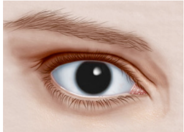
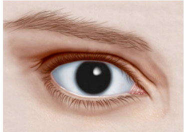




Hands-On Nine
The PAX6 Gene and Protein

Main Purpose of Hands-On Activity:

Using bioinformatics tools to examine the sequences, homology, and disease relevance of the Pax6: a master gene of eye formation.

Summary:

In this hands-on exercise we will explore some current bioinformatics tools. We will use different databases and other Internet resources to learn about the PAX6 gene in different organisms. We will learn that mutations in this gene affect the normal development of the eye in humans.
We will use a multiple sequence alignment tool to find conserved regions in a number of proteins and attempt to predict the phylogenetic relationship of PAX6 proteins. Finally, we will learn about protein families and sub-families and how to find them on the Internet.

Objective

The objective of this hands-on exercise is to introduce students to:

- Some of the biological databases on the Internet
- Basic bioinformatics tools for
  - Aligning multiple nucleotide or protein sequences
  - Constructing phylogenetic trees from sequences
- Conserved regions in proteins

A) Using BLAST at NCBI for finding orthologs

We are going to use BLAST at NCBI to find a human ortholog of the zebrafish Pax6 protein.

- Go to the main page of NCBI: http://www.ncbi.nlm.nih.gov
- Click on "BLAST" under "Popular Resources" on the right hand side of the page
- In the new page, click on "protein blast" under "Basic BLAST"

We are now ready to enter the pax6 protein of the zebrafish.

- Retrieve the pax6 protein of the zebrafish: zebrafish_pax6_protein_fasta.txt.
- Paste the sequence in the BLAST window under "Enter Query Sequence"

The Pax6 sequence can now be compared to various datasets of protein sequences in various databases.

- Scroll down and choose "UniProtKB/Swiss-Prot(swissprot)" from the dropdown window next to "Database"
- Start typing "Homo sapiens" in the "Organism" window (to limit the search to human proteins) and choose "Homo sapiens (taxid:9606)"
- Make sure that "blastp (protein-protein BLAST)" is chosen under "Program Selection"
- Choose "Show results in a new window"
- Click on the blue "BLAST" button to start the search

The graphic representation under "Show Conserved Domains" in the new page shows that two conserved domains have been detected in the Pax6 protein: PAX domain (**pa**ired bo**x** domain) and homeodomain (or **homeo**box **domain**).

The graphical view (under “Color key for alignment scores”) shows an overview of the results where the human sequences detected in SwissProt by the BLAST search (the “hits”) are aligned with the zebrafish Pax6 protein (represented as a red scale bar next to “Query”). The “Color key for alignement scores” shows the degree of similarity between the Query sequence and the results. Below the graphical overview, the detailed list of the sequences producing significant Alignments is given.

In the "Descriptions" section (under “Sequences producing significant alignments:”), you can examine the database matches in more details. Each database sequence has an identifier string, an accession number shown as a blue link. For example: **P26367.2** is an *accession number*:

- Click on **P26367.2** (under “Accession) and you will be taken to the database entry whose accession number is P26367.2. Note that it is the human Pax6.

1) When was the last time this record was updated?

- Go back to the BLAST output page.

Next to the the accession number you will find a one-line, short description of the protein, and the Total score that shows the level of similarity to the QUERY sequence and the E value assigned to each “hit”. The Total score and E value are special statistics that measure the degree of similarity between two sequences. Basically, the higher the Total score, the greater the similarity between the two sequences. The lower the E value, or the closer it is to zero, the more "significant" the match is.

Below the list of hits, the individual “Alignments” for each hit are shown. For each alignment, the query sequence (“Query” – the zebrafish protein) is shown at the top and the hit (“Sbjct” – the human protein returned by BLAST) underneath it, with the position of the amino acids indicated on the right and left of the alignment.

Go to the first alignment and answer the following questions:

2) a) Which protein in the human dataset is the closest to the zebrafish Pax6?
b) How long is this protein?

3) What is the degree of similarity between the query and the hit?

4) What is the probability that the similarity between the query and the hit occurs only by chance?

5) In the first alignment, what do you think the stretches “---“ represent?

6) Look at the second and third most relevant hits. How similar are they to the zebrafish Pax6 protein sequence?

The human sequence most similar to our QUERY is the protein Pax6. It has the highest Total score and the lowest E value in the list of hits. It is the human ORTHOLOG of the zebrafish Pax6 protein. Its accession number in the SwissProt database is P26367.2. Let us study the information available about it in several relevant biological databases.

B) The SwissProt Database

By going to http://www.uniprot.org/, you would have accessed the "Swiss-Prot Protein knowledgebase" database hosted by the "Swiss Institute of Bioinformatics". Note that the Protein Knowledgebase database is one of several UniProt (Universal Protein Resource) databases.

7) What is the mission of UniProt?

- Type the accession number "P26367" in the "Query" field at the top of and click on "**Search**".

The result of your search is a page for the human Pax6 protein.
The page contains information grouped in categories [Name and origin], [Protein attributes], [General annotation (Comments)], [Ontologies], [Binary interactions], [Alternative products], [Sequence annotation (Features)], [Sequences], [References], [Web resources], [Cross-references], [Entry information], and [Relevant documents] easily identified by light blue headers.
Scroll up and down the page to study the different categories of information available and answer the following questions.

8) In which tissues is the protein found?
9) How many diseases are described in relation with defects in the Pax6 protein? Which organs are affected by mutations in the PAX6 gene?
10) What is the function of Pax6?
11) Scroll down to the "References" section of the SwissProt Report. How many bibliographic references are quoted in this entry? Which paper describes the evolutionary conservation of PAX6 gene?

C) Studying the architecture of proteins with SMART

Let us use SMART to study the 3-Dimensional structure of Pax6.

- Go to SMART at http://smart.embl-heidelberg.de/

SMART (**S**imple **M**odular **A**rchitecture **R**esearch **T**ool) is based on the principle that proteins are modular in nature, i.e. they contain functional modules (or domains) that are detectable because they are conserved between species. SMART allows the identification of protein domains and the analysis of domain architectures. More than 500 domain families found in signaling, extra-cellular and chromatin-associated proteins are detectable. These domains are extensively annotated with respect to phylogenic distributions, functional class, TERTIARY STRUCTURES and functionally important residues.

- Click on the blue "SMART M5ODE:" under "Normal mode"
- Type "P26367" in the "Sequence ID or ACC" window
- Click on "Sequence SMART"

The resulting page, "Domains within *Homo sapiens* protein **PAX6_HUMAN** (P26367)", shows the two conserved domains detected by SMART: the PAX domain and the HOX domain. The figure also shows low complexity regions (LCRs) as pink bars.

Study the SMART result page and answer the following questions:
12) Name the two conserved domains found in PAX6 and write down their start and end positions.
13) Is the function of the paired box domain known?
14) Are paired box genes found in plants? In fungi?
15) What is the function of the HOX domain?

Note that the page also includes information on proteins that interact with PAX6 under "Network interaction"

- Click on the graph under "Network interaction" and you will be taken to string-db.org/version_9_0
- Click on "Home" to be taken to the main page of STRING.

We learn that "**S**earch **T**ool for the **R**etrieval of **In**teracting **G**enes/Proteins (STRING) is a database of known and predicted protein interactions. The interactions include direct (physical) and indirect (functional) associations."

- Type "PAX6" in the search window and select "Danio rerio" from "organism" and click on "GO!"
- Click on "Continue" from the new page to get the network of protein interaction with the *Danio.rerio* "pax6a".

16) Which protein has the strongest interaction with PAX6 of the zebra fish?

D) Comparing pax6 proteins of different species

It is often useful to compare multiple sequences, for example PAX6 proteins from different sequences. In general, similar protein sequences yield similar protein structures and functions. To find similarity between multiple proteins, bioinformaticians build multiple sequence alignments.

Consider the file "PAX6_seqs.txt".

It contains 5 pax6 protein sequences in FASTA format:

- **P26630** PAX_BRARE: Danio rerio (Zebrafish)
- **Q26046** PAX6_PARLI: Paracentrotus lividus (Sea urchin)
- **P26367** PAX6_HUMAN: Homo sapiens (Human)
- **P63015** PAX6_MOUSE: Mus musculus (Mouse)
- **O73917** PAX6_ORYLA: Oryzias latipes (Medaka fish) (Japanese ricefish)

To compare multiple sequences, we will use CLUSTAL Omega at the EMBL-EBI website in the United Kingdom.

- Go to http://www.ebi.ac.uk/Tools/msa/clustalo/ and paste the five sequences (in FASTA format) from "PAX6_seqs.txt" in the window.
- Use the default conditions provided by the program
- Click on "Submit" under STEP 3, at the bottom of the page to align the five sequences
- In the "Clustal Omega" result page, click on the "Show Colors" button.

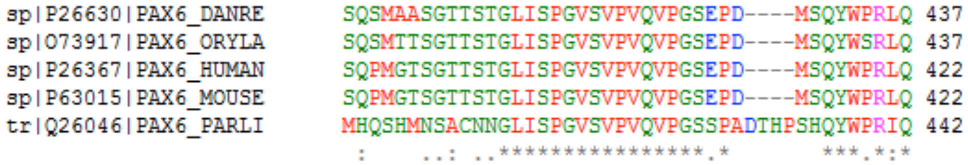
Examine the alignment and answer the following questions.

17) What are the significances of the symbols "." ":" and "*" ?
Hint: Consult: "http://www.ebi.ac.uk/Tools/msa/clustalw2/help/faq.html#23".

18) Which sequence appears to be the closest to the human sequence?

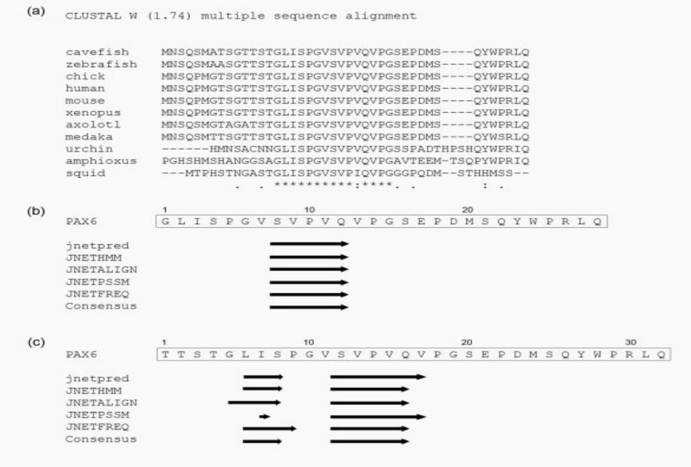
19) Can you identify conserved regions in the five proteins?

Optional: Note for example that one of the conserved regions is “GLISP….GS”



```
sp|P26630|PAX6_DANRE      SQSMAASGTTSTGLISPGVSVPVQVPGSEPD----MSQYWPRLQ 437
sp|O73917|PAX6_ORYLA      SQSMTTSGTTSTGLISPGVSVPVQVPGSEPD----MSQYWSRLQ 437
sp|P26367|PAX6_HUMAN      SQPMGTSGTTSTGLISPGVSVPVQVPGSEPD----MSQYWPRLQ 422
sp|P63015|PAX6_MOUSE      SQPMGTSGTTSTGLISPGVSVPVQVPGSEPD----MSQYWPRLQ 422
tr|Q26046|PAX6_PARLI      MHQSHMNSACNNGLISPGVSVPVQVPGSSPADTHPSHQYWPRIQ 442
                           :    ..: ..****************.*       ***.*:*
```

You can read more about this conserved region in “A screen for proteins that interact with PAX6: C-terminal mutations disrupt interaction with HOMER3, DNCL1 and TRIM11” by Cooper and Hanson, BMC Genetics, 2005. The following figure is taken from their article.



(a)

```
CLUSTAL W (1.74) multiple sequence alignment


cavefish      MNSQSMATSGTTSTGLISPGVSVPVQVPGSEPDMS----QYWPRLQ
zebrafish     MNSQSMAASGTTSTGLISPGVSVPVQVPGSEPDMS----QYWPRLQ
chick         MNSQPMGTSGTTSTGLISPGVSVPVQVPGSEPDMS----QYWPRLQ
human         MNSQPMGTSGTTSTGLISPGVSVPVQVPGSEPDMS----QYWPRLQ
mouse         MNSQPMGTSGTTSTGLISPGVSVPVQVPGSEPDMS----QYWPRLQ
xenopus       MNSQPMGTSGTTSTGLISPGVSVPVQVPGSEPDMS----QYWPRLQ
axolotl       MNSQSMGTAGATSTGLISPGVSVPVQVPGSEPDMS----QYWPRLQ
medaka        MNSQSMTTSGTTSTGLISPGVSVPVQVPGSEPDMS----QYWSRLQ
urchin        ------HMNSACNNGLISPGVSVPVQVPGSSPADTHPSHQYWPRIQ
amphioxus     PGHSHMSHANGGSAGLISPGVSVPVQVPGAVTEEM-TSQPYWPRIQ
squid         ---MTPHSTNGASTGLISPGVSVPIQVPGGGPQDM--STHHMSS--
                       .  . **********:****. .        : .
```

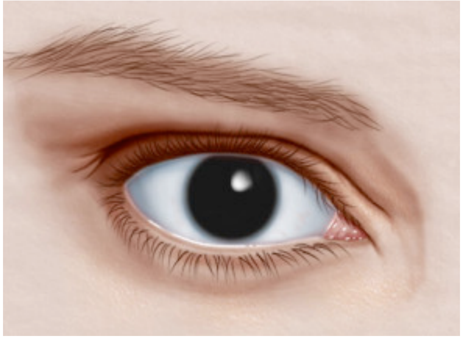
[**End of Optional**]

Note that although “PAX6_seqs.txt” contains Pax6 proteins from vertebrates and inverterbrates which have evolved separately, the proteins have long stretches of similar amino acids.

- Scroll to the top of the “Clustal Omega" result page and click on "Phylogenetic Tree".

Examine the “Phylogram” tree and answer the following questions.
20) Which organism is the closest to humans?
21) Compare your answer to the one you gave based on the alignment of the sequences. Do they agree?
22) Do you think that these sequences are orthologous?

Additional Remarks:



The term *aniridia* means, literally, “without iris.” Some unfortunate people are born missing part or all of the iris, the colored part of the eye. This uncommon condition, also known as *iris hypoplasia*, occurs in one out of every 50,000 to 100,000 infants born worldwide (although incidence varies from one region to another). Aniridia presents a striking, distinctive appearance. The eyes will appear black, with no color separating the white from the pupil. Because we depend on the iris to regulate the amount of light that enters the eye, the absence of the iris causes extreme sensitivity to light, (aka photophobia). www.eyehealthweb.com/aniridia/.

If you want to know more about PAX6, go to http://ghr.nlm.nih.gov/gene/PAX6