
2. On the left side of the page, click on “Resource List (A-Z) and select” Open Reading Frame Finder” from the alphabetized list.

We will begin by analyzing those sequences, identified by BLAST, in which you are confident in their status as functional genes.

3. Choose one of these sequences and enter it into the query input box in ORF-Finder.
4. Click on the “ORFfind button”.

5. You will get a window which shows all of the ORF's in six reading frames. You will see many ORF's because random start and stop sequences will exist in all six reading frames. The largest ORF is most likely to be the correct one – the one that can be translated into the intact protein.
6. On the right is a table of the ORF's in order of decreasing size.
7. Select the largest ORF. Click on the BLAST button at the top of the screen.
8. Click on the "View Report" button.
9. The BLAST results will indicate whether the ORF that you chose can be translated into the intact GULO protein. You may want to choose one or two additional ORF's from the table to see what happens when these are translated and "BLASTed".

We will now analyze those sequences, returned by nucleotide BLAST, that you suspect represent GULO pseudogenes. Remember that these are sequences that appear to be broken up and poorly match the query sequence (the mouse GULO sequence). These occurred in organisms in which BLASTp did not identify a matching protein product.

For each sequence, follow the same procedure (steps 3-9) outlined above. If the sequence is a pseudogene, when translated, it will not produce a protein product that matches the GULO enzyme sequence.