Discovering Bioinformatics

Sami Khuri
Natascha Khuri
Alexander Picker
Aidan Budd
Sophie Chabanis-Davidson
Julia Willingale-Theune

English version

ELLS – European Learning Laboratory for the Life Sciences

EMBL

Sami Khuri, Natascha Khuri, Alexander Picker, Aidan Budd,
Sophie Chabanis-Davidson and Julia Willingale-Theune

Discovering Bioinformatics

Following a protein in the World Wide Web

Table of contents

1 Objective 4

2 **Activities** 5

A.1 Compare a protein with a collection of sequences in a database by performing a BLAST search 5

A.2 The SwissProt database: (almost) all you need to know about your favourite protein 11

A.3 Studying the architecture of proteins with the SMART resource 13

A.4 Visualizing 3D-structures using the PDBsum resource 15

A.5 The function of the Pax6 protein and its relationship to human diseases: the OMIM database 17

A.6 Exploring the scientific literature in PubMed 19

3 Glossary 22

4 References 25

Appendix I 26

1 Objective

In this activity we are going to search for information about a protein using databases of biological information on the World Wide Web. These databases collect and store information about genes and proteins (sequence, STRUCTURE, expression) about human inherited diseases for which the genetic cause is known, scientific literature, etc. Many databases that are accessible via the World Wide Web offer so called QUERY interfaces: special web pages on which you can enter and combine search terms and restrict them to special sections or fields of the database. In a text search you can enter a search term, (the name of a protein, a disease, a cell type) which is subsequently compared to the textual content of the database. You can also compare the sequence of a protein or gene to the collection of known, annotated sequences stored in a protein or gene database. In other words, you can search these databases to find out what is already known about your favourite protein. As we will see the main biological databases are interconnected (through so-called cross-references), providing links with one another and allowing the user to access different types of information from the result of a single QUERY.

We are now going to look at the **Pax6** protein from zebrafish which is involved in eye development. By "following" this protein on the World Wide Web we can find the human protein corresponding to the zebrafish **Pax6** (its ORTHOLOG), information about its function, STRUCTURE, sub-cellular localization, and the molecular basis of diseases linked to mutations in its sequence.

The standard conventions to denote genes and their products (proteins) are as follows:

- ***PAX6*** = human gene
- ***pax6*** = gene of any other species
- **Pax6** = protein

In the following text, actions are indicated in **bold**, and glossary terms are indicated in SMALL CAPS.

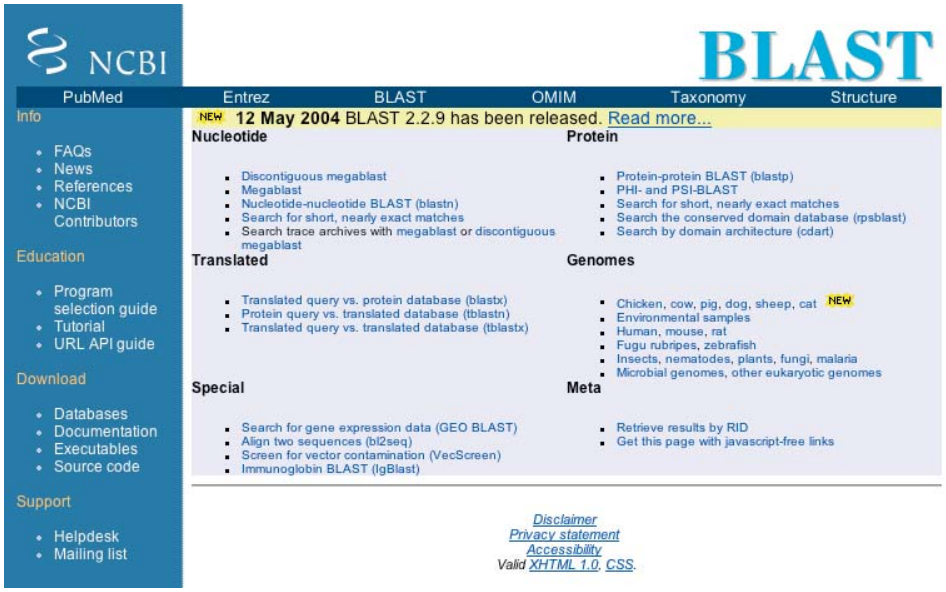
Activities

A.1 Compare a protein with a collection of sequences in a database by performing a BLAST search

BLAST ("Basic local alignment search tool") is an interactive program maintained by the National Centre for Biotechnology Information (NCBI) that allows a rapid comparison of a nucleotide or protein sequence against a database of sequences using ALIGNMENTS.

Start a web browser and open a new window by clicking on the following link: http://www.ncbi.nlm.nih.gov:80/BLAST/ (alternatively you can copy and paste it in the URL address bar).

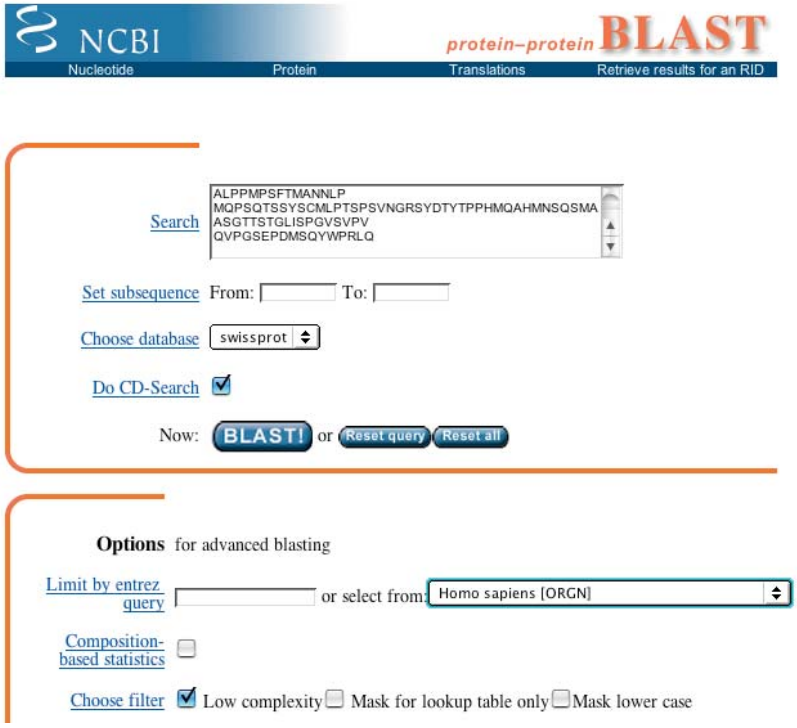
For our purpose, we will perform a protein-protein BLAST: under *Protein*, click on **Protein-protein BLAST (blastp)**. The following window will appear: it shows the submission form that you will use to search a protein database called SwissProt (see references) using the **Pax6** protein sequence.



The sequence of letters below represents the amino acid (AA) sequence of the zebrafish **Pax6** protein. Each letter corresponds to a single AA (e.g. M = methionine).

We are going to use this sequence to perform a BLAST search. From now on it will be referred to as the QUERY sequence. Copy it and paste it in the **search** field of the BLAST window.

MPQKEYYNRATWESGVASMMQNSHSGVNQLGGVFVNGRPLPDSTRQKIV
ELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEV
VGKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSINRVLRNLASEKQQ
MGADGMYEKLRMLNGQTGTWGTRPGWYPGTSVPGQPNQDGCQQSDGG
GENTNSISSNGEDSDETQMRLQLKRKLQRNRTSFTQEQIEALEKEFERTHYP
DVFARERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQASNSSSHIPI
SSSFSTSVYQPIPQPTTPVSFTSGSMLGRSDTALTNTYSALPPMPSFTMANN
LPMQPSQTSSYSCMLPTSPSVNGRSYDTYTPPHMQAHMNSQSMAASGTT
STGLISPGVSVPVQVPGSEPDMSQYWPRLQ



The **Pax6** sequence can now be compared to various datasets of protein sequences in various databases. Select **swissprot** from the drop-down list. For more details on the databases available for BLAST search, click on **Choose database**.

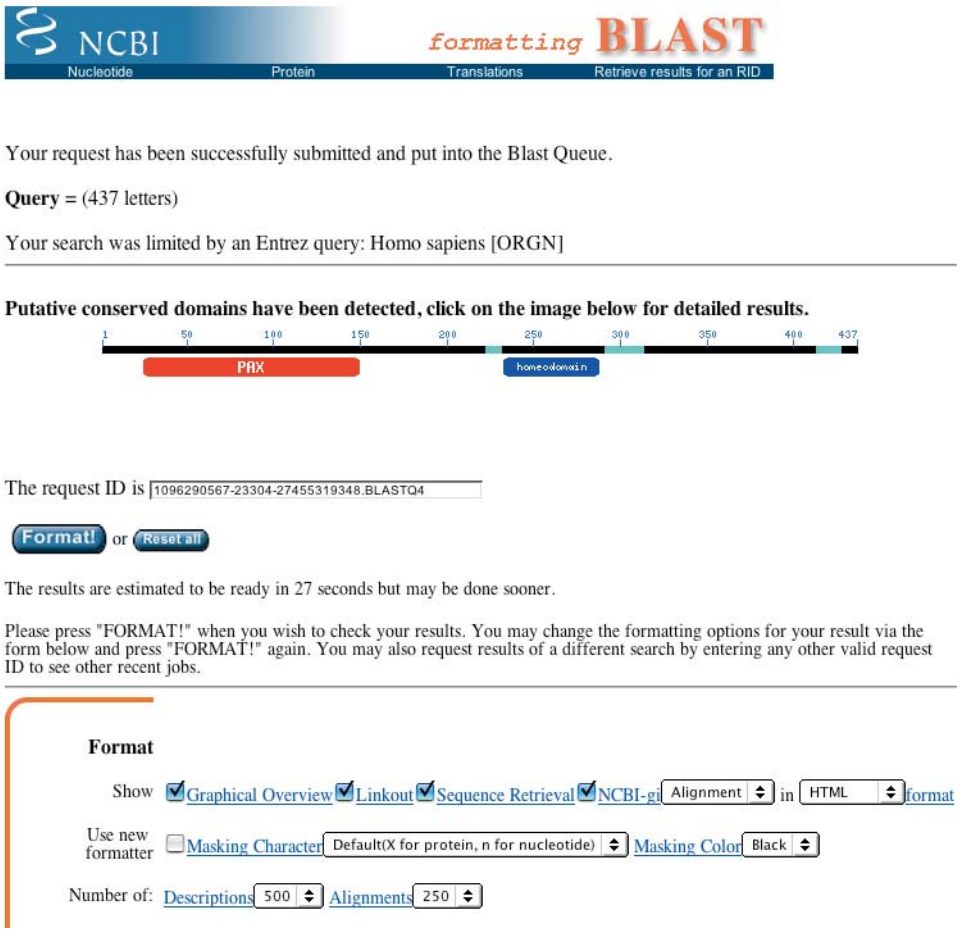
With the **Options for advanced blasting** we have the possibility to limit the BLAST search to a specific parameter or combination of parameters (also called terms). This will limit the search to a subset of the chosen database: only the entries containing the terms entered will be searched. Change from **all organisms** to **Homo sapiens [ORGN]** in the drop-down list to limit the search to human proteins. Click on **Limit Entrez Query** for details.

The other options define the parameters for the search, such as the matrix, the FILTERING parameters. They are set on default parameters—suitable for most basic searches—which we will use now. Likewise, we will use the default settings for the output options (under **Format**). If time allows, you can explore each parameter or option by clicking on the corresponding links.

Finally, click on **BLAST** at the bottom of the page to start the search.

When the new page appears: Click on **Format!** to continue.

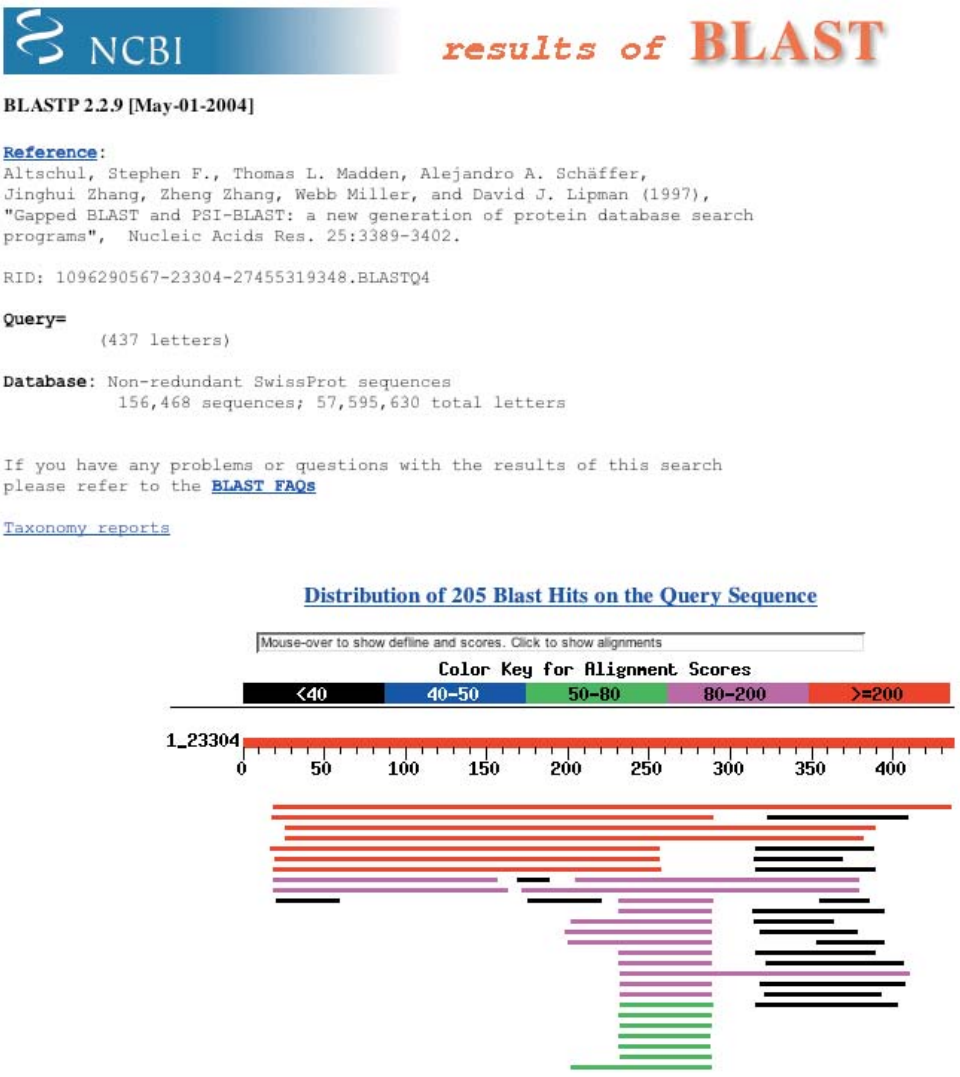
It will take a few minutes to complete the search. Then the results of the LAST page will appear.



In the window ”formatting BLAST” you will see some basic information about the sequence you have just submitted:

- The protein is 437 AA long.
- The graphic representation shows that two conserved domains (see *Glossary*) have been detected in the **Pax6** protein: PAX domain (paired box domain) and homeodomain (or homeobox domain). Three low complexity regions (LCRs) were detected as well. An LCR is a region of biased composition including homopolymeric runs, short-period repeats and subtle overrepresentation of one or a few residues.

Go to the **results of BLAST** page, which is divided into three sections:



First, the graphical view 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). The Color key for

ALIGNMENT SCORES shows the degree of similarity between the QUERY sequence and the results.

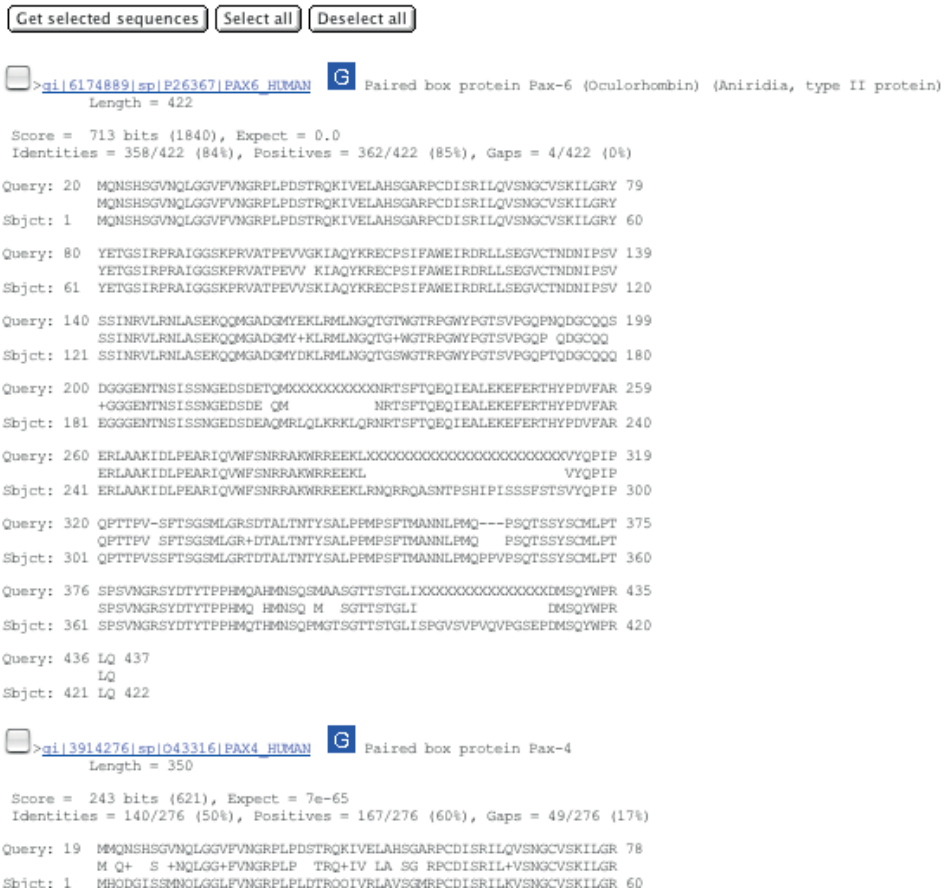
Below the graphical overview, the detailed list of the sequences producing significant ALIGNMENTS is given.

Sequences producing significant alignments:		Score (bits)	E Value	
gi|6174889|sp|P26367|PAX6_HUMAN	Paired box protein Pax-6 (O...	713	0.0	G
gi|3914276|sp|O43316|PAX4_HUMAN	Paired box protein Pax-4	243	7e-65	G
gi|1172022|sp|P23760|PAX3_HUMAN	Paired box protein Pax-3 (H...	240	8e-64	G
gi|8247951|sp|P23759|PAX7_HUMAN	Paired box protein Pax-7 (H...	235	2e-62	G
gi|2506538|sp|Q02962|PAX2_HUMAN	Paired box protein Pax-2	227	7e-60	G
gi|417449|sp|Q02548|PAX5_HUMAN	Paired box protein Pax-5 (B-...	220	8e-58	G
gi|548459|sp|Q06710|PAX8_HUMAN	Paired box protein Pax-8	219	1e-57	G
gi|8247950|sp|P15863|PAX1_HUMAN	Paired box protein Pax-1 (H...	182	2e-46	G

In blue you will find an identifier containing the accession numbers of each sequence found in the SwissProt database. Click on some identifiers to explore the details of the listed proteins: you will have access to detailed entries of these proteins in the database.

Next to the accession number you will find a short description of the protein, and the BIT SCORE that shows the level of similarity to the QUERY sequence and the E VALUE assigned to each “hit”. The BIT SCORE and E VALUES are calculated from the ALIGNMENT. Basically, the higher the BIT SCORE the greater the similarity between the two sequences. The lower the E VALUE, or the closer it is to „0“ the more „significant“ the match is. (For more details see *Glossary*).

Below the list of hits, the individual ALIGNMENTS for each hit are shown. For each ALIGNMENT, the QUERY sequence (“Query”) is shown at the top and the hit (“Sbjct”)underneath it, with the position of the AAs indicated on the right and left.



Get selected sequences | Select all | Deselect all

>gi|6174889|sp|P26367|PAX6_HUMAN Paired box protein Pax-6 (Oculorhombin) (Aniridia, type II protein)
Length = 422

Score = 713 bits (1840), Expect = 0.0
Identities = 358/422 (84%), Positives = 362/422 (85%), Gaps = 4/422 (0%)

```
Query: 20  MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRY 79
           MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRY
Sbjct: 1   MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRY 60

Query: 80  YETGSIRPRAIGGSKPRVATPEVVGKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSV 139
           YETGSIRPRAIGGSKPRVATPEVV KIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSV
Sbjct: 61  YETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSV 120

Query: 140 SSINRVLRNLASEKQQMGADGMYEKLRMLNGQTGTWGTRPGWYPGTSVPGQPNQDGCQQS 199
           SSINRVLRNLASEKQQMGADGMY+KLRMLNGQTG+WGTRPGWYPGTSVPGQP QDGCQQ
Sbjct: 121 SSINRVLRNLASEKQQMGADGMYDKLRMLNGQTGSWGTRPGWYPGTSVPGQPTQDGCQQQ 180

Query: 200 DGGGENTNSISSNGEDSDETQMXXXXXXXXXXXNRTSFTQEQIEALEKEFERTHYPDVFAR 259
           +GGGENTNSISSNGEDSDE QM           NRTSFTQEQIEALEKEFERTHYPDVFAR
Sbjct: 181 EGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEQIEALEKEFERTHYPDVFAR 240

Query: 260 ERLAAKIDLPEARIQVWFSNRRAKWRREEKLXXXXXXXXXXXXXXXXXXXXXVYQPIP 319
           ERLAAKIDLPEARIQVWFSNRRAKWRREEKL                    VYQPIP
Sbjct: 241 ERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQASNTPSHIPISSSFSTSVYQPIP 300

Query: 320 QPTTPV-SFTSGSMLGRSDTALTNTYSALPPMPSFTMANNLPMQ---PSQTSSYSCMLPT 375
           QPTTPV SFTSGSMLGR+DTALTNTYSALPPMPSFTMANNLPMQ   PSQTSSYSCMLPT
Sbjct: 301 QPTTPVSSFTSGSMLGRTDTALTNTYSALPPMPSFTMANNLPMQPPVPSQTSSYSCMLPT 360

Query: 376 SPSVNGRSYDTYTPPHMQAHMNSQSMAASGTTSTGLIXXXXXXXXXXXXXXXXDMSQYWPR 435
           SPSVNGRSYDTYTPPHMQ HMNSQ M  SGTTSTGLI               DMSQYWPR
Sbjct: 361 SPSVNGRSYDTYTPPHMQTHMNSQPMGTSGTTSTGLISPGVSVPVQVPGSEPDMSQYWPR 420

Query: 436 LQ 437
           LQ
Sbjct: 421 LQ 422
```

>gi|3914276|sp|O43316|PAX4_HUMAN Paired box protein Pax-4
Length = 350

Score = 243 bits (621), Expect = 7e-65
Identities = 140/276 (50%), Positives = 167/276 (60%), Gaps = 49/276 (17%)

```
Query: 19  MMQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGR 78
           M Q+  S +NQLGG+FVNGRPLP  TRQ+IV LA SG RPCDISRIL+VSNGCVSKILGR
Sbjct: 1   MHQDGISSMNQLGGLFVNGRPLPLDTROOIVRLAVSGMRPCDISRILKVSNGCVSKILGR 60
```

Question Set A:

A.1- Which protein in the human dataset is the closest to the zebrafish **Pax6**? How long is this protein?

A.2- What is the degree of similarity between the query and the hit?

A.3- What is the probability that the similarity between the query and the hit occurs only by chance?

A.4- In the first alignment, what do you think the stretches XXX represent? And the stretch "---"?

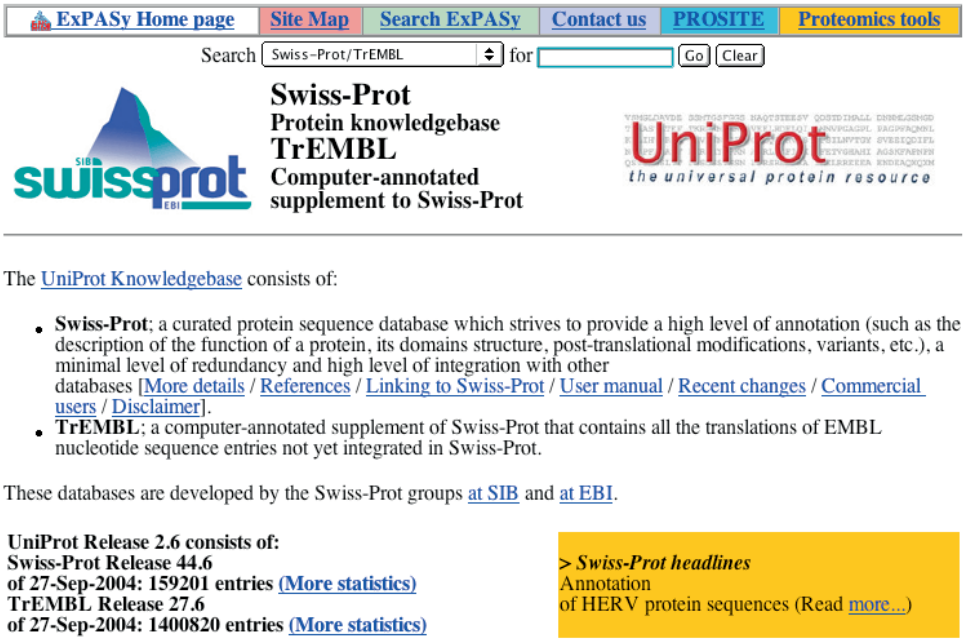
A.5- Look at the second and third most relevant hits. How similar are they to the zebrafish **Pax6** sequence?

The human sequence most similar to our QUERY is the protein **Pax6**. It has the highest BIT 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**. Now let's study the information available about it in several relevant biological databases.

A.2 The SwissProt database: (almost) all you need to know about your favourite protein

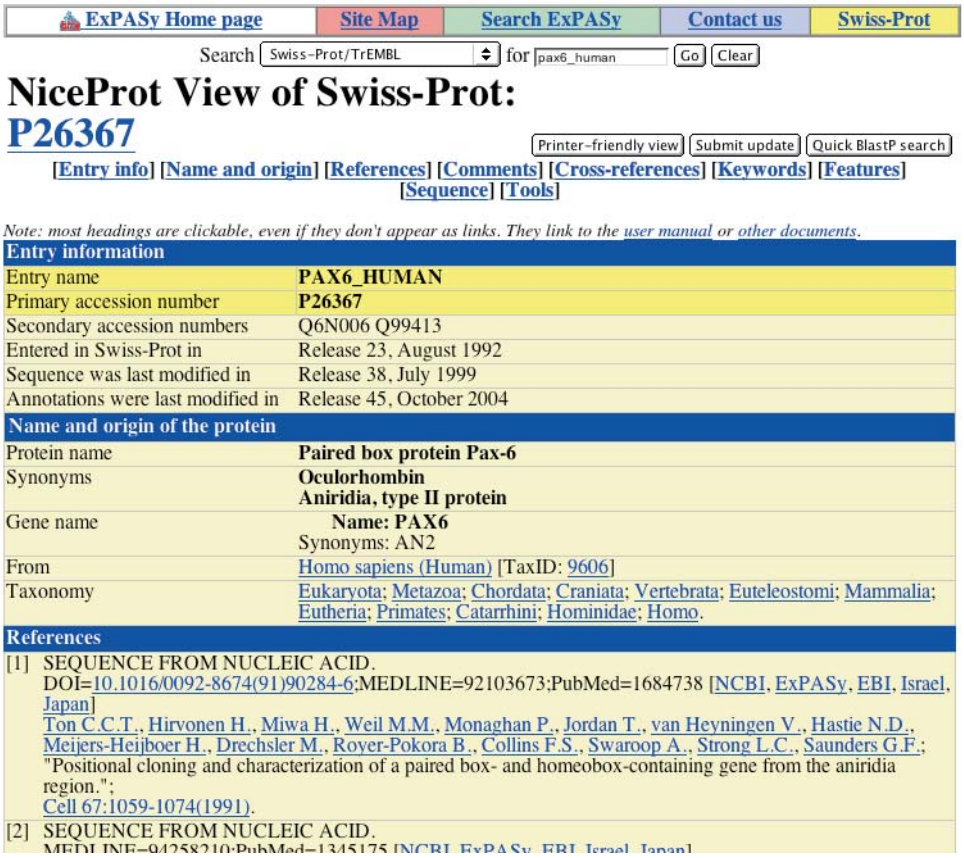
Open a new window at http://www.expasy.org/sprot/sprot-top.html by clicking on the link (or copy and paste the URL into the URL address bar). You have accessed the SwissProt-Protein (curated Protein Knowledgebase) and TrEMBL (computer annotated supplement to SwissProt) databases hosted by ExPaSy (Expert Protein Analysis System) proteomics server of the Swiss Institute of Bioinformatics (SIB).

These two databases are grouped in the Universal Protein resource Uniprot. More detailed information about the databases is available through the links on this page.



At the top of the page, a search field is available to submit a QUERY term: type either the accession number (**P26367**) or the identifier of the human **Pax6** protein (**Pax6_HUMAN**) in the search field, select **SwissProt-TrEMBL** as database and click **GO**.

The result of your search is the NiceProt View of Swiss-Prot for the human **Pax6** protein.



ExPASy Home page | Site Map | Search ExPASy | Contact us | Swiss-Prot

Search Swiss-Prot/TrEMBL for pax6_human Go Clear

NiceProt View of Swiss-Prot: P26367

Printer-friendly view | Submit update | Quick BlastP search

[Entry info] [Name and origin] [References] [Comments] [Cross-references] [Keywords] [Features] [Sequence] [Tools]

Note: most headings are clickable, even if they don't appear as links. They link to the user manual or other documents.

Entry information	
Entry name	**PAX6_HUMAN**
Primary accession number	**P26367**
Secondary accession numbers	Q6N006 Q99413
Entered in Swiss-Prot in	Release 23, August 1992
Sequence was last modified in	Release 38, July 1999
Annotations were last modified in	Release 45, October 2004
Name and origin of the protein	
Protein name	**Paired box protein Pax-6**
Synonyms	**Oculorhombin** **Aniridia, type II protein**
Gene name	**Name: PAX6** Synonyms: AN2
From	Homo sapiens (Human) [TaxID: 9606]
Taxonomy	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
References	
[1]	SEQUENCE FROM NUCLEIC ACID. DOI=10.1016/0092-8674(91)90284-6;MEDLINE=92103673;PubMed=1684738 [NCBI, ExPASy, EBI, Israel, Japan] Ton C.C.T., Hirvonen H., Miwa H., Weil M.M., Monaghan P., Jordan T., van Heyningen V., Hastie N.D., Meijers-Heijboer H., Drechsler M., Royer-Pokora B., Collins F.S., Swaroop A., Strong L.C., Saunders G.F.; "Positional cloning and characterization of a paired box- and homeobox-containing gene from the aniridia region."; Cell 67:1059-1074(1991).
[2]	SEQUENCE FROM NUCLEIC ACID. MEDLINE=94258210;PubMed=1345175 [NCBI, ExPASy, EBI, Israel, Japan]

This page contains information grouped in categories [Entry info], [Name and origin], [References], [Comments], [Cross-references], [Keywords], [Features], [Sequence], [Tools], easily identified by blue horizontal bars.

Scroll up and down the page to study the different categories of information available and answer the Question Set B.

Question Set B:

B.1- In which tissues is the protein found and at what stage of fetal development? Look under *Comments*.

B.2- How many diseases are described in relation with defects in the **Pax6** protein? Which organs are affected by mutations in the ***PAX6*** gene?

B.3- Why does “3D-structure” appear under keywords in this entry?

B.4- What is the molecular function of Pax6 and its cellular localization? Look under *Comments*.

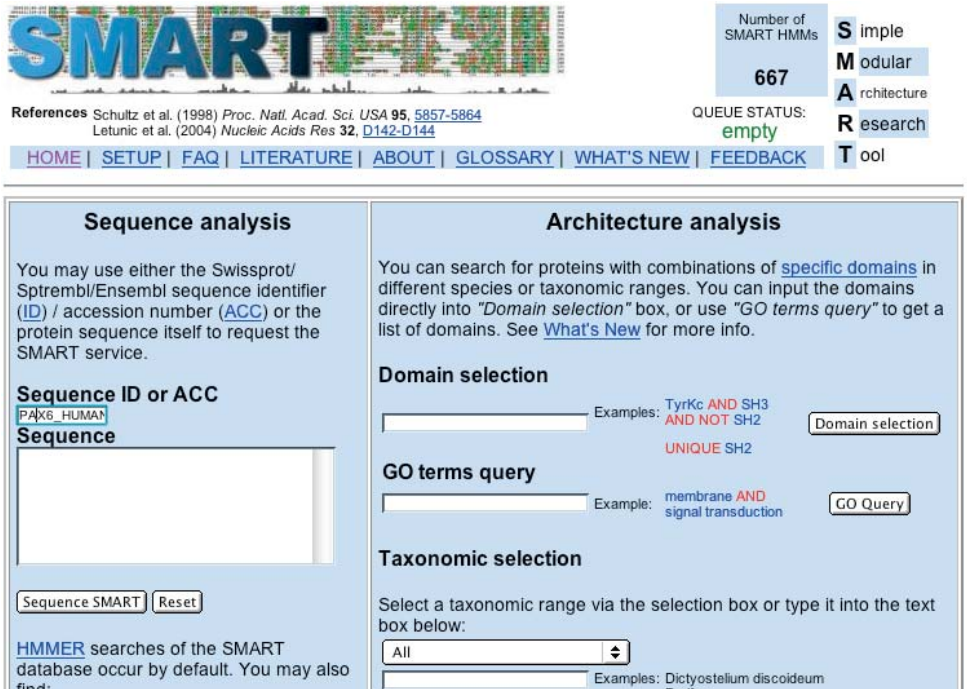
B.5- How many bibliographic references are quoted in this entry? What are the main topics published in these papers? Which paper describes the evolutionary conservation of ***PAX6*** gene? Look under *References*.

The information centralized in the Cross-references section of this SwissProt entry provides links to other databases that contain additional information about **Pax6**. They are directly accessible by clicking on the accession numbers or identifiers of the Cross-references. For example, the gene sequence is available from EMBL Genbank, the coordinates of the 3D STRUCTURE from PDB, the domain composition from SMART, Prosite, InterPro and Pfam.

At the bottom of the page the sequence of the **Pax6** protein is shown under "Sequence information". From the SwissProt entry, you could now perform a new BLAST search or other sequence-based searches using this sequence (see *Tools*).

A.3 Studying the architecture of proteins with the SMART resource

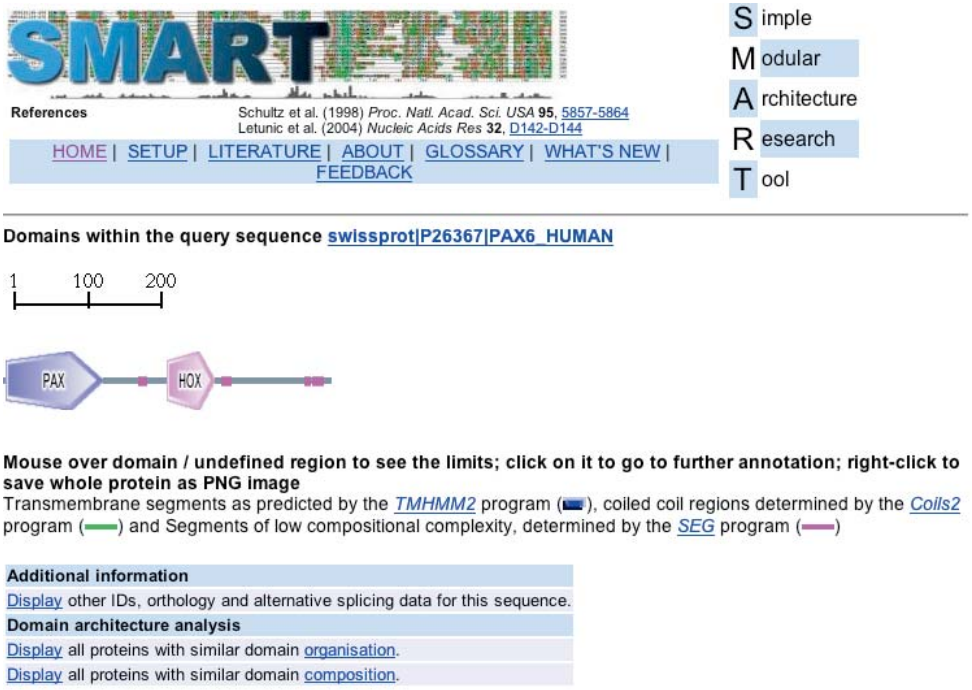
Now open the SMART home page at http://smart.embl-heidelberg.de/.



SMART (Simple Modular Architecture Research Tool) 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 signalling, 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. As you can see on the QUERY interface, SMART can be searched using a sequence ("sequence analysis" QUERY), or used to retrieve all the proteins containing a type of domain or a combination of domain the "architecture analysis" QUERY.

We will use the sequence QUERY to study the architecture of the human Pax6 protein: type the SwissProt accession number of the human Pax6_human protein (**P26367**) in the **Sequence ID or ACC** field and click on **Sequence SMART**.



In the schematic representation of the **Pax6** protein shown on the result page, the conserved domains detected by SMART (PAX domain and HOX domain) are depicted as boxes and the LCRs as coloured bars. The details of the results can be found in a table below the graphic.

Study the SMART result page and answer the Question Set C:

Question Set C:

C.1- Is the function of the paired box domain known?

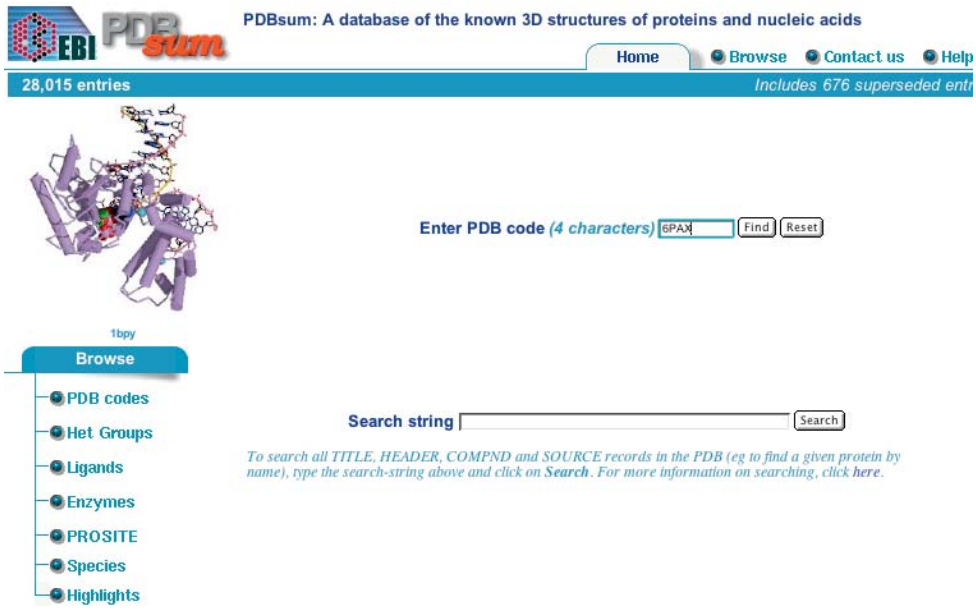
C.2- Are paired box genes found in plants? In fungi?

C.3- What is the function of the HOX domain?

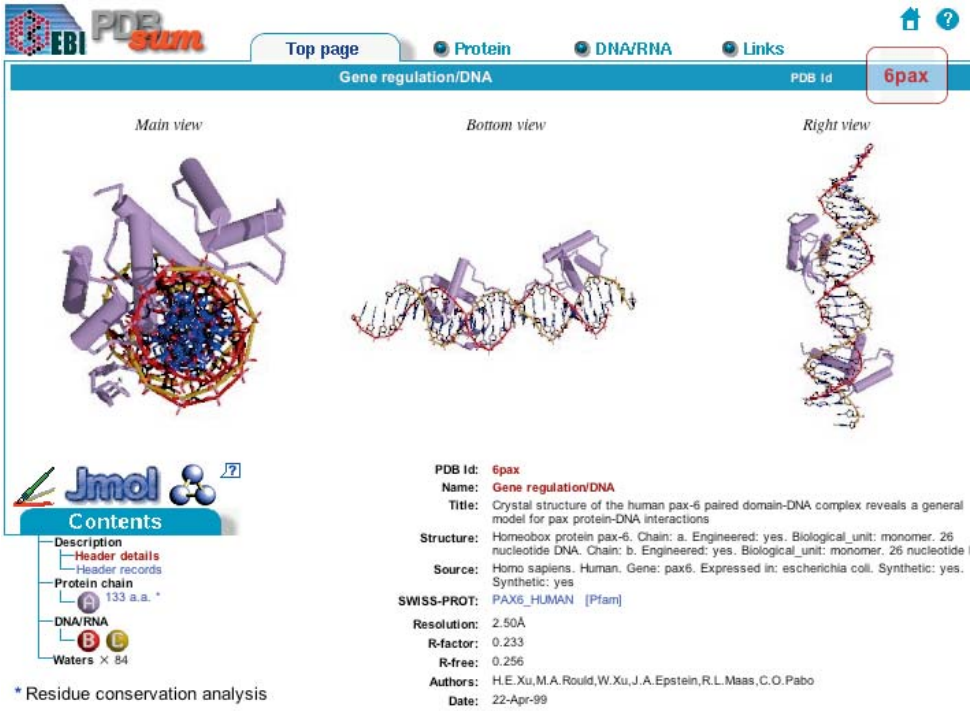
C.4- Are the structures of the domains resolved? How does the structure of the PAX domain help us to understand the structural basis of the mutations known to be linked to diseases?

A.4 Visualizing 3D-structures using the PDBsum resource

The PDBsum database (http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/) is a pictorial database that provides an at-a-glance overview of the contents of each 3D-STRUCTURE deposited in the Protein Data Bank (PDB). It shows the molecule(s) that make up the STRUCTURE (i.e. protein chains, DNA, ligands and metal ions) and schematic diagrams of their interactions. Entries are accessed either by their 4- character PDB code, by the simple text search provided on the PDBsum home page, or via any of the Browse options.



In the *Search field type* the PDB code for the X-ray STRUCTURE of **Pax6** (**6PAX**) and click **FIND**.



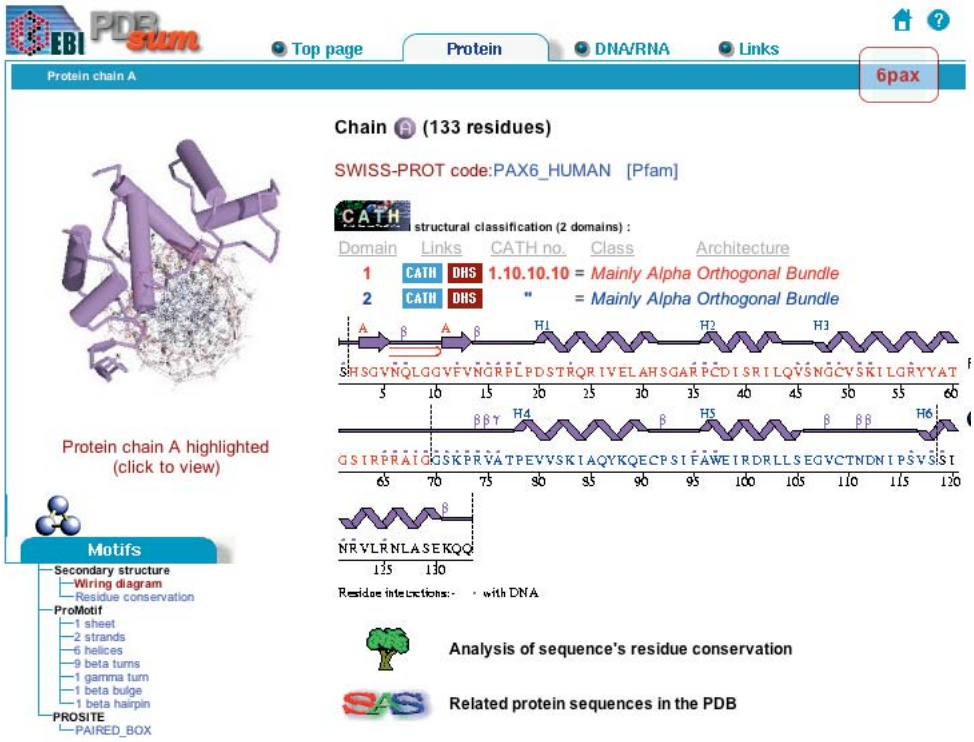
The result page shows the crystal STRUCTURE of the complex between the paired domain of **Pax6** and DNA. The top figures show static representations of the TERTIARY STRUCTURE of the domain in a complex with DNA—the protein is represented as a solid purple object.

Click on **Jmol** to get an interactive view of the protein-DNA complex (it rotates and has a zoom option—rotate with mouse and zoom with "alt").

The main components of SECONDARY STRUCTURES are ALPHA-HELICES, BETA STRANDS, RANDOM COILS (see *Glossary* for definitions).

Note that in Jmol, ALPHA HELICES are depicted as pink spirals (purple cylinders in the static view). In general, BETA STRANDS are depicted as arrows (yellow in Jmol), and RANDOM COILS as threads. DNA chains are depicted as the typical double-helix.

Click on **Protein chain A 133 a.a** under **Contents**. The following page will be displayed.

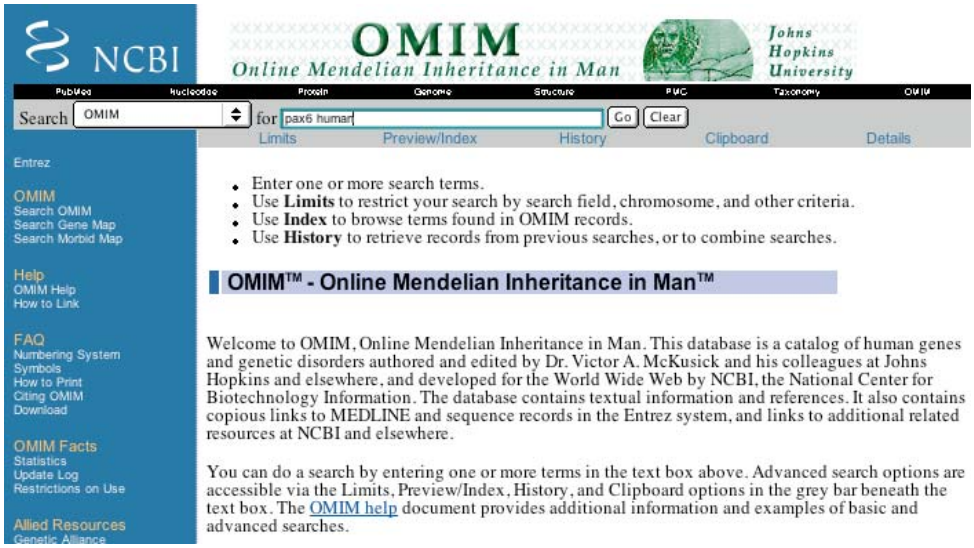


This other graphical output shows details of the SECONDARY STRUCTURE of the paired box domain such as ALPHA HELICES, BETA STRANDS and RANDOM COILS, aligned on the sequence itself. The elements of the SECONDARY STRUCTURE fold together to build the TERTIARY STRUCTURE of a protein.

A.5 The function of the Pax6 protein and its relationship to human diseases: the OMIM database.

The OMIM (Online Mendelian Inheritance in Man) database is a catalogue of human genes and genetic disorders. It contains textual information and references. It also contains links to literature and sequence records, and links to additional related resources at NCBI and elsewhere.

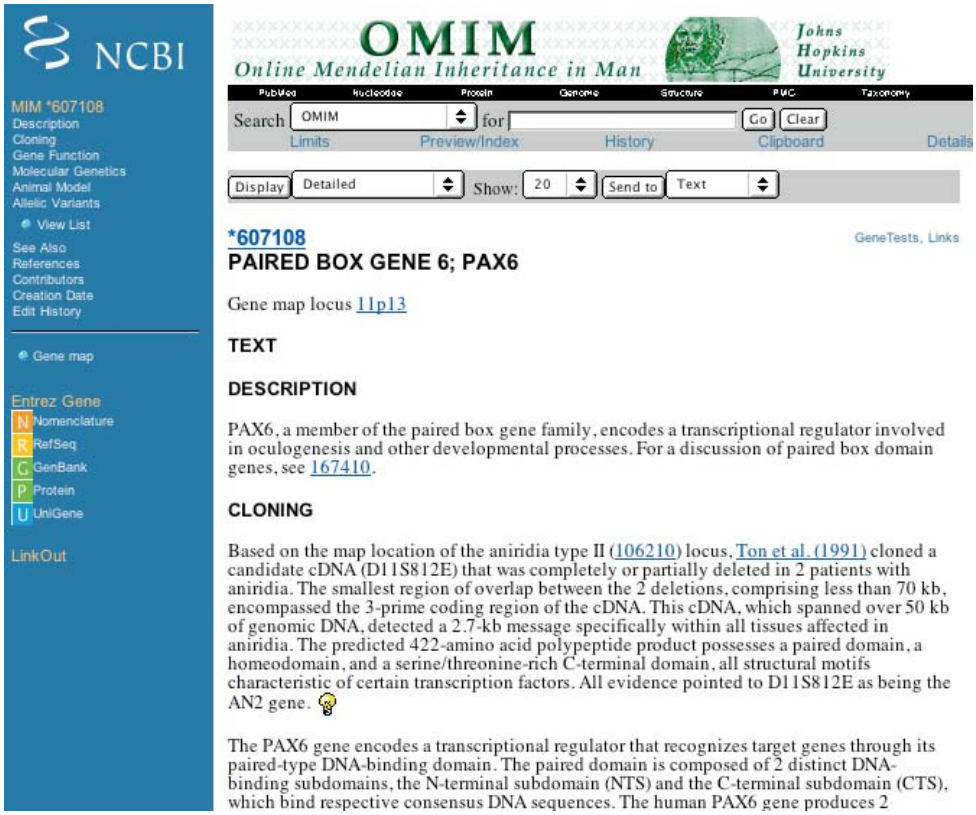
Open the NCBI home page at http://www.ncbi.nlm.nih.gov and click on **OMIM**.



OMIM can be searched by entering one or more terms in the text field at the top of the page. Advanced search options are accessible in the grey bar beneath the text box.

Type **Pax6 AND human** into the Search field to search for entries containing both terms. Click **GO**.

Click on the first entry:
1: *607108
PAIRED BOX GENE 6; PAX6
Gene map locus 11p13



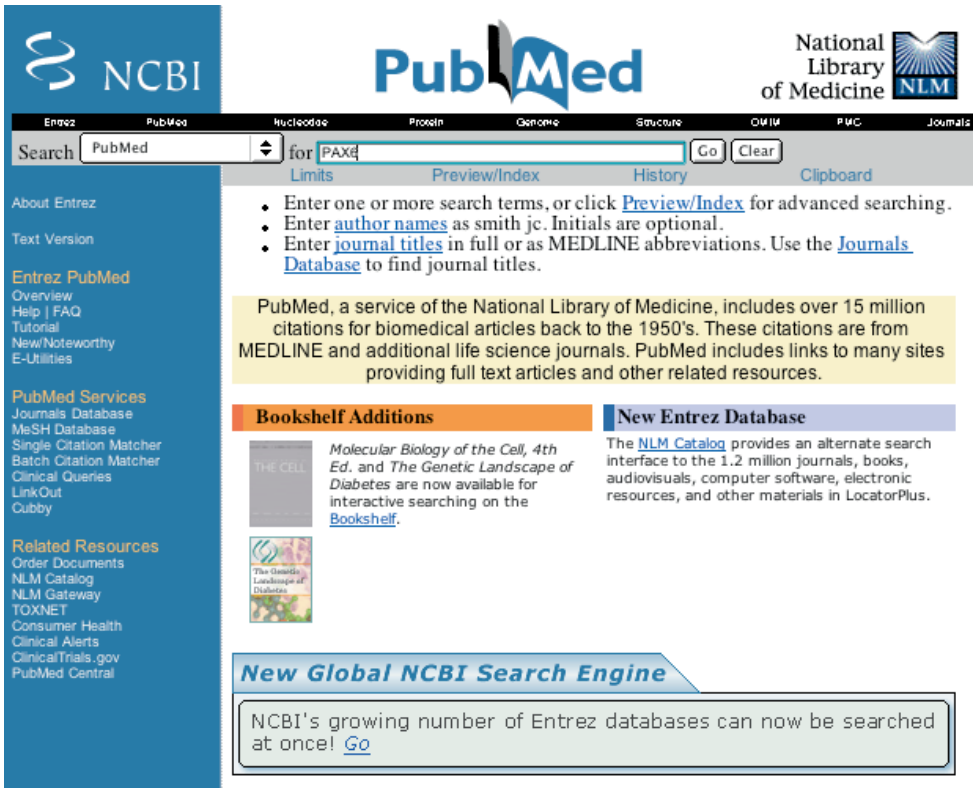
The following page will appear: it contains relevant information about the diseases associated with defects in the **Pax6** protein.

A.6 Exploring the scientific literature in PubMed

PubMed, a service of the National Library of Medicine, includes over 15 million citations for biomedical articles back to the 1950s. PubMed was designed to provide access to citations from biomedical literature. LinkOut provides access to full text articles at journal web sites and other related web resources. PubMed also provides access and links to the other Entrez molecular biology resources.

Open the NCBI home page at http://www.ncbi.nlm.nih.gov/.

Click on **PubMed**.



PubMed can be searched by entering one or more term(s) in the Search field; for example, the name of a protein, author or journal. The grey Features bar provides additional search options to limit your search to a specific type of publication and/or language, to the publication date, etc. The terms are searched in various fields of the citation. Your search may include Boolean operators (see *Glossary*).

In the Search field type in **Pax6** and click **GO**.

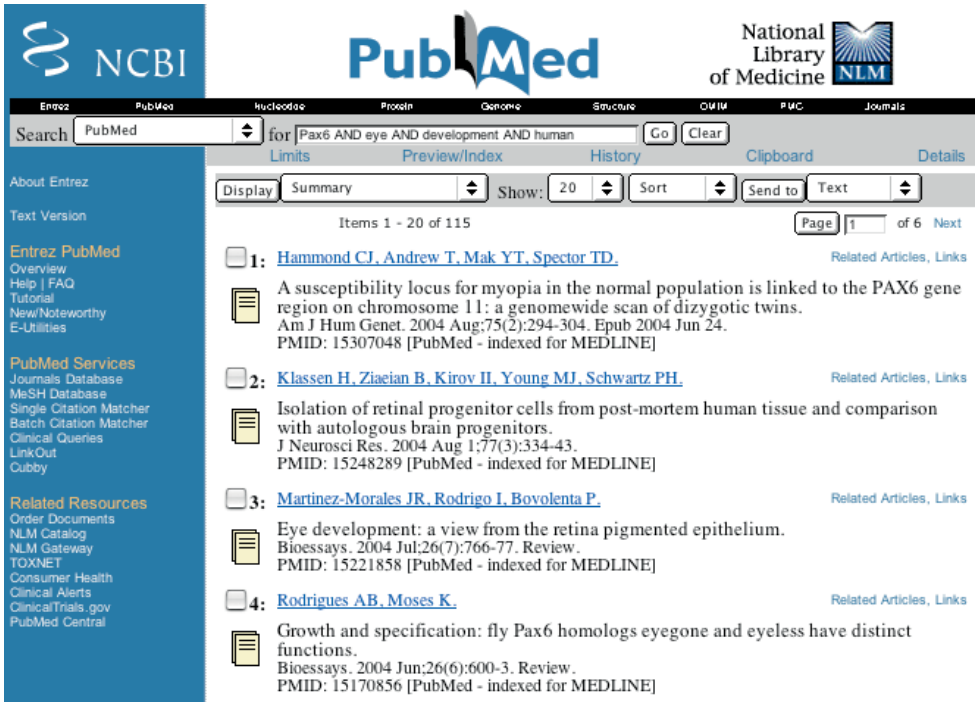
The result page shows the papers containing the terms used to search the database. By default the most recent papers are usually shown at the top of the page.

Your search will retrieve over 770 articles (the number will vary as new articles are added). Click on the yellow icon on the left to retrieve the abstract and, when available, the full text of the article.

The abstract view displays the information about the journal in which the article is published, the authors names and the laboratory, company or institute where they work. The title of the article and a summary of the paper are shown. The abstract is written by the author(s) and is part of the original paper. Finally, each PubMed entry has a unique identifier, the PMID.

By connecting to publishers web sites, it is possible to access the full text version of some articles. The article can then be downloaded for printing. Some useful links are also provided by the publisher (which articles quoted this paper for example).

Type **Pax6 AND eye AND development AND human** in the Search field.



Now you will have retrieved 115 papers or more focused on eye development. Click on the **yellow icon** to read one or two abstracts.

For each entry in PubMed, links to related articles and other databases (for genes, proteins, etc.) are provided on the right-hand side of the result list.

Of course, your search could go on forever, but we have now reached the end of our bioinformatics tour…

3 Glossary

Alignment:
The process of lining up two or more sequences to achieve maximal levels of identity (and conservation, in the case of amino acid sequences) for the purpose of assessing the degree of similarity and the possibility of homology.

Alignment Score:
The raw score S for an alignment is calculated by summing the scores for each aligned position and the scores for gaps. In AA alignments, the score for an identity or a SUBSTITUTION is given by the specified substitution matrix, e.g. BLOSUM62 (see NCBI tutorial for more details).
http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Alignment_Sc ores2.html

Alpha helices:
a common secondary structure in proteins where the polypeptide backbone is folded into a spiral that is held in place by hydrogen bonds between the oxygen and hydrogen atoms of the backbone. The outer surface of the helix is covered by AA side-chain groups.

Beta sheet:
formed by the hydrogen bonding between backbone atoms of adjacent beta strands, belonging either to the same chain or to different chains. The beta strands can be oriented in the same (parallel) or opposite (anti-parallel) direction (as defined by the orientation of the peptide bond) with respect to each other.

Beta strand:
short (5 to 8 AA) polypeptide segment, nearly fully extended.

Bit score:
The bit score shown on the result page (S') is derived from the raw alignment score S in which the statistical properties of the scoring system used have been taken into account. Because bit scores have been normalized with respect to the scoring system, they can be used to compare alignment scores from different searches.

Boolean:
Boolean is a logic system. Using the „AND" operator between terms retrieves documents containing both terms. „OR" retrieves documents containing either term. „NOT" excludes the retrieval of terms from your search. Use „NOT" with caution, in particular, for PubMed searches.

Conservation:
Changes at a specific position of an amino acid or (less commonly, DNA) sequence that preserve the physico-chemical properties of the original residue.

Domain:
A discrete portion of a protein assumed to fold independently of the rest of the protein and possessing its own function.

E value:
The Expect (E) value is a parameter that describes the number of hits one can „expect“ to see just by chance when searching a database of a particular size. It decreases exponentially with the Score (S) that is assigned to a match between two sequences. Essentially, the E value describes the random background noise that exists for matches between sequences. For example, an E value of 1 assigned to a hit can be interpreted as meaning that in a database of the current size one might expect to see 1 match with a similar score simply by chance.

Filtering:
Also known as Masking. The process of hiding regions of (nucleic acid or amino acid) sequence having characteristics that frequently lead to spurious high scores.

Homology:
Similarity attributed to descent from a common ancestor. Identity: The extent to which two (nucleotide or amino acid) sequences are invariant.

Orthologous:
Homologous sequences in different species that arose from a common ancestral gene during speciation; may or may not be responsible for a similar function.

Paralogous:
Homologous sequences within a single species that arose by gene duplication.

Primary structure (of a protein):
its linear arrangement of amino acids.

Query:
The input sequence (or other type of search term) with which all of the entries in a database are to be compared.

Random coil:
in the absence of stabilizing non-covalent interactions, apolypeptide adopts a random coil structure. This flexible region can be rich in functionally important determinants like short linear motifs.

Secondary structures:
various spatial arrangements from the folding of localized parts of a polypeptide chain.

Substitution:
The presence of a non-identical amino acid at a given position in an alignment. If the aligned residues have similar physico-chemical properties the substitution is said to be „conservative“.

Tertiary structure:
refers to the overall conformation of a polypeptide chain, i.e. the 3D-arrangement of all its AAs. In contrast with secondary structures, which are stabilized by hydrogen bonds, tertiary structure is primarily stabilized by hydrophobic interactions between the non-polar side chains, hydrogen bonds between polar side chains and peptide bonds. These stabilizing forces hold elements of the secondary structure compactly together.

4 References

Programme or Web based Resource	References
BLAST	1- Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410. 2- Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.
SWISS PROT/UNIPROT	1- Bairoch A., Apweiler R. The SWISS-PROT protein sequence database: its relevance to human molecular medical research J. Mol. Med. 75:312-316(1997). 2- Apweiler R., Gateau A., Contrino S., Martin M.J., Junker V., O'Donovan C., Lang F., Mitaritonna N., Kappus S., Bairoch A. Protein sequence annotation in the genome era: the annotation concept of SWISS-PROT + TREMBL. (In) ISMB-97; Proceedings 5th International Conference on Intelligent Systems for Molecular Biology, pp33-43, AAAI Press, Menlo Park, (1997).
SMART	Schultz,J., Milpetz,F., Bork,P. & Ponting,C.P. (1998). SMART, a simple modular architecture research tool: Identification of signaling domains. PNAS, 95, 5857-5864
PDB sum	1- Laskowski R A (2001). PDBsum: summaries and analyses of PDB structures. Nucleic Acids Res., 29, 221-222. 2- Laskowski R A, Hutchinson E G, Michie A D, Wallace A C, Jones M L, Thornton J M (1997). PDBsum: A Web-based database of summaries and analyses of all PDB structures. Trends Biochem. Sci., 22, 488-490.

Appendix I:
Answers to Questions

Question Set A

A.1- Which protein in the human dataset is the closest to the zebrafish **Pax6**? How long is this protein?

The human **Pax6** *is the closest to the zebrafish protein. It has the highest score (713 bits). It is 422 AA long. and its gene identifier is: gi 6174889. Its SwissProt accession number is P26367 and identifier PAX6_HUMAN.*

A.2- What is the degree of similarity between the query and the hit?

The 2 sequences share 84% identity (358 AA out of 422 are identical).

A.3- What is the probability that the similarity between the query and the hit occurs only by chance?

The E value is 0.0. This means that the 2 sequences are orthologs.

A.4- In the first alignment, what do you think the stretches XXX represent? And the stretch "---"?

XXX represents the low complexity regions (LCRs) which are taken into consideration during the alignment because they are masked by the low complexity FILTERING selected in the search. There are 3 LCRs, as depicted on the graphical output on the formatting BLAST page.
The "---" stretch represents "gaps" in one of the sequences, i.e. regions present in only one of the two aligned sequences. There are 4 gaps (i.e. 4 AA are missing in the zebrafish proteincompared to the human one).

A.5- Look at the second and third most relevant hits. How similar are they to the zebrafish **Pax6** sequence?

The next hits are human **Pax4** *and* **Pax3** *proteins, which are only 50 and 39% identical to zebrafish Pax6, respectively. These proteins belong to the PAX family of proteins. Their sequences are more divergent than that of Pax6, hence their lower Scores and higher E values.*

Question Set B

B.1- In which tissues is the protein found and at what stage of fetal development? Look under ***Comments***.

Pax6 *is expressed in the eye, brain, spinal cord and olfactory epithelium during foetal development.*

B.2- How many diseases are described in relation with defects in the **Pax6** protein? Which organs are affected by mutations in the ***PAX6*** gene?

Nine diseases are associated with defects in **Pax6** *protein function, affecting the eye or optic nerve.*

B.3- Why does “3D-structure” appear under keywords in this entry?

Because the tri-dimensional structure of **Pax6** *has been resolved experimentally by X-ray crystallography. Its coordinates are available in the PDB database.*

B.4- What is the molecular function of Pax6 and its cellular localization? Look under ***Comments***.

Pax6 *is a transcription factor (defined as any protein required to initiate or regulate transcription; includes both gene regulatory proteins as well as the general transcription factors). It is localized in the cell nucleus.*

B.5- How many bibliographic references are quoted in this entry? What are the main topics published in these papers? Which paper describes the evolutionary conservation of ***PAX6*** gene? Look under ***References***.

The entry contains 25 references. They contain information about the sequence of the nucleic acids encoding the protein, alternatively spliced isoforms of the protein (2 articles), DNA-binding properties (1 article), variants linked to diseases (17 articles) and three-dimensional structure of **Pax6** *(1 article). Ref [2] describes the genomic structure, evolutionary conservation and aniridia mutations in the human* ***PAX6*** *gene.*

Question Set C

C.1- Is the function of the paired box domain known?

The exact function of the PAX domain is unknown.

C.2- Are paired box genes found in plants? In fungi?

PAX genes are not found in plants or fungi. They are restricted to animals.

C.3- What is the function of the HOX domain?

It is involved in the regulation of transcription. It has DNA-binding properties

C.4- Are the structures of the domains resolved? How does the structure of the PAX domain help us to understand the structural basis of the mutations known to be linked to diseases?

Yes. 4 structures are collected in PDB for PAX domain representatives, and more than 40 for HOX genes. Under “Literature”, you can find an article showing that all known developmental miss-sense mutations in the paired box of mammalian ***pax*** *genes map to the N-terminal sub-domain, and most of them are found at the protein-DNA interface. Thus, the mutations affecting the development of the organs expressing* ***pax*** *genes are located in a region of the protein involved in an important function namely interaction with DNA.*

Acknowledgements

The cover image from the EMBL Photolab archive;

Layout design by Nicola Graf;

Edited by Corinne Kox.