

1. Go to the NCBI website at: <http://www.ncbi.nih.gov/>
2. Drop down the search box to **CoreNucleotide**, then enter the accession number (L03394) [L 'zero' 3394] in the box at the top of the page. Click on the L03394 link that appears on the results page.
3. The resulting page is a sequence report page. The DNA sequence reported here is for the entire cloned insert. We are really only interested in the protein coding region (labeled "CDS"). Scroll through the Features section and find the annotation for "goosecoid homeodomain protein". Below this is an ID number link to the protein sequence. Click on this to load the protein sequence.
4. At the far right near the top of the page is a tiny link that says "BLink". Click on this. BLink summarizes the results of protein BLAST searches that have been pre-run on the protein. You can see that there are many similar protein sequences in the database. A quick examination of several hits shows that most include "goosecoid", "homeobox", or "transcription factor" as a part of their names. The information retrieved can be examined in a number of ways. Click on the Common Tree button to see the hits arranged according to taxonomic groups.
5. We would like to see what the goosecoid homeodomain protein does. Return to the BLink screen. Click on the 3D structures button. The resulting list shows hits for which 3D structures are available in the database. To view these structures, your computer must have the Cn3D browser plugin installed. This plugin is already installed on all of the lab computers. On other computers you may need to download it from the NCBI web site. To load the 3D structure of a protein into Cn3D, click on the small blue circle to the right of its score. Click on the circle for the hit entitled: " Chain **B**, Homeodomain From The Drosophila Paired Protein Bound To A Dna Oligonucleotide". In the resulting window, click on the "Get 3D Structure Data" button. Two windows will appear: the 3D viewer window and a sequence/alignment window.

Examine the sequence/alignment viewer window. Notice that the amino acids shown in red are the same in *D. rerio* and in *Drosophila*. Amino acids shown in gray do not correspond between the two species. In the 3D window, the color coding is the same. The default view is pretty boring. It shows only the backbone chain of the protein. In the Show/Hide menu, select "Show Everything". Now you can see other parts of the protein outside the *goosecoid* homeodomain (in dark gray) as well as the two strands of DNA that are being bound by the homeodomain (the green and **gray** helices). From the Style menu, select "Rendering Shortcuts", then "Toggle Sidechains". Although the display is much "busier" now you can see the basepairs in the middle of the DNA helix, as well as the sidechains of the *goosecoid* protein. Look carefully at the amino acid sidechains that project toward the DNA helix. Are the side chains that come the closest to the helix mostly coming from red amino acids (i.e. ones that are conserved in both zebrafish and fruitflies) or blue amino acids (ones that have changed over evolutionary time since the divergence of fish and insects)?

Why is this called a "binding domain"? To find out, go to the Style menu, select "Rendering Shortcuts", then "spacefill". Then go to the Style menu, select "Coloring Shortcuts", then "Molecule". Notice how the brown *goosecoid* homeodomain fits snugly in the major groove of the DNA helix, while a small protrusion sticks in the minor groove. These parts of the *goosecoid* transcription factor apparently interact with the bases in the specific promoter sequence of the gene that *goosecoid* "turns on".

6. Close the 3D window. Return to the BLink page and click on the "CDD - search" button to identify domains based on a comparison to the Conserved Domain Database (CDD). On the results page, click on the PSSMLD link for the hit " pfam00046, homeobox, Homeobox domain". From the Conserved Domain Database window you can explore many aspects of homeodomains, including references available online through PubMed. Click on "+ structure" then the "structure view" button. Three windows will appear: the 3D viewer window, a sequence/alignment window, and a CDD Descriptive Items window. Scroll down the CDD Descriptive Items window. Notice that the query sequence is from *Mus musculus* (mouse). Dismiss the CDD Descriptive Items window. Examine the Sequence/Alignment Viewer. By default, the parts of the sequence that align well are centered in the viewer and shown in capital letters. If you mouse over the accession numbers on the left, you should recognize other familiar model organisms: *Dictyostelium discoideum* (a slime mold), *Homo sapiens* (human), and even *Arabidopsis thaliana*, a plant! Clearly DNA binding domains have a long evolutionary history. The amino acids that are the most conserved are shown in red. Those that are the least conserved are blue. In the 3D window, select Show Everything from the Show/Hide menu. Notice that the highly conserved red area is in the part of the domain that sits in the major groove. Presumably those amino acids have a crucial role in binding to the DNA helix.